Actes du Seizième Colloque sur l'Optimisation et les Systèmes d'Information

COSI'2019

24 - 26 Juin 2019, Tizi-Ouzou, Algérie

Organisation

Laboratoire LAROMAD, la Faculté des Sciences et et l'association AADEMTI
en partenariat avec
l'Agence Thématique de Recherche en Sciences et Technologie (ATRST)

Présidents d'honneur

Pr. Ahmed TESSA, Recteur de l'Université Mouloud Mammeri de Tizi-Ouzou
Dr. S. HASSANI, Directeur Général de l'ATRST
Pr. Smain HOCINE, Doyen de la Faculté des Sciences

Président du colloque

Pr Brahim OUKACHA, Directeur du Laboratoire LAROMAD, UMMTO (Algérie)

Membres

ABBAD Moussa (ATRST),
SADI Bachir, AIDENE Mohamed, OUANES Mohand, BELLAHCENE Fatima
KASDI Kamel, BOUARAB Ouiza, LESLOUS Fadila, BELHADJ Abdelaziz
AOUANE Mohouhend, HICINI Allaoua, TALEB Youcef, GOUMEZIANE Lynda
NOURI Naima, SELLAM Rachid, CHEBBAH Mohamed, SMAIL Rabah
OUBAKOUK Lynda, SADI Aris, MEDJOUTI Norredine, HAMOUTENE Ouafa, SIFAOUI Thiziri

Comité de Pilotage

Mohamed AIDENE, Univ. Mouloud Mammeri de Tizi-Ouzou, Algérie
Mohand Saïd HACID, Univ. Claude Bernard, Lyon I, France
Nadjet KAMEL, Université Ferhat Abbas Sétif 1, Algérie
Kahina LOUADJ, Université Akli Mohand Oulhadj de Bouira, Algérie
Lhouari NOURINE, Univ. Clermont Auvergne, Clermont Ferrand, France

Samia OURARI, CDTA, Alger, Algérie
Jean Marc PETIT, INSA de Lyon, France
Mohamed Said RADJEF, Université de Béjaia, Algérie
Bachir SADI, Univ. Mouloud Mammeri deTizi-Ouzou, Algérie
Lakhdar SAÏS, Univ. Artois, Lens, France
Rachid NOURINE, Université d'Oran 1, Algérie

Comité de programme

Président

Frédéric SAUBION, Professeur à l'Université d'Angers (France)

Membres

Mohamed Ahmed-Nacer, USTHB (Algérie)
Meziane Aider, USTHB, Alger (Algérie)
Hacène Ait Haddadene, USTHB Alger (Algérie)
Otmane Ait Mohamed, Université Concordia (Canada)
Hassan Aït-Kaci, Université Claude Bernard Lyon 1 (France)
Zaia Alimazighi, USTHB (Algérie)
Makhlouf Aliouat, UFAS1, Sétif (Algérie)
Zibouda Aliouat, UFAS1, Sétif (Algérie)
Adel Alti, UFAS1, Sétif (Algérie)
Tassadit Amghar, UniversitÈ d'Angers (Angers)
Kamel Barkaoui, CNAM-Paris (France)
Moussa Benaissa, LITIO (Oran 1)
Salima Benbernou, Université Paris Descartes (France)
Nacéra Benamrane, USTO Mohammed Boudiaf, Oran (Algérie)
Fatiha Bendali, LIMOS, Clermont-Ferrand (France)
Fatima Bendella, USTO Mohammed Boudiaf, Oran (Algérie)
Belaid Benhamou, Université d'Aix-Marseille I (France)
Djamel Benterki, UFAS1, Sétif (Algérie)
Abdelkader Benyettou, USTO Mohammed Boudiaf, Oran (Algérie)
Mohamed Benyettou, USTO Mohammed Boudiaf, Oran (Algérie)
Abdelhafid Berrachedi, USTHB Alger (Algérie)
Mokrane Bouzeghoub, Université de Versailles - CNRS, Paris (France)
Mohand Ouamer Bibi, Université de Béjaia (Algérie)
Isma Bouchemakh, USTHB (Algérie)
Mourad Boudhar, USTHB (Algérie)
Mohand Boughanem, IRIT, Toulouse (France)
Kamel Boukhalfa, USTHB (Algérie)
Brice Chardin, Lyon(France)
Bruno Cremilleux, Caen (France)
Guillaume Damiand, CNRS / LIRIS / UniversitÈ de Lyon 1 (France)
Laurent D'Orazio, Université Blaise Pascal, (France)
Fedoua Didi, Abou Bekr Belkaid Tlemcen (Algérie)
Habiba Drias, USTHB (Algérie)
Fizazi Hadria, USTO Mohammed Boudiaf, Oran (Algérie)
Jean-Marie Favreau, LIMOS, Université d'Auvergne (France)
Frederic Flouvat, Nouvelle calédonie(France)
Pierre Fouilhoux, Université Pierre et Marie Curie (France)
Zahia Guessoum, Lip6, Université de Paris 6 (France)
Michel Habib, Université Paris 7 (France)

Allel Hadjali, ENSSAT, (Lannion)
Said Hanafi, Université de Valenciennes (France)
Youssef Hamadi, LIX Ecole Polytechnique (France)
Said Jabbour, CRIL - Université d'Artois (France)
Hao Jin-Kao, Université d'Angers (France)
Souhila Kaci, LIRMM(France)
Okba Kazar, Université de Biskra (Algérie)
Hamamache Kheddouci, Université Claude Bernard Lyon 1 (France)
Nacima Labadie, Université de Technologie de Troyes (France)
Philippe Lacomme, Université Blaise Pascal (France)
Arnaud Lallouet, Huawei technolgies Ltd, Paris (France)
Sylvain Lamprier, Université Pierre et Marie Curie - UPMC (France)
Nadjib Lazaar, Lirmm, Université de Montpellier 2 (France)
Vincent Limouzy, Université Blaise Pascal (France)
Samir Loudni, Université de Normandie, Caen (France)
Sofian Maabout, LABRI(France)
Philippe Mahey, LIMOS(France)
Arnaud Mary, LBBE - Université de Lyon 1(France)
Zoulikha Mekkakia Mazaa, USTO Mohammed Boudiaf, Oran (Algérie)
Nouredine Melab, LIFL(France)
Engelbert Mephu, Université Blaise Pascal (France)
Rokia Missaoui, Université de Quebec en Outaouais (Canada)
Safia Nait Bahloul, LITIO (Oran 1)
Mohand Ouanes, Université Mouloud Mammeri de Tizi-Ouzou (Algérie)
Brahim Oukacha, Université Mouloud Mammeri de Tizi-Ouzou (Algérie)
Hacène Ouzia, Université Pierre et Marie Curie - Paris 6, Paris (France)
Badran Raddaoui, Télécom SudParis, Paris (France)
Michael Rao, ENS Lyon (France)
Allaoua Refoufi, UFAS1, sétif (Algérie)
Yakoub Salhi, Université d'Artois, Lens (France)
Hanafi Said, Université de Valenciennes (France)
Yacine Sam, Université de Tours (France)
Michel Schneider, Université Blaise Pascal, Clermont-Ferrand (France)
Hachem Slimani, UniversitÈ de Béjaia (Algérie)
Pierre Spiteri, INP- Toulouse (France)
Abdelkamel Tari, Université de Béjaia (Algérie)
Antoine Vacavant, Université Clermont Auvergne, IUT Le Puy-en-Velay (France)
Djemel Ziou, Université de Scherbrooke (Canada)

Relecteurs additionnels

Mohamed Bachir Belaid, Amine Kechid

Table des matières

Articles Courts (Posters)

Préface

Le Colloque sur l'Optimisation et les Systèmes d'Information (COSI) rentre cette année dans sa seizième édition, après celles de Tizi-Ouzou (2004), Bejaia (2005), Alger (2006), Oran (2007), Tizi-Ouzou (2008), Annaba (2009), Ouargla (2010), Guelma (2011), Tlemcen (2012), Alger (2013), Bejaia (2014), Oran (2015), Sétif (2016), Bouira (2017) et Oran*(2018).

L'édition de cette année est organisée par le Laboratoire LAROMAD, la Faculté des Sciences et et l'association AADEMTI en partenariat avec l'Agence Thématique de Recherche en Sciences et Technologie (ATRST).

Cette rencontre annuelle pluridisciplinaire est un modèle du genre, elle permet à des chercheurs Algériens et étrangers travaillant sur des thématiques transversales comme les l'intelligence artificielle, les systèmes d'information, l'optimisation combinatoire et la théorie des graphes de se rencontrer et de s'imprégner des dernières avancées technologiques. Les divers thèmes abordés admettent souvent des connexions à la fois théoriques - les problèmes posés sont souvent de nature combinatoire et difficiles - et pratiques - les applications cibles font souvent appel à diverses techniques issues des différents thèmes abordés durant le colloque.

COSI est un événement chaleureux et convivial favorisant les échanges scientifiques et humains ! Cette ambiance de travail fait que tout participant en garde un souvenir mémorable.

L'édition COSI'2019 a un caractère particulier, en plus des articles acceptés cette année, le comité de pilotage a pris la décision d'inclure dans les actes et dans le programme scientifique de COSI'2019, les articles et posters acceptés n'ayant pas pu être présentés lors de l'édition de COSI 2018 à Oran. Le programme scientifique comporte une trentaine d'articles acceptés et neuf posters provenant de différentes villes d'Algérie, de Tunisie, de France, du Canada, d'Ireland et du Royaume-Uni. Le programme du Colloque comporte également trois conférenciers invités. La première plénière, du Professeur Mourad Baiou, chargé de recherche au CNRS, est intitulée "*On Some Network Security Games*". Le Professeur Enjelbert Mephu Nguifo de l'Université Clermont Auvergne (France) répondra dans la seconde plénière à la question "*L'IA est-elle multiforme ? Regard croisé entre intelligence humaine et intelligence machine*". La troisième, du Professeur Mohand Ouamer Bibi, Professeur à l'Université de Béjaia (Algérie), mettra en exergue les "*Liens dialectique existant entre l'Optimisation, l'Analyse et l'Algèbre*".

Cette année, le colloque est précédé d'une école d'été d'une journée sur l'Intelligence Artificielle permettant ainsi aux jeunes chercheurs de s'imprégner de cette thématique d'actualité. Cette école d'IA, soutenue est co-organisée par l'Agence Thématique de Recherche en Sciences et Technologie (ATRST) et le laboratoire Laromad. Elle comporte trois cours :

- *Introduction to statistical and machine learning*
  Karim LOUNICI, Professeur à l'école Polytechnique, Palaiseau, Paris (France)
- *Things to Know about Machine Learning and Data Mining*
  Engelbert MEPHU NGUIFO, Professeur à l'Université Clermont Auvergne (France)
- *Towards Cross-Fertilization between Data Mining and Artificial Intelligence*
  Lakhdar SAÏS, Professeur à l'Université d'Artois, Lens (France)

Nous tenons à remercier très chaleureusement l'équipe organisatrice et plus particulièrement le Président du colloque Professeur Brahim Oukacha pour son dévouement et son travail remarquable pour la préparation de cette seizième édition. Nous tenons à remercier tous les auteurs qui ont soumis des articles montrant ainsi la vitalité de cette rencontre scientifique. Nous tenons également à exprimer notre profonde reconnaissance aux membres du comité de programme et aux rapporteurs supplémentaires qui ont fait un excellent travail. Enfin, n'oublions pas que l'organisation de ce colloque ne serait pas possible sans les aides de nos partenaires institutionnels et industriels. COSI Qu'ils reçoivent ici notre profonde reconnaissance.

Tous nos voeux de réussite pour cette rencontre scientifique et longue vie à COSI !

Frédéric Saubion (Président du comité scientifique), Lhouari Nourine et Lakhdar Saïs (Comité de Pilotage)

DevOps Workflow Specification based on Non-Markovian Stochastic Petri Nets with Enhanced Expressiveness

Walid Ben Mesmia[1], Mohamed Escheikh[2], and Kamel Barkaoui[3]

[1] SySCom-ENIT, Tunis, Tunisia
Benmesmiawalid77@yahoo.fr
[2] SySCom-ENIT, Tunis, Tunisia
mohamed.escheikh@gmail.com
[3] CEDRIC-CNAM, Paris, France
kamel.barkaoui@cnam.fr

Abstract. In this paper, we suggest to extend Non-Markovian Stochastic Petri Nets paradigm to specify the $DevOps\ Workflow$ steps reliability and performance. In order to detect $Devops\ Workflow$ steps execution failures. The proposed extension is based on the decomposition of each transition into three mciro-transitions, having specific firing conditions.

Keywords: Non-Markovian WSPN · Workflow · DevOps · Specification.

1 Introduction

Analytical modeling has an important role in computer systems design, verification and analysis. Indeed, the management, evaluation and verification of a company's business processes are fundamental for its success in a market's competitive environment [1, 2]. In this respective, the association of the business processes with the corporate objectives is not only strategically important, but the day-to-day operations and the business process support are also needed to ensure sound operations [3]. The business process is defined by [4], as any collection of activities that have one or more input types and provide the client with a satisfying output. Equally important, the same definition shows the relationship-semantics between the inputs (pre-condition) and the outputs (post-condition) of a business process. In this definition, the process concept abstraction is obvious. Therefore, we come to adopt a standard definition, which is proposed by BPMI [5], of the Workflow domain to cover the business process life cycle (design, deployment, execution, maintenance and optimization).
This paper focuses on the Workflow performance aspect, through the use a formal tool: Non-Markovian Stochastic Petri Nets. Hence, we propose an $Non-Markovian\ SPN$ extension favoring a $DevOps$ Workflow process specification, verification and evaluation.
In other words, our contribution consists in the Non-Markovian SPN extension

proposal that helps to specify the *DevOps* business process expressive performance. In order to closing the *DevOps* gap, we adopt a culture fostering collaboration; besides, we employ learning-based realization methods and a significant method in the automation process.
The article remaining parts are organized as follows. The next section is meant to gather some previous researches related to our topic. In Section 3, we define the our suggested model basic concepts. Section 4 illustrates the proposed $Non - Markovian\ WSPN$ model. Section 5 is devoted to presenting the *DevOps* process specification. The last section concludes our work and mirrors the future researches.

2 Relatedworks

The research works in [6] suggest a verification method for the collaborative $workflow$ processes based on the checking techniques model. The approach proposes a way to check the coherence properties of these workflow processes by using the SPIN model checker. Also, the model developed by [6] is similar to the model defined by [7] because both of them use the SPIN model checker in the workflow process verification phase. In [8–10], they study the synthesis under partial observability for logical specifications on finite marking expressed in $LTLf$ / $LDLf$. Actually, they prove that this form of synthesis can be considered as a generalization of the partial observation planning in non-deterministic domains. They also show that the usual "belief-state construction" used in partial observability planning is also useful for the general $LTLf$ / $LDLf$ synthesis. They, then, indicate that the belief state construction the can be avoided in the direct automation construction favor that uses the projection to hide unobservable propositions. To ensure business process control, they propose, in [11], a precise prediction remaining time given case and missing estimate a delay risk. To do so, they propose a specific Stochastic Petri Nets type capable to capturing the arbitrary duration distributions. Thus, they are able to achieve higher predictive accuracy than the related approaches.

3 Definitions and basic concepts

To introduce our model and the associated formal tools for $Workflow$ specifying, we define the following concepts.

3.1 Definition 1 (Workflow)

According to BPMI [4], a $Workflow$ is any activity choreography involving an interaction between the participants in the form of information exchange. The participants are applications, information system services, human actors or business processes. To meet the standardization goal, $BPMI$ proposes a set standards as shown in [5, 12, 13]:

- BPMN (Business Process Modeling Notation): defines the graphical notation that allows the designers to specify, in a standard way, the socio-economic organization processes.
- BPML (Business Process Modeling Language): is a meta-language based on XML(Extensible Markup Language) ensuring the business processes definition in or between the socio-economic organizations. The $BPML$ language stems from the $WSFL$ efforts [14], $WSCL$ [15], $WLANG$ [16].

3.2 Definitions 2 (DevOps)

DevOps [17] is a neologism representing a ICT movement professionals approaching a different attitude towards software delivery through the collaboration between the software development and the operation-functions based on a principles and practices set such as culture, automation, measurement and sharing. Another definition in [18] shows that *DevOps* is a development method aimed at bridging the gap between development and operations by focusing on communication and collaboration, continuous integration, quality and delivery assurance with automated deployment through the use of a development practices set.

3.3 Definition 3 (Stochastic Petri Net)

By introducing a delay in the PNs (Petri Nets) which aim to design a unique model allowing for both of the system verification and a qualitative and quantitative analysis, the *Stochastic Petri Nets* ($SPNs$) have been developed. An SPN is a temporized PN with a probability measure over the trajectories space. The firing sequences are measurable by considering a random space. Formally, an SPN is described in [19]: An SPN is a couple $S =< R, \phi >$, such that:

- $R =< P, T, Pre, Post, M0 >$ is the underlying Petri net,
- $\phi : T \rightarrow R^+$, the function that associates with each transition a fixed firing rate,
- $M0$: the Petri Nets initial marking.

In order to analyze the complex systems performance, many stochastic Petri nets classes are proposed:

- Generalized Stochastic Petri Nets ($GSPN$): we distinguish between transitions with a zero delay called immediate transitions and transitions with exponentially distributed random delay called stochastic transitions [20], [21].
- Deterministic and Stochastic Petri nets($DSPN$): an generalized stochastic Petri nets extension. It is characterized by immediate transitions, deterministic temporal transitions and stochastic delay transitions distributed according to exponential laws [22–25].

- Extended Stochastic Petri nets ($ESPN$): The model is formed only by random timed transitions. The time follows distribution a given law. The stochastic process underlying the marking chart is, with some restrictions, a semi-Markovian process [26].

We find that the Stochastic Petri Nets do not hold a firing memory. Therefore, we resort to the Non-Markovian Stochastic Petri Nets.

3.4 Definition and concept (Non-Markovian Stochastic Petri Nets)

Since the 1980s, some extensions have been proved so effective that they are now considered PN standard definition part. They are:

- The inhibitor arcs: They join a place at a transition and are drawn with a small circle on their destination. An p_i place inhibitor arc has a transition t_k which deactivates t_k when p_i is not empty.
- The priority transitions: Are integer numbers assigned to transitions. A transition in marking is active only if and only if no higher priority transition is active.
- The Marking-dependent arc Multiplicity: It has been presented in [27, 28] to model the situations in which the tokens number to be transferred along the arcs (or activating a transition) depends on the system state.

According to the PN model basic definition, to define an SPN with generally distributed transitions, the following entities must be specified for each transition $t_g \in T$: $thecdf$ $(G_g(t))$ of the random firing time γ_g, and the execution policy to determine (a_g, t_g). Since the 1980s, many SPN models classes have been developed. They incorporate non-exponential features in their specifications and favor the individual memory semantics which is discussed in [29]. In order to specify $Non - Markovian\ SPN$ models which are analytically dealt with: There are three research lines that can be considered [30, 31]:

- An approach based on the Markov regenerator theory [6, 32],
- An approach based on the additional variables use [6],
- An approach based on the state-space expansion [31].

The proposed model covers the $workflow$ life cycle by taking the Stochastic Petri Net approach advantages and the Non-Markovian approach to the systems' specification. The following section introduces the proposed model with its adjacent tools for $workflow$ specifying.

4 Non-Markovian Stochastic Petri Net Workflow Model (Non-Markovian WSPN)

Our ($Non - Markovian\ WSPN$) model is an stochastic Petri nets (SPN) extension which introduces new components (places, tokens and transitions) dedicated to the $Workflow$ context. This section is devoted to the $Non -$

$Markovian\ WSPN$ model basic formal concepts. After defining the model formally, its operating rules are formulated, interpreted and illustrated. The suggested model adopts the $WFMS$ concepts [27] and the refinements attributed by Van der Aslst and Van Her Hee in [28]. The activity $Workflow$ states possible belong to the set:
{"initiated", "in-execution", "active", "suspended", "achieved", "completed", "archived"}.

4.1 Definition of the model

The **Non-Markovian Stochastic Petri Net Workflow**, denoted **Non-Markovian WSPN**, is the 15-uplets:
Non-Markovian WSPN =< P, T, A, AC, SAC, Pre, Post, Disting, δ**,** φ**, M0, CardM, Frag, Com, Execute**>
Where:

- P is a non-empty finished set of places,
- $T = T_i \cup T_d \cup T_e = \{t_1, t_2, \ldots, t_n\}$: a non-empty finished transitions set,
- T_i : immediate transitions set,
- T_d : deterministic transitions set,
- T_e : stochastic transitions set,
- $A : Workflow - actors$ non-empty finished set,
- AC: $Workflow\ activities$ finished set,
- SAC: $sub - activities$ finished set associated with a Workflow activity,
- Pre: $P \times T \to \mathbb{N}$ an incidence application before,
- $Post$: $P \times T \to \mathbb{N}$ a back incidence application corresponds to the arcs,
- $Disting$: a function to distinguish between the three transitions types: immediate, deterministic and exponential,

$$\textbf{Disting: } T \to \{1,2,3\}\,;\, t \mapsto \begin{cases} 1, & t \in T_i \\ 2, & t \in T_d \\ 3, & t \in T_e \end{cases} \tag{1}$$

- δ : a transitions priority function such as:

$$\delta\textbf{: } T \to \mathbb{N}\,;\, t \mapsto \begin{cases} 0, & \forall t \in (T_e \cup T_d) \\ \delta(t) \geq 0, & \forall t \in T_i \end{cases} \tag{2}$$

  The function δ is useful to solve the conflicts problems between two immediate transitions. After firing an immediate transition, this value will be automatically decreased.
- φ : function that determines the transitions firing rate such that:

$$\varphi\textbf{: } T \times \mathbb{N}^{\times} \to \mathbb{R}^{+}\,;\, t \mapsto \begin{cases} 0, & \forall t \in (T_i) \\ \varphi(t) \geq 0, & \forall t \in T_d \\ \varphi(t) : follows\, the\, exponential\, laws, & \forall t \in T_e \end{cases} \tag{3}$$

,

- $M0$: $Non-Markovian\ WSPN$ initial marking is defined as follows:

$$M0\textbf{:}\ P \rightarrow A \cup SAC \cup \mathbb{N}^* \tag{4}$$

,

- $CardM$: function that determines the markings number in each mark accessible from $M0$. Formally $CardM$ is written as :

$$CardM\textbf{:}\ P \rightarrow \mathbb{N}\,;\, p \mapsto \begin{cases} \text{CardM(p)}=Card(M(p)),\quad \forall p \in P \\ \text{CardM:determines the tokens number in each place} \end{cases} \tag{5}$$

,

- $Frag$: function specifying an activity fragmentation into several sub-activities to ensure the execution-parallelism. This function is formally illustrated as follows :

$$Frag\textbf{:}\ AC \rightarrow SAC \times \mathbb{N}^*\,;\, a \mapsto \begin{cases} \text{Frag(a)}=< b, nbsac > \\ \forall a \in AC,\quad \exists b \in SAC \\ \text{nbsac}\quad \in \mathbb{N}^* \\ \text{nbsac=Card(b)} \end{cases} \tag{6}$$

- Com: function determines the value $< Wcom, \phi, r >$ which comes from communication between two or more actors involved in the sub-activity operational phase in a Workflow process. Com is formally defined as:

$$Com\textbf{:}\ (A \times A_c^2 \times P^2 \times T) \rightarrow \{0,1\}^2 \times \mathbb{R}^+ \times \{0,1\} \tag{7}$$

Such that:
$Com(a, A_i, A_j, P_c, P_d, T_k) =< Wcom, \phi, r >$
$Wcom$: Boolean Values sequence,
r : Boolean Value,
(ϕ : communication duration between the actors),
The Com is a pre-condition for transition firing by analyzing the $< Wcom,\ \phi,\ r >$ values.

- $Execute$: function specifying the Workflow sub-activity execution. The formal definition function is presented as follows:

$$Execute\textbf{:}\ (SAC \times A_c \times P^2 \times T) \rightarrow \{0,1\} \times \mathbb{R}^+ \times \{0,1\} \tag{8}$$

Such that:
$Execute(a, A_i, P_c, P_d, T_k) =< Wex, \phi, a >$
Wex : Boolean Values sequence
a : a Workflow sub-activity
(ϕ : a Workflow sub-activity execution time)

4.2 Semantics firing in the Non-Markovian WSPN model

We indicate with ${}^{e}t$:all transition input places $t \setminus^{e} t \subset P$. t^{s}: all transition output places $t \setminus t^{s} \subset P$. The model imposes that any transition goes through three phases, whatever their type, at the firing time:

- The transition t starts firing t^{c} by consuming a token from each ${}^{e}t$ place.
- The firing transition time t^{w} is equal to the duration $\varphi(t^{w})$. In the period restricted by τ and $\tau+ \varphi(t^{w})$, the tokens are stored in the transition t. So the tokens cannot participate in another firing in the time interval $\tau..\varphi(t^{w})$. As soon as the time counter reaches the value $\tau+\varphi(t^{w})$ and the returned values interpretation by the functions specific to the Workflow context ($Execute$, $Com\ and\ Frag$) is favorable, the $\varphi(t^{w})$ value will be automatically neutralized and, then, t can reach its final phase.
- The transition t is fulfilled by the t^{p} firing, by producing a token in t^{s} each place.

Fig 1 illustrates the notion T^{cwp}=$\{t^{c}, t^{w}, t^{p} \setminus t \in T\}$, ${}^{e}t$ and t^{s}. see(Fig. 1).

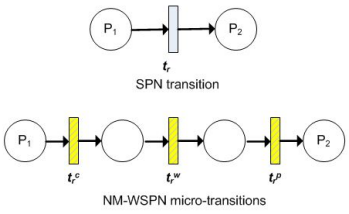


Fig. 1. SPN transition in to Non-Markovian WSPN transition.

In the next section, we apply the techniques adopted by our model to specify a software development $workflow$ according to the *DevOps* paradigm.

5 Case study (Geldzin development workflow)

5.1 Geldzin business process

We are inclined to the *Geldzin* software development $workflow$ [34] which adopts the *DevOps* culture. Fig 2 shows see(Fig. 2)the software development business process according to Geldzin. Our specification is restricted to the DevOps process steps (11, 12, 13 and 14) which is realized by using *Bizagi* Modeler [35] as illustrated by fig 3 see(Fig. 3).
The following section illustrates the $Non-Markovian\ WSPN$ model specifying process.

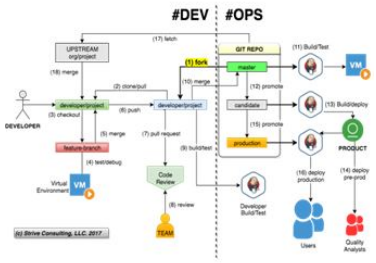


Fig. 2. Geldzin development business process[34].

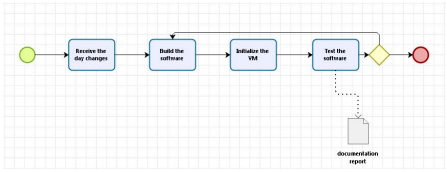


Fig. 3. The business sub-process represented with Bizagi Modeler.

5.2 Sub-process Specification with Non-Markovian WSPN model

- The set $A = \{A_1, A_2, A_3, A_4, A_5\}$ the developers and the operators.
- Specification AC represents the Geldzin development $workflow$ software tool.
- Specification $SAC = \{Sac_1, Sac_2\}$ two sub-activities resulting from the Ac activity fragmentation.
- Specification of sub-process model places :
  The set $P = \{P_2, P_6, P_{13}, P_{14}, P_{15}, P_{16}, P_{17}, P_{18}, P_{19}\}$
  $\bigsqcup \{P_{20}, P_{21}, P_{22}, P_{23}, P_{24}, P_{25}, P_{26}, P_{27}, P_{28}, P_{29}, P_{30}, P_{31}, P_{32}\}$.
- Specification of transitions:
  $Non - Markovian\ WSPN$ model is liable to distinguish between three transitions types ($immediate$, $temporized$, $deterministic\ and\ stochastic$). Each stochastic or temporized transition model is divided into three micro-transitions (T^c, T^w and T^p). An formal firing mechanism is elucidated by fig 4 see(Fig. 4).

Before the firing, the model detects the transition type via the function call $Distinct\ (T_{14})= 2$. The transition T_{14} firing, which is in sub activity executing charge $Sac1$, consists in sequentially firing the three micro-transitions $T_{14}{}^{c}$, $T_{14}{}^{w}$ and $T_{14}{}^{p}$. The $T_{14}{}^{c}$ transition pre-condition is the $uplet < 11, time, 1 >$ result provided by the function call $Com(A_2, A_3, P_{14}, P_{20}, T_{14})$ and the stochastic value $(0.33 * \varphi_{11})$. The micro-transition firing is influenced by the stochastic value $(0.67 * \varphi_{11})$ and formally by the $uplet < 1, time, Sac_1 >$ resulting from the function $Execute\ (Sac_1,$

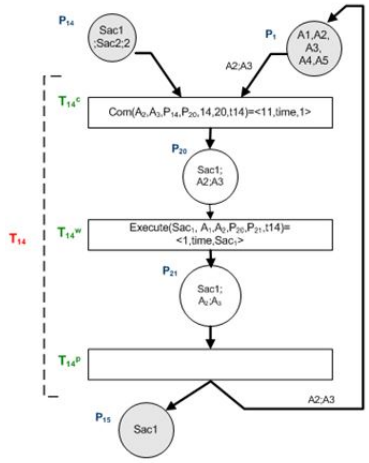


Fig. 4. Execution of the sub-activity " Test the software and publish reports or documentation " by the actors A_2 and A_3 .

A_2, A_3, P_{20}, P_{21}, T_{14}). The $T_{14}{}^{p}$ firing is immediate; i.e, the token Sac_1 is moving towards the place P_{19} and the two tokens A_1 and A_2 will be released to the place P_1 (the original place before firing T_{14}).

6 Conclusion

Our $Non-Markovian\ WSPN$ proposed model is a combination Stochastic Petri Nets properties and Non-Markovian-approach rules as well as $WFMC$ $workflow$ concepts and refinements attributed by Van der Aslst and Van Her Hee [12].
In this paper, we applied the $Non-Markovian\ WSPN$ model to specify some *Geldzin DevOps* $Workflow$ steps, where we illustrate the transition decomposition semantics and the formal firing mechanism. The formal transitions firing mechanism, is not based only on the stochastic and deterministic aspect, but on other formal tools adopted by the proposed model. Proposed model shows the dynamic interactions between $Workflow$ actors.
Workflow-oriented semantics, which we propose, enables detailed different steps description involved in the Workflow activity execution. These steps are formally represented by micro-transitions and concern respectively tokens consumption, the task execution (sub-activity) and the new tokens production. Such a representation makes it possible to distinguish and locate in a precise and fine manner the possible failure type.
The verification process adopted by $Non-Markovian\ WSPN$ model can be improved by a state space based on the markings cardinality to control and locate the shared resources conflicts.

In us future works, $Non-Markovian\ WSPN$ model can be enriched to adopt the analysis and verification concepts of the Fluid Stochastic Petri Net models $FSPN$ [35] in order to estimate and predict the $workflow$ execution delays.

References

1. Weske, M.: Business Process Management: Concepts, Languages, Architectures, Springer(2), (2012).
2. Dumas, M.,La Rosa, M.,Mendling, J., Reijers, H. A.: Fundamentals of Business Process Management, Springer(2013).
3. Mutschler, B., Reichert, M., Bumiller, J.: Unleashing the effectiveness of process-oriented information systems: Problem analysis, critical success factors, and implications, Systems, Man, and Cybernetics. In: PartC: Applications and Reviews, IEEE Transaction son , pp. 280–291 (2008).
4. Hammer, M., Champy, J.: Reengineering the Corporation: A Manifesto for Business Revolution. Harper business Essentials (2003).
5. WFMC: Workflow Management Application Programming Interface (interface 2 and 3) specification,Wfmc-tc-1009, vol. 2.0, (1998).
6. Sbai, Z., Barkaoui, K.: Vrification Formelle des Processus Workflow Collaboratifs. Confrence francophone sur les Systmes Collaboratifs (SysCo12) Conference 2012, pp. 197–20 (2012).
7. Yamaguchi, S., Yamaguchi, M., Tanaka, M.: A soundness verification tool based on the SPIN model checker for acyclic workflow nets. In the proceeding of ITC-CSCC (2008).
8. De Giacomo, G., Vardi, M. Y.: Ltlf and ldlf synthesis under partial observability. In Proc. of IJCAI, pp. 10441050 (2016).
9. Torres, J., Baier, J. A.: Polynomial-time reformulations of LTL temporally extended goals into final-state goals. In Proc. of IJCAI, pp. 16961703 (2015).
10. Camacho, A., Triantafillou, E., Muise, Ch., Baier, J. A., McIlraith, Sh: Non-deterministic planning with temporally extendedgoals: LTL over finite and infinite traces. In Proc. of AAAI (2017).
11. Andreas Rogge, S. A., Mathias, W.: Prediction of business process durations using non-Markovian stochastic Petri nets. In Journal Information Systems, vol. 54, pp. 01-14 (2015).
12. Van der Aalst, W. M. P., Van Hee, K.: Workflow Management: Models, Methods, and System. The MIL Press, ISBN 0-262-01189-1 (2002).
13. Axenath, B.: The Aspects of Business Processes: An Open and Formalism Independent Ontology, Technical report, Department of the Unversity of Paderborm (2005).
14. Akazi Technologies: Le Business Process Management et Administration Electronique, Akazi Technologies, Confrence Technoforum Novembre 2002, http://www.akazi.com/.
15. Brathauget, T. A, Evjen, T. A.: Enterprise Modeling, STF 38 A96302, (1996).
16. Chauvet, J.M.: Web Services avec SOAP, WSDL, UDDI, ebXML, Editions Eyrolles (2002)
17. Jabbari, R., bin Ali, N., Petersen, K., Tanveer, B: What is DevOps?.In Proc, Scientific Workshop Proceedings on Agile Conference 2016, ACM Press, pp. 0111 New York (2016).

18. Frana, d. B., Jeronimo, H., Travassos, GH.: Characterizing DevOps by Hearing Multiple Voices.In Proc. Proceedings of the 30th Brazilian Symposium on Software Engineering: ACM Press, pp. 5362 New York (2016).
19. Peterson, J.L: Petri net theory and the modeling of systems. Prentice Hall. Englewood Clis. (1981).
20. AjmoneMarsan, M., Balbo, G., Conte, G.: A Class of Generalized Stochastic Petri Nets for the Performance Analysis of Multiprocessor Systems. In ACM Transaction Computer, Systems, 2(2), pp. 93-122, (1984).
21. AjmoneMarsan, M., Balbo, G., Conte, G., Donatelli, S., Franceschinis, G.: Modelling with Generalized Stochastic Petri Nets. In John Wiley and Sons, (1995).
22. Lindemann, C.: Performance Modelling with Deterministic and Stochastic Petri Nets.In John Wiley and Sons,(1998)
23. Choi, I., Song, M., Park, C., Park, N.: An XML-based process definition language for integrated process management. In Computers in Industry, 50, pp. 85-102, (2003).
24. Choi, H.: Performance and reliability modelling using Markov regenerative stochastic Petri nets, Phd thesis, Graduate school of Duke University, (1993).
25. Ciardo, G., Lindemann, C.: Analysis of deterministic and stochastic Petri nets. In Proc. 5th International Workshop on Petri Nets and Performance Models (PNPM'93), IEEE Comp. Soc. Press., pp. 160-169, France (1993).
26. Dugan, J., Trivedi, K., Geist, R., Nicola, V.: Extended stochastic petri nets: Applications and analysis. In Proceeding Performance 84, France (1984).
27. Ciardo, G.: Petri nets with marking-dependent arc cardinality: properties and analysis. In Proceedings of the 15-th International Conference on Application and Theory of Petri Nets, pp 179-198. Lectures Notes in Computer Science 815 . Springer Verlag, (1994).
28. Ciardo, G.: Discrete-time markovian stochastic Petri nets. In Proceedings of the 2-nd International Workshop on Numerical Solution of Markov Chains, pp. 339-358, (1995).
29. AjmoneMarsan, M., Balbo, G., Bobbio, A., Chiola, G., Conte, G., Cumani, A.: The effect of execution policies on the semantics and analysis of stochastic Petri nets. In IEEE Transactions on Software Engineering, SE-15, pp. 832-846. (1989).
30. Ciardo, G., German, R., Lindemann, C.: A characterization of the stochastic process underlying a stochastic Petri net. In IEEE Transactions on Software Engineering 20, pp. 506-515. (1994).
31. Bobbio, A., Telek, M.: Non-exponential stochastic Petri nets: an overview of methods and techniques. In Computer Systems Science and Engineering. (1997).
32. Cox, D.R.: The analysis of non-markovian stochastic processes by the inclusion of supplementary variables. In Proceedings of the Cambridge Phylosophical Society 51, pp. 433- 440. (1955).
33. Trivedi, K., Kulkarn, V.i: FSPNs: fluid stochastic Petri nets. In Proceedings 14-th International Conference on Application and Theory of Petri Nets, pp. 24-31. Chicago (1993).
34. Geldizin softwhere development, http://blog.strive-ltd.com/category/software-development/devops/. Last accessed 07 April 2019.
35. Bizagi Modeler, https://www.bizagi.com/en/products/bpm-suite/modeler. Last accessed 01 April 2019.

Unscented Kalman Filter et observateurs exponentiels pour des systèmes non linéaires

DAID Assia[1,2], AIDENE Mohamed[1] et BUSVELLE Eric[2]

[1] Laboratoire de Conception et Conduite des Systèmes de Production,
Université Mouloud MAMMERI,Tizi-Ouzou, ALGÉRIE
aidene@ummto.dz
[2] Laboratoire d'Informatique et des Systèmes, LIS UMR 7020,
Université de Toulon, FRANCE
busvelle@univ-tln.fr, assia.daid@yahoo.fr

Résumé : Dans cet article, nous étudions les propriétés du filtre UKF ("Unscented Kalman Filter") en tant qu'observateur non linéaire. Sous des hypothèses d'observabilité, le filtre de Kalman étendu (EKF) est un observateur exponentiel sous réserve d'être écrit dans une forme canonique d'observabilité et sous sa forme grand-gain. On montre que contrairement à EKF, UKF n'est pas un observateur à convergence exponentielle. On propose une modification d'UKF qui est un meilleur candidat en tant qu'observateur. Finalement, les propriétés étudiées dans l'article sont illustrées sur un exemple de géolocalisation d'un navire.
Mots clés : Observateurs non-linéaires. Grand gain. Unscented Kalman filter. High-gain Unscented Kalman filter.

1 Introduction

Le filtre de Kalman étendu (EKF, "Extended Kalman Filter") est très utilisé par les ingénieurs pour estimer l'état d'un système à partir de mesures fournies par des capteurs [10]. Le modèle non-linéaire du système est établi d'après les connaissances physiques du système, de même que la relation entre les variables d'état internes et les mesures.

$$\begin{cases} \dfrac{dx(t)}{dt} = f(x(t), t) \\ y(t) = h(x(t), t) \end{cases} \tag{1}$$

Usuellement, $x(t) \in \mathbb{R}^n$ représente l'état du système et $y(t) \in \mathbb{R}^p$ représente les p mesures. Pour justifier l'utilisation du filtre de Kalman étendu, deux approches sont possibles, que nous allons rappeler dans les deux sections suivantes.

1.1 Filtre gaussien

Une première approche est de supposer l'existence de bruits (d'état, de mesure, conditions initiales) gaussiens, additifs, et c'est le cadre habituel pour utiliser le filtre de Kalman étendu. Ce dernier repose sur une linéarisation du système

autour de la trajectoire estimée. En raison des nonlinéarités, les variables aléatoires perdent leur caractère gaussien mais elles sont approchées par des variables aléatoires gaussiennes. Pour établir les équations du filtre, il suffit de savoir approximer par une loi normale la probabilité d'un couple de vecteurs aléatoires $X \in \mathbb{R}^n$ et $Y \in \mathbb{R}^p$ telles que X soit un vecteur aléatoire gaussien de loi $N(m, P)$ et $Y = g(X)$ où g est une transformation non-linéaire. L'approximation habituelle de la loi du couple (X, Y), qui conduit à EKF, utilise l'approximation au premier ordre :

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m \\ g(m) \end{pmatrix}, \begin{pmatrix} P & PA^t \\ AP & APA^t \end{pmatrix} \right) \tag{2}$$

où A est la matrice Jacobienne de g au point m. Dans le cas particulier où g est une application linéaire, $g(X) = AX$, cette approximation est une égalité et conduit aux équations de Kalman qui, rappelons-le, donnent la moyenne et la matrice de covariance de la loi de l'état conditionnellement aux mesures. Dans le cas plus général, cette approximation conduit au célèbre filtre de Kalman étendu (EKF).

Une autre approche a été développée plus récemment, permettant de mieux prendre en compte la transformation non linéaire d'un vecteur gaussien. Elle est basée sur la transformation dite "unscented" [3]. L'approximation dite "unscented" consiste à écrire

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m \\ \mu_U \end{pmatrix}, \begin{pmatrix} P & C_U \\ C_U^t & S_U \end{pmatrix} \right) \tag{3}$$

où le vecteur μ_U et les matrices C_U et S_U sont approximées à partir de $g(m)$ et de l'image par g de $2n$ σ-points judicieusement placés autour de m. Nous expliciterons plus loin cette construction, voir aussi [6,9]. Cette approximation conduit à une variante du filtre de Kalman connue sous le nom de "unscented Kalman filter" (UKF). Cette version non linéaire du filtre de Kalman présente plusieurs avantages, le premier étant donc une meilleure prise en compte de la propagation d'un bruit gaussien dans un système non linéaire. Le deuxième avantage est que cette version non linéaire ne nécessite pas de calculer deux matrices Jacobiennes qui sont parfois complexes et sources d'erreurs numériques (voir [4] pour un exemple concret).

1.2 Observateurs

Une seconde approche au problème d'estimation de l'état d'un système est purement déterministe. Elle consiste à construire un autre système dynamique ("un observateur") qui utilise comme entrée les mesures disponibles et dont l'état converge asymptotiquement (et généralement exponentiellement) vers l'état du système. Dans le cas linéaire, le plus simple est l'observateur de Luenberger, mais on peut aussi utiliser le filtre de Kalman dont les propriétés sont bien

3. La traduction littérale serait "sans parfum" mais elle n'est presque jamais utilisée. Nous laisserons donc le terme en anglais

supérieures (c'est la même différence qu'entre le placement de pôle et la commande linéaire-quadratique en contrôle). Dans cette approche, les matrices Q et R qui représentaient les matrices de covariance du bruit d'état et du bruit de mesure sont interprétées comme des matrices de coût quadratique. Le lien entre les deux approches est assuré par le fait qu'une espérance conditionnelle dans le cas gaussien n'est autre qu'une projection orthogonale et donc une minimisation quadratique. Une condition nécessaire importante pour la convergence d'un observateur est la condition d'observabilité (qui est la condition de Kalman dans le cas linéaire). L'observabilité permet d'établir que le problème est bien posé. C'est une condition qui n'intervient pas dans le cas stochastique mais qui est pourtant nécessaire [4] pour que le filtre de Kalman étendu ait une chance de donner de bons résultats. Sous cette hypothèse, le filtre de Kalman étendu dit "grand gain" est un observateur exponentiel du système [5]. Le filtre de Kalman étendu grand-gain (HG-EKF) est un filtre de Kalman étendu dont les matrices Q et R sont judicieusement choisies, appliqué au système dans des coordonnées canoniques dont l'existence est assurée par l'hypothèse d'observabilité. Puisque le filtre de Kalman étendu permet de construire un observateur exponentiel, il est légitime de se demander ce que le "unscented Kalman filter" peut apporter en tant qu'observateur. La réponse à cette question est simple et sera donnée dans la section suivante : contrairement à ce qui se passe avec EKF, la version grand-gain de UKF n'est pas un observateur, car elle ne converge pas vers l'état du système.

L'utilisation d'un filtre en tant qu'observateur s'est révélée très efficace et le filtre de Kalman est très apprécié des ingénieurs, en l'absence de toute hypothèse sur le bruit. Les matrices Q et R sont utilisées comme des paramètres de réglage, ce qu'elles sont si on pose le problème d'observation comme un problème d'estimation optimale avec critère quadratique. Mais la propriété fondamentale d'un observateur est la convergence, si possible exponentielle. On montre facilement que le filtre de Kalman est un estimateur exponentiellement convergent. Dans le cas non linéaire, on peut vérifier que le filtre de Kalman étendu ne converge pas toujours et peut même donner une estimation très éloignée de la valeur réelle de l'état. En revanche, on a montré que le filtre de Kalman étendu grand-gain est un observateur exponentiellement convergent. Dans la suite, nous désignerons les filtres et observateurs considérés par leurs acronymes qu'il est utile de rappeler ici :

- EKF pour "Extended Kalman filter", le filtre de Kalman étendu ;
- UKF pour "Unscented Kalman filter", le filtre de Kalman "unscented" ;
- UKO pour "Unscented Kalman observer", l'observateur de Kalman "unscented".

et on fera précéder ces acronymes du préfixe "HG-" pour désigner leur version "high-gain" (grand-gain, que nous détaillerons dans la section 4). Nous ne parlons pas d'observateur de Kalman étendu (EKO) car il coïnciderait exactement avec EKF. Nous présenterons UKF dans la Section 2, puis dans la Section 3, nous montrerons que UKF ne converge pas et nous proposerons donc une version

4. éventuellement dans une version affaiblie : la détectabilité

améliorée (du point de vue observateur) d'UKF que nous appellerons UKO pour "Unscented Kalman Observer". Dans la Section 4, nous proposerons la version grand-gain de ce nouvel observateur, que nous désignerons par HG-UKO. Les performances de cet observateur seront étudiées dans la Section 5 et comparées aux précédents.

2 De EKF à UKF

Nous allons présenter brièvement UKF dans le cas continu [7]. Ce filtre est donc basé sur l'utilisation de "σ-points" et d'une transformation non linéaire par la dynamique du système. Le nombre de σ-points dépend de la taille du vecteur d'état. Il convient de choisir adéquatement ces σ-points, notés X, ainsi que les poids W qui leur sont associés.

2.1 La transformation Unscented

Dans cette section, nous rappelons les bases de la transformation "unscented" et les équations de UKF [7,8]. Nous nous plaçons dans $\mathbb{R}^n$.

1. Choisir $2n+1$ σ-points :

$$X = [m \cdots m] + \sqrt{c}\left[0 \ \sqrt{P} \ -\sqrt{P}\right] \tag{4}$$

   X est la matrice des σ-points, $c = \alpha^2(n+k)$, avec $k \geq 0$, $\alpha \in (0,1]$.
   c, k et α sont des paramètres de réglage. La matrice P est symétrique définie positive. Elle peut donc se décomposer sous la forme de Cholesky $P = BB^t$ et on note $B = \sqrt{P}$.

2. Calculer les poids associés aux σ-points :
   $W_m = (W_m^0, W_m^1, \dots W_m^{2n})^t$ où

$$W_m^{(0)} = \frac{\lambda}{n+\lambda};$$
$$W_m^{(i)} = \frac{1}{2(n+\lambda)}, \quad i = 1, \dots 2n;$$

   et $W_c = (W_c^0, W_c^1, \dots W_c^{2n})^t$ où

$$W_c^{(0)} = \frac{\lambda}{n+\lambda} + (1 - \alpha^2 + \beta);$$
$$W_c^{(i)} = \frac{1}{2(n+\lambda)}, \quad i = 1, \dots 2n;$$

   λ est un paramètre scalaire défini par $\lambda = c - n$.

3. Transformer chaque σ-point par la transformation non linéaire g,

La moyenne et la covariance de $g(X)$ sont données par :

$$E[g(X)] \approx m = g(X)W_m = \sum_{i=0}^{2n} W_m^i g(X_i);$$

$$\mathrm{Cov}(g(X)) \approx \sum_{i=0}^{2n} W_m^i (g(X_i) - m)(g(X_i) - m)^t;$$

On a noté X_i la $i^{\text{ième}}$ colonne de X, et lorsque g est appliquée à la matrice X, $g(X)$ représente la matrice $[g(X_0) \cdots g(X_{2n})]$.

La matrice W est définie comme suit

$$W = \left(I - \begin{bmatrix} W_m & \cdots & W_m \end{bmatrix}\right) \times \mathrm{diag}\left(W_c^0 \cdots W_c^{2n}\right) \times \left(I - \begin{bmatrix} W_m & \cdots & W_m \end{bmatrix}\right)^t \quad (5)$$

2.2 Algorithme de UKF dans le cas continu

Les équations correspondant à UKF dans le cas continu pour le système (1) sont données par

$$\begin{cases} K(t) = X(t)Wh^t(X(t),t) \\ \dfrac{dm(t)}{dt} = f(X(t),t)w_m + K(t)\left(y(t) - h(X(t),t)w_m\right) \\ \dfrac{dP(t)}{dt} = X(t)Wf^t(X(t),t) + f(X(t),t)WX^t(t) + Q(t) - K(t)R(t)K^t(t) \end{cases} \quad (6)$$

Dans cet algorithme Q et R sont des matrices de covariance de l'état et de bruit de mesure respectivement, elles sont symétriques définies positives. Dans le cas déterministe, ces deux matrices seront considérées comme des paramètres de réglage. Voir [7,8] pour le passage de la transformation "unscented" (3) aux équations UKF.

3 De UKF à UKO

3.1 Non convergence de UKF en tant qu'observateur

Pour les même raisons que le filtre de Kalman étendu, les filtres de type "unscented" non grand-gain ne convergent pas dans le cas non linéaire. Il est donc légitime de se demander s'ils peuvent converger dans une version grand-gain. Malheureusement, HG-UKF ne peut pas converger exponentiellement pour une raison assez simple.

L'équation qui régit l'évolution de l'estimation de l'état est de la forme

$$\frac{dm(t)}{dt} = f(X(t),t)W_m + K(t)\left(h(x(t),t) - h(X(t),t)W_m\right) \quad (7)$$

où X est une matrice dont la première colonne est $m(t)$, l'estimation de l'état à l'instant t donnée par l'observateur et les colonnes suivantes sont les σ-points. Le vecteur de pondération W_m est défini dans la Section 2.

Afin de montrer pourquoi UKF ne peut converger, il suffit de considérer l'exemple très simple $f(x,t) = -x(1+(2x-1)^2)$ et $h(x,t) = x$ avec $x(0) = 0$. Cet exemple est construit de sorte que 0 soit un point globalement asymptotiquement stable du système mais que $x \longrightarrow f(x,t)$ ne soit pas impaire (et bien sûr non linéaire).

Par l'absurde, si $m(t) = 0$, c'est à dire si l'observateur estime parfaitement l'état, alors (7) devient

$$\frac{dm(t)}{dt} = f(X(t),t)W_m = 4cPW_m^1 = 2P \tag{8}$$

où P est la solution asymptotique strictement positive d'une équation de Riccati. Ainsi, ce résultat contredit l'hypothèse que $m(t)$ converge asymptotiquement vers la solution. Le résultat d'une simulation numérique est montré Figure 1. Puisqu'il s'agit d'observateurs à convergence exponentielle, on a utilisé une échelle logarithmique pour l'axe des temps.

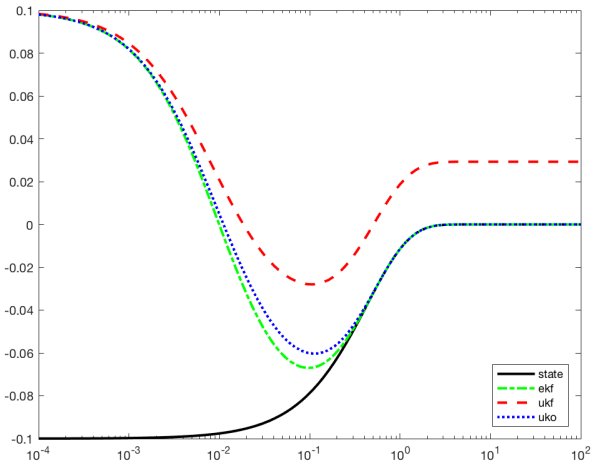


FIGURE 1: L'estimation biaisée de UKF

Nous voyons que UKF ne converge pas et qu'il reste un biais d'estimation, conforme à l'équation (8). Nous allons maintenant expliciter UKO dont on voit Figure 1 que les performances sont bonnes.

3.2 "Unscented Kalman Observer"

Pour cela, nous allons introduire une variante à mi chemin entre l'approximation du premier ordre (2) et l'approximation "unscented" (3) qui consiste à

utiliser

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m \\ g(m) \end{pmatrix}, \begin{pmatrix} P & C_U \\ C_U^t & S_U \end{pmatrix}\right) \tag{9}$$

Comme les deux précédentes, cette transformation permet d'établir les équations d'un filtre que nous appellerons "unscented Kalman observer". En effet, ce filtre n'apporte rien de plus à EKF (contrairement à UKF) mais du point de vue des observateurs, UKO permet de construire un observateur grand-gain (HG-UKO) exponentiellement convergent. C'est un observateur de ce type qui est utilisé sur la Figure 1. Les équations de cet observateur, dans sa version grand-gain, sont données Section 4. Nous montrerons l'intérêt de ce nouvel observateur par rapport à HG-EKF dans la Section 5

4 De UKO à HG-UKO

4.1 Observateur Grand-Gain

L'observateur grand-gain présenté dans cette section est celui de Gauthier et al [5]. Il est montré dans cette monographie que si un système à la propriété d'être observable pour toutes les entrées, alors il existe une transformation de coordonnées tel qu'un système à plusieurs entrées et une seule sortie (MISO) puisse s'écrire sous forme canonique d'observabilité. L'observateur grand-gain consiste à écrire l'observateur dans ces nouvelles coordonnées et à redéfinir les matrices Q et R de sorte qu'elles dépendent d'un paramètre qui sera supposé "assez grand" (la notion "assez grand" étant explicitée dans le théorème). Ainsi, le système (1) s'écrit dans les nouvelles coordonnées :

$$\begin{cases} \dot{x}(t) = Ax(t) + b(x(t), t) \\ y(t) = Cx(t) \end{cases} \tag{10}$$

ou $x(t) \in \mathbb{R}^n$ et $y(t) \in \mathbb{R}$ sont l'état et la mesure respectivement. Pour simplifier, nous n'avons pas considéré de variable d'entrée $u(t)$ (ce qui ne change rien pour notre propos qui est de construire un observateur) et nous avons supposé que la dimension des sorties est 1, ce qui facilite la définition de la forme normale. Pour le cas général, voir par exemple [2].

Les matrices A,C sont définies comme suit :

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & & 1 \\ 0 & \cdots & & & 0 \end{pmatrix} \text{ et } C = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}$$

Le vecteur $b(x(t), t)$ est supposé a support compact et ayant une forme triangulaire :

$$b(x(t), t) = \begin{pmatrix} b_1(x_1(t), t) \\ b_2(x_1(t), x_2(t), t) \\ \vdots \\ b_n(x(t), t) \end{pmatrix}$$

On note L_b la borne de la matrice Jacobienne $b^*(x(t),t)$ de $b(x(t),t)$, c'est à dire que $\|b^*(x(t),t)\| \leq L_b$ ($b^*(x,t)$ représente la matrice jacobienne de b par rapport à x).

La fonction $b(x(t),u(t))$ est supposée uniformément Lipschitz par rapport à $x(t)$: $\|b(x_1(t),t) - b(x_2(t),t)\| \leq L_b \|x_1(t) - x_2(t)\|$

Alors le système (1) admet un observateur de la forme suivante :

$$\begin{cases} \dfrac{dm}{dt} = Am + b(m,t) + PC^tR^{-1}(y(t) - Cm) \\ \dfrac{dP}{dt} = (A + b^*(m,t))P + P(A + b^*(m,t))^t + Q^\theta - PC^tR^{-1}CP \end{cases} \tag{11}$$

où la ligne i et la colonne j de Q^θ est égale à $Q^\theta_{i,j} = \theta^{i+j+1}Q_{i,j}$.

Cet observateur est décrit, parmi d'autres, dans [3]. Il s'agit d'un classique EKF, mais écrit dans les coordonnées canoniques et avec une matrice Q particulière (Q^θ) et c'est cet observateur que nous appelons HG-EKF.

4.2 Algorithme HG-UKO dans le cas continu

Les équations correspondant à HG-UKO dans le cas continu pour le système (1) sont données par

$$\begin{cases} \dfrac{dm}{dt} = Am + b(m,t) + PC^tR^{-1}(y(t) - Cm) \\ \dfrac{dP}{dt} = XW(AX + b(X,t))^t + (AX + b(X,t))WX^t \\ \qquad\qquad + Q^\theta - XWX^tC^tR^{-1}CXWX^t \end{cases} \tag{12}$$

La matrice Q^θ est définie comme dans (11). Ces équations sont simplement obtenues comme dans [7,8] mais en utilisant la transformation "unscented" modifiée (9).

5 Exemple de géolocalisation

5.1 Énoncé

On considère un navire qui se déplace dans le demi-plan $x_2 > 0$ (le demi-plan $x_2 \leq 0$ correspondant à la terre) à une vitesse inconnue dans un repère orthonormé (x_1, x_2). Pour se repérer il utilise deux amers (par exemple, deux phares notés A et B) et il mesure en continu la direction de ces phares par rapport au nord, donné par une boussole électronique. En utilisant l'angle θ entre sa position et celle de chaque phare, il peut estimer sa position instantanée. L'état du système est noté $x(t) = (x_1(t), x_2(t), \theta(t))$, θ étant l'angle entre la direction du navire et le nord.

Le problème de géolocalisation s'exprime alors sous la forme d'équation d'état et de mesure suivante :

$$\begin{cases} \dot{x_1}(t) = u(t)\cos(\theta(t)) \\ \dot{x_2}(t) = u(t)\sin(\theta(t)) \\ \dot{\theta}(t) = v(t) \end{cases} \tag{13}$$

Les sorties sont modélisées par les équations

$$\begin{cases} y_1(t) = \arctan \dfrac{x_2(t)}{x_1(t)} \\ y_2(t) = \arctan \dfrac{x_2(t)}{(x_1(t)-1)} \end{cases} \tag{14}$$

en supposant que les deux phares ont comme positions respectives $(0,0)$ pour A et $(1,0)$ pour B, ce qui ne nuit pas à la généralité du problème posé (quitte à faire un changement de coordonnées linéaire).

Pour pouvoir appliquer le HG-UKO on met le système (13)-(14) sous une forme canonique observable en utilisant un changement de variable approprié. Nous avons choisit :

$$\phi(x(t)) = \begin{pmatrix} \arctan \frac{x_2(t)}{x_1(t)} \\ \arctan \frac{x_2(t)}{(x_1(t)-1)} \\ \frac{x_1(t)sin\theta(t)-x_2(t)cos\theta(t)}{x_1(t)^2+x_2(t)^2} \\ \frac{(x_1(t)-1)sin\theta(t)-x_2(t)cos\theta(t)}{(x_1(t)-1)^2+x_2(t)^2} \end{pmatrix} = \xi(t)$$

Le système (13) s'écrit dans ces coordonnées canoniques :

$$\begin{pmatrix} \dot{\xi_1}(t) \\ \dot{\xi_2}(t) \\ \dot{\xi_3}(t) \\ \dot{\xi_4}(t) \end{pmatrix} = \begin{pmatrix} u\xi_3(t) \\ u\xi_4(t) \\ (v-2u\xi_3(t))\frac{x_1(t)\cos\theta(t)+x_2(t)\sin\theta(t)}{x_1(t)^2+x_2(t)^2} \\ (v-2u\xi_4(t))\frac{(x_1(t)-1)\cos\theta(t)+x_2(t)\sin\theta(t)}{(x_1(t)-1)^2+x_2(t)^2} \end{pmatrix}$$

et les sorties sont $y_1(t) = \xi_1(t)$ et $y_2(t) = \xi_2(t)$.

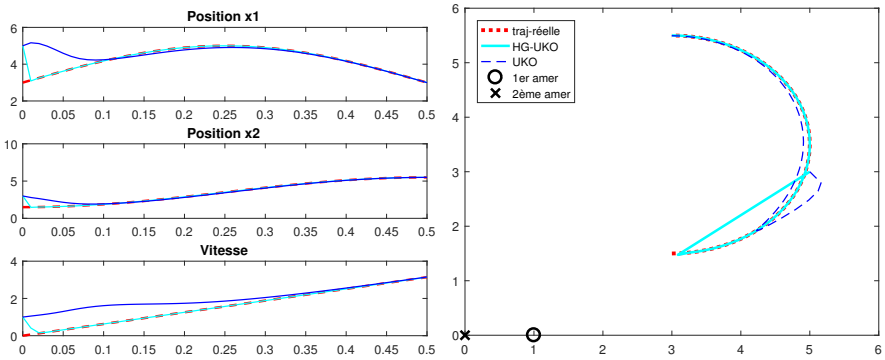
5.2 Etude comparée des performances de UKO avec HG-UKO, HG-UKO avec HG-UKF et HG-EKF

Pour les tests, on va supposer que le navire suit une trajectoire circulaire à vitesse constante V, ce qui se fait en prenant les deux contrôles constants $u(t) = V$ et $v(t) = \omega$ où $\omega = \frac{V}{R}$. En choisissant $x_1(0) = c_1$, $x_2(0) = c_2 - R$, $\theta(0) = 0$, la solution est

$$\begin{cases} x_1(t) = R\sin(\omega t) + c_1 \\ x_2(t) = -R\cos(\omega t) + c_2 \\ \theta(t) = \omega t \end{cases} \tag{15}$$

et on a bien $\sqrt{\dot{x_1}(t)^2 + \dot{x_2}(t)^2} = V$.

Pour commencer, on compare UKO et HG-UKO qui sont tous deux appliqués pour la géolocalisation du navire. L'un comme l'autre sont utilisés comme observateurs, mais le second est appliqué dans les coordonnées canoniques d'observabilité et avec un grand gain ce qui lui assure de meilleures performances.



(a) Etat en fonction du temps (b) Trajectoire

FIGURE 2: UKO et HG-UKO

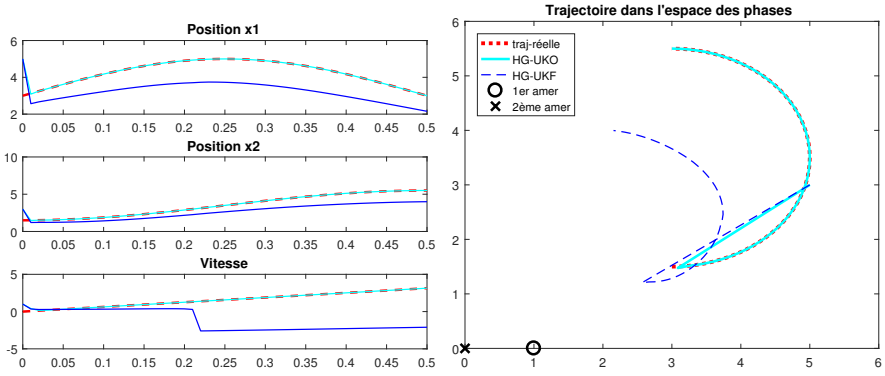
La figure 2 montre les résultats de cette première simulation. La figure de gauche, Fig. 2a, montre l'évolution de l'état et des estimations en fonction du temps, ce qui permet de vérifier que la convergence de HG-UKO peut-être très rapide (en fait, arbitrairement rapide). Il faut cependant signaler que nous n'avons pas mis de bruits dans ces simulations dans la mesure ou nous testons uniquement la performance d'UKO en tant qu'observateur. On peut constater Fig. 2b la performance des deux observateurs dans l'espace physique.

On compare maintenant HG-UKO avec HG-UKF (Fig. 3)

On voit, comme cela a été montré dans la Section 3, que le HG-UKF ne peut pas converger, la figure 3a montre très clairement une erreur statique inévitable et particulièrement importante pour la version de base HG-UKF (amplifiée par le grand-gain, voir Fig. 3b), alors que HG-UKO converge tout aussi rapidement mais vers le bon état.

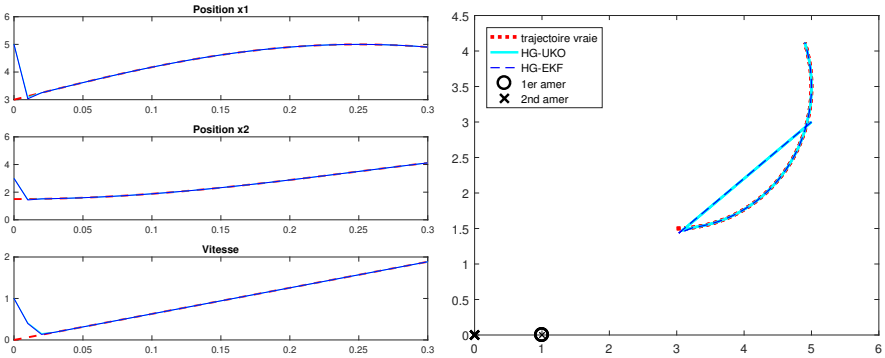
Nous avons ensuite comparé HG-UKO avec HG-EKF(Fig.4). On sait que ce dernier est un observateur exponentiel qui converge à une vitesse arbitrairement grande. La différence entre les deux algorithmes se situe essentiellement sur les calculs effectués (calcul des Jacobiennes dans le cas de HG-EKF, propagation des σ-points dans le cas de HG-UKO).

On peut constater que les résultats sont similaires, ce qui est très intéressant car on retrouve les bonnes performances de HG-UKO (en tant qu'observateur) sans les calculs fastidieux des Jacobiennes. L'algorithme HG-UKO ne prend donc en entrée que les mesures et le modèle entrées-sorties du système.



(a) État en fonction du temps (b) Trajectoire

FIGURE 3: HG-UKO et HG-UKF



(a) Etat en fonction du temps (b) Trajectoire

FIGURE 4: HG-UKO et HG-EKF

6 Conclusion

Parmi les filtres/observateurs non linéaires présentés, seuls HG-EKF et HG-UKO convergent en simulation, alors que HG-UKF n'est pas un observateur. L'avantage de HG-UKO est une relative simplicité d'écriture puisqu'il n'est pas nécessaire de calculer les Jacobiennes du système. Cette propriété est particulièrement intéressante lorsque le système est écrit dans sa forme canonique, les expressions étant souvent plus complexes.

Dans cet article, nous n'avons pas démontré la convergence de HG-UKO qui nécessiterait de borner la solution de l'équation de Riccati dans (12), comme cela a été fait pour HG-EKF dans [5]. Ce travail fera l'objet d'une prochaine publication.

Ces résultats ont été établis pour des systèmes continus. De même que pour le filtre de Kalman étendu grand-gain, on peut imaginer que ces résultats se généralisent pour des systèmes continus/discrets. De même, une version "square-root" doit pouvoir être écrite, améliorant la stabilité numérique de UKO.

Enfin, nous n'avons pas étudié les performances de HG-UKO en présence de bruit. Nous pouvons facilement imaginer qu'elles seront sensiblement équivalentes à celles de HG-EKF et qu'il sera donc préférable d'utiliser une version à gain adaptatif de HG-UKO (comme dans [1]).

Références

1. N. Boizot, E. Busvelle, and J.-P. Gauthier, An adaptive high-gain observer for nonlinear systems,. *Automatica,* vol. 46, no. 9, pp. 1483–1488, Sep. 2010.
2. Nicolas Boizot, Adaptative high-gain extended Kalman filter and applications, *Ph. D. Université de Bourgogne, Université du Luxembourg*, 2010
3. Nicolas Boizot and Eric Busvelle, Adaptive-gain observers and applications, in *Nonlinear Observers and Applications*, Springer Berlin Heidelberg, 2007, pp. 71–114.
4. M. Doumiati, A. Victorino, A. Charara, and D. Lechner, Unscented Kalman filter for real-time vehicle lateral tire forces and sideslip angle estimation, *IEEE Intelligent Vehicles Symposium*, Jun. 2009.
5. J.-P. Gauthier and I. A. K. Kupka, *Deterministic Observation Theory and Applications,* Cambridge University Press, 2001.
6. Simon J. Julier and Jeffrey K. Uhlmann, New extension of the Kalman filter to nonlinear systems, *Signal processing, sensor fusion, and target recognition* VI. Vol. 3068. International Society for Optics and Photonics, 1997
7. S. Särkkä, On Unscented Kalman Filtering for State Estimation of Continuous-Time Nonlinear Systems, *IEEE Transactions on Automatic Control,* vol. 52, no. 9, Sep. 2007, pp. 1631–1641
8. S. Särkkä, Bayesian Filtering and Smoothing, *Cambridge University Press*, 2009.
9. E. A. Wan and R. Van Der Merwe The unscented Kalman filter for nonlinear estimation. In Adaptive Systems for Signal Processing, *Communications, and Control Symposium 2000* AS-SPCC, 2000, pp. 153–158
10. Z. Yacine, Observateurs pour l'Estimation de la Dynamique Latérale du véhicule et Application à la Détection de Situations Critiques, *Thèse de Doctorat*, Université Mouloud Mammeri, Tizi Ouzou, 2016.

A New Method For Solving Multi-Objective Multi-Item Solid Transportation Problem With Interval-Valued Trapezoidal Fuzzy Numbers

Sifaoui Thiziri[1], Aïder Méziane[2], and Aidene Mohamed[1]

[1] L2CSP, Fac. Sciences UMM-Tizi-Ouzou, Algeria
[2] LaROMaD, Fac. Maths, USTHB, PB 32, 16111 Bab Ezzouar, Algeria.

Abstract. In real-life situations, the decision maker mostly involves to optimize many conflicting objectives and the parameters are usually under uncertainty. In this paper, we consider the multi-objective multi-item solid transportation problem which is a generalization of the classical transportation problem since in addition to the source and destination constraints it also considers conveyance and different item constraints. To solve its uncertain version, we propose a method based on a new ranking formula by representing the parameters in terms of interval-valued fuzzy numbers. We illustrate this method and analyze it on a numerical example.

Keywords: multi-objective · interval valued fuzzy numbers · transportation problem.

1 Introduction

The transportation problem (TP) is a classical mathematical programming problem in operational research developed by Hitchcock [6]. It consists to transport a homogeneous product from m sources to n destinations by one mode of conveyance. The solid transportation problem (STP) appears as a continuation of TP with three-dimensional properties, modes of conveyance are taken into account in the objective and constraint set, instead of two-dimensional properties. When considering different items to transport we talk about multi-item solid transportation problem.

Most of the real-life application is often modeled as a multi-objective optimization problem and the objectives are measured in different scales and are in conflict at the same time, it is a difficult task to find a solution that simultaneously optimizes all the objectives under the same restrictions. The optimal decisions need to be taken in the presence of trade-offs with two or more conflicting objectives. Bit et al. [1] used a fuzzy linear programming approach for solving multi-objective STP. Moreover, the associated parameters for such problems may be imprecise because of insufficient or inexact information due to incompleteness or lack of evidence, statistical analysis, etc. Kundu et al. [7] studied a Multi-objective multi-item solid transportation problem in which the parameters are taken as trapezoidal fuzzy numbers.

Dalman et al. [3] presented a multi-objective multi-item STP with the parameters considered as trapezoidal fuzzy. In contrast, there are a few papers that deal with the problems involving interval-valued fuzzy numbers. Chiang [2] illustrated that it is better to represent the availabilities and demands as (γ, δ) interval-valued triangular fuzzy numbers instead of normal fuzzy numbers. Gupta and Kumar [5] pointed out the shortcomings of the Chiang´s method and to overcome them, presented a new method for finding the solution of a linear multi-objective transportation problem by representing the value of cost, supply and demand as (γ, δ) interval-valued fuzzy numbers. Ebrahimnejad [4] proposed a new method based on fuzzy linear programming approach for solving a single transportation problem with interval-valued trapezoidal fuzzy numbers.

In this paper, we propose a new method based on a new ranking formula for solving multi-objective multi-item solid transportation problem by representing the parameters (γ, δ) in terms of interval-valued fuzzy numbers. This paper is organized as follows: In Section 2, some basic definitions and arithmetic operations are reviewed. Section 3 presents a multi-objective multi-item solid transportation problem in terms of (γ, δ) interval-valued fuzzy numbers. Section 4, develops the proposed method. Section 5, gives an illustration of the proposed method. In Section 6, we present our conclusions.

2 Preliminaries

In this section, we review some basic definitions, the arithmetic operations and the comparison of interval valued fuzzy numbers.

Definition 1. *A level γ-trapezoidal fuzzy number $\widetilde{A}$ or a generalized trapezoidal fuzzy number $\widetilde{A}$, denoted by $\widetilde{A} = (a_1, a_2, a_3, a_4; \gamma)$,$0 < \gamma \leq 1$, is a fuzzy number with the membership function as follows:*

$$\mu_{\widetilde{A}} \begin{cases} \gamma \dfrac{x - a_1}{a_2 - a_1}, & a_1 \leq x \leq a_2, \\ \gamma, & a_2 \leq x \leq a_3, \\ \gamma \dfrac{a_4 - x}{a_4 - a_3}, & a_3 \leq x \leq a_4, \\ 0 & \textit{otherwise.} \end{cases}$$

Let $F_{TN}(\gamma)$ be the family of all level γ-trapezoidal fuzzy numbers, that is:

$$F_{TN}(\gamma) = \{\widetilde{A} = (a_1, a_2, a_3, a_4; \gamma), a_1 \leq a_2 \leq a_3 \leq a_4\}, \;\; 0 < \gamma \leq 1.$$

Definition 2. *Let $\widetilde{A}^L \in F_{TN}(\gamma)$ and $\widetilde{A}^U \in F_{TN}(\delta)$. A level (γ, δ)-interval-valued trapezoidal fuzzy number $\widetilde{\widetilde{A}}$, denoted by $\widetilde{\widetilde{A}} = [\widetilde{A}^L, \widetilde{A}^U] = \langle (a_1^L, a_2^L, a_3^L, a_4^L; \gamma), (a_1^U, a_2^U, a_3^U, a_4^U; \delta) \rangle$ is an interval-valued fuzzy set on $\mathbb{R}$ with the lower trapezoidal*

fuzzy number $\widetilde{A}^L$ *expressed by:*

$$\mu_{\widetilde{A}^L}\begin{cases}\gamma\dfrac{x-a_1^L}{a_2^L-a_1^L}, & a_1^L\le x\le a_2^L,\\ \gamma, & a_2^L\le x\le a_3^L,\\ \gamma\dfrac{a_4^L-x}{a_4^L-a_3^L}, & a_3^L\le x\le a_4^L,\\ 0 & \textit{otherwise.}\end{cases}$$

and the upper trapezoidal fuzzy number $\widetilde{A}^U$ *expressed by:*

$$\mu_{\widetilde{A}^U}\begin{cases}\delta\dfrac{x-a_1^U}{a_2^U-a_1^U}, & a_1^U\le x\le a_2^U,\\ \delta, & a_2^U\le x\le a_3^U,\\ \delta\dfrac{a_4^U-x}{a_4^U-a_3^U}, & a_3^U\le x\le a_4^U,\\ 0 & \textit{otherwise.}\end{cases}$$

where $a_1^L\le a_2^L\le a_3^L\le a_4^L$, $a_1^U\le a_2^U\le a_3^U\le a_4^U$, $0<\gamma\le\delta\le 1, a_1^U\le a_1^L$, *and* $a_4^L\le a_4^U$. *Moreover,* $\mu_{\widetilde{A}^L}(x)\le\mu_{\widetilde{A}^U}(x)$.

This means that the least and greatest grades of membership of x in the interval $\widetilde{\widetilde{A}}=[\mu_{\widetilde{A}^L}(x),\mu_{\widetilde{A}^U}(x)]$, are $\mu_{\widetilde{A}^L}(x)$ and $\mu_{\widetilde{A}^U}(x)$ respectively.

Let $F_{IVTN}(\gamma,\delta)$ be the family of all level (γ,δ)-interval-valued trapezoidal fuzzy numbers, that is,

$$F_{IVTN}(\gamma,\delta)=\{\ \widetilde{\widetilde{A}}=[\widetilde{A}^L,\widetilde{A}^U]=\langle(a_1^L,a_2^L,a_3^L,a_4^L;\gamma),(a_1^U,a_2^U,a_3^U,a_4^U;\delta)\rangle:\\ \widetilde{A}^L\in F_{TN}(\gamma),\widetilde{A}^U\in F_{TN}(\delta),\quad a_1^U\le a_1^L\le a_4^L\le a_4^U\}\}$$

where $0<\gamma\le\delta\le 1$.

Definition 3. *A triangular fuzzy number denoted by*

$$\widetilde{\widetilde{A}}=[\widetilde{A}^L,\widetilde{A}^U]=\langle(a_1^L,a_2^L,a_3^L;\gamma),(a_1^U,a_2^U,a_3^U;\delta)\rangle$$

is an interval-valued fuzzy set on $\mathbb{R}$ *with the lower triangular fuzzy number* $\widetilde{A}^L$ *expressed by :*

$$\mu_{\widetilde{A}^L}\begin{cases}\gamma\dfrac{x-a_1^L}{a_2^L-a_1^L}, & a_1^L\le x\le a_2^L,\\ \gamma\dfrac{a_3^L-x}{a_3^L-a_2^L}, & a_2^L\le x\le a_3^L,\\ 0 & \textit{otherwise.}\end{cases}$$

and the upper triangular fuzzy number $\widetilde{A}^U$ *expressed by:*

$$\mu_{\widetilde{A}^U}\begin{cases}\delta\dfrac{x-a_1^U}{a_2^U-a_1^U}, & a_1^L\le x\le a_2^U,\\ \delta\dfrac{a_3^U-x}{a_3^U-a_2^L}, & a_2^U\le x\le a_3^U,\\ 0 & \textit{otherwise.}\end{cases}$$

where $a_1^U \leq a_1^L \leq a_2^U \leq a_2^L \leq a_3^L \leq a_3^U$, $0 < \gamma \leq \delta \leq 1, a_1^U \leq a_1^L$, *and* $a_4^L \leq a_4^U$. *Moreover,* $\mu_{\widetilde{A}^L}(x) \leq \mu_{\widetilde{A}^U}(x)$.

Definition 4. *Let* $\widetilde{\widetilde{A}} = [\widetilde{A}^L, \widetilde{A}^U] = \langle (a_1^L, a_2^L, a_3^L, a_4^L; \gamma), (a_1^U, a_2^U, a_3^U, a_4^U; \delta) \rangle$ *and* $\widetilde{\widetilde{B}} = [\widetilde{B}^L, \widetilde{B}^U] = \langle (b_1^L, b_2^L, b_3^L, b_4^L; \gamma), (b_1^U, b_2^U, b_3^U, b_4^U; \delta) \rangle$ *belong to* $F_{IVTN}(\gamma, \delta)$ *and* k *be a non-negative real number. Then the exact formulas for the extended multiplication are defined as follows:*

- $\widetilde{\widetilde{A}} \oplus \widetilde{\widetilde{B}} = \langle (a_1^L + b_1^L, a_2^L + b_2^L, a_3^L + b_3^L, a_4^L + b_4^L; \gamma), (a_1^U + b_1^U, a_2^U + b_2^U, a_3^U + b_3^U, a_4^U + b_4^U; \delta) \rangle$
- $\widetilde{\widetilde{A}} \otimes \widetilde{\widetilde{B}} = \langle (a_1^L b_1^L, a_2^L b_2^L, a_3^L b_3^L, a_4^L b_4^L; \gamma), (a_1^U b_1^U, a_2^U b_2^U, a_3^U b_3^U, a_4^U b_4^U; \delta) \rangle$
- $\langle (ka_1^L, ka_2^L, ka_3^L, ka_4^L; \gamma), (ka_1^U, ka_2^U, ka_3^U, ka_4^U; \delta) \rangle, \quad k > 0$
- $\langle (ka_4^L, ka_3^L, ka_2^L, ka_1^L; \gamma), (ka_4^U, ka_3^U, ka_2^U, ka_1^U; \delta) \rangle, \quad k < 0$
- $\langle (0, 0, 0, 0; \gamma), (0, 0, 0, 0; \delta) \rangle, \quad k = 0.$

Definition 5. *([8]) Consider the* F_{IVTN}

$$\widetilde{\widetilde{A}} = [\widetilde{A}^L, \widetilde{A}^U] = \langle (a_1^L, a_2^L, a_3^L, a_4^L; \gamma), (a_1^U, a_2^U, a_3^U, a_4^U; \delta) \rangle$$

The Centroid of a trapezoidal into three plane figures namely a triangle, a quadrilateral and a triangle respectively. Let G_1, G_2, G_3 *be the Centroids of these three plane figures. The Centroid of these Centroids* G_1, G_2, G_3 *is considered as the point of reference to define the ranking of generalized Interval valued fuzzy numbers. As the Centroid of these three plane figures is their balancing points, the Centroid of these Centroid points is a much better balancing point.*

The Centroids of these plane figures are:

$$G_1 = (\frac{a_1^L + 2a_2^L}{3}, \frac{\gamma}{3}), G_2 = (\frac{a_2^L + a_3^L}{2}, \frac{\gamma}{2}), G_3 = (\frac{2a_3^L + a_4^L}{3}, \frac{\gamma}{3})$$

respectively.
Thus G_1, G_2 and G_3 are not collinear and they form a triangle. Thus the Centroid of these Centroids is

$$G(x_0, y_0) = (\frac{2a_1^L + 7a_2^L + 7a_3^L + 2a_4^L}{18}, \frac{7\gamma}{18}).$$

Now we define:

$$S(\mu_{\widetilde{A}^L}) = x_0.y_0 = (\frac{2a_1^L + 7a_2^L + 7a_3^L + 2a_4^L}{18}, \frac{7\gamma}{18})$$

This is the area between the Centroid of the Centroids and the original point.
Similarly the trapezoidal corresponding to the upper membership function is divided into three plane figures. In similar fashion, the Centroid of the three plane figures and the Centroid of these Centroids is evaluated.

The Centroids of these plane figures are:

$$G_1 = (\frac{a_1^U + 2a_2^U}{3}, \frac{\delta}{3}), G_2 = (\frac{a_2^U + a_3^U}{2}, \frac{\delta}{2}), G_3 = (\frac{2a_3^U + a_4^U}{3}, \frac{\delta}{3}).$$

They are collinear and they form a triangle. Thus the centroid of these Centroids is

$$G(x_0, y_0) = (\frac{2a_1^U + 7a_2^U + 7a_3^U + 2a_4^U}{18}, \frac{7\delta}{18}).$$

Now we define

$$S(\mu_{\widetilde{A}^U}) = x_0.y_0 = (\frac{2a_1^U + 7a_2^U + 7a_3^U + 2a_4^U}{18}, \frac{7\delta}{18}).$$

Using the above definitions, the rank of $\widetilde{\widetilde{A}}$ is defined as follows:

$$R(\widetilde{\widetilde{A}}) = \frac{\gamma S(\mu_{\widetilde{A}^L}) + \delta S(\mu_{\widetilde{A}^U})}{\gamma + \delta}.$$

Definition 6. *([8]) Consider the* F_{IVTN} $\widetilde{\widetilde{A}} = [\widetilde{A}^L, \widetilde{A}^U] = \langle (a_1^L, a_2^L, a_3^L; \gamma), (a_1^U, a_2^U, a_3^U; \delta) \rangle$
The Centroid of a triangle is considered to be the balancing point of the triangle.
The Centroid of the triangle is: $(\frac{a_1^U + a_2^U + a_3^U}{3}, \frac{\delta}{3})$.
Now we define

$$S(\mu_{\widetilde{A}^U}) = x_0.y_0 = (\frac{a_1^U + a_2^U + a_3^U}{3}, \frac{\delta}{3}).$$

This is the area between the Centroid of the Centroid and the original point.
Similarly

$$S(\mu_{\widetilde{A}^L}) = x_0.y_0 = (\frac{a_1^L + a_2^L + a_3^L}{3}, \frac{\gamma}{3}).$$

Using the above definitions, the rank of $\widetilde{\widetilde{A}}$ *is defined as follows:*

$$R(\widetilde{\widetilde{A}}) = \frac{\gamma S(\mu_{\widetilde{A}^L}) + \delta S(\mu_{\widetilde{A}^U})}{\gamma + \delta}.$$

Definition 7. *Let* $\widetilde{\widetilde{A}}$ *and* $\widetilde{\widetilde{B}} \in F_{IVTN}(\gamma, \delta)$*.Then the ranking of level* (γ, δ)*-interval-valued trapezoidal fuzzy numbers in* $F_{IVTN}(\gamma, \delta)$*. is defined as follows:*

$$\widetilde{\widetilde{A}} \prec \widetilde{\widetilde{B}}, \qquad R(\widetilde{\widetilde{A}}) < R(\widetilde{\widetilde{B}}).$$
$$\widetilde{\widetilde{A}} \succ \widetilde{\widetilde{B}}, \qquad R(\widetilde{\widetilde{A}}) > R(\widetilde{\widetilde{B}}).$$
$$\widetilde{\widetilde{A}} = \widetilde{\widetilde{B}}, \qquad R(\widetilde{\widetilde{A}}) = R(\widetilde{\widetilde{B}}).$$

3 Multi-objective multi-item solid transportation problem in terms of (γ, δ) interval valued fuzzy numbers

A multi-objective multi-item solid transportation problem in terms of (γ, δ) interval valued fuzzy numbers is formulated as follows:

$$P\begin{cases} \min Z^o = \sum_{p=1}^{P}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K}\widetilde{\widetilde{C^{po}_{L_{ijk}}}}x^p_{ijk} \\ \sum_{j=1}^{n}\sum_{k=1}^{K}x^p_{ijk} \le \widetilde{\widetilde{A^p_i}},\ \forall\, i,p, \\ \sum_{i=1}^{m}\sum_{k=1}^{K}x^p_{ijk} \ge \widetilde{\widetilde{B^p_j}},\ \forall\, j,p, \\ \sum_{p=1}^{P}\sum_{i=1}^{m}\sum_{j=1}^{n}x^p_{ijk} \le \widetilde{\widetilde{E_k}},\ \forall\, k, \\ x^p_{ijk} \ge 0, \quad \forall\, i,j,k,p, \end{cases}$$

with

- x^p_{ijk}: the unknown quantity to be transported from i^{th} origin to j^{th} destination by k^{th} conveyance of p^{th} item (decision variable);
- $\forall\, i$: means "for $i = 1, \ldots, m$";
- $\forall\, j$: means "for $j = 1, \ldots, n$";
- $\forall\, k$: means "for $k = 1, \ldots, K$";
- $\forall\, p$: means "for $p = 1, \ldots, P$";
- m : number of sources of the transportation problem;
- n : number of destinations;
- K : number of conveyances (modes of transportation);
- P: number of items;
- $\widetilde{\widetilde{A^p_i}}=[\widetilde{A^p_{L_i}}, \widetilde{A^p_{R_i}}]= \langle(\widetilde{a^p_{L_i}}, \gamma), (\widetilde{a^p_{R_i}}, \delta)\rangle$: The amount of product available at i^{th} origin for p^{th} item;
- $\widetilde{\widetilde{B^p_j}}= [\widetilde{B^p_{L_j}}, \widetilde{B^p_{R_j}}]=\langle(\widetilde{b^p_{L_j}}, \gamma), (\widetilde{b^p_{R_j}}, \delta)\rangle$: The demand of product at j^{th} destination for p^{th} item;
- $\widetilde{\widetilde{E_k}} =[\widetilde{E_{L_k}}, \widetilde{E_{R_k}}] =\langle(\widetilde{e_{L_k}}, \gamma), (\widetilde{e_{R_k}}, \delta)\rangle$: The amount of product which can be carried by the k^{th} conveyance;
- $\widetilde{\widetilde{C^{po}_{L_{ijk}}}}=[\widetilde{C^{po}_{L_{ijk}}}, \widetilde{C^{po}_{R_{ijk}}}]= \langle(\widetilde{C^{po}_{L_{ijk}}}, \gamma), (\widetilde{C^{po}_{R_{ijk}}}, \delta)\rangle$: The cost for the transportation problem from i^{th} origin to j^{th} destination by k^{th} conveyance of p^{th} item for o objectives o=1,...,O;

4 Proposed method

Step 1: Solve the multi-objective transportation problem as a single objective transportation problem, taking each time only one objective as objective function and ignoring all others.

Step 2: Calculate:

$$L_o = \min Z_o, \qquad \forall\ o$$
$$U_o = \max Z_o, \qquad \forall\ o$$

Step 3: Define the membership function.

$$\mu_o(Z_o(x)) = \begin{cases} 1, & \text{if } Z_o(x) \le L_o, \\ 1 - \dfrac{Z_o - L_o}{U_o - L_o}, & \text{if } L_o \le Z_o(x) \le U_o, \\ 0, & \text{if } Z_o(x) \ge U_o \end{cases}$$

Step 4: Define the weights for each deviation.

Step 5: Develop the proposed model as follows:

$$\begin{cases} \min \sum_{o=1}^{O}(t_{o1} + t_{o2}) + \sum_{o=1}^{O} t_o \quad , & \forall\ o \\ \sum_{p=1}^{P}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} R(\widetilde{\widetilde{C^{po}_{L_{ijk}}}})x^p_{ijk} - (1 - W_o)t_o = L_o, & \forall\ o \\ \dfrac{U_o - Z_o}{U_o - L_o} - (1 - W_{o1})t_{o1} + (1 - W_{o2})t_{o2} = 1, & \forall\ o \\ \sum_{j=1}^{n}\sum_{k=1}^{K} x^p_{ijk} \le R(\widetilde{\widetilde{A^p_i}}), & \forall\ i, p, \\ \sum_{i=1}^{m}\sum_{k=1}^{K} x^p_{ijk} \ge R(\widetilde{\widetilde{B^p_j}}), & \forall\ j, p, \\ \sum_{p=1}^{P}\sum_{i=1}^{m}\sum_{j=1}^{n} x^p_{ijk} \le R(\widetilde{\widetilde{E_k}}), & \forall\ k, \\ x^p_{ijk} \ge 0, & \forall\ i, j, k, p, \end{cases}$$

where

- t_{o1}, t_{o2}, t_o are deviational variables from the membership functions and the ideal objectives $\forall\ o$.
- $\mu_o(Z^o(x)$ are the membership functions $\forall\ o$.
- L_o are the ideal objectives $\forall\ o$.
- W_o, W_{o1}, W_{o2} are the weights assigned to o membership functions and the ideal objectives.

5 Application example

In order to illustrate the proposed method, let us consider three examples.

5.1 Example 1

Table 1. Data of Example 1

Origin	Destination			Supply
	1	2	3	
1	$\langle(10,20,30,40;\frac{2}{3}),(5,15,35,45;1)\rangle$	$\langle(50,60,70,90;\frac{2}{3}),(45,55,75,95;1)\rangle$	$\langle(80,90,110,120;\frac{2}{3}),(75,85,115,125;1)\rangle$	$\langle(70,90,90,100;\frac{2}{3}),(65,85,95,105;1)\rangle$
2	$\langle(60,70,80,90;\frac{2}{3}),(55,65,85,95;1)\rangle$	$\langle(70,80,100,120;\frac{2}{3}),(65,75,105,125;1)\rangle$	$\langle(20,30,50,60;\frac{2}{3}),(15,25,55,65;1)\rangle$	$\langle(40,60,70,80;\frac{2}{3}),(35,55,75,85;1)\rangle$
Dm	$\langle(30,40,50,70;\frac{2}{3}),(25,35,55,75;1)\rangle$	$\langle(20,30,40,50;\frac{2}{3}),(15,25,45,55;1)\rangle$	$\langle(40,50,50,80;\frac{2}{3}),(35,45,55,85;1)\rangle$	

The following results were obtained by applying the Ebrahimnjad's method to the proposed example:

$$Z = \langle(1700, 3550, 5850, 8950; \frac{2}{3}), (1325, 3350, 6400, 9450; 1)\rangle$$

Applying our method we get

$$Z^* = \langle(843.6, 1189.1, 1669.9, 2106.1; \frac{2}{3}), (670.85, 1016.4, 1842.7, 2278.8; 1)\rangle$$

$$x_{11} = 11.95, x_{12} = 9.07, x_{23} = 13.53$$

By applying the formula of ranking, our method gives a solution better than those found by Ebrahimnjad [4]

$$R(Z) > R(Z^*), \qquad Z \succ Z^*$$

5.2 Example 2

Let us consider the proposed example by Gupta and Kumar [5] as follows (Tables 2-3):

$$a_1 = \langle(6, 7, 12; 0.6), (5, 7, 15; 0.9)\rangle, a_2 = \langle(17, 20, 21; 0.6), (11, 20, 22; 0.9)\rangle$$

$$a_3 = \langle(15, 16, 21; 0.6), (14, 16, 24; 0.9)\rangle$$

Table 2. Data of objective 1

Origin	Destination			
	1	2	3	4
1	$\langle(0.5,1,1.5;0.6),$ $(0.25,1,1.75;0.9)\rangle$	$\langle(0.5,1,5.5;0.6),$ $(0.25,1,7.75;0.9)\rangle$	$\langle(4,6,12;0.6),$ $(2,6,16;0.9)\rangle$	$\langle(6,7,8;0.6),$ $(5,7,9;0.9)\rangle$
2	$\langle(0.4,0.5,2.6;0.6),$ $(0.25,0.5,3.75;0.9)\rangle$	$\langle(6,8,12;0.6),$ $(5,8,18;0.9)\rangle$	$\langle(2,3,4;0.6),$ $(1,3,5;0.9)\rangle$	$\langle(2,3,4;0.6),$ $(1,3,13;0.9)\rangle$
3	$\langle(5,7,13;0.6),$ $(3,7,17;0.9)\rangle$	$\langle(8.5,9,9.5;0.6),$ $(7,9,11;0.9)\rangle$	$\langle(2,3,4;0.6),$ $(1,3,13;0.9)\rangle$	$\langle(5,6,7;0.6),$ $(3,6,9;0.9)\rangle$

Table 3. Data of objective 2

Origin	Destination			
	1	2	3	4
1	$\langle(2,3,4;0.6)$ $(1,3,13;0.9)\rangle$	$\langle(3,4,5;0.6),$ $(2,4,6;0.9)\rangle$	$\langle(2.5,3,3.5;0.6),$ $(1,3,5;0.9)\rangle$	$\langle(1.5,2,4.5;0.6),$ $(1,2,10;0.9)\rangle$
2	$\langle(3,5,7;0.6),$ $(2,5,8;0.9)\rangle$	$\langle(6,7,8;0.6),$ $(5,7,17;0.9)\rangle$	$\langle(7,10,11;0.6),$ $(1,10,12;0.9)\rangle$	$\langle(9,10,11;0.6),$ $(8,10,12;0.9)\rangle$
3	$\langle(4,5,8;0.6),$ $(3,5,14;0.9)\rangle$	$\langle(0.5,1.5,5;0.6),$ $(0.25,1,7.75;0.9)\rangle$	$\langle(3,5,7;0.6),$ $(2,5,8;0.9)\rangle$	$\langle(0.5,1,1.5;0.6),$ $(0.25,1,1.75;0.9)\rangle$

$$b_1 = \langle(10,11,12;0.6),(9,11,13;0.9)\rangle, b_2 = \langle(1.5,2,4.5;0.6),(1,2,10;0.9)\rangle$$

$$b_3 = \langle(13,14,15;0.6),(11,14,17;0.9)\rangle, b_4 = \langle(14,15,20;0.6),(13,15,23;0.9)\rangle$$

The following results were obtained by applying the Gupta and Kumar method on this example [5]:

$$Z^1 = \langle(116.699,152.0250,218.611;0.6),(66.71,152.025,292.23;0.9)\rangle$$

$$Z^2 = \langle(134.765,201.95,245.555;0.6),(43.21,201.95,308.48;0.9)\rangle$$

Applying our method, we get

$$Z^{1^*} = \langle(81.33,105.46,152.742;0.6),(46.81,105.46,203.96;0.9)\rangle$$

$$Z^{2^*} = \langle(94.09,141.190,172.5;0.6),(31.74,141.19,221.16;0.9)\rangle$$

$$Z^{1^*} \prec Z^1 \qquad R(Z^{1^*}) < R(Z^1)$$

$$Z^{2^*} \prec Z^2 \qquad R(Z^{2^*}) < R(Z^2)$$

By applying the formula of ranking, our method gives results that dominate the results found by Gupta and Kumar [5].

5.3 Example 3

A company has two different products to transport from two origins to two destinations using two different conveyances. By taking all parameters as interval-valued trapezoidal fuzzy numbers (Tables 4-7) and applying our method we get the following results:

$$Z^1 = \langle(568.67, 1103.3, 1482.2, 2509; \frac{2}{3}), (287.53, 761.34, 1673, 3184.7; 1)\rangle$$

$$Z^2 = \langle(730.1, 1233.2, 1612.2, 2638.9; \frac{2}{3}), (417.44, 749.41, 1802.9, 3341.7; 1)\rangle$$

$$x^1_{111} = 11.95, x^1_{122} = 9.07, x^2_{111} = 13.53, x^2_{222} = 15.75$$

There is not a method for solving Multi-Objective Multi-Item Solid Transportation Problem With Interval-Value.

Table 4. costs C^{11}_{ijk}.

i	j		j	
	1	2	1	2
1	$\langle(10, 20, 30, 40; \frac{2}{3}),$ $(5, 15, 35, 45; 1)\rangle$	$\langle(50, 60, 70, 90; \frac{2}{3}),$ $(45, 55, 75, 95; 1)\rangle$	$\langle(25, 30, 40, 50; \frac{2}{3}),$ $(25, 35, 45, 75; 1)\rangle$	$\langle(15, 25, 30, 60; \frac{2}{3}),$ $(5, 15, 35, 80; 1)\rangle$
2	$\langle(60, 70, 80, 90; \frac{2}{3}),$ $(55, 65, 85, 95; 1)\rangle$	$\langle(60, 80, 90, 100; \frac{2}{3}),$ $(45, 65, 75, 105; 1)\rangle$	$\langle(25, 30, 40, 50; \frac{2}{3}),$ $(25, 35, 45, 75; 1)\rangle$	$\langle(15, 25, 30, 60; \frac{2}{3}),$ $(5, 15, 35, 80; 1)\rangle$
k	1		2	

Table 5. costs C^{12}_{ijk}.

i	j		j	
	1	2	1	2
1 1	$\langle(8, 18, 28, 40; \frac{2}{3}),$ $(3, 12, 32, 43; 1)\rangle$	$\langle(45, 55, 65, 85; \frac{2}{3}),$ $(40, 55, 75, 95; 1)\rangle$	$\langle(25, 30, 40, 50; \frac{2}{3}),$ $(20, 35, 45, 75; 1)\rangle$	$\langle(15, 25, 30, 60; \frac{2}{3}),$ $(10, 25, 35, 85; 1)\rangle$
2	$\langle(60, 70, 80, 90; \frac{2}{3}),$ $(55, 65, 85, 95; 1)\rangle$	$\langle(55, 80, 90, 100; \frac{2}{3}),$ $(45, 65, 75, 105; 1)\rangle$	$\langle(25, 30, 40, 50; \frac{2}{3}),$ $(15, 35, 45, 75; 1)\rangle$	$\langle(13, 25, 30, 60; \frac{2}{3}),$ $9, 18, 32, 85; 1)\rangle$
k	1		2	

$$a^1_1 = \langle(70, 90, 90, 100; \frac{2}{3}), (65, 85, 95, 105; 1)\rangle,$$

$$a^1_2 = \langle(72, 92, 97, 100; \frac{2}{3}), (64, 85, 92, 105; 1)\rangle$$

$$a^2_1 = \langle(40, 60, 70, 80; \frac{2}{3}), (35, 55, 75, 85; 1)\rangle,$$

Table 6. costs C^{21}_{ijk}.

i	j		j	
	1	2	1	2
1	$\langle((12,22,32,42;\frac{2}{3}),$ $(7,17,37,47;1)\rangle$	$\langle((52,62,72,92;\frac{2}{3}),$ $(47,57,77,97;1)\rangle$	$\langle((27,32,42,52;\frac{2}{3}),$ $(27,37,47,77;1)\rangle$	$\langle((17,27,32,62;\frac{2}{3}),$ $(7,17,37,82;1)\rangle$
2	$\langle((62,72,82,92;\frac{2}{3}),$	$\langle((62,82,92,102;\frac{2}{3}),$	$\langle((27,32,42,52;\frac{2}{3}),$	$\langle((17,27,32,62;\frac{2}{3}),$
2	$(57,67,87,97;1)\rangle$	$(47,67,77,107;1)\rangle$	$(27,37,47,77;1)\rangle$	$(7,17,37,82;1)\rangle$
k	1		2	

Table 7. Costs C^{22}_{ijk}.

i	j		j	
	1	2	1	2
1	$\langle((11,21,31,43;\frac{2}{3}),$ $(6,15,35,48;1)\rangle$	$\langle((48,58,68,88;\frac{2}{3}),$ $(43,58,78,98;1)\rangle$	$\langle((28,33,43,53;\frac{2}{3}),$ $(23,38,48,78;1)\rangle$	$\langle((18,28,33,63;\frac{2}{3}),$ $(13,28,38,88;1)\rangle$
2	$\langle((63,73,83,93;\frac{2}{3}),$	$\langle((58,83,93,103;\frac{2}{3}),$	$\langle((28,33,43,53;\frac{2}{3}),$	$\langle((18,28,33,63;\frac{2}{3}),$
2	$(58,68,88,98;1)\rangle$	$(48,68,78,108;1)\rangle$	$(18,38,48,78;1)\rangle$	$(12,12,35,88;1)\rangle$
k	1		2	

$$a^2_2 = \langle(42,62,72,82;\frac{2}{3}),(35,57,75,88;1)\rangle$$

$$b^1_1 = \langle(30,40,50,70;\frac{2}{3}),(25,35,55,75;1)\rangle,$$

$$b^1_2 = \langle(20,30,40,50;\frac{2}{3}),(15,25,45,55;1)\rangle$$

$$b^2_1 = \langle(40,50,50,80;\frac{2}{3}),(35,45,55,85;1)\rangle,$$

$$b^2_2 = \langle(45,52,52,82;\frac{2}{3}),(35,57,75,88;1)\rangle$$

$$e_1 = \langle(146,149,151,159;\frac{2}{3}),(140,150,160,190;1)\rangle,$$

$$e_2 = \langle(145,152,105,182;\frac{2}{3}),(130,159,168,220;1)\rangle$$

5.4 Results and discussion

- It is obvious from both Example 1 and Example 2 that our method gives better results than the existing method.
- From all Example 1, Example 2 and Example 3, it is easy to see that our method can treat all kinds of transportation problems.
- In Example 3, we considered other not yet studied variants of transportation problems by considering different items and conveyances.

6 Conclusion

In our study, we investigate in multi-objective multi-item solid transportation problem in terms of (γ, δ) interval-valued fuzzy numbers in which we consider the unit transportation costs, supplies at origins, demands at destinations and conveyances capacities are expressed as (γ, δ) interval-valued fuzzy numbers. We have proposed a new method based on a new ranking formula. Our results can be extended to apply in the real-life application.

References

1. Bit, A.K., Biswal, M.P., Alam, S.S.: *Fuzzy Programming approach to multi-objective solid transportation Problem*, Fuzzy Sets Syst, 57(2), 183-194 (1993).
2. Chiang, J.: *The optimal solution of the transportation problem with fuzzy demand and fuzzy product*, J. Inf. Sci. Eng, 21, 439451 (2005).
3. Dalman, H., Güzel, N., Sivri, M.: *A Fuzzy Set-Based Approach to Multi-Objective Multi-item Solid Transportation Problem Under uncertainty* , Int. J. Fuzzy Syst., 18(4), 716-729 (2016).
4. Ebrahimnejad, A.: *Fuzzy linear programming approach for solving transportation problems with interval-valued trapezoidal fuzzy numbers*, Sadhana, 41(3), 299316 (2016).
5. Gupta, A., and Kumar, A.: *A new method for solving linear multi-objective transportation problems with fuzzy parameters*, Appl. Math. Model., 36, 14211430 (2012).
6. Hitchcock, F.L.: *The distribution of a product from several sources to numerous loclities*, Journal of Mathematical Physics, 20(1-4), 224-230 (1941).
7. Kundu, P., Kar, S., Maiti, M.: *Multi-objective multi-item solid transportation problem in fuzzy environment*, Applied Mathematical Modeling, 37(4), 2028-2038 (2013).
8. Shunmugapriya, S., Uthra, G.,: *Ranking Interval valued Fuzzy Numbers*, International Journal of Scientific Research and Reviews, 7(4), 1232-1242 (2018).

Dynamic Optimization Based on the Variational Iteration Method for Predictive Control

Rima Terkmani[0000−0001−6610−8864], Ahmed Maidi[0000−0002−0987−5896], Saïd Guermah[0000−0002−1414−3710], and Mohamed Aidene[0000−0001−6248−4838]

Laboratoire de Conception et Conduite des Systèmes de Production
Université Mouloud Mammeri de Tizi-Ouzou, 15 000 Tizi-Ouzou, Algérie
ahmed.maidi@gmail.com

Abstract. In this paper, a new approach is proposed for predictive control of dynamical systems using the minimum principle. The idea consists in solving using the variational iteration method the optimality conditions over a prediction horizon. These conditions are solved iteratively using a correction functional that yields both the optimal state and costate. Then by imposing, at each sampling time as the boundary conditions the current state and the desired reference, a system of algebraic equations is obtained. The solution of these equations allows to define the piecewise optimal control to be applied at the current sampling time. The proposed approach is illustrated by an application example, which shows its effectiveness.

Keywords: Optimal control · Model predictive control · Minimum principle · Variational iteration method.

1 Introduction

Model predictive control (MPC) has been developed at the end of 1970's and since then several successful practical applications are reported in the literature [14, 12, 19]. MPC is a very effective control approach that optimizes the performances and handles the process constraints with ease [4, 5].

In predictive control, an optimal control problem (OCP) is solved over a prediction horizon with moving initial time and moving terminal time [4]. Carrying out the solution of an optimal control reduces to solve the optimality conditions, that is, a set of ordinary differential equations with appropriate boundary conditions [11, 21]. Thus, considerable efforts are spent on developing efficient techniques that solve the optimality conditions [3, 6]. Analytically solution can be obtained in some rare cases, that is, for fairly simple optimal control problems. In general, the use of appropriate numerical methods is needed[3, 24, 6]. These methods can be categorized into two classes [21]: indirect and direct methods. The indirect methods (Euler-Lagrange equation, Minimum principle, Differential Ricatti equation and Hamilton-Jacobi-Bellman equation) consists in reducing the optimal control problem to a boundary value problem (BVP) and they are based on the variational calculus theory [11, 15]. In direct methods the optimal

is transformed to a nonlinear programming problem [3, 24, 6] and an optimization algorithm is used to determine the solution [3]. Popularity of the direct approaches can be argued by the availability of efficient optimization algorithms that can handle different types of optimization problems [16]. Nevertheless, compared to indirect method, direct methods are less attractive because seldom they produce sub-optimal or approximate solutions and need fast optimization algorithms for on-line implementation. In addition, the performances depend on the quality of the optimum achieved (local or global optimum).

Among the iterative methods developed for solving initial or boundary value problems (differential equations), the variational iteration method (VIM) [9, 23, 17] is widely used for solving different kinds of differential equations. The solution is carried out using a correctional functional [9, 25]. The method can yield approximate analytical or numerical solutions of differential equations with a desired accuracy [25]. Recently, the VIM has been successfully applied to solve optimal control problems [2, 13, 1].

In this paper, an indirect approach is proposed to solve a predictive control problem based on the VIM. The idea consists in deriving an approximate solution of the Hamilton-Pontryagin equations (optimality conditions) using variational iteration method over a prediction horizon. Then, by imposing the boundary conditions, a system of linear algebraic equations is obtained. The solution of these algebraic equations provides the sequence of the optimal controls over the prediction horizon and only the first control is applied to the system to be controlled. This process is then repeated at each sampling time.

The paper is organized as follows: Section 2 gives the principle of the predictive control. Section 3 is devoted to Pontryagin's minimum principle used for solving optimal control problems. The VIM is presented in Section 4. Section 5 presents the proposed approach for solving a predictive control problem. In Section 6, an application example is given to demonstrate the effectiveness of proposed approach. The paper ends with a conclusion.

2 Predictive Control

The principle of predictive control or receding control consists in solving an optimal control over a prediction horizon [4, 20]. Thus, at each sampling time, an optimal control sequence is carried out by solving an optimization problem but only the first control of the sequence is actually applied to the system to be controlled. Then, by moving both control and prediction horizons forward, the process prediction and optimization is repeated. The whole procedure to be repeated at each sampling time can be summarized as follows:

1. Get the current measurement of the controlled variable,
2. Solve the optimal control problem, using either direct or indirect methods, over a prediction control to obtain the optimal control sequence,
3. Apply the first control of the sequence to the system.

Model predictive control problem consists in solving in real time the following optimal control problem [4, 5]:

$$\min_{u(t)} J(u(t)) = \int_0^{\infty} (x^d(t) - x(t))^T \, Q \, (x^d(t) - x(t)) + u(t)^T \, R \, u(t) \, dt \qquad (1)$$

subject to:

$$\dot{x}(t) = f(x(t), u(t), d(t), t) \qquad (2)$$

$$x(0) = x_0 \qquad (3)$$

In the formulated optimal control problem (1)–(3), $t \in \mathbb{R}^+$ is the time variable, $x(t) \in \mathbb{R}^n$ is the state, $x^d(t) \in \mathbb{R}^n$ is the desired reference trajectory for the state $x(t)$, $u(t) \in \mathbb{R}^m$ is the control variable, $d(t) \in \mathbb{R}^r$ is the external disturbance, $x_0 \in \mathbb{R}^n$ is the initial state. $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{m \times m}$ are weighting matrices assumed to be positive definite. $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \times \mathbb{R}^+$ is a continuous vector-valued function. In the sequel, it is assumed that the dynamical system (2) is controllable.

To solve the optimal control problem (1)–(3) in the framework of predictive control, we propose to determine the optimal control sequence $u^*(t_i)\,(i = k, \ldots, k + T_p)$ by solving the following optimal control problem

$$\min_{u(t)} J(u(t)) = \int_{t_k}^{t_{k+T_p}} (x^d(t) - x(t))^T \, Q \, (x^d(t) - x(t)) + u(t)^T \, R \, u(t) \, dt, \qquad (4)$$

subject to:

$$\dot{x}(t) = f(x(t), u(t), d(t), t), \qquad (5)$$

$$x(t_k) = x_k, \qquad (6)$$

$$x(t_{k+T_p}) = x^d(t_{k+T_p}), \qquad (7)$$

where x_k is the state at sampling time $t_k = k \, \Delta t$ ($k \in \mathbb{N}$ and Δt is the sampling period).

The solution of the optimal control problem (5)–(6) can be achieved using either a direct or an direct method [21, 24]. In this work, the minimum principle [15, 11], classified as an indirect method, is adopted to solve the optimal control problem. The optimal control sequence is obtained by solving a system of ordinary differential equations termed Hamilton-Pontryagin equations. The minimum principle is reviewed in the following section.

3 Minimum Principle

Let us consider the following optimal control problem

$$\min_{u(t)} J(u(t)) = \int_{t_0}^{t_f} \psi(x(t), u(t), t) \, dt, \qquad (8)$$

subject to:

$$\dot{x}(t) = f(x(t), u(t), t) \tag{9}$$

$$x(t_0) = x_0 \tag{10}$$

$$x(t_f) = x_f \tag{11}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ is the control vector, ψ, $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ are scalar and vector-valued, respectively. These functions are assumed differentiable with respect to their arguments.

Using the minimum principle of Pontryagin, the optimal control $u^*(t)$ is obtained by minimizing the Hamilton function given as follows [11, 15]:

$$H(x(t), p(t), u(t), t) = f(x(t), u(t), t) + p^T(t)\, \psi(x(t), u(t), t) \tag{12}$$

with $p(t) = [p_1(t),\, p_2(t), \ldots,\, p_n(t)]^T$ is the co-state vector.

Hence, the expression of the optimal control law $u^*(t)$ is obtained by solving the following algebraic equation

$$\nabla_{u(t)} H(x(t), p(t), u(t), t) = 0, \tag{13}$$

with respect to the control variable $u(t)$. Consequently, the optimal control law $u^*(t)$ will be a function of the state $x(t)$ and the co-state $p(t)$, that is,

$$u^*(t) = \Phi(x(t), p(t), t) \tag{14}$$

that verifies the following condition

$$\nabla^2_{u(t)} H(x(t), p(t), u(t), t) > 0 \tag{15}$$

The optimal trajectories $x(t)$ and $p(t)$ are determined by solving the following Hamilton-Pontryagin equations (optimality conditions) [11, 15]

$$\dot{x}(t) = +\nabla_{p(t)} H^*(x(t), p(t), t) \tag{16}$$

$$\dot{p}(t) = -\nabla_{p(t)} H^*(x(t), p(t), t) \tag{17}$$

Eqs. (16) and (17) are the state and co-state equations, respectively and H^* is the the optimal value of the Hamiltonian (12), i.e.,

$$H^*(x(t), p(t), t) = H(x(t), p(t), u(t), t)|_{u(t)=u^*(t)} \tag{18}$$

The solution of the system of differential equations (16)-(17), with boundary conditions (9)–(11), yields the optimal trajectories $x^*(t)$ and $p^*(t)$. Then, the open-loop optimal control $u^*(t)$ is obtained by replacing these optimal trajectories into the optimal control law (14).

The solution of the two-point boundary value problem (16)–(17) is seldom difficult to be carried out analytically. Thus, numerical methods or semi-analytical methods are used [10, 13, 1]. Among the semi-analytical methods that solves differential equations, the variational iteration method represents an efficient mathematical tool that can provides numerical [10, 22] or approximate analytical [9, 25] solutions with a desired accuracy. The solution is obtained iteratively using a correction functional [9, 25]. The principle of the variational iteration method will be discussed in the next section.

4 Variational Iteration Method (VIM)

Variational Iteration Method (VIM) is a reliable, an accurate and an effective method for solving both analytically and numerically different kinds of differential and integro-differential equations [8, 9, 25, 7].

For the ordinary differential equation:

$$L\,x(t) + N\,x(t) = g(t) \tag{19}$$

where L and N are linear and nonlinear operators, respectively and $g(t)$ is the inhomogeneous term. The VIM provides successive approximations of the solution using the following correction functional [9, 25, 17, 23] :

$$x^{(j+1)}(t) = x^{(j)}(t) + \int_0^t \lambda(\tau)(L\,x^{(j)}(\tau) + N\,\tilde{x}^{(j)}(\tau) - g(\tau))\,d\tau, \tag{20}$$

where $\lambda(t)$ is the Lagrange multiplier to be identified optimally via the variational theory [25] and $\tilde{x}^{(j)}(t)$ is a restricted variation, that is, $\delta\widetilde{x_{(j)}}(t) = 0$.

Thus starting by any selective continuous function $x^{(0)}(t)$, the successive approximations $x^{(j+1)}(t)$ of the solution $x(t)$ can be readily obtained. Consequently, the exact solution [9, 25, 17, 23]

$$x(t) = \lim_{j\to\infty} x^{(j)}(t) \tag{21}$$

Thus assuming a given time interval $[0,\ t]$, an accurate approximate solution $x^{(j)}(t)$ with a threshold ε can be obtained by checking the following condition

$$\|x^{(j)}(t) - x^{(j-1)}(t)\|_{L^2[0,\,t]} \leq \varepsilon \tag{22}$$

hence

$$x(t) \approx x^{(j)}(t), \quad t \in [0,\ t] \tag{23}$$

In Eq. (22), $L^2\left([0,\ t]\right)$ denotes the space of square-integrable functions on $[0,\ t]$ defined as follows [26]

$$L^2([0,\ t]) = \left\{ v(t) : [0,\ t] \to \mathbb{R} \mid \int_0^t v^2(t)\,dt < \infty \right\} \tag{24}$$

equipped with the inner product

$$\langle v(t)\,, w(t)\rangle = \int_0^t v(t)\,w(t)\,dt, \quad v(t)\,, w(t) \in L^2([0,\ t]) \tag{25}$$

and the norm

$$\|v(t)\|_{L^2([0,\,t])} = \sqrt{\int_0^t v^2(t)\,dt}, \quad v(t) \in L^2([0,\ t]) \tag{26}$$

5 Predictive Control Based on VIM

Optimization play a key role in predictive control to determine the sequence of the optimal controls $u^*(t_i)$ $(i = k, \ldots, k + T_p)$. In this section, it is proposed to solve the optimal control problem (4)–(6) using the minimum principle (indirect method), that is, to solve the optimality conditions (16)–(17) using VIM, that is, the correction functional (20).

Predictive control needs fast optimization algorithms for real time implementation [18, 5]. In the proposed approach, instead of solving an optimization problem, a system of linear algebraic equations is solved at each sampling time t_k, which is easy to achieve compared to an optimization problem. The system of algebraic equations is obtained off-line by solving the optimality conditions (16)–(17), with the boundary conditions $x(t_k) = x_k$ and $x(t_{k+T_c}) = x^d(t_{k+T_c})$, using the correction functional (20).

As it is mentioned in Section 4, assuming starting by $x^{(0)}(t) = x_k$ and $p^{(0)}(t) = p_k$, an approximate analytical solution of (16)–(17) can obtained under the following form

$$x(t) \approx x^{(j)}(t) = f_x(x_k, p_k, t) \tag{27}$$

$$p(t) \approx p^{(j)}(t) = f_p(x_k, p_k, t) \tag{28}$$

where f_x and f_p are vector-valued functions. Note that the variable x_k is available and represents the current measurement while $p_k \in \mathbb{N}$ is unknown vector of parameters to be determined as explained below. Note that the approximate solutions $x^{(j)}(t)$ and $p^{(j)}(t)$ can be obtained off-line using a computer algebra system.

Now, by imposing the boundary condition (7), the following system of algebraic equations follows

$$h_x(x_k, p_k, t_{k+T_c}) = x^d(t_{k+T_c}) \tag{29}$$

and the solution yields the unknown vector p_k. If equation (29) admits several solutions, we keep the solution that yields better optimal performance index (5). In the following this solution is denoted by p_k^* and the corresponding optimal trajectories are $x^*(t) = h_x(x_k, p_k^*, t)$ and $p^*(t) = h_p(x_k, p_k^*, t)$

The receding control approach proposed is summarized as follows:

1. Get the current measurement $x(t_k) = x_k$,
2. Solve the system of algebraic equations (29) to get p_k^*,
3. Calculate the values of the optimal control trajectories $x^*(t)$ and $p^*(t)$ at sampling time t_k, that is, $x^*(t_k)$ and $p^*(t_k)$
4. Calculate the optimal control $u^*(t_k)$ to apply to the dynamical system using equation (14).

These four steps are repeated at each sampling time t_k.

6 Application example

Consider the following predictive control problem

$$\min_{u(t)} J(u(t)) \int_0^{\infty} (x^d(t) - x(t))^2 + u^2(t)\, dt, \tag{30}$$

subject to:

$$\dot{x}(t) = (e^{-2t} + 1)\, x(t) + u(t) + d(t), \tag{31}$$

$$x(0) = 0. \tag{32}$$

Assuming that the disturbance $d(t) = 0$, the minimum principle yields the following optimal control law

$$u^*(t) = -\frac{p(t)}{2} \tag{33}$$

and the co-state $p(t)$ is determined by solving the following optimality conditions (Hamilton-Pontryagin equations):

$$\dot{x}(t) = (e^{-t} + 1)\, x(t) - \frac{p(t)}{2} \tag{34}$$

$$\dot{p}(t) = -(e^{-t} + 1)\, p(t) + 2\,(x^d(t) - x(t)) \tag{35}$$

The proposed approach summarized at the end of Section 5 is implemented to solve the predictive control. Both tracking and regulation performances are evaluated via numerical simulation for three profiles of the desired reference $x^d(t)$ defined as follows:

1. $x^d(t) = 6$,
2. $x^d(t) = t$,
3. $x^d(t) = \sin(t)$.

For all cases, an impulse of magnitude $d(t) = 2.5$ is applied at $t = 3\,\mathrm{s}$ as a disturbance.

The obtained results (Figs. 1–6) show clearly the effectiveness of the proposed approach. It can be seen that the proposed predictive controller ensures the tracking and completely reject the effect of the disturbance. The control moves of the control $u^*(t)$ are acceptable.

7 Conclusion

In this paper, an indirect method based on the variational iteration method is proposed to solve the optimal control problem in a predictive control strategy. In the proposed approach, the optimal control sequence over an horizon control is determined by solving a system of algebraic equations instead of solving an

optimization problem. Thus it is proposed to solve the optimality conditions, derived following the minimum principle, using the variational iteration method. This method provides an approximate solution, with a desired accuracy, of both state and co-state variables. Then by assuming a prediction horizon and imposing the terminal boundary condition a system of algebraic equations is obtained. The solution of this system provides the optimal value of the co-state vector. Then having the measurement and the optimal value of the co-state vectors at the current sampling time, the optimal control to be applied to the system can be determined using the analytical expression of the optimal control derived by minimizing the Hamiltonian.

Both tracking and regulation performances of the proposed approach are illustrated through numerical simulation by an application example.

References

1. Akkouche, A., Maidi, A., Aidene, M.: Optimal control of partial differential equation based on the variational iteration method. Computers and Mathematics with Applications **68**(5), 622–631 (2014)
2. Berkani, S., Manseur, F., Maidi, A.: Optimal control based on the variational iteration method. Computers and Mathematics with Applications **64**(4), 604–610 (2012)
3. Betts, J.T.: Survey of numerical methods for trajectory optimization. Journal of Guidance, Control and Dynamics **21**(2), 193–207 (1998)
4. Camacho, E.F., Alba, C.B.: Model Predictive Control. Springer-Verlag, London (2007)

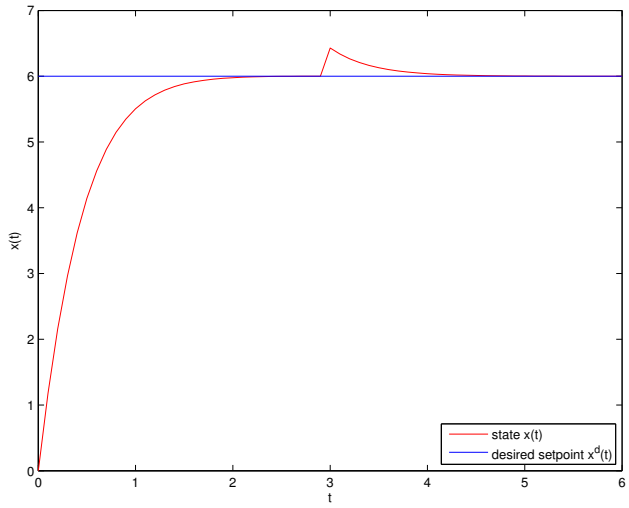


Fig. 1. Case 1: evolution of the optimal state trajectory $x(t)$.

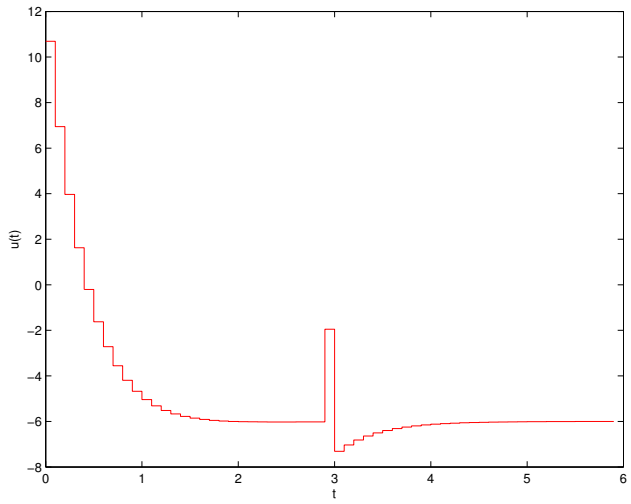


Fig. 2. Case 1. evolution of the optimal control $u^*(t_k)$.

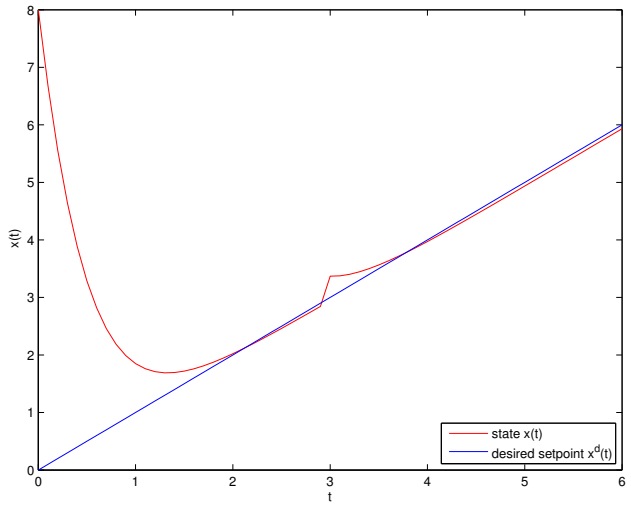


Fig. 3. Case 2. evolution of the optimal state trajectory $x(t)$.

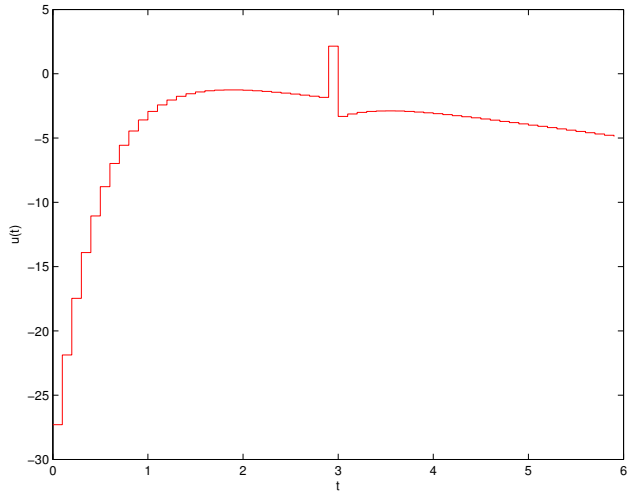


Fig. 4. Case 2. evolution of the optimal control $u^*(t_k)$.

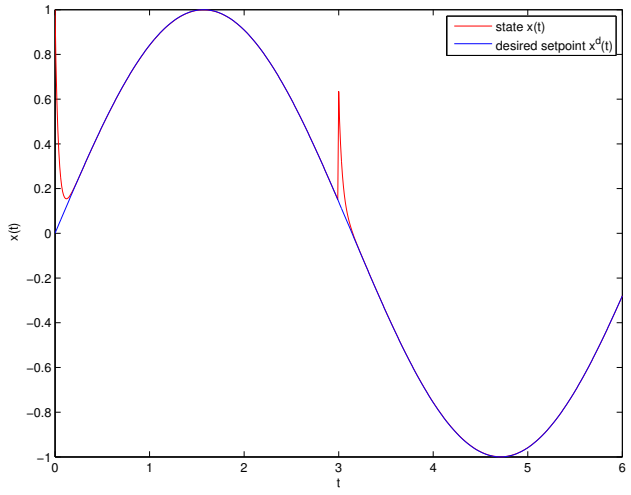


Fig. 5. Case 3. evolution of the optimal state trajectory $x(t)$.

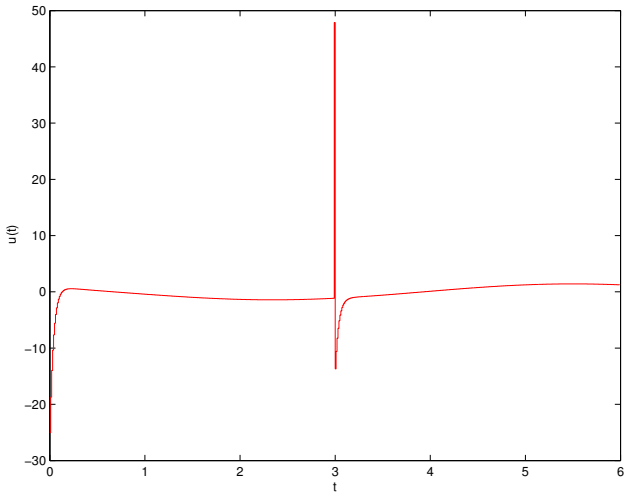


Fig. 6. Case 3. evolution of the optimal control $u^*(t_k)$.

5. Cannon, M.: Efficient nonlinear model predictive control algorithms. Annual Reviews in Control **28**(2), 229–237 (2004)
6. Conway, B.A.: A survey of methods available for the numerical optimization of continuous dynamic systems. Journal of Optimization Theory and Applications **152**(2), 271–306 (2012)
7. Ghaneai, H., Hosseini, M.M.: Variational iteration method with an auxiliary parameter for solving wave-like and heat-like equations in large domains. Computers & Mathematics with Applications **69**(5), 363–373 (2015)
8. He, J.H.: Variational iteration method for delay differential equations. Communications in Nonlinear Science & Numerical Simulation **2**(4), 230–235 (1997)
9. He, J.H.: Variational iteration method a kind of non-linear analytical technique: some examples. International Journal of Non-Linear Mechanics **34**(4), 699–708 (1999)
10. Kierzenka, J., Shampine, L.F.: A BVP solver based on residual control and the MATLAB PSE. ACM Transactions on Mathematical Software **27**(3), 299–316 (2001)
11. Kirk, D.E.: Optimal Control Theory. An Introduction. Prentice-Hall, New Jersey (1970)
12. Lee, J.H.: Model predictive control: Review of the three decades of development. International Journal of Control, Automation and Systems **9**(3), 415–424 (2011)
13. Maidi, A., Corriou, J.P.: Open-loop optimal controller design using variational iteration method. Applied Mathematics and Computation **219**(16), 8632–8645 (2013)
14. Morari, M., Lee, J.: Model predictive control: past, present and future. Computers & Chemical Engineering **23**(4–5), 667–682 (1999)
15. Naidu, D.S.: Optimal Control Systems. CRC Press, Boca Raton (2003)
16. Nocedal, J., Wright, S.: Numerical Optimization. Springer, New York (2006)
17. Odibat, Z.M.: A study on the convergence of variational iteration method. Mathematical and Computer Modelling **51**(9-10), 1181–1192 (2010)

18. Ohtsuka, T., Fujii, H.A.: Real-time optimization algorithm for nonlinear receding horizon control. Automatica **33**(6), 1147–1154 (1997)
19. Qin, S.J., Badgwell, T.A.: A survey of industrial model predictive control technology. Control Engineering Practice **11**(7), 733–764 (2003)
20. Rawlings, J.B.: Tutorial overview of model predictive control. IEEE Control Systems Magazine **20**, 38–52 (2000)
21. Sargent, R.W.H.: Optimal control. Journal of Computational and Applied Mathematics **124**(1–2), 361–371 (2000)
22. Shampine, L.F., Gladwell, I., Thompson, S.: Solving ODEs with MATLAB. Cambridge University Press, Cambridge (2003)
23. Tatari, M., Dehghan, M.: On the convergence of He's variational iteration method. Journal of Computational and Applied Mathematics **207**(1), 121–128 (2007)
24. TrÃ©lat, E.: Optimal control and applications to aerospace: some results and challenges. Journal of Optimization Theory and Applications **154**(3), 713–758 (2012)
25. Wazwaz, A.M.: Partial Differential Equations and Solitary Waves Theory. Springer, Berlin (2009)
26. Zeidler, E.: Applied Functional Analysis. Main Principles and Their Applications. Springer-Verlag, New York (1995)

Allocation Dynamique des Ressources Radio Cognitive Basée sur la Négociation Multi-Agents

Djamila Boukredera[1][0000−0003−4318−7701], Karima Adel-Aissanou[2], Amine Ziane[3], and Chafâa Kherib[3]

[1] Laboratoire des Mathématiques Appliquées, Faculté des Sciences Exactes, Université de Bejaia
boukredera@hotmail.com
[2] Unité de recherche LaMOS (Modélisation et Optimisation des Systèmes), Faculté des Sciences Exactes, Université de Bejaia
ak_adel@yahoo.fr
[3] Faculté des Sciences Exactes, Université de Bejaia

Résumé La radio cognitive (RC) est actuellement l'une des technologies de transmission de l'information les plus prometteuses pour résoudre le problème de la pénurie et de la sous-utilisation du spectre dans les systèmes de communication sans fil. En effet, les technologies se succèdent, les débits sont de plus en plus élevés et les besoins en bandes fréquentielles connaissent une croissance vertigineuse. Ceci a intensifié les risques d'interférences d'une part et la saturation du spectre électromagnétique d'autre part. Des études récentes ont montré que les réseaux radio actuels n'exploitent pas l'intégralité de la bande de fréquence disponible étant donné que la plupart du spectre est allouée et reste inutilisée en l'absence de transmission par les utilisateurs licenciés. Pour une gestion optimale et efficace du spectre de fréquences, la radio cognitive est apparue comme une solution originale et prometteuse et une technologie clé capable de relever ce défi. Le principe de base de la RC est de permettre aux utilisateurs non licenciés d'accéder et d'exploiter, de manière dynamique, les bandes spectrales inoccupées grâce à leur faculté d'écoute et leur capacité d'adaptation à l'environnement. C'est dans cette optique que nous proposons d'aborder le problème de l'allocation des ressources radio par une approche intelligente basée sur les systèmes multi-agents qui convient parfaitement aux spécificités des RCs. A cet effet, nous proposons de résoudre ce problème par un modèle de négociation multi-agents basé sur le principe de médiation et dirigé par un nouveau protocole de négociation qui régit les interactions entre les différents agents du système.

Keywords: Radio cognitive · Systèmes multi-agents · Protocole de négociation · Décision multi-critères.

1 Introduction

Au cours des deux dernières décennies, la prolifération sans cesse croissante des appareils mobiles et le développement rapide d'applications très exigeantes

en bande passante ont fait que les demandes de spectre sont de plus en plus accrues, entraînant ainsi une pénurie grandissante de la ressource spectrale. Ceci est dû principalement à l'utilisation de la politique d'attribution statique des radio-fréquences à des utilisateurs primaires spécifiques titulaires d'une licence pendant une longue période afin d'éviter les interférences et les collisions. De nombreuses études ont alors révélé que dans le cadre de telles stratégies d'accès statique, l'utilité du spectre est très faible car il est extrêmement sous-utilisé. Par exemple, la bande de fréquences du "Global System for Mobile communications" (GSM) est très utilisée alors que les systèmes radio militaires possèdent des bandes qui sont rarement exploitées. Un autre exemple est celui des espaces blancs TV qui représentent les bandes de fréquences de télévision inutilisées à tout moment et en tout lieu. Afin de contourner ces limites de bande passante et de maintenir le développement durable de nouvelles applications sans fil, il est impératif d'explorer des politiques d'attribution du spectre plus efficaces et plus optimales.

La technologie radio cognitive (RC) s'est alors avérée comme une solution viable et une technologie efficace permettant d'apporter des solutions intelligentes et prometteuses au problème de la congestion, la sous-utilisation voire la pénurie du spectre [13,23]. Une RC est une radio intelligente qui permet de scruter son environnement, de détecter les canaux de communication vacants et de les exploiter de manière dynamique et intelligente. L'objectif d'une telle gestion dynamique est de maximiser le taux d'exploitation du spectre radio tout en minimisant les interférences avec les autres utilisateurs [18,20].

Dans les réseaux RC, nous distinguons les utilisateurs primaires (Primary Users, PUs), qui représentent les titulaires de licence du spectre, et les utilisateurs secondaires ou cognitifs (Secondary Users, SUs), qui accèdent dynamiquement au spectre pour se servir des portions inutilisées. Le principe est de permettre à ces utilisateurs sans licence d'accéder temporairement et dynamiquement à la bande passante inutilisée ou peu utilisée tout en assurant qu'ils n'interfèrent pas et ne dégradent pas la performance des titulaires de licence en place. Cela signifie que les SUs sont des entités reconfigurables capables de détecter et d'estimer l'occupation du spectre et d'adapter dynamiquement et de manière autonome leurs paramètres opérationnels et leurs protocoles en exploitant les connaissance acquises afin d'atteindre des objectifs prédéfinis.

Il est à noter que la coexistence des deux types de réseaux primaires et secondaires est très avantageuse pour l'un comme pour l'autre. En effet, les SUs peuvent utiliser les bandes de spectre libres sans avoir besoin d'acheter une licence à long terme ce qui augmente leur profit. Par ailleurs, les PUs peuvent aussi augmenter leur profit en allouant les bandes non utilisées au SUs pour un tarif préétabli. Cependant, plusieurs défis ont émergé suite à cette gestion dynamique du spectre tels que la gestion d'interférences, les problèmes de sécurité et de tarification, la complexité des algorithmes d'allocation et de partage des ressources spectrales, etc [6,5,4].

Dans ce travail, nous nous intéressons particulièrement au problème de l'allocation dynamique des ressources radio. Pour parvenir à une allocation efficace

et dynamique du spectre entre des dispositifs CR hautement distribués, une approche équilibrée, simple et coopérative est nécessaire. Cette problématique a fait l'objet d'une panoplie de travaux dans la literature où plusieurs techniques ont été proposées. Nous distinguons celles qui sont basées sur les enchères [19,26], la théorie des jeux [24], les approches de Markov [9] et les systèmes multi agents (SMA)[7,12,14,25,21]. Il convient cependant de noter que les systèmes multi-agents représentent des candidats aux qualités idéales pour modéliser les réseaux RC. Ceci est dû principalement aux fortes similitudes qui existent entre un agent et un nœud à RC. En effet, les SMA sont des systèmes composés d'un ensemble d'entités autonomes en interaction répondant à des mécanismes de décision individuels et sont à juste titre considérés comme des prototypes pour comprendre le comportement global des sociétés élémentaires formées par des membres intelligents et distribués tels que les réseaux de RC. Ainsi, déployer un agent intelligent sur chaque nœud du reseau RC lui confère la capacité d'interagir avec les autres agents représentant les RC voisines pour former un réseau collaboratif dynamique, faiblement couplé et sans infrastructure.
La question de savoir comment partager les ressources radioélectriques dans des scénarios coopératifs ou compétitifs est un axe de recherche très important pour les chercheurs actuels. C'est dans cette optique que s'inscrit notre travail qui propose une approche intelligente de l'allocation des ressources spectrales basée sur un modèle de négociation multi-agents avec médiation. Notre contribution principale peut être résumée comme suit :
- Nous proposons une architecture multi-agents basée sur trois types d'agents : les agents PUs soumissionnaires d'offres, les agents SUs demandeurs de ressources radio et un agent Médiateur qui représente la pièce maîtresse de l'architecture car c'est à lui qu'incombe de décider de l'allocation dynamique du spectre. Ce dernier est doté d'un processus de décision multi-critères.
- Nous proposons de réguler l'interaction entre les PUs et le Médiateur par un protocole d'interaction simple, par contre nous proposons un protocole de négociation plus complexe entre le Médiateur et les SUs. L'objectif de ce protocole est de permettre aux SUs d'avoir un comportement plus élaboré et plus flexible.
Nous avons utilisé la plateforme JADE (Java Agent Development Framework) [8] pour implémenter le modèle proposé et nous avons également évalué les résultats obtenus pour montrer l'intérêt de notre proposition. La suite de ce papier est organisée comme suit : la section 2 présente quelques travaux antérieurs qui ont utilisé les SMA dans le contexte RC. La section 3 détaille le modèle de négociation multi-agents proposé. L'implémentation et l'evaluation des résultats de simulation sont décrites dans la section 4, puis nous concluons et apportons des perspectives dans la section 5.

2 Revue de Littérature

Les systèmes multi-agents ont déjà fait leurs preuves dans l'allocation et le partage des ressources dans différents domaines tels que les réseaux et les télécommunications. C'est l'un des concepts les plus populaires dans le milieu de la

recherche notamment dans le nouveau domaine de la radio cognitive. Dans ce contexte, plusieurs travaux ont récemment eu recours au paradigme SMA pour une gestion efficace, équitable, optimale et décentralisée des ressources radio partagées entre les PUs et les SUs. Ces travaux peuvent être classés dans trois grandes catégories : les approches basées sur la négociation, sur l'apprentissage et sur la coopération. Dans [1], les auteurs proposent une approche basée sur les SMA coopératifs où les agents ont des intérêts en commun. Ils collaborent en partageant leurs connaissances pour augmenter aussi bien leurs gains individuels que le gain collectif. Dans [11], les auteurs proposent une approche basée sur le protocole Contract Net où les agents SUs sont considérés comme les managers alors que les agents PUs représentent les contractants. Par contre dans [10], les auteurs ont proposé que les agents SUs interagissent et forment plusieurs coalitions sur les bandes spectrales libres afin de fournir un accès au spectre plus coopératif et moins conflictuel. D'autres chercheurs ont préconisé des SMA coopératifs basés sur l'apprentissage par renforcement [27,2,25].
Outre les approches basées sur la coopération et l'apprentissage, les approches basées sur la négociation ont aussi fait l'objet de travaux de recherche récents pour l'allocation du spectre dans les réseaux RC. Dans ce contexte, les PUs et SUs négocient pour atteindre un accord qui répond le mieux à leurs besoins et maximise l'utilité de chacun. Selon le type du protocole utilisé, nous distinguons les approches basées sur la négociation, les heuristiques et le enchères. Dans [16], des techniques issues de la théorie des graphes sont utilisées pour trouver une allocation valide et optimale quand le nombre de canaux est limité. Vu que le problème est NP complet, les auteurs proposent une nouvelle heuristique distribuée qui a donné de bons résultats. Néanmoins, les heuristiques sont rarement utilisées dans les réseaux RC car elles sont complexes à définir. Dans [22], les auteurs ont abordé, en plus du partage du spectre, le problème du changement dynamique du spectre (handoff) pour les SUs mobiles. Ils ont proposé un nouvel algorithme basé sur un mécanisme de négociation pour le partage et le changement de spectre. Dans [15], les auteurs proposent un modèle de négociation du spectre basé principalement sur un agent courtier qui se veut plus intelligent que les agents PUs et SUs. En effet, l'agent courtier centralise quasiment toute l'intelligence à son niveau vu qu'il procède, à la place des SUs, à la détection du spectre et à l'achat des portions spectrales de chez les PUs à un prix déterminé. L'agent courtier revend le spectre acheté aux SUs demandeurs tout en assurant la maximisation de leurs profits. Dans ce travail, nous proposons un modèle de négociation basée sur une intelligence répartie sur l'ensemble des agents composants le système. De ce fait, les agents PUs et SUs négocient l'allocation des canaux et décident de manière autonome d'accepter ou de rejeter les propositions selon leurs fonctions d'utilité. Nous supposons que l'objet de la négociation est multi-critères incluant le prix, la durée de l'allocation et le nombre de canaux demandés. Afin d'éviter l'encombrement du réseau par les messages échangés entre les nombreux PUs et SUs, nous proposons d'inclure un agent Médiateur dont le rôle est de collecter les offres des PUs et les demandes des SUs, d'évaluer les différentes propositions pour envoyer aux SUs celle(s) qu'il aurait jugé

la meilleure pour lui. Le SU a la latitude d'accepter ou de rejeter l'offre émise par le Médiateur.

3 Description du Modèle de Négociation Proposé

Nous considérons un réseau de RC composé de N SUs et M PUs. Chaque PU est propriétaire d'une bande de spectre subdivisée en canaux et chaque canal ne peut être utilisé que par un seul utilisateur. A tout instant t, les PUs soumettent des offres d'allocation des canaux libres aux SUs intéressés. Les SUs, de leur côté, émettent des appels à proposition pour allouer un ou plusieurs canaux pour une transmission de durée limitée. Dans notre modèle, nous proposons d'utiliser un Médiateur entre les PUs et les SUs pour une gestion plus efficace du réseau. Nous proposons une architecture multi-agents pour modéliser le réseau à RC.

3.1 Architecture Multi-Agents Proposée

Comme le montre la Figure 1, notre architecture est basée sur trois types d'agent : Agent PU, Agent SU et Agent Médiateur.
Agent PU : Représente un utilisateur PU qui est fournisseur de ressources.

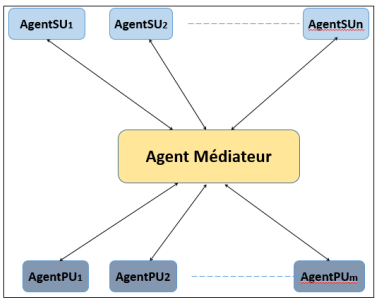


Figure 1. Architecture du modèle de négociation proposé

Il interagit avec l'agent Médiateur en lui soumettant ses offres d'allocation de canaux. Il spécifie dans chaque offre le prix et le temps d'allocation de chaque canal libre. Il met à jour sa liste de canaux alloués, de canaux en attente d'allocation et ceux qu'il se réserve.
Agent SU : Chaque utilisateur SU est représenté par un agent SU. Les agents SUs envoient leurs demandes à l'agent Médiateur et se mettent en attente de propositions. Dans notre modèle, l'agent SU peut accepter ou refuser une proposition si celle-ci ne lui convient pas en terme de prix ou de temps. Par ailleurs, il peut aussi annuler une demande s'il juge qu'il n'est plus opportun d'attendre.
Agent Médiateur : Il représente la pièce maîtresse de notre architecture. C'est

à lui que revient la résolution du problème de l'allocation dynamique et intelligente du spectre. Il interagit avec les agents PUs et les agents SUs afin de parvenir à une allocation du spectre équitable, efficace et qui convient au mieux aux deux types d'agents. Il permet ainsi une collaboration indirecte entre les PUs et les SUs. De part son rôle d'intermédiaire ou d'arbitre, il dispose des informations nécessaires pour décider des meilleures allocations de ressources radio qui satisfont les SUs et maximisent le profit des PUs tout en assurant une exploitation optimale du spectre.
Il convient de noter que le choix d'une négociation avec médiation permet de répondre plus efficacement aux différentes demandes des SUs. En effet, les canaux blancs sont alloués indépendamment des PUs propriétaires, c'est-à-dire qu'un SU peut recevoir une proposition de canaux en provenance de plusieurs PUs distincts. Ainsi, sans interaction directe, des PUs peuvent fournir une solution combinée qu'un PU tout seul ne pourrait pas avoir. Ceci permet plus de flexibilité, de souplesse et un plus grand éventail de solutions pour un meilleur partage du spectre.

3.2 Protocole de Négociation Proposé

Le protocole de négociation proposé représente la deuxième contribution de ce travail. Alors que l'interaction entre l'agent Médiateur et les agents PUs se limite à de simples échanges de messages qui prennent la forme de propositions et de notifications, le protocole de négociation avec les SUs se veut plus complexe et plus élaboré. Notre objectif est de permettre aux SUs d'exhiber un comportement plus flexible en leur permettant d'accepter, de rejeter ou d'annuler une demande auprès du Médiateur. Celui-ci a aussi la possibilité de privilégier, dans la mesure du possible, l'envoi d'une proposition incomplète plutôt qu'un rejet d'une demande. La Figure 2 montre une description semi-formelle de notre protocole ansi que les messages échangés en utilisant le diagramme d'interaction AUML (Agent UML) [3]. Comme illustré par la Figure 2, le protocole est décrit par une série d'échanges de message entre les différents agents. Initialement, l'agent Médiateur réceptionne toutes les offres émises par les agents PUs via le message (Inform) qui définit un ensemble de n canaux libres, le prix d'allocation par unité de temps et la durée d'allocation. De même, chaque agent SU formule sa requête où il exprime ses besoins en canaux libres et le temps d'allocation souhaité qu'il soumet via le message (Cfp) à l'agent Médiateur. Celui-ci procède alors à l'analyse des données dont ils dispose. Il commence par trier les différentes offres des PUs en considérant le prix et le temps d'allocation proposés. Etant donnée que l'offre est caractérisée par plusieurs critères, nous proposons que l'agent Médiateur classe les offres en utilisant l'algorithme TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) [17] en considérant les poids attribués au temps et au prix d'allocation. De la même manière, l'agent Médiateur classe les demandes SUs selon le temps d'allocation demandé. Pour ce faire, nous expérimentons plusieurs méthodes de tri des demandes SUs à savoir FIFO (premier arrivé premier servi) et par ordre croissant ou décroissant du temps

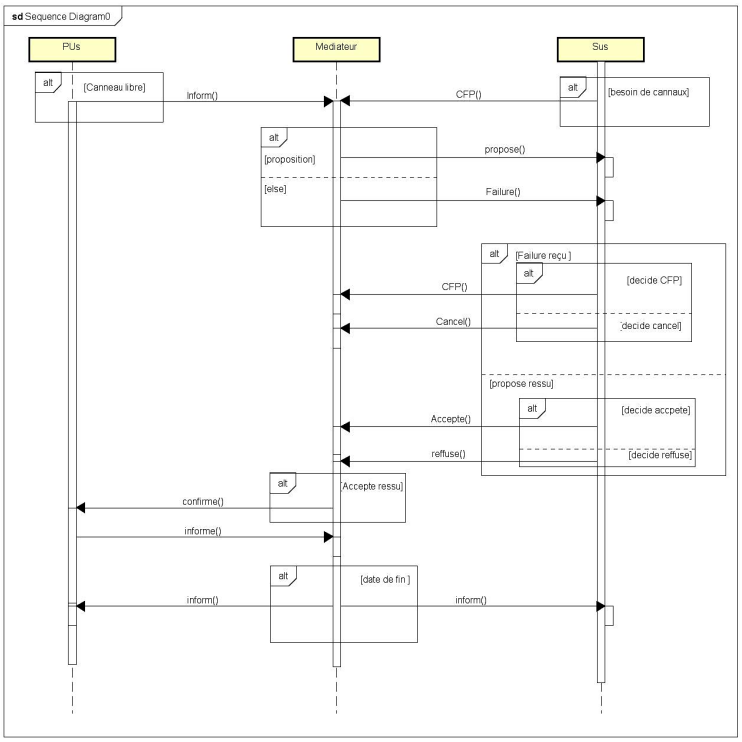


Figure 2. Diagramme d'interaction AUML du protocole proposé

d'allocation demandé. L'impact de ces méthodes de tri, des poids attribués au prix et au temps offert, ainsi que le rapport offre/demande sera mis en exergue lors de l'évaluation des résultats des différentes configurations proposées.
Se basant sur les résultats de l'évaluation des offres et des demandes, le Médiateur exécute la procédure d'allocation puis décide de la meilleure allocation (si elle existe) qui satisferait l'ensemble des SUs. Dans le cas où une demande SU ne peut être satisfaite avec les offres disponibles et dans le souci d'augmenter la satisfaction des SUs, nous proposons de doter l'agent Médiateur avec la capacité de prendre l'initiative de ne pas rejeter une telle demande s'il existe une solution alternative qui se rapproche de la solution complète mais qui ne satisfait qu'en partie la demande en question. Ainsi, Pour les SUs dont les demandes peuvent être satisfaites par une allocation possible (complète ou incomplète), il envoie un message (Propose) qui contient tous les détails de la proposition. Par contre pour les demandes SUs qui ne peuvent être satisfaites, il envoie un message d'échec (Failure). A la réception d'une proposition, l'agent SU peut soit l'accepter ou la refuser notamment s'il s'agit d'une proposition incomplète qui ne lui convient pas. Il informe, par conséquent, le Médiateur de sa décision via les messages (Accept) et (Refuse) respectivement. Par contre, s'il reçoit un message d'échec

Table 1. Les différentes configurations de simulation dans scénario1.

Configuration	Poids du Temps	Poids du Prix	Priorité des demandes SUs
A	0.2	0.8	FIFO
B	0.8	0.2	Tri décroissant du temps demandé
C	0.2	0.8	Tri décroissant du temps demandé
D	0.8	0.2	FIFO

(Failure), il peut soit reformuler sa demande et relancer le processus ou envoyer un message d'annulation (Cancel) au Médiateur.
Après réception d'un message d'acceptation de la part d'un SU, le Médiateur envoie une confirmation (Confirm) aux PUs concernés par les canaux alloués, et enregistre cet accord dans une table de contrats.
Les PUs ayant reçus un message de confirmation (Confirm) mettent à jour l'état des canaux alloués.
Le Médiateur consulte périodiquement la table des contrats, et envoie pour tout contrat expiré un message de terminaison (Inform) aux PUs et SUs concernés pour les aviser que leur contrat a pris fin.

4 Implémentation et Evaluation du Modèle Proposé

Afin de valider l'architecture et le modèle de négociation proposé, nous avons implémenter les différents algorithmes puis nous avons conduit quelques tests et simulations en considérant différents scénarios d'exécution. Pour ce faire, nous avons utilisé la plateforme JADE qui est une plateforme Java de développement et d'exécution des systèmes multi-gents conformes à la norme FIPA. Les simulations que nous avons effectuées portent sur deux scénarios principaux : le cas où l'offre est supérieure à la demande et le cas où la demande dépasse l'offre.
Scénario 1 : offre supérieure à la demande Rappelons qu'une offre et une demande sont caractérisées respectivement par :
offre(prix unitaire (canal) par unité de temps, nombre de canaux à allouer, durée d'allocation) et *demande(nombre de canaux demandés, temps d'allocation désiré)*. Dans ce scénario, nous avons considéré 3 PUs et 3 SUs dont les offres et les demandes sont respectivement décrites comme suit :
- Offre(PU1)=(15,5,12), Offre(PU2)=(10,3,16), Offre(PU1)=(5,4,9).
- Demande(SU1)= (4,6), Demande(SU2)= (2,9) et Demande(SU3)= (2,3). La Table 1 présente les quatre configurations expérimentées où nous varions les poids attribués aux critères du temps et du prix dans l'algorithme TOPSIS.
La Figure 3 montre le déroulement d'une instance de notre protocole avec la configuration décrite dans le scénario 1. Les résultats de simulation obtenus après exécution des 4 configurations sont illustrés dans la Figure 4. Ces résultats représentent le prix payé par chaque agent SU selon chaque configuration. Il est clairement montré que la configuration C donne les meilleurs résultats car elle minimise le prix unitaire payé par les SUs ayant des demandes avec une grande durée d'allocation et par conséquent minimise aussi le prix total payé par ces SUs. Cela veut dire que le Médiateur alloue les meilleurs canaux en termes de

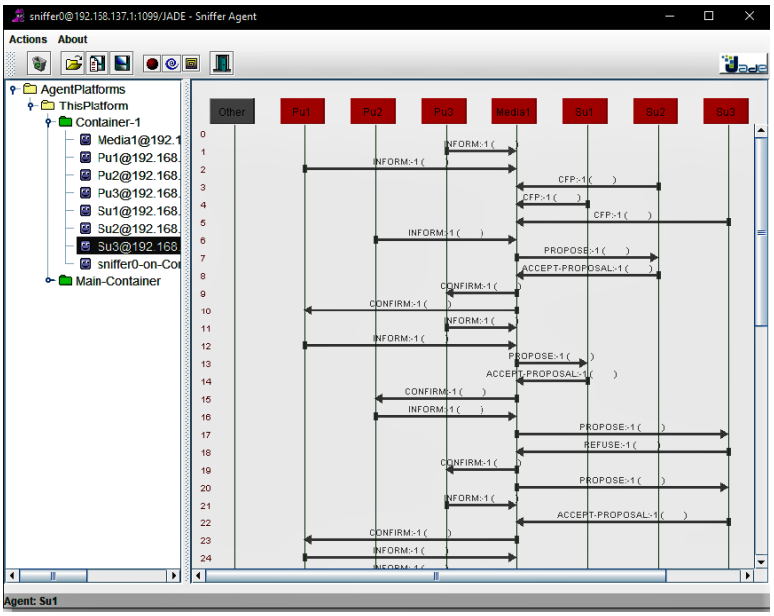


Figure 3. Messages échangés entre les différents agents et illustrés par l'agent Sniffer JADE.

prix aux demandes SUs exprimant une grande durée d'allocation, ce qui permet ainsi de maximiser l'utilisation du spectre libre. Nous concluons donc que dans le cas où l'offre est supérieure à la demande, le Médiateur devrait choisir la configuration C pour optimiser l'exploitation du spectre en favorisant les grandes demandes SUs.

Scénario 2 : offre inférieure à la demande Dans ce scénario, nous allons étudier le taux d'occupatio du spectre. Nous considérons les deux configurations B et D détaillées dans le Table 1. Les offres et les demandes considérées dans ce scénario sont comme suit :

- Offre(PU1)=(15,4,10), Offre(PU2)=(10,3,19), Offre(PU3)=(5,5,15).
- Demande(SU1)= (4,6), Demande(SU2)= (2,9) et Demande(SU3)= (2,3).

Les résultats de simulation obtenus sont illustrés dans les Figures 5(a), 5(b) et 5(c). La Figure 5(a) montre que la Configuration B donne un meilleur taux d'occupation du spectre par rapport à la configuration D. Elle permet de remplir les espaces blancs du spectre en 20 unités de temps, tandis que la configuration D nécessite 29 unités de temps pour remplir ces mêmes espaces. Dans la Figure 5(b), nous remarquons que le nombre de demandes satisfaites est très proche du nombre de canaux libres, ce qui prouve que le spectre est très bien géré. Par contre, la configuration D ne réussit pas ce challenge, comme nous pouvons le remarquer dans la Figure 5(c). En effet, le nombre de canaux libres et le nombre de demandes satisfaites sont plus éloignés d'où une gestion moins bonne

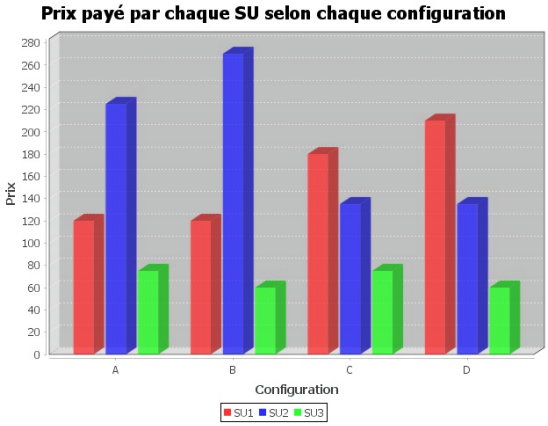


Figure 4. Prix payé par chaque SU selon chaque configuration.

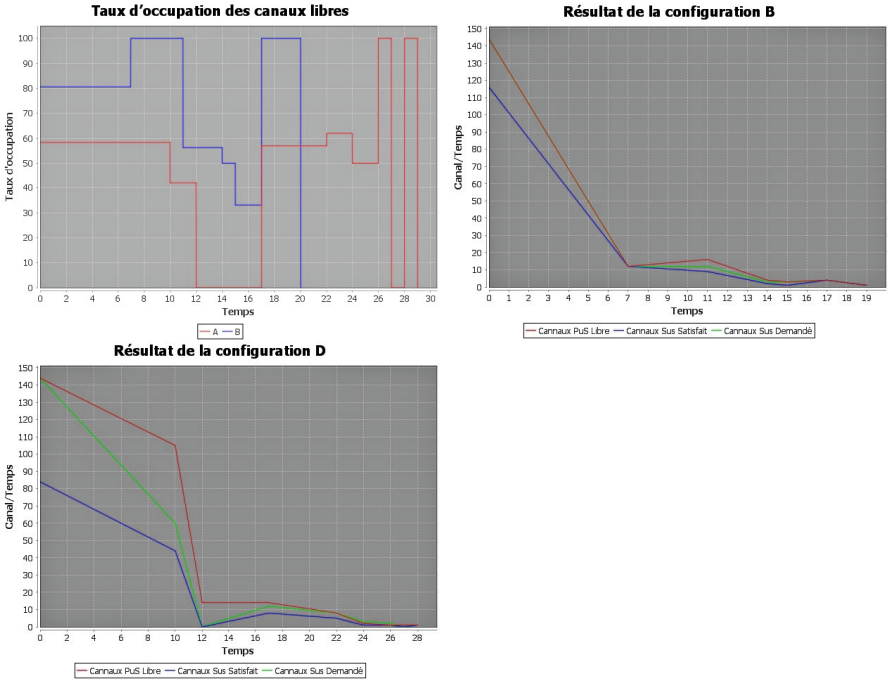


Figure 5. Résultats obtenus avec le scénario 2.

du spectre. Nous concluons donc que dans le cas où la demande est supérieure à l'offre le Médiateur devrait choisir la configuration B pour une meilleure gestion du spectre.

5 Conclusion

Dans ce papier, nous avons proposé un modèle de négociation multi-agents avec médiation pour modéliser un réseau RC. Nous avons proposé une architecture basée sur 3 types d'agents (PUs, SUs et Médiateur) en plus d'un protocole de négociation pour une allocation optimale des ressources spectrales de sorte à maximiser le taux d'utilisation du spectre tout en satisfaisant les utilisateurs secondaires (SUs). Nous avons conduit plusieurs simulations en faisant varier divers paramètres. Les résultats obtenus nous ont permis de déduire quelle est la meilleure configuration que le Médiateur devrait adopter afin d'atteindre les objectifs tracés. Comme perspectives, il serait intéressant de considérer une topologie avec clustering moyennant plusieurs Médiateurs en particulier dans le cas d'un grand nombre d'agents PUs et SUs. Une autre perspective serait de proposer un algorithme de recouvrement en cas de pannes du Médiateur.

Références

1. Ahmed, A., Mubashir Hassan, M., Sohaib, O., Hussain, W., Qasim Khan, M. : An agent based architecture for cognitive spectrum management. Australian Journal of Basic and Applied Sciences (2011)
2. Amraoui, A., Baghli, W., Benmammar, B. : Improving video conferencing application quality for a mobile terminal through cognitive radio. In : 2012 IEEE 14th International Conference on Communication Technology. pp. 1–5. IEEE (2012)
3. AUML : Agent Unified modeling language. `http://www.AUML.org`, retrieved January, 2016
4. Ben Dhaou, A. : Allocation dynamique des bandes spectrales dans les réseaux sans-fil à radio cognitive (2011)
5. Benmammar, B., Amraoui, A. : Réseaux de radio cognitive : Allocation des ressources radio et accés dynamique au spectre. arXiv preprint arXiv :1407.2705 (2014)
6. Das, D., Das, S. : Intelligent resource allocation scheme for the cognitive radio network in the presence of primary user emulation attack. IET Communications **11**(15), 2370–2379 (2017)
7. El-Sisi, A.B., Mousa, H.M. : Argumentation based negotiation in multiagent system. In : 2012 Seventh International Conference on Computer Engineering & Systems (ICCES). pp. 261–266. IEEE (2012)
8. JADE : JADE homepage. `https://jade.tilab.com`, retrieved January, 2018
9. Li, J., Yang, C. : A markovian game-theoretical power control approach in cognitive radio networks : A multi-agent learning perspective. In : 2010 International Conference on Wireless Communications & Signal Processing (WCSP). pp. 1–5. IEEE (2010)
10. Mir, U. : Utilization of cooperative multi-agent systems for spectrum sharing in cognitive radio networks. Ph.D. thesis, Troyes (2011)
11. Mir, U., Merghem-Boulahia, L., Gaïti, D. : Comas : a cooperative multiagent architecture for spectrum sharing. EURASIP Journal on Wireless Communications and Networking **2010**(1), 987691 (2010)

12. Mir, U., Merghem-Boulahia, L., Gaïti, D. : Dynamic spectrum sharing in cognitive radio networks : a solution based on multiagent systems. International Journal on Advances in Telecommunications **3**(3) (2010)
13. Mitola, J. : Cognitive radio : an integrated agent architecture for software defined radio (2000)
14. Pourpeighambar, B., Dehghan, M., Sabaei, M. : Multi-agent learning based routing for delay minimization in cognitive radio networks. Journal of Network and Computer Applications **84**, 82–92 (2017)
15. Qian, L., Ye, F., Gao, L., Gan, X., Chu, T., Tian, X., Wang, X., Guizani, M. : Spectrum trading in cognitive radio networks : An agent-based model under demand uncertainty. IEEE Transactions on Communications **59**(11), 3192–3203 (2011)
16. Rao, V.S., Prasad, R.V., Yadati, C., Niemegeers, I. : Distributed heuristics for allocating spectrum in cr ad hoc networks. In : IEEE Global Telecommunications Conference. pp. 1–6. IEEE (2010)
17. Roszkowska, E. : Multi-criteria decision making models by applying the topsis method to crisp and interval data. Multiple Criteria Decision Making/University of Economics in Katowice **6**, 200–230 (2011)
18. Saleem, Y., Rehmani, M.H. : Primary radio user activity models for cognitive radio networks : A survey. Journal of Network and Computer Applications **43**, 1–16 (2014)
19. Teng, Y., Zhang, Y., Niu, F., Dai, C., Song, M. : Reinforcement learning based auction algorithm for dynamic spectrum access in cognitive radio networks. In : 2010 IEEE 72nd Vehicular Technology Conference-Fall. pp. 1–5. IEEE (2010)
20. Tragos, E.Z., Zeadally, S., Fragkiadakis, A.G., Siris, V.A. : Spectrum assignment in cognitive radio networks : A comprehensive survey. IEEE Communications Surveys & Tutorials **15**(3), 1108–1135 (2013)
21. Trigui, E., Esseghir, M., Boulahia, L.M. : On using multi agent systems in cognitive radio networks : A survey. International Journal of Wireless & Mobile Networks **4**(6), 1 (2012)
22. Trigui, E., Esseghir, M., Merghem-Boulahia, L. : Multi-agent systems negotiation approach for handoff in mobile cognitive radio networks. In : 5th International Conference on New Technologies, Mobility and Security (NTMS). pp. 1–5. IEEE (2012)
23. Wang, B., Liu, K.R. : Advances in cognitive radio networks : A survey. IEEE Journal of selected topics in signal processing **5**(1), 5–23 (2011)
24. Wang, B., Wu, Y., Liu, K.R. : Game theory for cognitive radio networks : An overview. Computer networks **54**(14), 2537–2561 (2010)
25. Wu, C., Chowdhury, K., Di Felice, M., Meleis, W. : Spectrum management of cognitive radio using multi-agent reinforcement learning. In : Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems : Industry track. pp. 1705–1712. International Foundation for Autonomous Agents and Multiagent Systems (2010)
26. Yang, D., Xue, G., Zhang, X. : Group buying spectrum auctions in cognitive radio networks. IEEE Transactions on Vehicular Technology **66**(1), 810–817 (2017)
27. Yau, K.L.A., Komisarczuk, P., Paul, D.T. : Enhancing network performance in distributed cognitive radio networks using single-agent and multi-agent reinforcement learning. In : IEEE Local Computer Network Conference. pp. 152–159 (2010)

An Algorithm for Multiobjective Stochastic Problem Based on DC Programming

Ramzi Kasri and Fatima Bellahcene

LAROMAD, Mouloud Mammeri University, Faculty of Sciences, BP 17 RP, 15000 Tizi-Ouzou, Algeria.
kasriramzi1@gmail.com, bellahcene.fat@gmail.com

Abstract. In this paper we suggest an approach for solving a multiobjective stochastic linear programming problem with normal multivariate distributions. Our solution method is a combination between the multiobjective approach and a non convex technique. The problem is first transformed into a deterministic multiobjective problem introducing the expected value criterion and an utility function that represents the decision makers preferences. The obtained problem is reduced to a single objective quadratic problem using a weighting method. This last problem is solved by DC programming and DC algorithm. A numerical example is included for illustration.

Keywords: Multiobjective programming · Stochastic programming · DCA · DC programming · Utility function · Expected value criterion.

1 Introduction

Multiobjective stochastic linear programming (MOSLP) is a tool for modeling many concrete real-life problems because it is not obvious to have the complete data about problems parameters. Such a class of problems includes investment and energy resources planning [1, 20], production planning in manufacturing systems [7, 8], mineral blending [12], water use planning [2, 5] and multi-product batch plant design [23]. So, to deal with this type of problems it is required to introduce a randomness framework.

In order to obtain the solutions for these multiobjective stochastic problems, it is necessary to combine techniques used in stochastic programming and multiobjective programming. From this, two approaches are considered, both of them involve a double transformation. The difference between the two approaches is the order in which the transformations are carried out. Ben Abdelaziz qualified as *multiobjective approach* the perspective which transforms first, the stochastic multiobjective problem into its equivalent multiobjective deterministic problem, and *stochastic approach* the techniques that transforms in first the stochastic multiobjective problem into a mono-objective stochastic problem [4].

Several interactive methods for solving (MOSLP) problems have been developed. We can mention the Probabilistic Trade-off Development Method or PROTRADE by Goicoechea et al. (1976) [10], The Strange method proposed by

Teghem et al. (1986) [21] and the interactive method with recourse which uses a two stage mathematical programming model by Klein et al. (1990) [11].

In this paper, we propose another approach which is a combination between the *multiobjective approach* and a nonconvex technique (Difference of Convex functions), to solve the multiobjective stochastic linear problem with normal multivariate distributions. The DC programming and DC Algorithm have been introduced by Pham Dinh Tao in 1985 and developed by Le Thi Hoai An and Pham Dinh Tao since 1994 [13–16]. This method has proved its efficiency in a large number of non-convex problems [17–19].

The paper is structured as follows: In section 2, the problem formulation is given. Section 3, shows how to reformulate the problem by introducing utility functions and applying the weighting method. Section 4 presents a review of DC programming and DCA. Section 5 illustrates the application of DC programming and DCA for the resulting quadratic problem. Our, experimental results are presented in the last section.

2 Problem statement

Let us consider the multiobjective stochastic linear programming problem formulated as follows:

$$\begin{array}{c} \min\ (\tilde{c}_1 x, \tilde{c}_2 x, ..., \tilde{c}_q x), \\ s.t.\ \ x \in S, \end{array} \tag{1}$$

where $x = (x_1, x_2, ..., x_n)$ denotes the n-dimensional vector of decision variables. The feasible set S is a subset of n-dimensional real vector space $\mathbb{R}^n$ characterized by a set of linear inequality constraints of the form $Ax \leq b$; where A is an $m \times n$ coefficient matrix and b an m-dimensional column vector. We assume that S is nonempty and compact in $\mathbb{R}^n$. Each vector $\tilde{c}_k$ follows a normal distribution with mean $\overline{c}_k$ and covariance matrix V_k. Therefore, every objective $\tilde{c}_k x$ follows a normal distribution with mean $\mu_k = \bar{c}_k x$ and variance $\sigma_k^2 = x^t V_k x$.

In the following section, we will be mainly interested in the main way to transform problem (1) into an equivalent multiobjective deterministic problem which in turn will be reformulated as a DC programming problem.

3 Transformations and Reformulation

First, we will take into consideration the notion of risk. Assuming that decision maker's preferences can be represented by utility functions, under plausible assumptions about decision maker's risk attitudes, problem (1) is interpreted as:

$$\begin{array}{c} \min\limits_x (E[U(\tilde{c}_1 x)], E[U(\tilde{c}_2 x)], ..., E[U(\tilde{c}_q x)]), \\ s.t.\ \ x \in S. \end{array} \tag{2}$$

The utility function U is generally assumed to be continuous and convex. In this paper, we consider an exponential utility function of the form $U(r) = 1 - e^{-ar}$,

where r is the value of the objective and a the coefficient of incurred risk (a large corresponds to a conservative attitude). Our choice is motivated by the fact that exponential utility functions will lead to an equivalent quadratic problem which encouraged us to design a DC method to solve it simply and accurately. Therefore, if $r \sim N(\mu, \sigma^2)$, we have:

$$E(U(r)) = \int_{-\infty}^{+\infty} (1 - e^{-ar}) \frac{e^{-(r-\mu)^2/2\sigma^2}}{\sqrt{2\pi}} \frac{dr}{\sigma} = 1 - e^{\frac{\sigma^2 a^2}{2} - \mu a}.$$

Minimizing $E(U(r))$ means maximizing $\frac{\sigma^2 a^2}{2} - \mu a$ or minimizing $\mu - \frac{\sigma^2 a}{2}$.

Our aim is to search for efficient solutions of the multiobjective deterministic problem (2) according to the following definition:

Definition 1. *[3] A feasible solution x^* to problem (1) is an efficient solution if there doesn't exist another feasible solution x such that $E[U(\tilde{c}_k x))] \geq E[U(\tilde{c}_k x^*)]$ with at least one strict inequality. The resulting criterion vector $E[U(\tilde{c}_k x^*)]$ is said to be non-dominated.*

Applying the widely used method for finding efficient solutions in multiobjective programming problems, namely the weighting sum method [3, 6], we assign to each objective function in(2) a non-negative weight w_k and aggregate the objectives functions in order to obtain a single function. Thus, problem (2) is reduced to:

$$\begin{array}{ll} \min\limits_{x} & \sum\limits_{k=1}^{q} w_k E[U(\tilde{c}_k x)], \\ s.t. & x \in S, \\ & w_k \in \Lambda \ \ \forall k \in \{1, \ldots, q\}, \end{array} \tag{3}$$

or equivalently

$$\begin{array}{ll} \min\limits_{x} & E[U(\sum\limits_{k=1}^{q} w_k \tilde{c}_k x)], \\ s.t. & x \in S, \\ & w_k \in \Lambda \ \ \forall k \in \{1, \ldots, q\}, \end{array} \tag{4}$$

where $\Lambda = \{w_k : \sum\limits_{k=1}^{q} w_k = 1, w_k \geq 0 \ \ \forall k \in \{1, \ldots, q\}\}$.

Theorem 1. *[9] A point $x^* \in S$ is an efficient solution to problem (2) if and only if $x^* \in S$ is optimal for problem (4).*

Given that the random variable $F(x, \tilde{c}) = \sum\limits_{k=1}^{q} w_k \tilde{c}_k x$ in (4) is a linear function of the random objectives $\tilde{c}_k x$; its variance depends on the variances of $\tilde{c}_k x$ and on their covariances. Since each $\tilde{c}_k x$ follows a normal distribution with mean μ_k and covariance σ_k^2, the function $F(x, \tilde{c})$ follows a normal distribution with mean μ and covariance σ^2 where,

$$\mu = \sum_{k=1}^{q} \mu_k = \sum_{k=1}^{q} w_k \bar{c}_k x, \tag{5}$$

$$\sigma^2 = \sum_{k=1}^{q} w_k^2 \sigma_k^2 + 2 \sum_{k,s=1}^{q} w_k w_s \sigma_{ks}, \tag{6}$$

where σ_{ks} denotes the covariance of the random objectives $\bar{c}_k x$ and $\bar{c}_s x$. Finally, we obtain the following quadratic problem:

$$\min_x \sum_{k=1}^{q} w_k \bar{c}_k^t x - \frac{a}{2} \left(\sum_{k=1}^{q} w_k^2 \sigma_k^2 + 2 \sum_{\substack{k,s=1 \\ k<s}}^{q} w_k w_s \sigma_{ks} \right), \quad s.t. \;\; x \in S, \tag{7}$$

or

$$\min_x \sum_{k=1}^{q} w_k \bar{c}_k^t x - \frac{a}{2} \left(\sum_{k=1}^{q} w_k^2 x^t V_k x + 2 \sum_{\substack{k,s=1 \\ k<s}}^{q} w_k w_s x^t V_{ks} x \right), \quad s.t. \;\; x \in S, \tag{8}$$

where $\bar{c}_k = (\bar{c}_{k1}, \bar{c}_{k2}, ..., \bar{c}_{kn})$ is the k-th component of the expected value of the random multinormal vector $\tilde{c}$, V_{ks} and V_k are elements of the positive definite covariance matrix V of $\tilde{c}$:

$$V = \left(\begin{array}{c|c|c|c|c|c} V_1 & V_{12} & ... & V_{1s} & ... & V_{1q} \\ \hline V_{21} & V_2 & ... & V_{2s} & ... & V_{2q} \\ \hline ... & ... & ... & ... & ... & ... \\ \hline V_{k1} & V_{k2} & ... & V_{ks} & ... & V_{kq} \\ \hline ... & ... & ... & ... & ... & ... \\ \hline V_{q1} & V_{q2} & ... & V_{qs} & ... & V_q \end{array} \right).$$

4 Review of DC programming and DCA

A general DC program has the form:

$$\alpha = inf\{f(x) = g(x) - h(x) : x \in \mathbb{R}^n\}, \tag{9}$$

where $g,\ h$ are lower semicontinuous proper convex functions on $\mathbb{R}^n$ called DC components of the DC function f while $g - h$ is a DC decomposition of f .
The duality in DC associates to problem (9) the following dual program:

$$\alpha = inf\{h^*(y) - g^*(y) : y \in \mathbb{R}^n\}, \tag{10}$$

where g^* and h^* are respectively the conjugate functions of g and h.
The conjugate function of g is defined by:

$$g^*(y) = sup\{\langle x, y\rangle - g(x) : x \in \mathbb{R}^n\}. \tag{11}$$

From [15], the most used necessary optimality conditions for problem (9), is:

$$\emptyset \neq \partial h(x^*) \subset \partial g(x^*), \tag{12}$$

where $\partial h(x^*) = \{y^* \in \mathbb{R}^n : h(x) \geq h(x^*) + < x - x^*, y^* >, \forall x \in \mathbb{R}^n\}$ is the subdifferential of h at x^*.
A point x^* is called critical point of $g - h$ if

$$\emptyset \neq \partial g(x^*) \cap \partial h(x^*). \tag{13}$$

DCA constructs two sequences $\{x^i\}$ and $\{y^i\}$ (candidates for being primal and dual solutions, respectively), such that their corresponding limit points satisfy the local optimality conditions (12) and (13). There are two forms of DCA: the simplified DCA and the complete DCA. In practice, the simplified DCA is most used than the complete DCA because it is less time cosuming [13]. The simplified DCA has the following scheme [13, 18]:

Simplified DCA Algorithm
Step 1 : Let $x^0 \in \mathbb{R}^n$ given. Set $i = 0$.
Step 2 : Calculate $y^i \in \partial h(x^i)$.
Step 3 : Calculate $x^{i+1} \in \partial g^*(y^i)$.
Step 4 : If a convergence criterion is satisfied, then **stop**, else set $i = i + 1$ and **goto** step 2.

We also can note that: [15, 18]

- DCA is a descent method without linesearch.
- If $g(x^{i+1}) - h(x^{i+1}) = g(x^i) - h(x^i)$, then x^i is a critical point of f and y^i is a critical point of $h^* - g^*$.
- DCA has a linear convergence for general DC programs, and has a finite convergence for polyhedral programs.
- If the optimal value of problem (8) is finite and the sequences $\{x^i\}$ and $\{y^i\}$ are bounded then every limit point x (resp. y) of the sequence $\{x^i\}$ (resp. $\{y^i\}$ is a critical point of $g - h$ (resp. $h^* - g^*$)

5 DCA Applied to Problem (8)

The function $f(x) = \min\limits_{x} \sum\limits_{k=1}^{q} w_k \bar{c}_k x - \frac{a}{2} \left(\sum\limits_{k=1}^{q} w_k^2 \sigma_k^2 + 2 \sum\limits_{\substack{k,s=1 \\ k<s}}^{q} w_k w_s \sigma_{ks} \right)$
in problem (8) will be decomposed in order to obtain a DC program of the form:

$$\min\{f(x) = g(x) - h(x) : x \in S\}, \tag{14}$$

with

$$g(x) = \chi_S(x) + \sum_{k=1}^{q} w_k \bar{c}_k^t x,$$

where $\chi_S(.)$ is the indicator function of the set S.

and

$$h(x) = \frac{a}{2}\left(\sum_{k=1}^{q} w_k^2 x^t V_k x + 2\sum_{\substack{k,s=1\\k<s}}^{q} w_k w_s x^t V_{ks} x\right).$$

After that, we will compute the two sequences $\{x^i\}$ and $\{y^i\}$ defined as follows:
$y^i \in \partial h(x^i)$ *and* $x^{i+1} \in \partial g^*(y^i)$.

Computation of y^i:
We choose $y^i \in \partial h(x^i) = \{\nabla h(x^i)\}$.
It is equivalent to calculate:

$$y^i = a\left(\sum_{k=1}^{q} w_k^2 V_k x^i + 2\sum_{\substack{k,s=1\\k<s}}^{q} w_k w_s V_{ks} x^i\right). \tag{15}$$

Computation of x^i:
We can choose $x^{i+1} \in \partial g^*(y^i)$ as the solution of the following convex problem

$$\min\left\{\sum_{k=1}^{q} w_k \bar{c}_k^t x - x^t y^i \ : \ x \in S\right\}. \tag{16}$$

The solution x^i is optimal for the problem (14) if one of the following conditions is verified

$$|(g-h)(x^{i+1}) - (g-h)(x^i)| \leq \epsilon, \tag{17}$$

$$\|(x^{i+1}) - (x^i)\| \leq \epsilon. \tag{18}$$

Finally, the DC Algorithm that we can apply to problem (8) with the decomposition (14) can be described as follows:

Algorithm DCAMOSLP

Step 1 : Initialization: Let $x^0 \in \mathbb{R}^n$, $\epsilon, k, w \in \mathbb{R}^+$,$a > 0$, V, A, b, $\bar{c}$ given. Set $i = 0$.
Step 2 : Calculate $y^i \in \partial h(x^i)$ using (15).
Step 3 : Calculate $x^{i+1} \in \partial g^*(y^i)$, solution of the convex problem (16).
Step 4 : If one of the conditions (17) or (18) is verified, then **stop** x^{i+1} is optimal for (14), else set $i = i + 1$ and **goto** step 2.

---

6 Experimental Results

To demonstrate the performances of our algorithm, two numerical examples will be given in this section. The first is taken from [6] to show the efficiency of the algorithm. The second example is given to present the performances of

DCAMOSLP according to the variation of certain parameters.
Let us consider the following stochastic bi-objective programming problem:

$$\begin{cases} \min\limits_{x} (\tilde{c}_{11}x_1 + \tilde{c}_{12}x_2, \tilde{c}_{21}x_1 + \tilde{c}_{22}x_2), \\ \quad s.t. \;\; x_1 + 2x_2 \geq 4, \\ \quad\quad x_1, x_2 \leq 3, \\ \quad\quad x_1, x_2 \geq 0, \end{cases} \tag{19}$$

with $\tilde{c} = (\tilde{c}_{11}, \tilde{c}_{12}, \tilde{c}_{21}, \tilde{c}_{22})^t$ being a random vector multinormal with expected value $\overline{c} = (0.5, 1, 1, 2.5)^t$ and with positive definite covariance matrix:

$$V = \begin{pmatrix} 25 & 0 & 0 & 3 \\ 0 & 25 & 3 & 0 \\ 0 & 3 & 1 & 0 \\ 3 & 0 & 0 & 9 \end{pmatrix}.$$

For this test, we will take $\epsilon = 10^{-6}$ and $x^0 = (0, 0)$ as initial point. The application of algorithm DCAMOSLP to this problem for different values of the coefficient of incurred risk a and a fixed weight vector $\mu = (0.8, 0.2)^t$ gives the results in Table 1 where *nbr_it* is the number of iterations.

Table 1. Results for different values of parameter a.

a	(x_1^*, x_2^*)	$\overline{c}_1 x^*$	$\overline{c}_2 x^*$	*nbr_it*
10^{-30}	$(3, 0.5)$	2	4.25	2
10^{-20}	$(3, 0.5)$	2	4.25	2
10^{-10}	$(3, 0.5)$	2	4.25	3
10^{-2}	$(3, 0.5)$	2	4.25	3
1	$(3, 3)$	4.5	10.5	5
10	$(3, 3)$	4.5	10.5	5
10^2	$(3, 3)$	4.5	10.5	5

The non dominated solution $(3, 0.5)$ is obtained for values of parameter $a \leq 10^{-2}$. The non dominated solution for $w = (0.8, 0.2)^t$ in Ref. [6] is $(3, 0.5)$. We also note that the number of iterations decreases with the decrease of the parameter a.

Now we will test the performance of the algorithm with a second problem which has a larger set of feasible solutions.

$$\begin{cases} \min\limits_{x} (\tilde{c}_{11}x_1 + \tilde{c}_{12}x_2, \tilde{c}_{21}x_1 + \tilde{c}_{22}x_2), \\ \quad s.t. \;\; 2x_1 + 3x_2 \geq 10, \\ \quad\quad x_1, x_2 \leq 5, \\ \quad\quad x_1, x_2 \geq 0, \end{cases} \tag{20}$$

with $\overline{c} = (6, -5, 3, 8)^t$ and positive definite covariance matrix:

$$V = \begin{pmatrix} 14 & 0 & 0 & 3 \\ 0 & 12 & 3 & 0 \\ 0 & 3 & 2 & 0 \\ 3 & 0 & 0 & 8 \end{pmatrix}.$$

The results of application of algorithm DCAMOSLP to this problem for different values of parameter a and the weight vector w are given in Table 2.

Table 2. Results for different values of a and vector w.

a	w	(x_1^*, x_2^*)	nbr_it
	$(0.2, 0.8)$	$(0.5524, 2.9651)$	2
	$(0.8, 0.2)$	$(0, 5)$	2
10^{-20}	$(0.6, 0.4)$	$(0, 3.3333)$	2
	$(0.5, 0.5)$	$(0, 3.3333)$	2
	$(0.9, 0.1)$	$(0, 5)$	2
	$(0.2, 0.8)$	$(0.5506, 2.9663)$	5
	$(0.8, 0.2)$	$(0, 5)$	2
10^{-10}	$(0.6, 0.4)$	$(0, 3.3333)$	3
	$(0.5, 0.5)$	$(0, 3.3333)$	3
	$(0.9, 0.1)$	$(0, 5)$	2
	$(0.2, 0.8)$	$(0, 3.3333)$	5
	$(0.8, 0.2)$	$(0, 5)$	3
10^{-2}	$(0.6, 0.4)$	$(0, 3.3333)$	4
	$(0.5, 0.5)$	$(0, 3.3333)$	3
	$(0.9, 0.1)$	$(0, 5)$	3
	$(0.2, 0.8)$	$(5, 5)$	5
	$(0.8, 0.2)$	$(5, 5)$	4
10	$(0.6, 0.4)$	$(5, 5)$	4
	$(0.5, 0.5)$	$(5, 5)$	4
	$(0.9, 0.1)$	$(5, 5)$	5
	$(0.2, 0.8)$	$(5, 5)$	4
	$(0.8, 0.2)$	$(5, 5)$	5
10^2	$(0.6, 0.4)$	$(5, 5)$	5
	$(0.5, 0.5)$	$(5, 5)$	5
	$(0.9, 0.1)$	$(5, 5)$	5

We observe from the results that the algorithm DCAMOSLP gives efficient solutions of the multiobjective stochastic problem for small values of the coefficient of incurred risk ($a \leq 10^{-2}$). The number of iterations decreases with the decrease of the parameter a.

7 Conclusion

We have presented a DC optimization approach for solving a multiobjective stochastic problem with multivariate normal distributions in which the objective functions should be minimized. The experimental results show the efficiency of the algorithm. However further experimental validation of this observation and comparison with existing methods is needed. As future works, an algorithm for a stochastic multiobjective maximization problem is planned.

References

1. Alarcon-Rodriguez, A., Ault, G., Galloway, S.: Multiobjective Planning of Distributed Energy Resources Review of the State-of-the-Art. Renewable and Sustainable Energy Reviews **14**(5), 1353-1366 (2010)
2. Ben Abdelaziz, F., Mejri, S.: Application of Goal Programming in a Multi-objective Reservoir Operation Model in Tunisia. European Journal of Operational Research **133**, 352-361 (2001)
3. Ben Abdelaziz, F., Lang, P., Nadeau, R.: Distributional Unanimity in Multiobjective Stochastic Linear Programming. J. Clmaco (ed.), Multicriteria Analysis Springer-Verlag Berlin Heidelberg 1997
4. Ben Abdelaziz, F., L'efficacité en programmation multi-objectifs stochastique. Ph.D. Thesis, Université de Laval, Québec,(1992)
5. Bravo, M., Gonzalez, I.: Applying Stochastic Goal Programming: A Case Study on Water Use Planning. European Journal of Operational Research. **2**(196), 1123-1129 (2009)
6. Caballero, R., Cerd, E., del Mar Muoz, M., and Rey, L.: Stochastic approach versus multiobjective approach for obtaining efficient solutions in stochastic multiobjective programming problems. European Journal of Operational Research, **158**(3), 633-648 (2004)
7. Caner, T.Z., Tamer, U.A.: Tactical Level Planning in Float Glass Manufacturing with Co- Production, Random Y ields and Substitutable Products. European Journal of Operational Research. **199**(1), 252-261 (2009)
8. Fazlollahtabar, H., Mahdavi, I.: Applying Stochastic Programming for Optimizing Production Time and Cost in an Automated Manufacturing System. In: International Conference on Computers & Industrial Engineering, 1226-1230, Troyes 6-9 July (2009)
9. Geoffrion, A.M.: Proper Efficiency and the Theory of Vector Maximimization. Journal of Mathematical Analysis and Applications **22**(3), 618-630 (1968)
10. Goicoechea, A., Dukstein, L., Bulfin, R.T.: Multiobjective Stochastic Programming. the PROTRADE-method. Operation Research Society of America (1976)
11. Klein, G., Moskowitz, H., Ravindran, A.: Interactive multiobjective optimization under uncertainty. Management Science. **36**(1), 58-75 (1990)
12. Kumral, M.: Application of Chance-Constrained Programming Based on Multiobjective simulated Annealing to Solve Mineral Blending Problem. Engineering Optimization. **35**(6), 661-673 (2003)
13. Le Thi, H.A., Pham Dinh, T.: Solving a class of linearly constrained indefinite-quadratic problems by dc algorithms. Journal of Global Optimization **11**(3), 253-285 (1997b)

14. Le Thi, H.A., Pham Dinh, T.: A continuous approach for globally solving linearly constrained quadratic zero-one programming problems. Optimization **50**, 93-120 (2001)
15. Le Thi, H.A., Pham Dinh, T.: The dc (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. Annals of Oper. Res. **133**, 23-46 (2005)
16. Le Thi, H.A., Pham Dinh, T., Huynh, V.N.: Exact penalty and error bounds in DC programming. J. of Glob. Opt. **52**, 509-535 (2012)
17. Le Thi, H.A., Pham Dinh, T., Nguyen, C.N., Nguyen, V.T.: DC programming techniques for solving a class of nonlinear bilevel programs. J. of Glob. Opt. **44**, 313-337 (2009)
18. Pham Dinh, T., Le Thi, H.A.: Convex analysis approach to DC programming: Theory, Algorithms and Applications (dedicated to Professor Hoang Tuy on the occasion of his 70th birthday). Acta Mathematica Vietnamica **22**, 289-355 (1997a)
19. Pham Dinh, T., Nguyen, C.N., Le Thi, H.A.: DC Programming and DCA for Globally Solving the Value-At-Risk, Comput. Manag. Sci. **6**, 477-501 (2009)
20. Teghem, J., Kunsch, P.: Application of Multiobjective Stochastic Linear Programming to Power Systems Planning. Engineering Costs and Production Economics **9**(13), 83-89 (1985)
21. Teghem, J., Dufrane, D., Thauvoye, M. and Kunsch, P.L.: Strange, an Interactive Method for Multiobjective Stochastic Linear Programming under Uncertainty. European Journal of Operational Research **26**(1), 65-82 (1986)
22. Vahidinasab, V., Jadid, S.: Stochastic Multiobjective Self-Scheduling of a Power Producer In Joint Energy & Reserves Markets. Electric Power Systems Research **80**(7), 760-769 (2010)
23. Wang, Z., Jia, X.P., Shi, L.: Optimization of Multi-Product Batch Plant Design under Uncertainty with Environmental Considerations. Clean Technologies and Environmental Policy **12**(3), 273-282 (2009)

Estimation à noyau discret dans le modèle de stock de type (R, s, S)

F. Afroun[1], D. Aïssani[2], and D. Hamadouche[3]

[1] Laboratoire de Mathématiques Pures et Appliquées (LMPA), Université de Tizi-Ouzou, 15000, Algérie. afrounfairouz@gmail.com
[2] Unité de recherche LaMOS (Modélisation et Optimisation des Systèmes), Université de Bejaia, 06000, Algérie. lamos_bejaia@hotmail.com
[3] Laboratoire de Mathématiques Pures et Appliquées (LMPA), Université de Tizi-Ouzou, 15000, Algérie. djhamad@yahoo.fr

Abstract. Dans ce travail, nous considérons l'estimation à noyau discret d'une matrice de transition associée à une chaîne de Markov décrivant un modèle de stock de type (R, s, S). Plus précisément, nous étudions l'effet du choix du paramètre de lissage sur les performances des estimateurs des caractéristiques stationnaires de ce modèle. A base d'exemples numériques, nous avons constaté que l'estimateur du paramètre de lissage choisi, en minimisant une certaine norme matricielle, donnait de meilleurs résultats, en termes de qualité des estimateurs des caractéristiques stationnaires du modèle, que celui choisi en minimisant le ISE.

Keywords: Noyau discret · Paramètre de lissage · Gestion de stock · Simulation · normes.

1 Introduction

Depuis le modèle de Harris, des milliers d'articles ont paru dans le domaine des sciences de gestion des stocks. Certainement, on se demande pourquoi une telle attention a été accordée aux modèles de gestion des stocks. L'explication est simplement qu'en pratique la constitution des stocks ainsi que leur gestion demeurent, dans la vie de toutes les entreprises aussi petites soient-elles et jusqu'à l'individu (le consommateur), incontournable. De plus, on rencontre de nombreuses situations différentes où chacune nécessite une analyse sur mesure.

Pour l'évaluation des mesures de performance d'un système de stock et la mise en place d'une politique optimale pour sa gestion, dans la littérature on trouve principalement deux approches, à savoir : l'approche déterministe et l'approche stochastique. Quoique l'approche déterministe nous fournit des résultats satisfaisants, les modèles stochastiques de gestion des stocks sont les plus réalistes, car ils prennent en considération le comportement incertain de certains paramètres de départ décrivant le système considéré.

Théoriquement pour l'évaluation des performances d'un système d'une manière générale et d'un système de gestion de stock en particulier, on se base sur les différents paramètres de départ le décrivant. Cependant, dans la pratique et en règle

générale, les valeurs des paramètres de départ d'un système ne sont connus que sous forme d'un échantillon de données. Dans ce sens, afin d'évaluer les performances du système considéré le recours aux techniques statistique d'estimation (paramétrique ou/et non paramétrique), qui visent à fournir une approximation pour les valeurs des paramètres (inconnus) en exploitant l'information apportée par l'échantillon, est inévitable.

Dans ce document, nous proposons de considérer un système de gestion de stock de type (R, s, S) modélisé par une chaîne de Markov sous l'hypothèse que la distribution des demandes est une fonction de masse générale et inconnue. En outre, notre objectif est d'estimer la matrice de transition associée à ce modèle qui est d'une grande importance dans son analyse transitoire et stationnaire, et qui nous permet également de déduire la totalité du reste de ses mesures de performance.

Dans la théorie classique de l'estimation paramétrique d'une matrice de transition associée à une chaîne de Markov, nous disposons de plusieurs méthodes, décrites dans [2], qui présentent l'avantage d'être simples à utiliser. Toutefois, il est difficile d'estimer avec précision des matrices de transition modélisant des phénomènes complexes. Pour pallier cette difficulté, nous faisons appel aux méthodes d'estimation non paramétriques. Ces dernières ont fait l'objet de travaux établis par Roussas (1969) [8] en utilisant la méthode du noyau. Les résultats obtenus par celui-ci ont été complétés par plusieurs autres auteurs mais ces résultats sont restreints dans le cadre théorique plus que pratique.

L'exploitation de la méthode du noyau dans les chaînes de Markov dans un cadre pratique revient initialement au travail de Bareche et Aïssani (2008) [1], où les auteurs ont prouvé l'applicabilité de la méthode du noyau dans les systèmes de files d'attente classiques lorsque l'une des lois les régissant est générale et inconnue. Par la suite, Gontijo et al. (2011) [4], ont appliqué la méthode de noyau pour estimer les mesures de performance du système $GI^{[X]}/M/C/N$. Récemment, Cherfaoui et al. (2015) [3] ont abordé le problème du choix du paramètre de lissage dans le contexte d'estimation à noyau d'une chaîne de Markov décrivant le système d'attente $GI/M/1/N$. Dans ce dernier travail, afin de prendre en considération l'interaction des différentes composantes du système les auteurs ont proposé une procédure de sélection du paramètre de lissage qui se base sur les normes matricielles où ils ont démontré que l'estimateur du paramètre de lissage choisi, par la minimisation de certaines normes matricielles, donne de meilleurs résultats que les méthodes classiques.

Il est à noter que la totalité des travaux cités auparavant ont été effectués via des estimateurs à noyaux continus asymétriques (pour estimer des distributions des durées de service ou des durées des inter-arrivées qui sont définies sur $\mathbb{R}^+$). Cependant, dans la pratique plusieurs situations sont modélisées par des chaînes de Markov régissant selon des distributions discrètes générales et inconnues. Dans ce cas, il est logique et naturel de procéder à l'estimation de ces lois générales

inconnues en utilisant les noyaux discrets. Dans ce sens, vu que la chaîne de Markov associée au modèle du stock de type (R, s, S) est discrète et à temps discret, alors pour estimer sa matrice de transition nous devons utiliser les noyaux discrets.

L'objectif du présent travail est de vérifier la validité des conclusions dégagées par Cherfaoui et al. [3] sur les chaînes de Markov continues lorsque nous considérons la chaîne de Markov discrète associée au modèle de stock (R, s, S). Autrement dit, nous allons considérer l'estimateur à noyau discret de la matrice de transition P correspondante au modèle en question et analyser le problème du choix du paramètre de lissage via les normes matricielles. Pour ce faire, nous allons développer des formes explicites des expressions, issues de trois normes matricielles $\|.\|_1$, $\|.\|_2$ et $\|.\|_\infty$, à minimiser afin de sélectionner le paramètre de lissage optimal lors de l'estimation de la matrice de transition P. De plus, dans le but d'appuyer et d'illustrer nos propositions, une application numérique comparative basée sur des échantillons simulés sera réalisée.

Le reste du document est organisé comme suit : Dans la deuxième section nous allons présenter brièvement le modèle stochastique de gestion de stock de type (R, s, S). Dans la troisième section, le problème du choix du noyau et du paramètre de lissage lors de l'estimation de la matrice de transition du modèle de stock en question sera présenté. Avant de conclure, dans la quatrième section nous allons présenter l'application numérique réalisée, les résultats obtenus ainsi que leurs discussions.

2 Description du modèle

Considérons le problème de gestion de stock de type (R, s, S) [7, 9–11], suivant : Au début de chaque période R, on décide si on doit ou non commander une quantité d'articles et dans le cas affirmatif, combien commander. On suppose que le fournisseur est parfaitement fiable et que les commandes arrivent immédiatement. Durant la $n$ième période, $n \geq 1$, la demande totale est une variable aléatoire discrète X_n. On suppose également que les variables aléatoires X_n, $n \geq 1$, sont indépendantes et identiquement distribuées, de loi commune :

$$f(x) = \Pr(X = x), \quad x = 0,\ 1,\ 2,\ 3,\ \ldots \tag{1}$$

Pour un tel problème de gestion de stock, l'état du stock L_n est inspectée aux dates $t_n = nR$ $(n \geq 1)$. Si le niveau du stock $L_n \leq s$, on passe une commande de manière à ramener le stock au niveau S, la taille de la commande est égale alors à $S - L_n$. Si par contre le niveau du stock est supérieur au seuil de risque s, on ne passe aucune commande et l'on attend jusqu'au prochain moment d'inspection.

L'état du stock L_{n+1} à la fin de la $(n+1)$ième période est alors donné par :

$$L_{n+1} = \begin{cases} (S - X_{n+1})^+, & \text{si } L_n \leq s; \\ (L_n - X_{n+1})^+, & \text{si } L_n > s; \end{cases} \tag{2}$$

où $(A)^+ = \max(A, 0)$.

On remarque que la variable aléatoire L_{n+1} ne dépend que de L_n et X_{n+1}, où X_{n+1} est indépendante de n et de l'état du système avant t_n. Donc L est une chaîne de Markov homogène, à espace d'état $E = \{0, 1, ..., S\}$. Enfin, la figure Fig. 1 illustre la forme générale du modèle de stock en question.

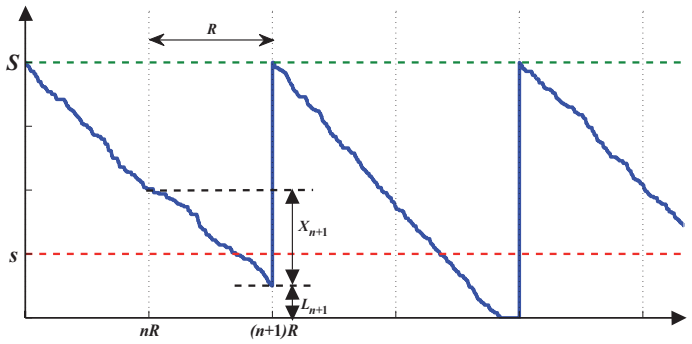


Fig. 1. Un schéma illustratif du processus du niveau du stock dans le système (R, s, S) avec un délai d'exécution nul.

Soit L_n la chaîne de Markov représentant le niveau du stock. Si on suppose que le niveau du stock à la date $t_n = nR$ et la date $t_{n+1} = (n+1)R$ est $L_n = i$ et $L_{n+1} = j$, respectivement. Alors les probabilités de transition, $P_{ij} = \Pr(L_{n+1} = j \mid L_n = i)$, de cette chaîne peuvent être résumées comme suit :

$$P_{ij} = \begin{cases} \sum_{x=S}^{+\infty} f(x), \text{ si } 0 \leq i \leq s \text{ et } j = 0; \\ f(S-j), \text{ si } 0 \leq i \leq s \text{ et } 1 \leq j \leq S; \\ \sum_{x=i}^{\infty} f(x), \text{ si } s+1 \leq i \leq S \text{ et } j = 0; \\ f(i-j), \text{ si } s+1 \leq i \leq S \text{ et } 1 \leq j \leq i; \\ 0, \qquad \text{si } s+1 \leq i \leq S \text{ et } j \geq i+1; \end{cases} \tag{3}$$

avec les éléments $f(x)$ sont donnés par l'expression (1).

3 Estimation à noyau de la matrice de transition du modèle

Rappelons que pour l'évaluation des performances d'un système de stock de type (R, s, S) on se base sur ses différents paramètres de départ le décrivant. Cependant, dans la pratique et en règle générale, les valeurs des paramètres de départ ne sont connus que sous forme d'un échantillon de données.

Dans ce sens, supposons qu'on est dans une situation où c'est la distribution des demandes qui est inconnue c'est-à-dire on ne dispose que d'une information

partielle sur la distribution des demandes qui est donnée sous forme d'un échantillon $X_1,\ X_2, ...,\ X_n$ de taille n. De plus, notre intérêt est l'estimation de la matrice de transition P associée à ce modèle tout en utilisant la méthode du noyau définie par :

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i), \quad x \in \mathbb{N}; \tag{4}$$

où h est le paramètre de lissage (fenêtre) et $K_{x,h}$ est le noyau discret de cible x et de fenêtre h sur le support $\aleph_{x,h} = \aleph_x$ (ne dépend pas de h).

Il est claire que la mise en œuvre de cette technique nécessite de fixer préalablement le noyau K et le paramètre de lissage h. Pour le choix du noyau K, le problème est a priori facile. En effet, il suffit de le sélectionner parmi les noyaux, les plus usités dans le cadre d'estimation d'une densité discrète, suivants : Noyau Poissonnien, Noyau Binomial, Noyau Binomial négatif et Noyau Triangulaire (pour plus de détails le lecteur peut se référer à [5, 6]). Tandis que, pour le choix du paramètre de lissage h, on peut envisager deux approches, à savoir : les techniques classiques et les normes matricielles.

3.1 Choix de paramètre de lissage via les techniques classiques

Soit $X_1, \ldots,\ X_n$ un n-échantillon iid de distribution des demandes inconnue f. L'idée dans ce cas est d'estimer les éléments $f(x) = \Pr(X = x)$ à partir de l'échantillon sans prendre en considération leurs répétitions dans la matrice P i.e. il suffit de quantifier les $\hat{f}(x)$ par

$$\hat{f}(x) = \hat{P}r(X = x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h_{opt}}(X_i), \qquad x \in \mathbb{N},$$

où h_{opt} est le paramètre de lissage optimal sélectionné par les procedures classiques et par la suite de remplacer $\hat{f}(x)$ dans la matrice P pour obtenir $\hat{P}$.

Ci-dessous quelques méthodes classiques pour le choix du paramètre de lissage dans l'estimation des fonctions discrètes.

1. **Minimisation du MISE**
   Un choix adéquat pour le paramètre de lissage peut être celui minimisant l'erreur quadratique intégrée (ISE) et qui est donné par :

   $$h^*_{ise} = \arg\min_h \left(\sum_{x\in\mathbb{N}} (\hat{f}(x) - f(x))^2 \right) = \arg\min_h ISE(X_1, ..., X_n, h, K, f), \tag{5}$$

   pour laquelle la mesure est sur un seul échantillon et la fenêtre optimale h^*_{opt} peut être obtenue, dans le cas de plusieurs échantillons, à travers

   $$h^*_{opt} = \arg\min_{h>0} E(ISE(X_1, ..., X_n, h, K, f)).$$

   Ces techniques ont été développées et détaillées par Kokonendji et al. [5] et Kokonendji et Kiessé [6].

2. **Validation croisée** Nous présentons ici deux techniques qui se basent sur la méthode de validation croisée. Plus précisément, l'idée est d'estimer la densité f au point X_i par la technique de validation croisée dont la forme est donnée par :

$$\hat{f}_{-i}(X_i) = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} K_{X_i,h}(X_j).$$

   (a) Validation croisée par les moindres carrées : le principe de cette méthode reste le même que celui de Scott et Terrell [12] dans le cas de variables continues où la technique consiste à estimer le ISE par la technique de validation croisée et par la suite de sélectionner le paramètre de lissage qui minimise l'estimateur en question. Ainsi, la fenêtre optimale, dans le cas de variables discrètes, s'obtient par :

$$h_{cv} = \arg\min_{h>0} \left[\sum_{x\in\mathbb{N}} \left\{ \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i) \right\}^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,\ j\neq i}^{n} K_{X_i,h}(X_j) \right].$$

   (b) Validation croisée par le maximum de vraisemblance. Ce critère consiste à choisir h qui maximise la fonctionnelle

$$LCV(h) = \prod_{i=n}^{n} \hat{f}_{-i}(X_i) \text{ ou encore } LCV(h) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\hat{f}_{-i}(X_i)\right).$$

3. **Excès des zéros [5]** Le choix de la fenêtre, par cette procédure, repose sur une particularité des données de comptage qui n'est autre que l'excès des zéros dans l'échantillon, c'est-à-dire de choisir une fenêtre adaptés $h_0 = h_0(X_1, ..., X_n, , K)$ tel que :

$$\sum_{i=1}^{n} \Pr(\mathcal{K}_{X_i,h_0} = 0) = n_0, \tag{6}$$

   où $\mathcal{K}_{x,h}$ est la variable aléatoire de loi $K_{x,h}$ définie sur $\aleph_{x,h}$ et $n_0 = card\{X_i = 0\}$ désigne le nombre des zéros dans l'échantillon $X_1, ..., X_n$ à condition que $n_0 > 0$. Ci-dessous quelques exemples de h_0 :
   - Cas noyau Poissonnien : $h_0 = \log\left(\frac{1}{n_0} \sum_{i=1}^{n} e^{-X_i}\right)$.
   - Cas noyau Binomial : h_0 est la solution de $n_0 = \sum_{i=1}^{n} \left(\frac{1-h_0}{X_i+1}\right)^{X_i+1}$.
   - Cas noyau Binomial négatif : h_0 est la solution de $n_0 = \sum_{i=1}^{n} \left(\frac{X_i+1}{2X_i+1+h_0}\right)^{X_i+1}$.

3.2 Le choix de paramètre de lissage par les normes matricielles

Dans ce cas l'idée est de prendre en considération le nombre de répétitions des éléments $\hat{f}(x)$ dans la matrice $\hat{P}$, notons que cette dernière est donnée comme suit :

$$\hat{P} = \begin{array}{c|cccc|cccc|c} & 0 & 1 & & & s & & s+1 & & S \\ \hline 0 & \sum_{x=S}^{+\infty}\hat{f}(x) & \hat{f}(S-1) & \cdots & \hat{f}(S-s) & \hat{f}(S-s-1) & & \cdots & & \hat{f}(0) \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & & & \vdots \\ s & \sum_{x=S}^{+\infty}\hat{f}(x) & \hat{f}(S-1) & \cdots & \hat{f}(S-s) & \hat{f}(S-s-1) & & \cdots & & \hat{f}(0) \\ \hline s+1 & \sum_{x=s+1}^{+\infty}\hat{f}(x) & \hat{f}(s) & \cdots & \hat{f}(1) & \hat{f}(0) & 0 & \cdots & \cdots & 0 \\ s+2 & \sum_{x=s+2}^{+\infty}\hat{f}(x) & \hat{f}(s+1) & \cdots & \hat{f}(2) & \hat{f}(1) & \hat{f}(0) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & & \ddots & & \vdots \\ S & \sum_{x=S}^{+\infty}\hat{f}(x) & \hat{f}(S-1) & \cdots & \hat{f}(S-s) & \hat{f}(S-s-1) & & \cdots & & \hat{f}(0) \end{array}$$

avec $\hat{f}(x) = \hat{P}r(X = x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)$, et $x \in \mathbb{N}$.

Dans le but de prendre en consideration les répétitions des éléments $\hat{f}(x)$, nous proposons d'utiliser les normes matricielles qui ont un impact sur la qualité de l'estimateur $\hat{P}$. En effet, l'utilisation des normes matricielles nous permet d'inclure les répétitions des quantités $\hat{f}(x)$ dans l'expression $\hat{P}$. Dans ce cas, le paramètre de lissage optimal peut être calculé selon l'une des trois expressions suivantes :

$$\begin{cases} h_1^* = \arg\min_h \left\|\hat{P} - P\right\|_1 = \arg\min_h \left[\max_{0\le i\le S}\left(\sum_{j=0}^{S}\left|\hat{P}_{ij} - P_{ij}\right|\right)\right], \\ h_2^* = \arg\min_h \left\|\hat{P} - P\right\|_2 = \arg\min_h \left[\sum_{i=0}^{S}\sum_{j=0}^{S}\left(\hat{P}_{ij} - P_{ij}\right)^2\right]^{\frac{1}{2}}, \\ h_3^* = \arg\min_h \left\|\hat{P} - P\right\|_\infty = \arg\min_h \left[\max_{0\le j\le S}\left(\sum_{i=0}^{S}\left|\hat{P}_{ij} - P_{ij}\right|\right)\right]. \end{cases}$$

En remplaçant les éléments de P et $\hat{P}$ par leurs expressions, ces trois dernières formules peuvent être réécrites comme suit :

$$\begin{aligned} h_1^* &= \arg\min_h \left[\max_{s+1\le i\le S}\left(\sum_{j=0}^{S}\left|\hat{P}_{ij} - P_{ij}\right|\right)\right] \\ &= \arg\min_h \left[\max_{s+1\le i\le S}\left(\sum_{x=0}^{i-1}\left|\hat{f}(x) - f(x)\right| + \left|\sum_{x=0}^{i-1}\left(\hat{f}(x) - f(x)\right)\right|\right)\right]. \end{aligned} \tag{7}$$

$$\begin{aligned} h_2^* = \arg\min_h \Bigg[&(S+1)\sum_{x=0}^{s}\left(\hat{f}(x) - f(x)\right)^2 + (s+2)\left(\sum_{x=0}^{S-1}\left(\hat{f}(x) - f(x)\right)\right)^2 \\ &+ \sum_{x=s+1}^{S-1}(S+s+1-x)\left(\hat{f}(x) - f(x)\right)^2 + \sum_{i=s+1}^{S-1}\left(\sum_{x=0}^{i-1}\left(\hat{f}(x) - f(x)\right)\right)^2\Bigg]^{\frac{1}{2}}. \end{aligned} \tag{8}$$

$$h_3^* = \arg\min_h \left[\max_{0\le j\le S} \left(\begin{array}{l} \left[(s+2) \left| \sum_{x=0}^{S-1} (\hat{a}_x - a_x) \right| + \sum_{i=s+1}^{S-1} \left| \sum_{x=0}^{i-1} (\hat{a}_x - a_x) \right| \right] \mathbf{1}_{\{j=0\}} \\ + \left[(s+2) \left| \hat{a}_{S-j} - a_{S-j} \right| + \sum_{i=s+1}^{S-1} \left| \hat{a}_{i-j} - a_{i-j} \right| \right] \mathbf{1}_{\{1\le j\le s\}} \\ + \left[(s+2) \left| \hat{a}_{S-j} - a_{S-j} \right| + \sum_{i=j}^{S-1} \left| \hat{a}_{i-j} - a_{i-j} \right| \right] \mathbf{1}_{\{s+1\le j\le S-1\}} \end{array} \right) \right]. \tag{9}$$

avec $\mathbf{1}_{\{.\}}$ est la fonction indicatrice, $a_x = f(x)$ et $\hat{a}_x = \hat{f}(x)$.

4 Application numérique

L'objectif de la présente section est d'analyser numériquement l'impact du choix du paramètre de lissage via les normes matricielles sur les performances de l'estimateur d'une matrice de transition associée à une chaîne de Markov décrivant un modèle de stock de type (R, s, S). Pour ce faire, nous avons implémenté un programme sous MATLAB dont les principales étapes sont comme suit :

Étape 1 Fixer la totalité des paramètres : $R,\ s,\ S,\ \lambda,\ f$.
Étape 2 Générer m échantillons de taille n de distribution f.
Étape 3 Estimer h_{opt} par les expressions (5), (7), (8) et (9) pour chaque échantillon.
Étape 4 Calculer $\hat{f}$, $\hat{P}$ et $\hat{Q}$ pour chaque h_{opt} obtenus dans l'Étape 3.
Étape 5 Calculer h^* et $\overline{Q}$ les moyennes des estimateurs h_{opt} et $\hat{Q}$.

Notons que, $\hat{Q}$ représente l'estimateur du niveau moyen du stock défini par :

$$\hat{Q} = \sum_{i=0}^{S} i\pi_i, \tag{10}$$

avec $\pi = (\pi_0,\ \pi_1, ..., \pi_S)$ est le vecteur des probabilités stationnaires du niveau du stock qu'on obtient par la résolution du système d'équations suivant :

$$\begin{cases} \pi * \hat{P} = \pi, \\ \sum_{i=0}^{S} \pi_i = 1. \end{cases} \tag{11}$$

Pour l'application numérique nous allons considérer les paramètres suivants:

- le nombre de réplications (échantillons) $m = 1000$,
- la taille de l'échantillon $n \in \{100;\ 500;\ 1000\}$,
- la période $R = 1$, le taux des demandes $\lambda R = 5$ et la distribution $f \in$ {Poisson(μ), Géométrique(p_1), Binomiale (N, p_2)} avec $\mu = 5$, $p_1 = 1/6$, $N = 12$ et $p_2 = 5/12$,

– le seuil de risque $s = 3$ et la capacité maximale du stock $S = 6$.
– le noyau $K \in \{Piosson;\ Binomial;\ Binomial\ Négatif;\ Triangulaire\}$, où :

Noyau Poissonnien

$$K_{P_o(x+h)}(y) = e^{-(x+h)} \frac{(x+h)^y}{y!}, \tag{12}$$

avec $\aleph_{x,h} = \mathbb{N}$, $(x, y) \in \mathbb{N}^2$ et le paramètre de lissage $h > 0$.

Noyau Binomial

$$K_{B\left(x+1, \frac{x+h}{x+1}\right)}(y) = \frac{(x+1)!}{y!(x+1-y)!} \left(\frac{x+h}{x+1}\right)^y \left(\frac{1-h}{1+x}\right)^{x+1-y} \mathbf{1}_{\{y \leq x+1\}}, \tag{13}$$

avec $x \in \mathbb{N}$, $\aleph_{x,h} = \{0, 1, ..., x+1\}$, $y \in \aleph_{x,h}$, $h \in]0\ ;\ 1]$ et $\mathbf{1}_{\{.\}}$ est la fonction indicatrice.

Noyau Binomial négatif

$$K_{BN\left(x+1, \frac{x+1}{2x+1+h}\right)}(y) = \frac{(x+y)!}{y!x!} \left(\frac{x+h}{2x+1+h}\right)^y \left(\frac{x+1}{2x+1+h}\right)^{x+1}, \tag{14}$$

avec $\aleph_{x,h} = \mathbb{N}$, $(x, y) \in \mathbb{N}^2$ et $h > 0$.

Noyau Triangulaire

$$K_{T_{(a,h,x)}}(y) = \frac{(a+1)^h - |y-x|^h}{(2a+1)(a+1)^h - 2\sum_{j=0}^{a} j^h} \mathbf{1}_{\{|y-x|<a\}}, \tag{15}$$

avec $x \in \mathbb{N}$, $\aleph_{x,h} = \{x, x \pm 1, \ldots, x \pm a\}$, $y \in \aleph_{x,h}$ $h > 0$, $a \in \mathbb{N}$ et $\mathbf{1}_{\{.\}}$ est la fonction indicatrice. Pour l'application numérique nous avons fixé $a = 5$.

Les résultats obtenus pour les différents paramètres précédents sont rangés dans les Tables 1 – 3. Les résultats obtenus montrent que :

– Lorsque f est Poissonnienne ou Binomiale, excepté le cas d'utilisation du noyau triangulaire la propriété de convergence du paramètre de lissage optimal h vers zéro lorsque n tend vers l'infinie n'est pas vérifiée (i.e. lorsque on utilise le noyau Poissonnien, Binomial ou Binomial négatif) et cela quelle que soit la procédure de sélection utilisée. Plus précisément, le paramètre de lissage h a tendance à être une constante dans ces cas (voir tables 1 et 3).
– Dans le cas où f est une distribution Géométrique, le paramètre de lissage h_{opt} est sensible à la taille de l'échantillon. De plus la propriété de convergence du paramètre de lissage optimal est vérifiée, et cela dans la quasi-totalité des situations considérées (voir table 2).
– Dans le cas où f est une distribution Poissonnienne, le paramètre de lissage qui nous fournis un estimateur, du niveau moyen du stock, plus proche à la valeur exacte $Q = 1.3936$ selon le noyau est : (h^*_{ise}, Poisson), (h^*_{ise}, Binomial), (h^*_1, Binomial négatif) et (h^*_3, Triangulaire) ou (h^*_2, Triangulaire) le fait que

ces deux derniers couples nous fournissent pratiquement les mêmes résultats. Enfin, dans ce cas, d'une manière générale il est préférable d'utiliser le noyau Triangulaire et de sélectionner le paramètre de lissage via la norme matricielle $\|.\|_2$ ou $\|.\|_\infty$ et d'éviter l'utilisation du noyau Binomial négatif (voir table 1).

- Dans le cas où f est une distribution Géométrique le paramètre de lissage qui nous fournis des estimateurs plus proche à la valeur exacte $Q = 2.4334$ selon le noyau est : (h_3^*, Poisson), (h_1^*, Binomial), (h_3^*, Binomial négatif) et (h_2^*, Triangulaire). Enfin, le couple qui nous fournis de meilleure résultats est bien que (h_1^*, Binomial), tandis que les pires estimateurs (les moins performants) sont obtenus lors de l'utilisation du noyau Binomial négatif et cela quelle que soit la procédure de sélection du paramètre de lissage utilisée (voir table 2).
- Dans le cas où f est une distribution Binomiale afin d'obtenir un bon estimateur (plus proche à la valeur exacte $Q = 1.2294$), du niveau moyen du stock, il est préférable de construire l'estimateur de la distribution des demandes à l'aide du noyau Triangulaire et le paramètre de lissage minimisant la norme matricielle $\|.\|_2$ (voir table 3).

Finalement, nos résultats nous permettent de pouvoir conclure d'une part que les paramètres de lissage sélectionnés via les normes matricielles nous fournissent, d'une manière générale, des estimateurs plus performants que ceux conçu à l'aide du h_{ise}^* et d'autre part, que le choix du noyau est d'une grande importance où il est préférable d'utiliser le noyau Triangulaire et d'éviter le noyau Binomial négatif.

Table 1: Estimateurs des paramètres de lissage h_{opt} et du niveau moyen du stock : cas f est une distribution poissonnienne.

Noyau	n	ISE		$\|.\|_1$		$\|.\|_2$		$\|.\|_\infty$	
		h_{ise}^*	$\widehat{Q}_{h_{ise}^*}$	h_1^*	$\widehat{Q}_{h_1^*}$	h_2^*	$\widehat{Q}_{h_2^*}$	h_3^*	$\widehat{Q}_{h_3^*}$
Poisson	100	0.2121	1.3269	0.5791	1.5125	0.5677	1.5088	1.0800	1.6441
	500	0.1855	1.3159	0.5796	1.5166	0.5667	1.5107	0.7765	1.5570
	1000	0.1857	1.3153	0.5816	1.5171	0.5676	1.5103	0.6091	1.5079
Binomial	100	0.0795	1.4219	0.1562	1.4692	0.1232	1.4489	0.1234	1.4487
	500	0.0543	1.4077	0.1307	1.4560	0.0943	1.4329	0.1102	1.4428
	1000	0.0451	1.4015	0.1298	1.4552	0.0885	1.4290	0.1109	1.4430
Binomial Négatif	100	0.3179	1.2398	0.4512	1.2925	1.0718	1.5126	1.3334	1.5654
	500	0.3028	1.2362	0.4179	1.2846	1.0723	1.5155	1.2952	1.5681
	1000	0.3049	1.2361	0.4153	1.2831	1.0750	1.5157	1.2906	1.5679
Triangulaire	100	0.1251	1.4290	0.0816	1.4170	0.0869	1.4178	0.1068	1.4145
	500	0.0298	1.4060	0.0222	1.4018	0.0215	1.4016	0.0222	1.4011
	1000	0.0163	1.3993	0.0134	1.3973	0.0130	1.3971	0.0121	1.3965

Table 2: Estimateurs des paramètres de lissage h_{opt} et du niveau moyen du stock : cas f est une distribution géométrique.

Noyau	n	ISE		$\|.\|_1$		$\|.\|_2$		$\|.\|_\infty$	
		h^*_{ise}	$\overline{Q}_{h^*_{ise}}$	h^*_1	$\overline{Q}_{h^*_1}$	h^*_2	$\overline{Q}_{h^*_2}$	h^*_3	$\overline{Q}_{h^*_3}$
Poisson	100	0.4489	2.3269	0.3022	2.3675	0.2975	2.3621	0.2922	2.3573
	500	0.2822	2.3015	0.1943	2.3488	0.1873	2.3486	0.1821	2.3467
	1000	0.2612	2.1396	0.1905	2.2752	0.1807	2.3018	0.1711	2.2885
Binomial	100	0.1625	2.3970	0.1873	2.4022	0.1915	2.3979	0.1870	2.3962
	500	0.0991	2.2738	0.0861	2.3523	0.0902	2.3619	0.0946	2.3391
	1000	0.0805	2.3706	0.0736	2.3965	0.0766	2.3980	0.0813	2.3990
Binomial	100	0.5982	2.3397	0.3978	2.3808	0.3875	2.3878	0.3793	2.3892
	500	0.4690	2.3817	0.3208	2.4042	0.2991	2.4075	0.2894	2.4097
Négatif	1000	0.4586	2.4154	0.3292	2.4172	0.3019	2.4158	0.2893	2.4142
Triangulaire	100	0.0993	2.3789	0.0452	2.3999	0.0339	2.4032	0.0391	2.3862
	500	0.0207	2.4222	0.0108	2.4234	0.0095	2.4225	0.0152	2.4219
	1000	0.0120	2.3523	0.0074	2.3886	0.0065	2.3984	0.0105	2.4008

Table 3: Estimateur des paramètres de lissage h_{opt} et du niveau moyen du stock : cas f est une distribution géométrique.

Noyau	n	ISE		$\|.\|_1$		$\|.\|_2$		$\|.\|_\infty$	
		h^*_{ise}	$\overline{Q}_{h^*_{ise}}$	h^*_1	$\overline{Q}_{h^*_1}$	h^*_2	$\overline{Q}_{h^*_2}$	h^*_3	$\overline{Q}_{h^*_3}$
Poisson	100	0.1750	1.1845	0.4686	1.3365	0.5145	1.3619	0.3985	1.2892
	500	0.1403	1.1735	0.4367	1.3279	0.5015	1.3625	0.3179	1.2656
	1000	0.1420	1.1726	0.4412	1.3288	0.5054	1.3630	0.3213	1.2656
Binomial	100	0.0808	1.2909	0.1264	1.3193	0.0939	1.2991	0.1145	1.3120
	500	0.0458	1.2779	0.0947	1.3090	0.0571	1.2851	0.1016	1.3134
	1000	0.0378	1.2703	0.0952	1.3070	0.0518	1.2792	0.1129	1.3183
Binomial	100	0.2443	1.0853	0.7299	1.2956	1.0341	1.4116	0.8426	1.3403
	500	0.2195	1.0801	0.7156	1.2970	1.0169	1.4129	0.8224	1.3401
Négatif	1000	0.2229	1.0799	0.7191	1.2972	1.0209	1.4132	0.8262	1.3405
Triangulaire	100	0.0555	1.2715	0.0464	1.2602	0.0498	1.2631	0.0815	1.2710
	500	0.0134	1.2457	0.0154	1.2459	0.0133	1.2443	0.0165	1.2455
	1000	0.0073	1.2361	0.0080	1.2362	0.0075	1.2357	0.0091	1.2368

5 Conclusion

Dans ce papier, nous avons considéré le choix du paramètre de lissage par des procédures qui se basent sur des normes matricielles dans le cadre d'estimation à noyaux discrets d'une matrice de transition d'une chaîne de Markov discrète, décrivant un modèle de stock de type (R, s, S).

L'étude de simulation réalisée montre d'une part l'intérêt de la minimisation des normes matricielles pour la sélection du paramètre de lissage, en particulier

la norme matricielle $\|.\|_2$ combinée avec le noyau triangulaire. D'autre part, elle nous suggère d'éviter l'utilisation du noyau Binomial négatif dans le contexte abordé.

Pour des travaux à venir, nous envisageons de développer les procédures proposées dans ce papier de telle sorte qu'elles soient exploitables en pratique et cela en utilisant par exemple le principe de la règle de référence ou la validation croisée.

References

1. Bareche, A. and Aïssani, D. : Kernel density in the study of the strong stability of the $M/M/1$ queueing system. Operations Research Letters **36**(5), 535–538 (2008)
2. Billingsley, P. : Statistical inference for Markov processes. University of Chicago Press (1961)
3. Cherfaoui, M., Boualem, M., Aïssani, D., Adjabi, S. : Choix du paramètre de lissage dans l'estimation à noyau d'une matrice de transition d'un processus semi-markovien. C. R. Acad. Sci. Paris, Ser. I **353**(3), 273–277 (2015)
4. Gontijo, G. M., Atuncar, G. S., Cruz, F. R. B., Kerbache, L. : Performance evaluation and dimensioning of $GI^{[X]}/M/c/N$ systems through kernel estimation. Mathematical Problems in Engineering **2011**, 1–20 (2011)
5. Kokonendji, C. C., Kiessé, T. S., Zocchi, S. S. : Discrete triangular distributions and non-parametric estimation for probability mass function. Journal of Nonparametric Statistics **19**(6-8), 241-254 (2007)
6. Kokonendji, C. C., Kiesse, T. S. : Discrete associated kernels method and extensions. Statistical Methodology **8**(6), 497–516 (2011)
7. Rabta, B., Aïssani, D. : Strong stability in an (R, s, S) inventory model. International Journal of Production Economics **97**(2), 159–171 (2005)
8. Roussas, G. G. : Nonparametric estimation in Markov processes. Ann. Inst. Statist. Math. **21**, 73–87 (1969)
9. Scarf, H. : Theoptimali ty of (S, s) policies in the dynamic inventory problem. In : Arrow, K.J., Karlin, S., Suppes, P. (Eds.), Mathematical Methods in the Social Sciences. Stanford University Press, Stanford, CA. 196–202 (1960)
10. Iglehart, D. : Dynamic programming and stationary analysis of inventory problems, multistage inventory models and techniques. In : Scarf, H., Gilford, D., Shelly, M. (Eds.), Multistage Inventory Models and Techniques. Stanford University Press, Stanford, CA. 1–31 (1963)
11. Veinott, A., Wagner, H. : Computing optimal (s, S) policies. Management Sciences **11**, 525–552 (1965)
12. Scott, D. W., Terrell, G. R. : Biased and unbiased cross-validation in density estimation. Journal of the American Statistical Association **82**, 1131–1146 (1987)

Performance evaluation of IP networks with differentiated services

LEKADIR Ouiza and ADEL-AISSANOU Karima

Laboratory of Modeling and Optimization of Systems (LaMOS)
University of Bejaia 06000, Algeria.
ouizalekadir@gmail.com ak_adel@yahoo.fr

Abstract. Initially, the Internet was a tool for a small community of agencies and organizations, where services such as file transfer "FTP", the E-mail and the procedures of remote connection dominated. However, with the appearance of new types of multimedia applications such as the videoconference, the telephony on IP and the e-commerce, new requirements such as quality of services appeared. Two approaches are currently proposed by the IETF (Internet Engineering Task Force) to assure the QoS in IP networks: the approach IntServ and the approach DiffServ which is the object of our study in this article. The approach DiffServ or services differentiated, allows introducing a new way of treating the streams of the data in the network and sharing the latter's resources. In this architecture, the bandwidth, the rate of loss and the delay of transmission are affected by the operations of conditioning of traffic during the entry to the network, as well as by the modifications brought to the behavior of the core routers. The objective of this study is the mathematical modeling of the architecture with differentiation of services (DiffServ) as defined by the workgroup of IETF by a $M/G/1/N$ queues with multiple vacations and exhaustive service. The study limits itself to the evaluation of the performances of the EF (Expedited Forwarding) class of the core routers which contains the most critical packages (voice, video). For the validation of the chosen analytical model, a simulation of a simple network under DiffServ was realized with NS2. This modeling (with $M/G/1/N$ queues) serves for the sizing of the parameters of the network: debit and size of buffer according to the load of the network.
Keywords: IP networks, DiffServ (differentiated services), $M/G/1/N$ Queue with multiple vacations, Simulation, Network Simulator (NS), Quality of Service (QoS).

1 Introduction

With the development of the multimedia applications, the IP network will have to allow the deployment of these applications to guarantee its success. However these applications have specific requirements in term of QoS (Quality of service). Certain services such as the vocal services will need weak delay and weak jitter (variation of the delays of crossing). There are many works in the literature which approaches the QoS of the real-time applications on IP networks. Our study is

one of these works which aims to determine the network parameters (debit and size of buffer) according to its load. QoS is all mechanisms that provide a good service level to specific flows. QoS refers to the ability of the network to transport in good conditions flows from different applications. This definition is reflected in the following techniques characteristics: Availability, Bandwidth, Delay, The jitter and Loss ratio.

2 Main mechanisms of Management of the QoS

To be able to assure the transported streams, it is necessary to treat in a differentiated way the various categories of traffic in the network organs as well as the protocols of road marking of the QoS to be able to assign resources according to the applications needs. There are two major problems for the management of QoS in IP network [6]:

2.1 The management of the congestion phenomena

Mechanisms to the QoS management have an impact only when the network is crowded, there are two basic approaches to the management of congestions, reactive methods and preventive methods.

2.2 Packets scheduling

Is also a fundamental mechanism to guarantee QoS to transported flows. This is evident if we consider heterogeneous flow: Gusts of certain connections can the disturb real time traffic even if there is no congestion.

The applications in IP have different needs in terms of quality of service. To answer the various requirements, the QoS defines three types of services(see [1,2,9,10,11]:

1. Guaranteed services: There are books of resources throughout the network for a flow. This type of service is reserved for critical applications.
2. Differentiated services: Statistically privilege certain types of flows. The guarantee is relative. This type of service is reserved for applications that require preferential treatment.
3. Best-Effort services: There is no guarantee; this is the normal operation of IP without QoS. This type of service is used for the rest of the applications: those with no constraints (mail, web, fttp ...).

The various congestion control mechanisms and packets scheduling are present in all architectures developed for the control of QoS in IP networks [12]. Historically, the first architecture which has been proposed combines QoS to each flow in the network. This is the model IntServ (Integrated Services). This model allows establishing a mode connected on a network; and establishes a road which will follow all the data of a stream and it by the booking of the necessary resources in routers. The Internet community has also proposed a model called DiffServ (Differentiated Services) in which QoS is associated with wave's aggregation.

3 DiffServ Description

Unlike the IntServ based on the booking of the resources in routers, the approach "differentiated services" tries to establish the QoS by a sorting of packages entering on the border of the network under DiffServ following various criteria (delay, bandwidth, address). The sorting is made at the level of the routers of the border. The streams of data are classified according to three categories of services (Expedited Forwarding, Assured Forwarding and Best Effort) and four classes of traffic predefined according to the performances asked for their transmission. Packages are "marked" and managed in routers by specific queues for every category or classes.
In DiffServ there are border routers which are connected to border routers of other field. They take charge of the classification of entering packages. We also find core routers which are connected to core routers or border routers of the same field. They take charge of treatment with differentiating the packages of various categories or class.

3.1 Border routers

The structure of a border router can be divided into 4 modules [4].

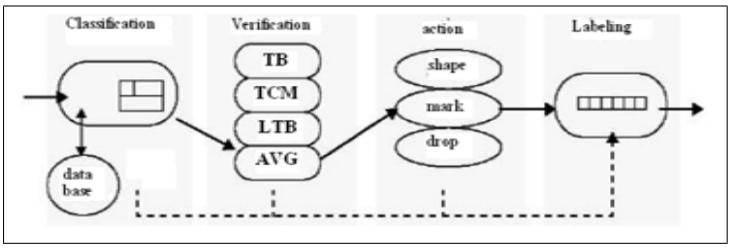


Fig. 1. Border (Edge) Router.

3.1.1. Classification To treat different flows generated by users with a differentiated way, the border router must first make a packet classification, which is based on the header IP.
3.1.2. Checking An auditor is responsible for determining the level of compliance for each packet flow coming into the router. This depends mainly on the instantaneous flow behavior and characteristics of the contract (SLA).
3.1.3. Actions (dropper & shaper) The corrective action to be taken for non-compliant packets varies depending on the service. Three types penalty can be identified:

1. The elimination is probably the most severe action, but necessary for the proper functioning of certain services.

2. The formatting is to delay, if necessary, the flow of a stream to conform it.
3. Marker: Before entering the network, the DSCP field (a part of header IP) of all packages that pass through the router entry is updated. It forms the label DiffServ and should not be changed by the heart of routers in the network.

3.2 Core routers

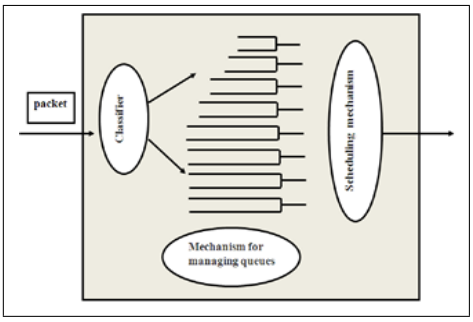


Fig. 2. Architecture of routers core (heart).

A PHB is describing characteristics of delivery that will be observed by all packets containing the same DSCP. Using a PHB or group of PHB added to the packaging operations performed at the input of the field form a DiffServ. Diffserv or PHB defines four classes of service:
3.2.1. Expedited Forwarding (EF) [**4**] Corresponding to the maximum priority and aims to ensure bandwidth with low loss ratio, delay and jitter, achieving the transfer of flow with severe constraints of time like IP telephony.
3.2.2. Assured Forwarding (AF) [**5**] Grouping several PHB guarantees delivery of IP packets with a high probability regardless of deadlines.
3.2.3. Best Effort (BE) [**6**] PHB default.
3.2.4 Default Forwarding (DF) Used only for Internet streams that do not require a real-time traffic.
At each queue, a management mechanism must decide how packets will be eliminated in case of congestion.
Several scheduling techniques (algorithms) have been developed to control the sharing of resources between the classes of service:

1. Fair Queuing (FQ) or Round Robin (RR)
2. Weighted Fair Queuing (WFQ)
   - Generalized Processor Sharing (GPS);
   - Weighted Fair Queuing (WFQ);
   - Weighted Round Robin (WRR);
   - Priority Queuing (PQ).

4 Modeling of core router

It focuses on core routers of the architecture DiffServ, because of the operations of these routers (queue management and scheduling), which will determine network performances.

In the context of DiffServ network, the queues are organized by classes of service which all share the same server (the same link output). A scheduling policy is then used to share the time service between classes of service. Although overall, a queue with sharing service behaves like classic queue, length of service are not the same for each service class, and thus the criteria for QoS differs from one class of service to another.

DiffServ architecture defines two big traffic classes: EF (Expedited Forwarding), AF (Assured Forwarding). The AF class is often divided into several subclasses, ordered by their higher expectations in terms of quality of service. In a DiffServ network, one solution is to use 2 successive schedulers: the first type PQ scheduler who selects priority packages of Class EF compared to other classes, and then a second type of scheduler WFQ (Weight Fair Queueing) to differentiate services AF1, AF2, AF3 and AF4. The WFQ algorithm ensures to provide for each service AF a minimum length of service, according to a weighting defined by the operator.

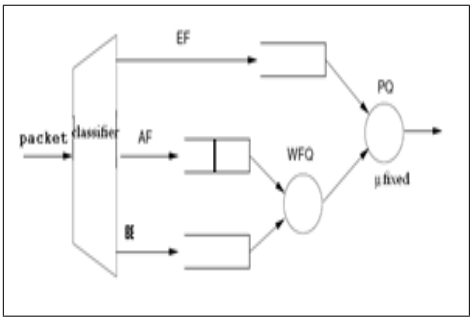


Fig. 3. Model.

4.1 The model

We consider that our model is represented by a queue $M/G/1/N$ with multiple vacations and exhaustive service.

- Packets of Class EF priority represent customers in the system.
- The output link is seen as the system server , it can get one packet only at a time.
- Packets of Class EF arrive according to a Poisson process with rate λ.

- The variation in the service length is due to the random length packets and not to server capacity, which is reflected in a general law service probability density $b(t)$, and cumulative distributing Function $B(t)$ or $\overline{X}$ the average time of service.
- The size of the queue is limited to $(N-1)$ packets.
- During the idleness period, the server begins to serve packets of other classes, which corresponds in our model to a vacation with random length(general), its probability density $f_v(t)$, and $F_v(t)$ its distribution function, let $\overline{V}$ the average duration of a vacancy.
- At the end of a vacation period, the server does not return in vacancy when there will be no packets to serve in this class (multiple vacancy and comprehensive service).

The performance evaluation of $M/G/1/N$ with finite capacity and server vacation has been the concern of Frey and Takahashi [5]. They are used the induced Markov chain method (included), where they observe the system at moments that are: either at the end of service or for the end of vacation period.
The state of the system at induced points is represented by the couple (n_i, ϕ_i) with:

- n_i: The traffic number (packets EF) in the system just after the $i-th$ induced point;
- $\phi_i = \begin{cases} 0, & \text{if the } i^{th} \text{ induced point corresponds to an end of period vacation;} \\ 1, & \text{if the } i^{th} \text{ induced point corresponds to an end of service.} \end{cases}$

Consider the system in the steady state.
Note:

- $q_k, \forall k = 0, \ldots, N$; the probability to be in the state $(k, 0)$;
- $r_k, \forall k = 0, \ldots, (N-1)$; the probability to be in the state $(k, 1)$;
- $f_j, j = 0, \ldots, +\infty$; the probability to have j packets of EF class in the system just after one vacation period, this probability is given by:

$$f_j = \int_0^{+\infty} \frac{(\lambda t)^j}{j!} e^{(-\lambda t)} f_v(t) d(t), \ j = 0, \ldots, +\infty. \tag{1}$$

- $\alpha_j, j = 0, \ldots, +\infty$; the probability that j packets of EF class arrives in the system during a service time, this probability is given by:

$$\alpha_j = \int_0^{\infty} \frac{(\lambda t)^j}{j!} e^{(-\lambda t)} b(t) d(t), \ j = 0, \ldots, +\infty. \tag{2}$$

The states probabilities of the system:

$$q_k = (q_0 + r_0) f_k,\ k = 0, \ldots, (N-1); \tag{3}$$

$$q_k = (q_0 + r_0) \sum_{(k=N)}^{\infty} f_k,\ k = N; \tag{4}$$

$$r_k = \sum_{(j=1)}^{(N+1)} (q_j + r_j)\alpha_(k - j + 1),\ k = 0, \ldots, (N-2); \tag{5}$$

$$r_{(N-1)} = q_N + \sum_{(j=1)}^{(N-1)} (q_j + r_j) \sum_{(k=N-j)}^{\infty} \alpha_k,\ k = N-1; \tag{6}$$

$$\sum_{(k=0)}^{N} q_k + \sum_{(j=0)}^{(N-1)} r_j = 1. \tag{7}$$

These probabilities are used to get some performance parameters of the system:

- The Load: $\rho_c = \dfrac{(1 - q_0 - r_0)\overline{X}}{(q_0 + r_0)\overline{V} + (1 - q_0 - r_0)\overline{X}}$;
- The offered charge: $\rho = \lambda \overline{X}$;
- Blocking probability (rejection): $P_B = \dfrac{(\rho - \rho_c)}{\rho}$;
- Average time D between successive included points:

$$D = (q_0 + r_0)\overline{V} + (1 - q_0 - r_0)\overline{X}.$$

To determine the usual system parameters, such as the mean number of customers in the system M and the mean sejourn time W, note by:

- $Q_k = Pk$Customers in the system, the server is in vacation; $k = 0, \ldots, N$;
- $R_k = Pk$Customers in the system, the server is busy; $k = 0, \ldots, (N-1)$;

The state probabilities are given by:

$$Q_k = \begin{cases} \dfrac{1}{\lambda D} \sum\limits_{(j=k+1)}^{N} q_j,\ k = 0, \ldots, (N-1); \\ 1 - \rho_c - \dfrac{1}{\lambda D} \sum\limits_{(j=1)}^{N} j q_j,\ k = N. \end{cases} \tag{8}$$

$$R_k = \begin{cases} \dfrac{1}{\lambda D}\Big(r_k - \sum\limits_{(j=k+1)}^{N} q_k\Big),\ k = 1, \ldots, (N-1); \\ \dfrac{\rho_c(\rho - 1)}{\rho} + \dfrac{1}{\lambda D} \sum\limits_{(j=1)}^{N} j q_j,\ k = N. \end{cases} \tag{9}$$

In general:

$$\begin{cases} P_0 = & Q_0,\ j = 0; \\ P_j = & Q_j + R_j = \dfrac{r_j}{\lambda D},\ j = 1, \ldots, (N-1); \\ P_N = & Q_N + R_N = \dfrac{(\rho - \rho_c)}{\rho},\ j = N. \end{cases} \tag{10}$$

The usual parameters of the system are:

- The mean number M of EF Class packets in the system is:

$$M = \frac{1}{\lambda D} \sum_{(j=1)}^{(N-1)} jr_j + N\Big(\frac{(\rho - \rho_c)}{\rho}\Big). \tag{11}$$

- The mean sojourn time of a packet in the system is:

$$W = \frac{M}{\lambda(1 - P_B)}. \tag{12}$$

- The probability that the server is busy is:

$$\sum_{(k=1)}^{N} R_k = \rho_c = \rho(1 - P_N). \tag{13}$$

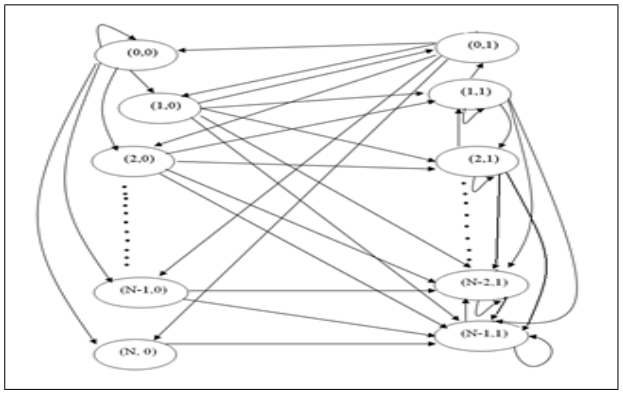


Fig. 4. **Graph of possibles transitions.**

5 Simulation

For the validation of our analytical model. A simulation of a simple network under DiffServ was realized. For this network two types of transactions are used: a CBR (Constant Bit Rate) stream based on the UDP protocol that models traffic audio, and an FTP flow based on TCP protocol models best-effort traffic. We focus our efforts on measuring the parameters of quality of service (time, number of packets lost) of CBR streams.
The architecture of our network consists of an UDP source, one TCP source,

two border routers, one router heart and one destination.
The simulation model incorporates features DiffServ architecture. Two different queues are managed. They model different classes of service: EF classes of traffic and BE. These files are served by a scheduler Priority Queuing (PQ) where the file of the class EF has the highest priority and BE file the lowest. All DiffServ domain files are managed by a mechanism RED, and other files are handled by DropTail, the latter wait for the filling of the buffers to reject packets. When

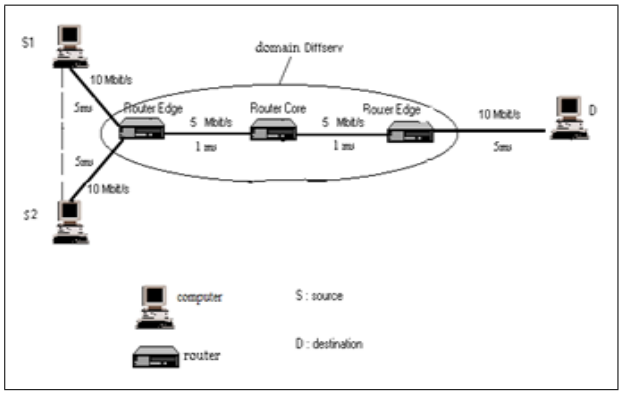


Fig. 5. Simulated Topology.

the number of packets in the queue is equal to N, all the packets received after, are rejected. In NS2, it is sufficient to compute the number of lost packets and divide it on the duration of the simulation. Analytically, it corresponds to the $\Big(1/\text{Probability(blocking)}\Big)$ given in the section 3.

6 Results and interpretation

The implementation of the analytical model (section 3) with Matlab allows us to obtain the analytical results of this system. By varying:

- The arrival rate of packets of priority Class EF (Packet / ms),
- The size of the system (packet),
- The service law and its parameters,
- The vacation law and its parameters.

The different analytical results obtained are shown in the tables given in the figures (Fig.6, Fig.7, Fig.8) with those of the simulated topology under NS2.

Service law	Service parameters	Vacation law	Vacation parametrs	Theoric results		Simulation results	
				Loss	Delay	Loss	Delay
Exponential	μ=3	Exp	μ=1	0.406	0.577	0.019	0.186
		Erlang	μ=1	0.589	0.751		
		Cox-2	μ_1=1 μ_2=0.1 a=0.7	1.468	1.949		
Erlang	μ=6	Exp	μ=1	3.08×10^{-6}	0.114		
		Erlang	μ=1	3.34×10^{-7}	0.114		
		Cox-2	μ_1=1 μ_2=0.1 a=0.7	2.59×10^{-6}	0.114		
Cox-2	μ_1=4 μ_2=7 a=0.7	Exp	μ=1	0.414	0.529		
		Erlang	μ=1	0.546	0.653		
		Cox-2	μ_1=1 μ_2=0.1 a=0.7	1.266	1.507		

Fig. 6. Obtained results for $N = 5$ and $\lambda = 3$.

Service law	Service parameters	Vacation law	Vacation parametrs	Theoric results		Simulation results	
				Loss	Delay	Loss	Delay
Exponential	μ=3	Exp	μ=1	0.019	0. 670	0.014	0.336
		Erlang	μ=1	0.293	0.715		
		Cox-2	μ_1=1 μ_2=0.1 a=0.7	0.1097	0.8136		
Erlang	μ=6	Exp	μ=1	0	0.260		
		Erlang	μ=1	0	0.260		
		Cox-2	μ_1=1 μ_2=0.1 a=0.7	9.9910^{-16}	0.258		
Cox-2	μ_1=4 μ_2=7 a=0.7	Exp	μ=1	0.356	0.801		
		Erlang	μ=1	0.154	0.750		
		Cox-2	μ_1=1 μ_2=0.1 a=0.7	0.85	0.791		

Fig. 7. Obtained results for $N = 10$ and $\lambda = 3$.

Service law	Service parameters	Vacation law	Vacation parametrs	Theoric results		Simulation results	
				Loss	Delay	Loss	Delay
Exponential	μ=3	Exp	μ=1	3.18×10^{-5}	1.523	0.004	0.404
		Erlang	μ=1	3.19×10^{-5}	1.523		
		Cox-2	μ_1=1 μ_2=0.1 a=0.7	1.93×10^{-4}	1.523		
Erlang	μ=6	Exp	μ=1	0	0.546		
		Erlang	μ=1	0	0.546		
		Cox-2	μ_1=1 μ_2=0.1 a=0.7	0	0.546		
Cox-2	μ_1=4 μ_2=7 a=0.7	Exp	μ=1	0.692	1.581		
		Erlang	μ=1	0.143	1.596		
		Cox-2	μ_1=1 μ_2=0.1 a=0.7	0.1429	1.596		

Fig. 8. **Obtained results for** $N = 20$ **and** $\lambda = 3$.

All laws tested give good results for a good choice of parameters for the service or vacation law:
• For a considerable packet traffic (burst) of Class EF, the occupancy rate of the server is of order 1, which does not allow it to go on vacation. Therefore, we will have the independence of performance metrics time vacation. Unlike the case when traffic is less intense, performance metrics also depends on the vacation law due to the fact that the server will have the opportunity to go on vacation. The latter's duration has a large influence on these metrics. Indeed, a great period vacation creates an increase in the period and the rate of loss.
• The length of the file EF is a parameter that also affects the loss of packets, because a bad dimensioning of this file may degrade the transmission of these packets with considerable losses. As can be expected over the length of the file unless it is a great package lost, and conversely, if the vacation periods are wider and the queue is smaller, we have more of lost packets and an acceptable delay.

7 Conclusion

We have studied quality of service parameters in a network where audio stream shares resources with Best Effort traffic. We analized two QoS parameters for audio flows : The mean time end to end and the number of packets lost. The measurements were made in an environment with a scheduler DiffServ with strict priority PQ (Priority Queueing) by modeling a system with $M/G/1/N$ queues with multiple vacation and exhaustive service as well as a simulation Network Simulator NS2. The results presented show that the proposed model ($M/G/1/N$ queue with vacation) allows us to evaluate the criteria of the quality of service

of real time applications in a satisfactory manner for a good choice of laws parameters (Service Or vacation). These criteria are used for dimensioning of system parameters: Debit and size of the buffer (queues), depending on the load on the system.

References

1. IntServ: http://www.ietf.org/html.charters/intserv-charter.html
2. DiffServ: http://www.ietf.org/html.charters/diffserv-charter.html
3. F. Baker, J. Heinanen, W. Weiss and J. Wroclawski. *Assured Forwarding PHB Group, Standards Track*, Request For Comments (RFC 2597), (1999).
4. Y. Bernet, S. Blake, D. Grossman and A. Smith. *An Informal Management Model for Diffserv Routers*, Standards Track, Request For Comments (RFC 3444), (2002).
5. A. Frey, and Y. Takahashi. *A note on a $M/G/1/N$ queue with vacation time and exhaustive service discipline*, Oper. Res. Lett., 21, 95-111, (1997).
6. D. Fuin. *Qualité de service : des réseaux IP à l'intégration dans les réseaux actifs, PhD Thesis, UFR des Sciences et Techniques de l'Université de Franche Comté*, Francen , (2004).
7. P. Hurley, J. Y. Le Boudec, and P. Thiran. *The Alternative Best-Effort Service*, In Proc. IEEE GLOBCOM'99, (1999).
8. V. Jacobson, K. Nichols and K. Poduri. *An Expedited Forwarding PHB, Standards Track*, Request For Comments (RFC 2598), (1999).
9. A. R. Mahlous. *Maximizing QoS for Video Streams*, IJCSNS International Journal of Computer Science and Network Security, VOL.18 No.10, October (2018).
10. A. R. Mahlous and R. Fretwell, *Differenciated service: A good choise for quality of services*, 2019, https://www.researchgate.net/publication/264886716_DIFFERENTIATED_SERVICES_A_GOOD_CHOICE_FOR_A_QUALITY_OF_SERVICES.
11. C. A. Martínez, D. López, J. J. Ramírez, and R. D. Gómez. *Performance of diffserv and intserv services in QoS on an academic network using NS2*, TECCIENCIA, Vol. 7 No. 14., 65-75, 2013, DOI: http:/dx.doi.org/10.18180/tecciencia.2013.14.9cc
12. L. Stephane, and D. Present, *Internet: Services et réseaux*, Edition Dunod, (2004).

Résolution du problème du sous-graphe de poids maximum des arêtes dans des réseaux biologiques

Youcef Djeddi[1], Hacene Ait Haddadene[2], and Nabil Belacel[3]

[1] Laboratoire LaROMad, Département de recherche opérationnelle, Faculté de mathématiques, Université USTHB, BP 32 Bab-Ezzouar, El-Alia 16111, Alger, Algerie.
djyoucef74@gmail.com

[2] Laboratoire LaROMad, Département de recherche opérationnelle, Faculté de mathématiques, Université USTHB, BP 32 Bab-Ezzouar, El-Alia 16111, Alger, Algerie.
aithaddadenehacene@yahoo.fr

[3] Technology Research Center, National Research Council, Ottawa, Ontario, Canada
nabil.belacel@nrc-cnrc.gc.ca

Résumé Soit $G = (V, E, W_e)$ un graphe avec des poids sur les arêtes, avec V l'ensemble des sommets, E l'ensemble des arêtes et W_e une fonction de pondération qui associe à chaque arête $e \in E$ un poids positif $w_e \geq 0$. Le problème du sous-graphe de poids maximum des arêtes (MEWS) vise à trouver un sous-graphe $H = (S, F)$ de G tel que la somme des poids des arêtes de F est la plus grande et le nombre des sommets de S égal à k, où k un entier ($k \geq 3$), il est prouvé que le MEWS est un problème NP-Complet. Dans ce papier, nous proposons un algorithme recherche tabou multistart pour le MEWS. Nous avons utilisé notre algorithme pour résoudre le MEWS dans des réseaux biologiques, qui ont des vrais poids sur les arêtes. Les résultats ont montré que notre algorithme surperforme les autres heuristiques de l'état de l'art.

Keywords: Le sous-graphe de poids maximum des arêtes · La recherche tabou· Réseaux biologiques · Heuristiques.

1 Introduction

Soit $G = (V, E, W_e)$ un graphe avec des poids sur les arêtes où $V = \{1, 2, ..., n\}$ l'ensemble des sommets, $E \subseteq V \times V$ l'ensemble des arêtes et $W_e : E \to Z^+$ une fonction de pondération qui associe à chaque arête $\{u, v\} \in E$ un poids positif w_{uv}, et soit k un entier tel que $k \geq 3$, le problème du sous-graphe de poids maximum des arêtes (the Maximum Edge-Weighted Subgraph problem (MEWS) en anglais) cherche à trouver un sous-graphe $H = (S, F)$ de G tel que la somme des poids des arêtes de F est la plus grande somme possible et le nombre des sommets de S égal à k, le MEWS est connu aussi sous le nom de problème du k-sous-graphe le plus lourd. Si tous les poids des arêtes sont égaux à 1, le MEWS devient équivalent au problème du k-sous-graphe le plus dense. La version de décision de MEWS est NP-Complet [1]. Le problème du sous-graphe de poids

maximum des arêtes a plusieurs applications dans des déférents domaines tel que les réseaux de télécommunications, les réseaux sociaux et les réseaux biologiques.

Vu l'importance du problème de sous-graphe de poids maximum des arêtes, un grand nombre de travaux ont été publiés sur ce problème. Macambira & de Meneses (1998) [9] ont proposé une heuristique gloutonne adaptative aléatoire (GRASP) pour le MEWS, Macambira (2002)[8] a proposé un algorithme recherche tabou pour résoudre le MEWS, Billionnet (2005) [2] a proposé différentes formulations pour résoudre le MEWS, Brimberg *et al.* (2009) [3] ont proposé une heuristique recherche à voisinage variable (VNS) pour le MEWS, Letsios *et al.* (2016) [6] ont proposé un algorithme exact pour le MEWS dans les médias sociaux, Singh *et al.* (2019) [11] ont proposé une approximation de la solution de MEWS, ils ont utilisé une heuristique gloutonne pour réduire la taille de graphe puis ils ont utilisé ce graphe comme une entrée dans un algorithme Branch and Bound.

Dans ce papier, nous proposons un algorithme recherche tabou multistart (MTSEWS) pour le MEWS, nous avons utilisé notre algorithme pour résoudre MEWS dans 12 réseaux biologiques avec des poids sur les arêtes, puisque l'identification des complexes protéiques et des modules fonctionnels a été formulée comme un problème du sous-graphe de poids maximum des arêtes [7]. Nous avons comparé les résultats de notre algorithme avec deux heuristiques, l'algorithme recherche tabou proposé par Macambira [8] et l'algorithme recherche à voisinage variable (VNS) proposé par Brimberg *et al.* [3]. Les résultats ont montré que notre algorithme surperforme les autres heuristiques dans la qualité des solutions. Dans le reste du papier, nous donnons une présentation détaillée de notre algorithme Section 2 et une étude comparative Section 3.

2 L'algorithme proposé MTSEWS

Le principe général de notre algorithme MTSEWS est résumé dans l'Algorithme 1. MTSEWS commence par une solution initiale générée par une procédure aléatoire (un sous-ensemble $S \subset V$ de taille k, Section 2.1). À partir de cette solution initiale MTSEWS utilise une procédure appelée BTS pour chercher à trouver une solution améliorée (Section 2.2). BTS utilise deux sous-ensembles critiques $A \subset S$ et $B \subset V \setminus S$, chaque itération BTS change un sommet de A avec un sommet de B (A et B sont des sous-ensembles contraints). Si BTS ne peut pas améliorer la solutions actuelle S pendant L itérations consécutives, MTSEWS redémarre (restart) BTS avec une nouvelle solution initiale générée suivant une stratégie de construction d'une nouvelle solution initiale pour un nouveau restart (voir Section 2.3). Après chaque appel de BTS, MTSEWS compare la solution S trouvée par BTS avec la meilleure solution trouvée jusqu'à présent S^* et elle fait le mis à jour de S^*. L'algorithme MTSEWS s'arrête quand le nombre d'itérations atteint It_{mx} (It_{mx} est le nombre maximum d'itérations autorisées).

L'espace de recherche Ω exploré par notre algorithme MTSEWS est l'ensemble de tous les sous-ensembles de taille k (l'ensemble de toutes les solu-

tions réalisables), soit $G = (V, E, W_e)$ un graphe avec des poids sur les arêtes, $\Omega = \{S \subset V : |S| = k\}$. L'évaluation de chaque solution S est basée sur la somme des poids des arêtes du sous-graphe induit par S, i.e., la fonction d'évaluation est définie comme suit :

$$f(S) = \sum\nolimits_{u,v \in S, \{u,v\} \in E} w_{uv}$$

La fonction $f(S)$ est à maximisée, on dit qu'une solution S' est mieux que S si $f(S') > f(S)$.

Algorithme 1 L'algorithme MTSEWS

Entres: Un graphe G avec des poids sur les arêtes, Un entier k (la taille du sous-graphe), Un entier L (profondeur de recherche), Un entier It_{mx} (le nombre maximum d'itérations).
Sorties: Le k-sous-graphe de poids maximum des arêtes.

```
Début
S ← solution initiale /*Générer une solution initiale, S est la solution actuelle*/
It ← 0 /* Initialisation du conteur global d'itérations*/
S* ← S /*Enregistrer la meilleure solution trouvée jusqu'à présent S**/
Tantque It < It_mx Faire
   S ← BTS(G, S, k, L, It)
   Si f(S) > f(S*) Alors
      S* ← S
   Finsi
   S ← nouvelle solution initiale /* Générer une nouvelle solution initiale suivant
   la stratégie de construction d'une nouvelle solution initiale*/
Tin tantque
Fin
Retourner(S*)
```

2.1 Solution initiale

La solution initiale utilisée par notre algorithme MTSEWS est générée de la manière suivante :

1. Initialiser un ensemble S à vide.
2. Sélectionner un sommet $v \in V \setminus S$ aléatoirement et ajouter v à S.
3. Ajouter à S un sommet $v \in V \setminus S$ tel que $\sum_{u \in S, \{u,v\} \in E} w_{uv}$ est la plus grande somme, s'il y a plusieurs sommets choisir un aléatoirement.
4. Tant que $|S| < k$ répéter (3) et retourner S.

2.2 L'algorithme BTS

L'algorithme BTS est un algorithme recherche tabou, le principe général de l'algorithme BTS est résumé dans l'Algorithme 2. BTS commence par enregistrer la meilleure solution trouvée jusqu'à présent S^* et sa valeur de la fonction

d'évaluation f^*, et initialisé le conteur d'itérations consécutives I à 0. S représente la solution actuelle. Pour améliorer la solution actuelle BST applique une série d'intensification et de diversification (Voir Section 2.2). BTS utilise deux listes tabous pour que les sommets qui sont entrés à (sorti de) la solution actuelle ne puissent pas sortir de (revenir à) la solution actuelle rapidement (Voir Section 2.2). Si BTS trouve une nouvelle solution S mieux que S^*, BTS met à jour S^* par S. Si BTS ne trouve pas une solution mieux que S^* pendent L itération consécutive, BTS s'arrête et elle retourne S^*.

Algorithme 2 L'algorithme BTS

Entres: Un graphe G avec des poids sur les arêtes, Un entier k (la taille de sous-graphe), Un entier L (profondeur de recherche), Un entier It (Conteur d'itérations).
Sorties: S^*

Début
$S^* \leftarrow S$ /*Enregistrer la meilleure solution trouvée jusqu'à présent*/
$f^* \leftarrow f(S^*)$ /*f^* enregistre la valeur de la fonction d'valuation de S^**/
$I \leftarrow 0$ /* Initialisation du compteur d'itérations consécutive à 0*/
Tantque $I < L$ **Faire**
  Si il existe une solution d'amélioration **Alors**
    $S \leftarrow Intensification(S)$
  Sinon
    $S \leftarrow Diversification(S)$
  Finsi
  Mis à jour de $Tlist\_u$ et $Tlist\_v$
  Si $f(S) > f^*$ **Alors**
    $S^* \leftarrow S$
    $I \leftarrow 0$
    $f^* \leftarrow f(S^*)$
  Sinon
    $I \leftarrow I + 1$
  Finsi
  $It \leftarrow It + 1$
Tin tantque
Fin
Retourner(S^*)

Intensification et Diversification Dans l'étape d'intensification, BTS met à jour (remplace) la solution actuelle S par une nouvelle solution S' tel que $f(S') \geq f(S)$. Dans l'étape de diversification, BTS remplace la solution actuelle S par une nouvelle solution S' tel que $f(S') < f(S)$. Dans cette section, nous expliquons l'étape d'intensification et de diversification, pour cela, nous commençons par définir quelques paramètres nécessaires. Soit S la solution actuelle ($S \in \Omega$), nous avons attribué à chaque sommet $v \in V$ un poids relatif à S noté par $wd(v)$

et un degré relatif à S noté $d(v)$, $wd(v)$ et $d(v)$ sont définis comme suit :

$$wd(v) = \sum_{u \in S, \{u,v\} \in E} w_{uv}$$

$$d(v) = |\{u \in S : \{u, v\} \in E\}|$$

Comme nous avons dit, BTS utilise deux sous-ensembles critiques $A \subset S$ et $B \subset V \setminus S$ pour définir le voisinage contraint. A et B sont définis comme suit :

Soit $Tlist\_u$ et $Tlist\_v$ les deux listes tabous utilisées par BTS (Voir Section 2.2).

Soit $PMinInS = min\{wd(u) : u \in S, u \notin Tlist\_u\}$ et

Soit $PMaxOutS = max\{wd(v) : v \in V \backslash S, v \notin Tlist\_v\}$

$$A = \{u \in S : u \notin Tlist\_u, wd(u) \leq PMinInS + Pmax\}$$

$$B = \{v \in V \backslash S : v \notin Tlist\_v, wd(v) \geq PMaxOutS - Pmax\}$$

Où $Pmax = max\{w_{uv} : u \in S, v \in V \backslash S, v \notin Tlist\_v, wd(v) = PMaxOutS\}$

Pour obtenir une solution voisine S' de la solution actuelle S, nous changeons un sommet de $u \in A$ avec un sommet de $v \in B$ i.e., $S' = S \bigoplus swap(u, v)$ ou $S' = S \backslash \{u\} \cup \{v\}$. Le voisinage contraint est composé de toutes les solutions obtenues par le changement d'un sommet de A avec un sommet B :

$$VC(S) = \{S' : S' = S \backslash \{u\} \cup \{v\}, u \in A, v \in B\}$$

Nous désignons un gain $\Delta_{u,v}$ pour chaque $swap(u, v)$, ce gain est défini par la variation de la fonction d'évaluation :

$$\Delta_{uv} = f(S') - f(S) = wd(v) - wd(u) - we_{uv}$$

avec $e_{uv} = w_{uv}$ si $\{u, v\} \in E$ sinon $we_{uv} = 0$.

Comme nous avons dit, dans l'intensification BTS cherche à trouver une solution mieux que la solution actuelle S ou une solution qui ne détruit pas la solution S ($f(S') \geq f(S)$), il est facile de voir que nous ne pouvons pas faire une intensification si et seulement s'il existe un $swap(u, v)$ tel que $\Delta_{u,v} > 0$. L'étape d'intensification est faite comme suit :

1. S'il existe un ou plusieurs couples de sommets (u, v) tel que $\Delta_{uv} > 0$, BTS choisir le couple de sommets (u, v) qui a le maximum de Δ_{uv} et il applique le changement $swap(u, v)$, s'il y a plusieurs couples choisir un couple aléatoirement.

2. S'il n'existe pas un couple de sommets (u, v) tel que $\Delta_{uv} > 0$, BTS continuer sa recherche par une étape de diversification.

Si BTS n'a pas pu appliquer une intensification donc elle est tombée dans un optimum local, pour sortir de cet optimum, BTS applique une diversification. L'étape de diversification est faite comme suit :

1. Avec une grande probabilité ($P = 0.9$), BTS applique une petite diversification. Choisir un couple de sommets (u, v) ($u \in A$ *et* $v \in B$) qui a le grand Δ_{uv} et appliquer $swap(u, v)$, s'il y a plusieurs le couple qui a le minimum de Δd_{uv} où $\Delta d_{uv} = d(v) - d(u) - e_{uv}$ et $e_{uv} = 1$ si $\{u, v\} \in E$ sinon $e_{uv} = 0$. S'il y a encore plusieurs choisir un couple aléatoirement.
2. Avec une petite probabilité ($1-P$), BTS applique une grande diversification. Choisir un sommet $u \in S$ aléatoirement et un sommet $v \in V \setminus S$ tel que $d(v) < \lfloor 0.4 \times k \rfloor$, s'il y a plusieurs choisir un aléatoirement, et appliquer le changement $swap(u, v)$

Après l'opération de changement ($swap(u, v)$), BTS met à jour les paramètres wd, d, $PMinInS$, $PMaxOutS$, A et B, pour préparer l'itération suivante.

Les listes tabous BTS utilise deux listes tabous pour éviter le retour rapide des sommets changés dans les étapes d'intensification et de diversification. Après chaque itération, BTS ajoute le sommet u qui est sorti de la solution actuelle à une liste tabou appelé $Tlist\_u$ pour qu'il ne puisse pas revenir la solution actuelle qu'après Tu itérations. BTS ajoute aussi le sommet v qui est entré à la solution actuelle à une liste tabou appelé $Tlist\_v$ pour qu'il ne puisse pas sortir de la solution actuelle qu'après Tv itérations (Tu et Tv sont appelés tabou tenure, voir Glover & Laguna (1997) [5]). Nous avons fixé $Tu = 15$ et $Tv = 9$ si $k \geq 30$, sinon $Tu = 8$ et $Tv = 5$.

2.3 La stratégie de construction d'une nouvelle solution pour un nouveau restart

Pour encourager l'algorithme MTSEWS à visiter des nouvelles régions de recherche, nous avons utilisé le mécanisme proposé par Wu & Hao (2013) [12] (Djeddi, Ait Haddadene & Belacel [4] ont aussi utilisé ce mécanisme). Wu & Hao ont attaché à chaque sommet v de V une fonction g_v appelée mémoire de fréquence à long terme (voir Wu & Hao (2013) [12] et Djeddi, Ait Haddadene & Belacel [4]). g_v enregistre le nombre de mouvements du sommet v, la fonction g_v est maintenue comme suit :

1. Initialement, $g_v = 0$, $\forall v \in V$.
2. Chaque fois qu'un sommet v est ajouté ou supprimé de la solution actuelle S, g_v est incrémenté par un ($g_v = g_v + 1$).
3. Si pour tout sommet $v \in V$, $g_v > k$ alors réinitialiser $g_v = 0$ pour tout $v \in V$.

La nouvelle solution est construite comme suit :

1. Initialiser un ensemble S à vide.

2. Ajouter à S le sommet v qui a la plus petite fréquence g_v.
3. Ajouter à S un sommet $v \in V \setminus S$ tel que $\sum_{u \in S, \{u,v\} \in E} w_{uv}$ est la plus grande somme, s'il y a plusieurs sommets choisir le sommet qui a la plus petite fréquence g_v, s'il y en a encore plusieurs choisir un aléatoirement.
4. Tant que $|S| < k$ répéter (3) et retourner S.

3 Résultats expérimentaux

Dans cette section, nous utilisons notre algorithme MTSEWS pour résoudre le problème du sous-graphe de poids maximum des arêtes dans quelques réseaux biologiques avec des poids sur arêtes, nous comparons aussi les résultats de notre algorithme avec les résultats de l'algorithme recherche tabou (TS_MEwS) proposé par Macambira [8] et les résultats l'algorithme recherche à voisinage variable (VNS) proposé par Brimberg *et al.* [3]. Nous avons téléchargé les réseaux biologiques à partir de référentiel des données des réseaux [10], ces réseaux sont à l'origine avec poids sur les arêtes.

Nous avons programmé notre algorithme (MTSEWS) et les autres algorithmes (TS_MEwS et VNS) sous MATLAB, et nous avons compilé les trois algorithmes sur la même machine, HP Workstation Z600 avec Intel Xeon X5550, 2.66GHz et 12 GB RAM. Nous avons exécuté MTSEWS avec les paramètres suivants $It_{mx} = 10^8$ et $L = 100$ pour quelques réseaux et $L = 1000$ les autres (voir Tableau 1), et TS_MEwS avec un nombre d'itérations maximum égal à 10^8. Nous avons fixé k à deux fois la borne inférieure de la taille de la clique maximum ω_{lb} ($k = 2 * \omega_{lb}$) sauf dans une instance (bio-CE-HT) $k = 4 * \omega_{lb}$ car ω_{lb} est très petit, nous avons obtenu les bornes inférieures de la taille de la clique maximum à partir de référentiel des données des réseaux [10]. Vu la nature stochastique des trois algorithmes, nous avons exécuté chaque algorithme 10 fois pour chaque réseau, et avons pris la valeur de la fonction objectif de la meilleure solution (notée par Meilleur) et la moyenne des valeurs des fonctions objectifs des 10 solutions (notée par Moyenne). Le temps de chaque exécution et de chaque algorithme est fixé à $1000s$. Les résultats sont résumés dans le Tableau 1.

Le Tableau 1 résume les résultats comparatifs de notre algorithme MTSEWS et les autres algorithmes (TS_Mews et VNS). Les colonnes 1, 2 et 3 présentent les informations des réseaux (le nom, le nombre des sommets $|V|$ et le nombre des arêtes $|E|$ respectivement). La colonne 4 montre la valeur de k. La colonne 5 montre la valeur de L utilisée dans l'exécution de MTSEWS. Les colonnes 6-11 présentent les valeurs des fonctions objectif des meilleurs solutions trouvées et la moyenne des solutions trouvées par chaque algorithme. Les résultats montrent que notre algorithme surperforme les autres algorithmes dans tous les réseaux sauf dans un seul cas (la moyenne du réseau bio-HS-CX). En outre, MTSEWS a obtenu des solutions strictement supérieures aux autres dans 11 sur 12 dans les moyennes et 12 sur 12 dans les meilleures solutions.

Table 1. Résumé des résultats des algorithmes MTSEWS, TS_Mews et VNS

Les instances					MTSEWS		TS_Mews		VNS	
Nom	$\|V\|$	$\|E\|$	k	L	Meilleur	Moyenne	Meilleur	Moyenne	Meilleur	Moyenne
bio-CE-CX	15229	245952	86	100	**6916.6**	**7940.0**	2570.4	2570.4	1126.1	1351.9
bio-CE-GN	2220	53683	32	100	**1095.5**	**1095.5**	192.4753	192.4753	99.0834	122.3925
bio-CE-GT	924	3239	16	100	**204.5353**	**204.7847**	126.9924	128.5918	161.9576	183.5881
bio-CE-HT	2617	2985	16	100	**58.9716**	**65.0661**	10.0054	12.1446	5.7250	18.6123
bio-DM-CX	4040	76717	64	100	**5723.7**	**7681.9**	3256.0	3256.0	1478.2	1834.4
bio-DR-CX	3289	84940	98	1000	**14648.0**	**15802.0**	14139.0	14139.0	1810.7	2831.8
bio-HS-CX	4413	108818	36	1000	1936.0	**2587.7**	**1975.6**	1975.6	440.9218	549.8944
bio-HS-HT	2570	13691	76	1000	**7324.6**	**7324.6**	5835.6	5835.6	2829.8	5766.1
bio-HS-LC	4227	39484	116	1000	**14421.0**	**14421.0**	11370.0	11370.0	4902.0	5226.0
bio-SC-GT	1716	33987	78	1000	**4064.1**	**5329.5**	3704.1	3707.3	4787.6	5266.4
bio-SC-HT	2084	63027	124	1000	**20047.0**	**20047.0**	19501.0	19501.0	4941.7	5347.6
bio-SC-LC	2004	20452	58	1000	**2650.7**	**3280.1**	1802.2	1822.9	950.8185	1422.3

4 Conclusion

Un algorithme recherche tabou multistart (MTSEWS) est proposé dans ce papier pour la résolution du problème de sous-graphe de poids maximum des arêtes. L'algorithme proposé utilise deux sous-ensembles critiques (un sous-ensemble de la solution courante S et un sous-ensemble de $V \setminus S$) pour construire un voisinage contraint. Pour une recherche efficace dans l'espace de recherche, MTSEWS utilise aussi une stratégie multistart pour générer des nouvelles solutions initiales pour des nouveaux redémarrages (restart).

Nous avons testé notre algorithme sur 12 réseaux biologiques avec des poids originaux sur les arêtes et comparé nos résultats avec deux algorithmes de l'état de l'art. Les résultats computationnels ont montré que notre algorithme donne une très bonne performance sur ces réseaux. De plus, notre algorithme domine largement les autres algorithmes dans presque tous les réseaux.

Références

1. Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti-Spaccamela, A., and Protasi, M. Complexity and approximation : Combinatorial optimization problems and their approximability properties. Springer Science & Business Media, 2012.
2. Billionnet, A. Different Formulations for Solving the Heaviest K-Subgraph Problem. INFOR : Information Systems and Operational Research 43, 3 (Aug. 2005), 171–186.

3. Brimberg, J., Mladenovic, N., Uroševic, D., and Ngai, E. Variable neighborhood search for the heaviest k-subgraph. Computers Operations Research 36, 11 (2009), 2885–2891.
4. Djeddi, Y., Ait Haddadene, H., and Belacel, N. An extension of adaptive multi-start tabu search for the maximum quasi-clique problem. Computers & Industrial Engineering 132 (June 2019), 280–292.
5. Glover, F., and Laguna, M. Tabu search (Vol. 22). Boston : Kluwer academic publishers, 1997.
6. Letsios, M., Balalau, O. D., Danisch, M., Orsini, E., and Sozio, M. Finding Heaviest k-Subgraphs and Events in Social Media. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) (Dec. 2016), pp. 113–120.
7. Li, W., Liu, Y., Huang, H.-C., Peng, Y., Lin, Y., Ng, W.-K., and Ong, K.-L. Dynamical systems for discovering protein complexes and functional modules from biological networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 4, 2 (2007), 233–250.
8. Macambira, E. M. An Application of Tabu Search Heuristic for the Maximum Edge-Weighted Subgraph Problem. Annals of Operations Research 117, 1 (Nov. 2002), 175–190.
9. Macambira, E. M., and de Meneses, C. N. A GRASP for the maximum edge-weighted subgraph problem. IX Confresso Latino-Iberoamericano de Investigacion Operativa (1998).
10. Rossi, R., and Ahmed, N. The network data repository with interactive graph analytics and visualization. In Twenty-Ninth AAAI Conference on Artificial Intelligence (2015).
11. Singh, H., Kumar, M., and Aggarwal, P. Approximation of Heaviest k-Subgraph Problem by Size Reduction of Input Graph. In Proceedings of 2nd International Conference on Communication, Computing and Networking (2019), C. R. Krishna, M. Dutta, and R. Kumar, Eds., Lecture Notes in Networks and Systems, Springer Singapore, pp. 599–605.
12. Wu, Q., and Hao, J.-K. An adaptive multistart tabu search approach to solve the maximum clique problem. Journal of Combinatorial Optimization 26, 1 (2013), 86–108.

Cooperation-Hierarchization based PSO for digital IIR filter design

Farid Hammou and Kamal Hammouche

Laboratoire Vision Artificielle et Automatique des Systemès (LVAAS). Mouloud Mammeri University of Tizi-Ouzou, Algeria.
faridhammou@yahoo.fr, kamal_hammouche@yahoo.fr

Abstract. In this paper, an improved version of the Particle Swarm Optimization (PSO) algorithm is proposed for the design of digital Infinite Impulse Response (IIR) filters. The proposed algorithm, called Cooperation- Hierarchization based PSO (CHPSO), introduces a new strategy based on cooperation and hierarchization concepts for the updating of the best positions of particles in order to improve the convergence of the PSO algorithm. Moreover, a specific mutation operator is embedded in the PSO algorithm in order to ensure the stability and a minimum phase of the designed IIR filter. Experimental results demonstrate that the proposed CHPSO gives better performance as compared to five others variants of PSO and to five others popular evolutionary optimization algorithms.

Keywords: IIR filter · Particle swarm optimization · Evolutionary optimization techniques · Stability operator

1 Introduction

Digital infinite impulse response (IIR) filters are used as an indispensable tool for a broad range of signal processing applications such as system identification, adaptive filtering, biomedical signal processing [13]. Design of a digital IIR filter consists in determining the coefficients of a realizable function transfer from the prescribed specifications. It can be formulated as a optimization problem with an objective function nonlinear and multi-modal with respect to the filter coefficients. Standard optimization methods like Gradient based algorithms are not suitable for the design of IIR filter. Evolutionary optimization techniques which are conceptually very simple and do not need the calculation of the gradient of the function to be optimize, are then proposed to solve efficiently the IIR filter design problem [7]. The most used ones include Differential Evolution (DE) [17], Real Genetic Algorithms (RGA) [3], Artificial Bee Colony (ABC) [1], Flower Pollination Algorithm (FPA) [14], Cuckoo Search Algorithm (CSA) [2] and Particle Swarm Optimization (PSO) [5,4].

The PSO algorithm is undoubtedly the one that arouses the greatest interest. This popularity is mainly due to its simple concept, its ease to be implemented and its speed to converge towards a high quality of solution. PSO is an iterative algorithm which uses a population of particles, where the position of each particle represents a potential solution. Each particle moves to find a optimal solution by simply adjusting its velocity

and its trajectory according to its personal best position experienced so far and the global best position of the entire swarm.

Although the PSO algorithm has advantage in solving global optimization problem compared with other evolutionary algorithms, it suffers from some problems related to the premature convergence towards the local minima, the stagnation and the revisiting of the same solution several times. Several modifications of the basic PSO algorithm involving different strategies have been proposed in the literature [18] and applied to IIR filter design [5,4,6,8,11]. Unlike the existing methods, we propose in this paper a new strategy for updating the personal best position of each particle in order to improve the convergence speed and the accuracy of the designed IIR filter. The idea behind this strategy is to allow a particle to use the positions experienced by its congeners as its own best position.

Another major problem in the IIR filter design is to ensure the stability of the designed filter. Techniques associated with evolutionary optimization algorithms integrate in the objective function the stability constraints, defined from magnitude of the poles [15] or from Jury criterion [9]. Besides the unstability problem the phase response of the designed filter also constitutes an important problem. To deal with this problem, minimum phase filters can be used because they can simultaneously meet delay and magnitude response constraints, yet generally require fewer computations and less memory than linear phase filters. In this work, we propose an stabilization operator which acts as a mutation operator in order to guarantee both the stability and a minimum phase of each IIR filter provided by each particle of PSO.

The rest of this paper is organized as follows. The digital IIR filter design problem is outlined in section 2. Section 3 gives a brief overview on the PSO algorithm. The modified PSO algorithm for digital IIR filter design is described in section 4. Section 5 presents the simulations results.

2 IIR filter design problem

A digital IIR filter is generally described by the input-output relationship defined by the following discrete difference equation:

$$y(k)+\sum_{i=1}^{P} a_i y(k-i)=\sum_{j=0}^{L} b_j x(k-j) \tag{1}$$

where $x(k)$ and $y(k)$ are the filter's input and output, respectively. P is the filter order which also corresponds to the number of delayed values of the output and L is the number of delayed values of the input such that $P \geqslant L$. a_i $(i = 1,\ 2,\ldots,\ P)$ and b_j $(j = 1,\ 2,\ldots,\ L)$ are the coefficients of IIR filter.
The transfer function of IIR filter is expressed as:

$$H(z)=\frac{B(z)}{A(z)}=\frac{\sum_{j=0}^{L} b_j z^{-j}}{1+\sum_{i=1}^{P} a_i z^{-i}} \tag{2}$$

The main problem of the IIR filter design is to determine the values of the coefficients b_j and a_i from desired characteristics. It can be formulated as an optimization

problem of the cost function $J(\alpha)$, stated as follows:

$$\alpha^* = argminJ(\alpha) \tag{3}$$

where $\alpha = [a_1, a_2, ..., a_P, b_0, b_1, ..., b_L]$ denotes the filter coefficient vector.
The cost function $J(\alpha)$ can be expressed as a error function between the magnitude response $H(\omega)$ and the desired response magnitude $D(\omega)$:

$$J(\alpha) = \|D(\omega) - H(\omega)\|_2, \tag{4}$$

$H(\omega) = H(z)$ with $z = e^{i\omega}$ and $\omega = 2\pi\left(\frac{f}{f_s}\right)$ in $[0\ \pi]$ is the digital frequency, f is analogue frequency and f_s the sampling frequency. $D(\omega)$, corresponding to a ideal transfer function, is defined by:

$$D(\omega) = \begin{cases} 1 \text{ if } \omega \in passband(PB) \\ 0 \text{ if } \omega \in stopband(SB) \end{cases} \tag{5}$$

In order to achieve a higher stop band attenuation, reduce the pass band and stop band ripples and to have more control on the transition width, the minimization of the error function below is adopted [9]:

$$J(\alpha) = \sum_{\omega \in PB} \big| |H(\omega)| - 1| - \delta_p \big| + \sum_{\omega \in SB} ||H(\omega)| - \delta_s| \tag{6}$$

δ_p and δ_s are the pass band ripple and stop band ripple, respectively.
Note that the pass band includes a portion of the transition band and the stop band includes the rest of the transition band. The portions of the transition band depend on the chosen pass band edge and stop band edge frequencies. The minimization of Eq. 6 is performed in this paper by the PSO algorithm.

3 Basic PSO algorithm

The PSO algorithm starts with a population of M particles characterized principally by their positions α_k, $(k = 1, 2, 3, \ldots, M)$ in Q dimension search space which correspond to the candidate solutions of an optimization problem. In the case of IIR filter design, each particle is a set of $Q = L + P + 1$ filter coefficients $\alpha_k = [a_{k1},\ a_{k2},\ a_{k3},\ \ldots,\ a_{kM},$-
$b_{k0},\ b_{k1},\ b_{k2},\ b_{k3},\ \ldots,\ b_{kL}]$. The quality of each particle is evaluated by the error function (*fitness*) $J(\alpha_k)$ (Eq. 6). At time t, each particle k moves in the search space with a velocity $v_k(t)$. The velocity and the position of the k^{th} particle are updated in each iteration according to its previous position $\alpha_k(t)$, its own best position $\beta_k(t)$ found so far and the global best position $\alpha^*(t)$ found by all particles, such that:

$$v_k(t+1) = wv_k(t) + c_1\varphi_1\left(\beta_k(t) - \alpha_k(t)\right) + c_2\varphi_2\left(\alpha^*(t) - \alpha_k(t)\right) \tag{7}$$

$$\alpha_k(t+1) = \alpha_k(t) + v_k(t+1) \tag{8}$$

Acceleration coefficients c_1 and c_2 are the two factors which control the influence of the attraction exerted by the best personal experience β_k of each particle and by the

attraction of the global best position α^*, respectively. φ_1 and φ_2 are two uniformly distributed random numbers, independently generated within [0, 1]. The inertia weight ω controls how much the particles tend to follow their current direction $\alpha(t)$ compared to the memorized positions $\beta_k(t)$ and $\alpha^*(t)$. In the original PSO algorithm, ω is set to 1. Several improved versions of the PSO algorithm are proposed in the literature following the basic steps which are summarized in Fig. 1.

4 Modified PSO algorithm for IIR filter design

In this section, we propose a new modification for the PSO algorithm, called Cooperation-Hierarchization PSO (CHPSO), to improve its performance in the framework of the digital IIR filter design. This modification concerns the updating way of the personal best positions of the particles. Moreover, in order to ensure the stability and the minimum phase of the IIR filter produced by each particle, we propose to embed in the CHPSO a specific mutation operator.

4.1 New updating strategy of the personal best position of particles

In the basic PSO as well as in all most of its variants, the new personal best position $\beta_k(t+1)$ of each particle k at the next time step, $t+1$, is updated by comparing its new position $\alpha_k(t+1)$ and its previous own best position $\beta_k(t)$:

$$\beta_k(t+1) = \begin{cases} \alpha_k(t+1) & \textit{if } J(\alpha_k(t+1)) < J(\beta_k(t)) \\ \beta_k(t) & \textit{else} \end{cases} \tag{9}$$

The global best position, $\alpha^*(t+1)$, at time step $t+1$, is updated as follows:

$$\alpha^*(t+1) = \begin{cases} \beta_{k^*}(t+1) & \textit{if } J(\beta_{k^*}(t+1)) < J(\alpha^*(t)) \\ \alpha_k(t) & \textit{else} \end{cases} \tag{10}$$

where $k^* = argmin_{k=1,...,M}\,[J(\beta_k(t+1))]$.

In CHPSO algorithm, we propose to update the personal best position of a particle by taking into account all old personal best positions and all new positions of all particles. For this purpose, we proceed as follows:
First, we gather into a vector $Y = [y_1,\ y_2,\ ...,\ y_{2M}]^T$ of size $(2M)$ all personal best positions $\beta_k(t)$ $(k = 1,\ 2,\ 3,\ ...,\ M)$ of all particles, at the time step t, and the new positions $\alpha_k(t+1)$ $(k = 1,\ 2,\ 3,\ ...,\ M)$ of all particles generated at the next time step $t+1$:

$$Y = [\beta_1(t), \beta_2(t), ..., \beta_M(t), \alpha_1(t+1), \alpha_2(t+1), ..., \alpha_M(t+1)]^T \tag{11}$$

The k^{th} element of this vector is:

$$y_k = \begin{cases} \beta_k(t) & \textit{if } k \le M \\ \alpha_{k-M}(t+1) & \textit{if } k > M \end{cases}, k = 1\ ,2\ ,\ ...,\ 2M \tag{12}$$

Then, the $2M$ elements of the vector $Y(t+1)$ are arranged in ascending order according to their fitness leading to the following vector:

$$Y = \left[y_{(1)},\ y_{(2)},\ ...,\ y_{(2M)}\right]^T \tag{13}$$

The M new personal best positions of the particles are finally chosen as the M first ordered elements of Y:

$$\beta_k(t+1) = y_{(k)},\ k = 1,\ 2,\ ...,\ M \tag{14}$$

In this updating strategy of personal best positions, the first particle memorizes the first best position found by all particles, the second particle memorizes the second best position found by all particles and so on until the last particle which stores the M^{th} best position found by all particles. This means that the particles cooperate with each other by exchanging their experiences, because a particle can consider a position occupied by one of its congeners as its own best experience, and that the positions experienced by a particle can serve as personal best experiences to other particles. Moreover, the particles are hierarchized according to their personal best experiences that have been transmitted to them, since a particle receives from its congeners one position to memorize according to the rank that it occupies in the population.
With this new strategy, the global best position $\alpha^*(t+1)$, at time step $t+1$, can be updated simply from the personal best position of the more experienced particle, i.e., the first particle in the population:

$$\alpha^*(t+1) = \begin{cases} \beta_1(t+1) & if\ J(\beta_1(t+1)) < J(\alpha^*(t)) \\ \alpha^*(t) & else \end{cases} \tag{15}$$

As CHPSO acts on the cognitive component of PSO, we apply the concept of time-varying acceleration coefficients [10]. Moreover, unlike the basic PSO algorithm, the random numbers φ_1 and φ_2 that determine the weight between the cognitive attraction and social attraction are dependently generated as $\varphi_2 = 1 - \varphi_1$.

4.2 Mutation operator

Some coefficients α_k provided by each particle does not guarantee that the corresponding IIR filter $H(z)$ is stable. To deal with this problem, we propose a stabilization operator able to maintain the poles of each filter $H(z)$ into unit circle.
Let $A(z) = 1 + \sum_{i=1}^{P} a_i z^{-i}$ be the characteristic polynomial of $H(z)$ where $a_1,\ a_2,\ ...,\ a_P$ are the coefficients found by a particle of the swarm and let $Z = [z_1,\ z_2,\ ...,\ z_P]^T$ be the roots of the corresponding characteristic polynomial $A(z)$. These roots can be put into unit circle according the following equation:

$$Z^{new} = [1 - (0.5\ sign\,(|Z| - 1) + 1)]\,z + \frac{(0.5\ sign\,(|Z| - 1) + 1)}{conj(Z)} \tag{16}$$

where $sign(.)$ is the signum function and $conj(Z)$ the complex conjugate of Z. This equation only consists in reflecting polynomial roots located outside the unit circle to the inside.
From the new roots $Z^{new} = [z_1^{new}, z_2^{new}, ..., z_P^{new}]^T$, we can construct a new characteristic polynomial:

$$A^{new}(z) = \left(z^{-1} - z_1^{new}\right)\left(z^{-1} - z_2^{new}\right)...\left(z^{-1} - z_P^{new}\right) \tag{17}$$

The new coefficients $a_1^{new}, a_2^{new}, ..., a_P^{new}$ of $H(z)$ are then identified from the polynomial $A^{new}(z)$. They replace the old coefficients of the considered particle. This operator acts as a mutation operator and has no effect on the poles located into unit circle or on an already stable filter. This operator is also applied with same manner on the coefficients b_j provided by each particle, which correspond to numerator polynomial $B(z) = \sum_{j=0}^{L} b_j z^{-j}$ of $H(z)$. With this operator, the poles and zeros of each filter provided by each particle are maintained into unit circle leanding to a stable filter with a minimum phase.

The flowchart of the proposed CHPSO algorithm for IIR filter design is illustrated in Fig. 1.

5 Experimental results

In order to evaluate the performance of the proposed CHPSO algorithm and demonstrate its effectiveness for the stable and minimum phase IIR filter design, two 8^{th} order filter types: Lowpass (LP) and Bandpass $(BP)$with design specifications, given in table 1, are considered [17].

Table 1. Design specifications of LP and BP IIR filters

Type of filter	Pass band ripple (σ_p)	Stop band ripple (σ_s)	Pass band normalized edge frequencies (ω_p)	Stop band normalized edge frequencies (ω_s)
LP	0.01	0.001	0.45	0.50
BP	0.01	0.001	0.35 and 0.65	0.30 and 0.70

Table 2. Control parameters of the algorithms used in the simulations

Algorithm	Parameters	Source
CHPSO	v^{min}=0.01; v^{max}=1; c^{min}=0.2; c^{max}=1.2; w=0.7298.	Our method
WPSO	v^{min}=0.01; v^{max}=1; $c_1 = c_2$=2.05; w^{max}=1; w^{min}=0.4.	[18]
CPSO	V_{Craz}=0.0001; P_{cr}=1; $c_1 = c_2 = 2.05$.	[11]
TVACPSO	v^{min}=0.01; v^{max}=1; c^{min}=0.5; c^{max}=2.5; w^{max}=0.9; w^{min}=0.4.	[10]
QPSO	α^{min}=0.5; α^{max}=1.	[6]
PSOGSA	$c_1 = c_2$=2.0; w^{max}=0.9; w^{min}=0.4; G_0=100; α=20.	[8]
DE	C_r=0.3; C_f=0.5.	[17]
RGA	Crossover rate=1; Crossover: Two-point crossover; Mutation rate= 0.01; Mutation: Roulette; Selection probability=1/3.	[16]
ABC	Limit=40; a=0.9; b=0.1.	[1]
FPA	P_{Switch}=0.8.	[14]
CSA	P_α=0.05; λ=1.	[2]

The results provided by the CHPSO algorithm are compared with those obtained by five improved versions of PSO, namely adaptive inertia Weight PSO (WPSO) [12],

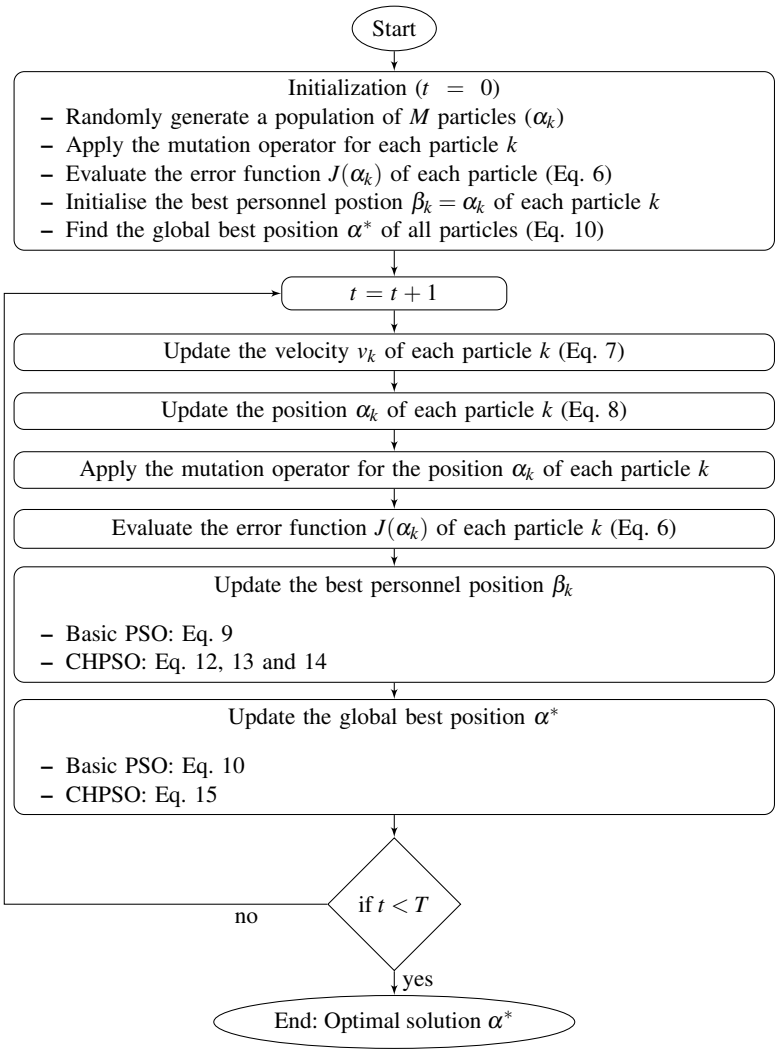


Fig. 1. Flowchart of the basic PSO and the CHPSO algorithm for IIR filter design

Time Varying Acceleration Coefficients PSO (TVACPSO) [10], Quantum-behaved PSO (QPSO) [6], Craziness based PSO (CPSO) [11] and the hybrid method based on PSO and GSA (PSOGSA) [8]. Five other popular evolutionary methods including RGA [3], DE [17], ABC [1], FPA [14], and CSA [2] are also used in this comparison. All algorithms are implemented with the parameters given in Table 2. These parameters are adjusted to achieve the best results. For a fair comparison, the size of population M and the total number of iterations T_{max} are the same for all algorithms. The mutation opera-

tor is also embedded in the ten algorithms used for comparison.
The value of the best fitness $J(\alpha^*)$ corresponding to the best solution is used as comparative criterion. Of course, the smaller the fitness value, the better the algorithm is. Note that, the fitness function (Eq. 6) is computed over 128 samples uniformly spaced within the frequency domain $[0, \pi]$. The maximal attenuation in the stop band is also used to evaluate the quality of the designed filter. The higher the maximal attenuation in the stop band, the better the filter is. Additional results are presented in order to investigate the convergence speed of each algorithm.

5.1 Analysis of magnitude responses

As the evolutionary algorithms are of stochastic type, they are run 30 times. Table 3 reports the best, the worst, the mean and the standard deviation of the fitness values over 30 runs, for the two types of IIR filters. We can see that CHPSO gives the best results both in terms of accuracy (mean fitness) and robustness (small standard deviation, i.e. similar results of repeated runs). Indeed, the mean fitness of CHPSO is always the smallest. In the case of LP filter, for example, it is about two and a half times smaller than the second best mean fitness, achieved by the WPSO. The standard deviation obtained by the proposed CHPSO is slightly larger than the values provided by the algorithm CSA. However, we note that the best results found by CHPSO during the 30 runs remain always better than the best results found by others algorithms and even its worst results often remain better than the best results of others algorithms. Figure 2 shows the plots

Table 3. Statistical results of the fitness for 8^{th} order LP and BP IIR filters

	CHPSO	WPSO	CPSO	QPSO	TVACPSO	PSOGSA	DE	RGA	ABC	FPA	CSA
LP filter											
Mean	**0.4644**	2.0452	6.4340	4.6960	1.9467	4.7263	6.6969	8.8257	3.8257	3.6897	2.0478
Std	**0.1359**	0.6930	1.5575	1.2420	0.7340	1.6799	0.4697	2.3013	0.3425	0.4337	0.1398
Best	**0.2233**	0.9927	4.1874	2.5150	0.7523	1.7126	5.4821	5.0692	3.0981	2.9816	1.7961
Worst	**0.7340**	3.9226	10.090	7.5844	4.2111	7.8954	7.4685	14.214	4.4191	4.5527	2.3457
PB filter											
Mean	**0.8999**	3.4004	5.8599	3.8985	2.7214	6.0018	5.5479	8.6711	4.0794	3.6447	2.4607
Std	0.2968	0.9134	1.9413	1.3315	0.4893	1.4061	0.3689	2.2474	0.4392	0.1826	**0.0474**
Best	**0.7831**	2.3166	3.5504	2.1248	2.2285	3.6627	4.9538	4.5995	3.3327	3.2857	2.3593
Worst	**2.1418**	5.5463	12.571	7.1047	4.8756	9.2385	6.4939	16.688	4.9024	4.2052	2.5735

of the normalized gains $H(\omega)$ computed with the coefficients corresponding to the best solution obtained by each algorithm. For a better illustration of the results, the first column of Fig. 2 displays the comparative results between CHPSO and the five variants of PSO, while the second column displays comparative results between CHPSO and the five other evolutionary algorithms.
The gain plots in dB of each designed 8^{th} order IIR filter are represented in Fig. 3. They reveal that for the two designed 8^{th} order IIR filters, CHPSO yields much higher attenuation in the stop band and much lower ripples in the pass band and the stop band.

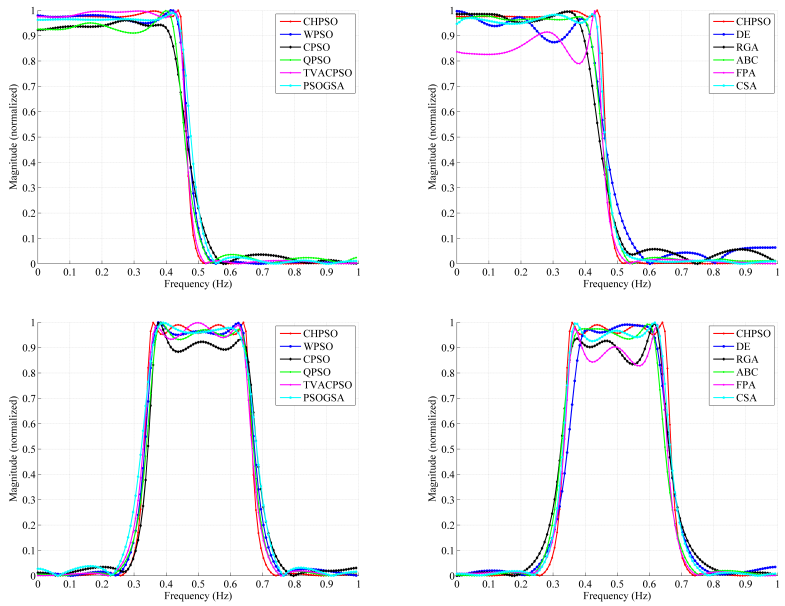


Fig. 2. Normalized gain pots of 8^{th} order LP and BP IIR filters.

Table 3 shows the maximum, the mean and the standard deviation of the maximal attenuation values in the stop band over 30 runs. It is observed that the greatest maximal attenuation and the greatest average value of the maximal attenuations in the stop band are always achieved by CHPSO.

Table 4. Statistics of maximum attenuations in the stopped band (dB) obtained during the 8^{th} LP and PB filter design.

Filter	CHPSO	WPSO	CPSO	QPSO	TVACPSO	PSOGSA	DE	GA	ABC	FPA	CSA
LP	**43.413**	41.632	29.016	28.868	38.664	31.539	27.087	25.009	31.836	35.250	37.625
PB	**40.148**	31.391	29.462	35.102	34.211	28.571	29.165	37.992	33.914	36.141	34.805

We can also easily verify that, thanks to mutation operator, the stability and minimum phase characteristics are ensured for the two filters since the poles and zeros provided by CHPSO are located inside the unit circle (Fig. 3).

5.2 Convergence speed

Figure 5 shows how the mean fitness varies with the iteration number. At each iteration, the fitness obtained by the CHPSO is much lower than those obtained by the other

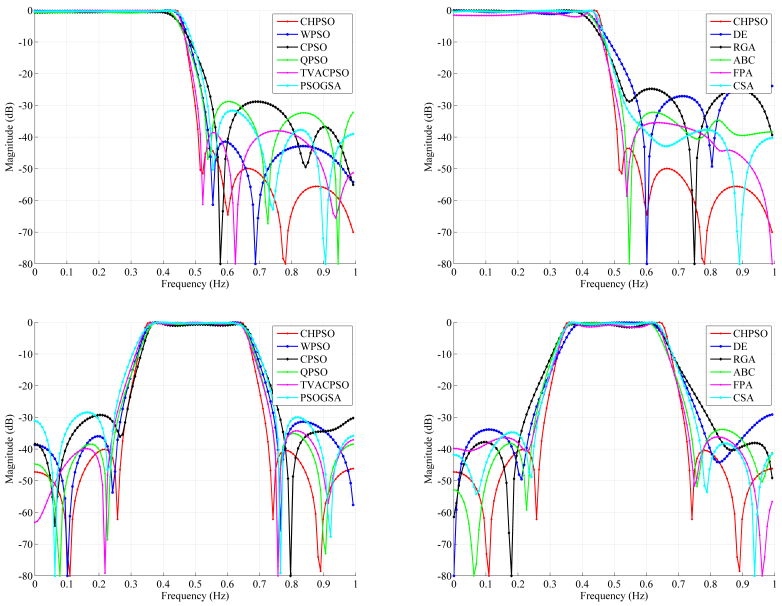


Fig. 3. Gain plots in dB of 8^{th} order LP and BP IIR filters.

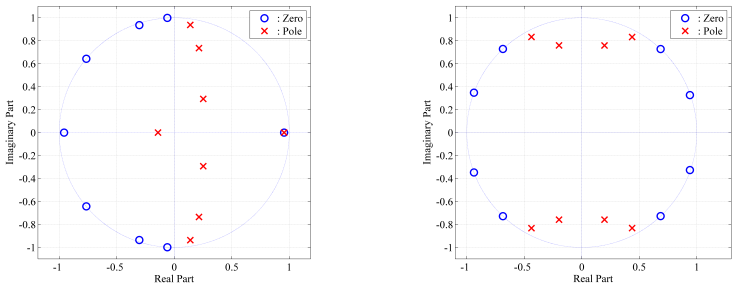


Fig. 4. Pole/Zero plots using CHPSO for designed 8^{th} order LP and BP IIR filters

algorithms. Moreover, the fitness curve decreases faster in the case of CHPSO comparatively to all other algorithms. Therefore, the iteration number necessary to ensure the convergence of CHPSO is lower than those of other algorithms.

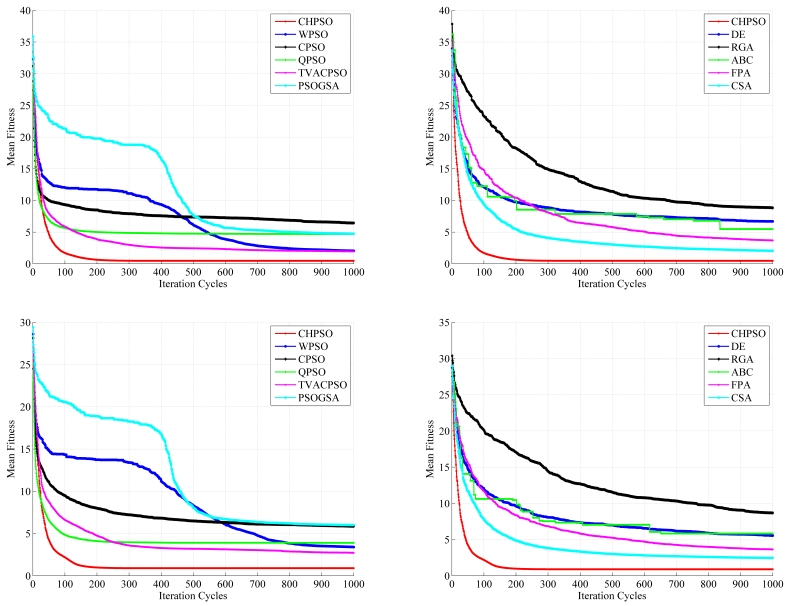


Fig. 5. Convergence profiles of 8^{th} order LP and BP IIR filters.

6 Conclusion

In this paper, a new modified PSO algorithm is proposed for the IIR filter design. This improved version of PSO stands out from the other existing variants by the way of updating the personal best position of the particles. It introduces a novel concept of co-operation between particles and that of their hierarchization according to their personal best positions. This new modification is easy to implement and does not require any additional control parameter with respect to basic PSO algorithm.
Moreover, a new mutation operator is proposed to guarantee the stability and the minimal phase of the designed filter. This operator is perfectly suitable for the IIR filter design by evolutionary algorithms.
Experiments on the design of 8^{th} order LP and BP IIR filters reveal the efficiency of the proposed PSO algorithm in terms of accuracy, robustness and convergence speed, and attest of the good behavior of the designed filters in terms of stability, minimum phase, maximal attenuation in the stop band and low ripples in the stop and pass bands.
The tests carried out also show that the proposed CHPSO algorithm outperforms five improved variants of the PSO algorithm and five other popular evolutionary algorithms in all respect digital filter design.

References

1. Agrawal, N., Kumar, A., Bajaj, V.: Optimized design of digital IIR filter using artificial bee colony algorithm. In: 2015 International Conference on Signal Processing, Computing and Control (ISPCC). pp. 316–321 (2015)
2. Agrawal, N., Kumar, A., Bajaj, V.: Digital IIR filter design with controlled ripple using cuckoo search algorithm. In: 2016 International Conference on Signal and Information Processing (IConSIP). pp. 1–5. IEEE (2016)
3. Chang, W.D.: Coefficient estimation of IIR filter by a multiple crossover genetic algorithm. Computers & Mathematics with Applications **51**(9-10), 1437–1444 (2006)
4. Chauhan, R.S., Arya, S.K.: An optimal design of IIR digital filter using particle swarm optimization. Applied Artificial Intelligence **27**(6), 429–440 (2013)
5. Chen, S., Luk, B.L.: Digital IIR filter design using particle swarm optimisation. International Journal of Modelling, Identification and Control **9**(4), 327–335 (2010)
6. Fang, W., Sun, J., Xu, W.: A new mutated quantum-behaved particle swarm optimizer for digital IIR filter design. EURASIP Journal on Advances in Signal Processing **2009**(1), 367465 (2010)
7. Gotmare, A., Bhattacharjee, S.S., Patidar, R., George, N.V.: Swarm and evolutionary computing algorithms for system identification and filter design: A comprehensive review. Swarm and Evolutionary Computation **32**, 68–84 (2017)
8. Jiang, S., Wang, Y., Ji, Z.: A new design method for adaptive IIR system identification using hybrid particle swarm optimization and gravitational search algorithm. Nonlinear Dynamics **79**(4), 2553–2576 (2014)
9. Kaur, R., Patterh, M.S., Dhillon, J.: A new greedy search method for the design of digital IIR filter. Journal of King Saud University - Computer and Information Sciences **27**(3), 278–287 (2015)
10. Ratnaweera, A., Halgamuge, S.K., Watson, H.C.: Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients. IEEE Transactions on Evolutionary Computation **8**(3), 240–255 (2004)
11. Saha, S., Kar, R., Mandal, D., Ghoshal, S.: IIR filter design with craziness based particle swarm optimization technique **60**, 1628–1635 (2011)
12. Shi, Y., Eberhart, R.C.: A modified particle swarm optimizer. In: Proceedings of IEEE International Conference on Evolutionary Computation. pp. 69–73. IEEE Computer Society, Washington, DC, USA (1998)
13. Singh, R., Verma, H.K.: Teaching–learning-based optimization algorithm for parameter identification in the design of iir filters. Journal of The Institution of Engineers (India): Series B **94**(4), 285–294 (Dec 2013)
14. Singh, S., Ashok, A., Rawat, T.K., Kumar, M.: Optimal IIR system identification using flower pollination algorithm. In: 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES). pp. 1–6. IEEE (jul 2016)
15. Słowik, A.: Comparative study on bio-inspired global optimization algorithms in minimal phase digital filters design. In: Intelligent Information and Database Systems, vol. 8398, pp. 217–226. Springer International Publishing (2014)
16. Suman Kumar Saha, Rajib Kar, D.M., Ghoshal, S.P.: Optimal IIR filter design using novel particle swarm optimization technique. International Journal of Circuits, Systems and Signal Processing **6**(2), 151–162 (2012)
17. Upadhyay, P., Kar, R., Mandal, D., Ghoshal, S.P.: An efficient differential evolution with wavelet mutation algorithm for optimal IIR filter design. International Journal of Bio-Inspired Computation **6**(5), 350–367 (2014)
18. Zhang, Y., Wang, S., Ji, G.: A comprehensive survey on particle swarm optimization algorithm and its applications. Mathematical Problems in Engineering **2015**, 1–38 (2015)

Métriques sous-Finslériennes en dimension trois : Un cas d'étude ⋆

F.Harrache[1,2], F.C.Chittaro[2], M.Aidene[1] et J.-P.Gauthier[2]

[1] Laboratoire de Conception et Conduite des Systèmes de Production, Université Mouloud MAMMERI,Tizi-Ouzou,Algérie
aidene@ummto.dz, fazia.harrache@gmail.com
[2] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
francesca.chittaro@univ-tln.fr, gauthier@univ-tln.fr

Résumé : Dans ce papier, nous étudions la structure locale de la géométrie sous-Finslérienne associée à deux champs de vecteurs dans $\mathbb{R}^3$ et à la norme L^1 des champs de contrôle. Nous construisons une forme normale pour les champs de vecteurs, inspirée par la symétrie du cas nilpotent. En utilisant le principe de maximum de Pontryagin, nous donnons une description des temps de commutation, des temps conjugués ainsi que des lieux de coupure *(cut loci)* pour des petites géodésiques dans le cas générique.

1 Introduction

Soit M une variété lisse connexe de dimension 3 ; considérons deux champs de vecteurs lisses f et g dans M, qui satisfont la condition de rang en tout point $x \in M$: $Lie_x(f,g) = T_xM$.

Soit $\|\cdot\|$ une norme quelconque dans $\mathbb{R}^2$ (par exemple, la norme euclidienne standard ou une p-norme) ; alors la paire $(\{f,g\}, \|\cdot\|)$ munit M d'une structure métrique, de la manière suivante : la distance entre deux points x_0 et $x_1 \in M$ est définie comme le minimum de la fonctionnelle

$$J(\boldsymbol{u}) = \int_0^1 \|\boldsymbol{u}(t)\|\, dt \tag{1}$$

sur toutes les fonctions $\boldsymbol{u} : [0,1] \to \mathbb{R}^2$ telles que les trajectoires du système de contrôle

$$\dot{\boldsymbol{\xi}}(t) = (u_1(t)f + u_2(t)g) \circ \boldsymbol{\xi}(t) \quad \boldsymbol{u} = (u_1, u_2) \tag{2}$$

satisfont $\boldsymbol{\xi}(0) = x_0$ et $\boldsymbol{\xi}(1) = x_1$ (si le minimum n'existe pas, la distance est fixée à $+\infty$). Quand $\|\cdot\|$ est une norme euclidienne standard, ce problème de contrôle optimale coïncide avec le très célèbre problème sous-Riemannien (voir par exemple [3,11] et les références qui y figurent).

⋆. Les auteurs ont été partiellement soutenus par le projet "Partenariat Hubert Curien - Tassili 2015" *PHC15MDU941 – Critères non-lisses en contrôle optimal.* F.C.C. a été partiellement soutenue par le projet *CARTT–IUT de Toulon.*

Dans ce papier, nous nous intéressons au problème L^1 sous-Finslérienne : plus précisément, nous étudions le cas où $\|\boldsymbol{u}\| = |u_1| + |u_2|$; dans ce cas, $J(\boldsymbol{u}(\cdot))$ correspond à la norme L^1 de la fonction de contrôle $\boldsymbol{u}(\cdot)$. La structure de la métrique induite sur M est appelée métrique *sous-Finslérienne*.

Afin de trouver les courbes qui minimisent J entre deux conditions initiales $\boldsymbol{x}_0$ et $\boldsymbol{x}_1$, nous faisons appel au Principe de Maximum de Pontryagin (PMP), qui fournit une famille de courbes minimisantes, appelées *géodésiques*, comme solutions d'un certain système Hamiltonien associé au problème. Comme le PMP est une condition nécessaire d'optimalité, les géodésiques joignant $\boldsymbol{x}_0$ à $\boldsymbol{x}_1$ ne réalisent pas toutes le minimum de (1). Pour sélectionner celles qui sont optimales, d'autres conditions d'optimalité s'imposent. Les conditions d'optimalité du second ordre sont très utiles pour prouver l'optimalité *locale* forte d'une géodésique ; d'autre part, pour trouver les géodésiques *globalement* optimales, il est nécessaire de comparer les coûts réalisés par chaque trajectoire joignant $\boldsymbol{x}_0$ à $\boldsymbol{x}_1$ qui satisfont le PMP, même si leurs graphes sont très éloignés les uns des autres.

Dans le cas nilpotent, ce problème a été déjà étudié par Barilari et. al en [6], par Breuillard et Le Donne en [7] et par Lokutsievskiy en [10].

Le même problème, dans le cas générique, a été étudié par Ali et Charlot en [5] ; les auteurs ont donné la synthèse optimale locale (c'est-à dire, à petite distance d'un certain point $\boldsymbol{x}_0$), sous l'hypothèse générique $span\{f(\boldsymbol{x}), g(\boldsymbol{x}), [f,g](\boldsymbol{x})\} = T_{\boldsymbol{x}}M$ pour tout $\boldsymbol{x} \in M$. Dans ce papier, nous traitons ce même problème ; par rapport à [5], nous proposons une autre forme normale de la distribution, qui prend en considération la symétrie de cas nilpotent. Cette nouvelle description a deux avantages : d'une part, elle réduit (au moins) à moitié les calculs qui nécessitent l'intégration du système Hamiltonien, pour trouver les temps de commutation, les temps conjugués et toutes les intersections entre les trajectoires ; (grâce au lemme 1-2), d'autre part, elle permet de classer efficacement les différentes formes des lieux de coupure, qui dépendent des valeurs de trois invariants de la distribution, qui sont notés A, D_1 et D_2.

Les étapes principales de la méthode sont les suivantes : en premier lieu, le problème de minimisation de la longueur est transformé en un problème en temps minimal ; ensuite, suivant la même approche utilisée par Agrachev et al. en [4] et Chakir et al. en [8] pour le cas sous-Riemannien, nous utilisons une reparamétrisation du temps qui nous permet d'écrire le système générique comme une perturbation de système nilpotent : il sera donc possible d'intégrer le système approché ; enfin, nous calculons les temps conjugués et étudions l'optimalité des géodésiques courtes. Nos allons obtenir une caractérisation des lieux de coupure *(cut locus)* des géodésiques *bang-bang*, dans le cas générique pour lesquels les trois invariants D_1, D_2 et A sont non nuls, et au moins un entre D_1 et D_2 est négatif.

La structure de ce papier est la suivante : dans la partie 2 nous énonçons le problème, et nous rappelons les résultats du cas nilpotent. Dans la partie 3, nous construisons la forme normale et écrivons le développement du système Hamiltonien perturbé, en fonction d'un petit paramètre ρ_0 (proportionnel au

rayon de la sphère). Dans la partie 4 les principaux résultats (calcul des temps conjugués et description du lieu de coupure) sont présentés.

En raison de la contrainte de la longueur de papier, nous ne fournissons pas de preuves et de calculs détaillés ; néanmoins, nous donnons les éléments principaux de notre construction.

2 Préliminaires

2.1 Enoncé du problème et analyse préliminaire

Soient f, g deux champs de vecteurs lisses sur M, tels que

$$\operatorname{span}\{f(\boldsymbol{x}), g(\boldsymbol{x}), [f,g](\boldsymbol{x})\} = T_{\boldsymbol{x}}M, \tag{3}$$

pour tout point $\boldsymbol{x} \in M$. Tout d'abord, nous remarquons que, grâce au théorème de Chow-Rashevsky, le système (2) est contrôlable sur M. En appliquant une reparamétrisation adéquate du temps, le problème en question peut être réécrit comme le problème en temps minimal suivant :

$$\text{minimiser } T \text{ sous les contraintes} \tag{4a}$$
$$\dot{\boldsymbol{\xi}}(t) = u_1(t) f(\boldsymbol{\xi}(t)) + u_2(t) g(\boldsymbol{\xi}(t)), \tag{4b}$$
$$\boldsymbol{\xi}(0) = \boldsymbol{x}_0, \quad \boldsymbol{\xi}(T) = \boldsymbol{x}_1, \tag{4c}$$
$$\boldsymbol{u}(\cdot) : [0,T] \to Q \quad \text{mesurable}, \tag{4d}$$

où $Q = \{(u_1, u_2) \in \mathbb{R}^2 : |u_1| + |u_2| \leq 1\}$ et T est le temps final libre. Une simple application du théorème de Filippov garantit que le minimum est atteint pour tout couple $(\boldsymbol{x}_0, \boldsymbol{x}_1)$ appartenant à un sous-ensemble suffisamment petit de M.

Pour trouver les trajectoires optimales, nous appliquons le Principe de Maximum de Pontryagin : on note $\boldsymbol{p} = (p, q, r)$ le vecteur adjoint (en coordonnées locales) et on écrit le Hamiltonien comme il suit :

$$h(\boldsymbol{p}, \boldsymbol{x}, \boldsymbol{u}) = u_1 F(\boldsymbol{p}, \boldsymbol{x}) + u_2 G(\boldsymbol{p}, \boldsymbol{x}),$$

avec $F(\boldsymbol{p}, \boldsymbol{x}) = \langle \boldsymbol{p}, f(\boldsymbol{x}) \rangle$ et $G(\boldsymbol{p}, \boldsymbol{x}) = \langle \boldsymbol{p}, g(\boldsymbol{x}) \rangle$. Le PMP affirme que, si $\widehat{\xi}$ est une solution optimale pour le problème en temps minimal, et $\widehat{\boldsymbol{u}}(\cdot)$ est sa fonction de contrôle associée, alors il existe une courbe Lipschitzienne non nulle $\widehat{\boldsymbol{p}}(\cdot)$ telle que :

$$\dot{\widehat{\boldsymbol{p}}}(t) = -\frac{\partial h}{\partial \boldsymbol{x}}(\widehat{\lambda}(t), \widehat{\boldsymbol{u}}(t)) \quad \dot{\widehat{\xi}}(t) = \frac{\partial h}{\partial \boldsymbol{p}}(\widehat{\lambda}(t), \widehat{\boldsymbol{u}}(t))$$
$$h(\widehat{\lambda}(t), \widehat{\boldsymbol{u}}(t)) = \max_{\boldsymbol{v} \in Q} h(\widehat{\lambda}(t), \boldsymbol{v}) \qquad \text{p.p. } t \tag{5a}$$
$$h(\widehat{\lambda}(t), \widehat{\boldsymbol{u}}(t)) \equiv \nu, \tag{5b}$$

où $\widehat{\lambda}(t) = (\widehat{\boldsymbol{p}}(t), \widehat{\xi}(t))$ et $\nu \in \{0, 1\}$. Les trajectoires qui satisfont le PMP sont appelées *géodésiques*, et les paires $\lambda(\cdot) = (\boldsymbol{p}(\cdot), \boldsymbol{\xi}(\cdot))$ satisfaisant le PMP sont

appelées *extrémales*. Par le PMP et l'équation (3), on voit que ν est différent de zéro, donc nous fixons $\nu = 1$. Dès maintenant, on considère que les trajectoires qui satisfont le PMP, sans rappeler qu'elles sont les projections d'une certaine extrémale.

Les équations (5a)-(5b) impliquent que le contrôle associé à une extrémale prend des valeurs au bornes de Q, i.e., les contrôles internes ne sont pas considérés. La valeur du contrôle est déterminée par les valeurs relatives à F et G : soit $I \subset [0, T]$ un intervalle et λ une extrémale de (4) ; alors

— si $|F(\lambda(t))| \neq |G(\lambda(t))|$ $\forall t \in I$, dans I le contrôle prend des valeurs sur l'un des sommets de Q (par exemple, si $F(\lambda(t)) > |G(\lambda(t))|$, alors $\boldsymbol{u} = (1, 0)$). Dans ce cas, $\lambda|_I$ est dit *arc bang régulier*.

— si $|F(\lambda(t)| = |G(\lambda(t))|$ $\forall t \in I$, le contrôle prend des valeurs sur l'un des côtés de Q, et, en particulier, il n'est pas déterminé. En effet, si par exemple $F(\lambda(t)) = G(\lambda(t)) > 0$ $\forall t \in I$, tout contrôle de la forme $(\alpha, 1-\alpha), \alpha \in [0, 1]$, réalise le maximum de l'équation (5a). Dans ce cas, $\widehat{\lambda}|_I$ est un *arc singulier*.

Quand une extrémale croise de façon transversale (non tangente) une des surfaces $\{F = G\}$ ou $\{F = -G\}$, le contrôle commute, autrement dit, sa valeur saute d'un sommet de Q à un autre ; en particulier, par continuité des extrémales et par le fait que F et G ne s'annulent pas simultanément le long d'une extrémale (par l'équation (5b)), un contrôle qui satisfait le PMP ne peut basculer que d'un sommet de Q à un sommet voisin. Pour ces raisons, F et G sont appelées *fonctions de commutation*. Leurs dérivées le long d'une extrémale sont données par

$$\frac{d}{dt}F(\lambda(t)) = -u_2(t)\Theta(\lambda(t)), \quad \frac{d}{dt}G(\lambda(t)) = u_1(t)\Theta(\lambda(t)),$$

où $\Theta(\boldsymbol{p}, \boldsymbol{x}) = \{F, G\}(\boldsymbol{p}, \boldsymbol{x}) = \langle \boldsymbol{p}, [f, g](\boldsymbol{x})\rangle$.

Notation : Dans ce qui suit, on va noter les géodésiques bang-bang selon leur vecteur adjoint initial : $\boldsymbol{\gamma_f}$ et $\boldsymbol{\gamma_{-f}}$ représentent les géodésiques associées aux extrémales λ qui satisfont $F(\lambda(0)) > G(\lambda(0)) \geq 0$ et $F(\lambda(0)) < G(\lambda(0)) \leq 0$, respectivement ; telles trajectoires commencent par un contrôle égale à $\boldsymbol{u} = (1, 0)$ et $\boldsymbol{u} = (-1, 0)$, respectivement. De façon analogue, $\boldsymbol{\gamma_{\pm g}}$ représente une géodésique dont l'extrémale λ satisfait $\pm G(\lambda(0)) > \pm F(\lambda(0)) \geq 0$; en particulier, ces trajectoires démarrent avec un contrôle égale à $\boldsymbol{u} = (0, \pm 1)$.

2.2 Système de Heisenberg

Le système le plus simple qui satisfait la condition (3) correspond au groupe de Heisenberg. Comme annoncé dans l'Introduction, le problème L^1 associé à cette distribution a été étudié en [6,7] ; puisqu'il représente un point de départ pour l'étude du cas générique, il convient de rappeler les propriétés essentielles de sa synthèse.

Soit $M = \mathbb{R}^3$ et soient f et g les champs de vecteurs

$$f = \begin{pmatrix} 1 \\ 0 \\ -y/2 \end{pmatrix} \qquad g = \begin{pmatrix} 0 \\ 1 \\ x/2 \end{pmatrix}.$$

Considérons le problème (4) avec $\boldsymbol{x}_0 = (0,0,0)$, pour tout point final. Puisque le système de contrôle ne dépend pas de z, le vecteur adjoint r est constant le long de chaque extrémale. En outre, $\Theta(\boldsymbol{p},\boldsymbol{x}) \equiv r$, ce qui a deux conséquences :

— Les arcs singuliers sont caractérisés par $r \equiv 0$; par conséquent, si une extrémale contient un arc singulier, alors l'extrémale entière est singulière.

— Étant donné un enchaînement quelconque d'arcs bang-bang, la séquence des contrôles associés est déterminée uniquement par le contrôle initial et le signe de r : en effet, si $r > 0$ (respectivement, $r < 0$), les contrôles suivent les sommets de Q dans le sens antihoraire (respectivement, sens horaire).

Nous remarquons de plus que le problème présente une symétrie discrète : il est invariant par rotations de $k\pi/2,\ k \in \mathbb{Z}$, dans le plan xy ; en plus, il s'avère que la sphère est symétrique par rapport au plan xy, et que les géodésiques atteignant les z négatifs sont les projections des extrémales avec $r \leq 0$. Pour ces raisons, dans cette partie on présente seulement les extrémales dont le vecteur adjoint initial $\boldsymbol{p}(0) = (p_0, q_0, r)$ satisfait $p_0 = 1$, $q_0 \in [-1,1]$ et $r \geq 0$; en effet, toutes les autres extrémales peuvent être récupérées de celles-ci en appliquant une transformation adéquate.

En premier lieu, nous considérons le cas où $q_0 \in [-1,1)$ et $r > 0$: telles extrémales sont associées à la trajectoire $\boldsymbol{\gamma_f}$; en effet, puisque $F(\lambda(0)) = 1 > |q_0| = |G(\lambda(0))|$, alors le contrôle associé à l'extrémale est $(1,0)$ dans l'intervalle $[0,T_1]$, où T_1 est le plus petit temps (positif) qui satisfait $F(T_1) = G(T_1)$; des calculs nous donnent $T_1 = \frac{1-q_0}{r}$. Après ce temps, le contrôle prend la valeur $(0,1)$, jusqu'au temps $T_2 > T_1$ satisfaisant $F(T_2) = -G(T_2)$; ce temps est donné par $T_2 = T_1 + \Delta T$, avec $\Delta T = 2/r$. Après T_2, le contrôle commute tous les ΔT, suivant les sommets de Q dans le sens antihoraire ; les temps $T_k, k \geq 1$, sont appelés *temps de commutation.*

Dans le cas particulier : $p_0 = q_0 = 1$, en intégrant explicitement le système Hamiltonien et en appliquant l'équation (5a), nous voyons que ces extrémales sont associées aux trajectoires $\boldsymbol{\gamma_g}$. En répétant l'analyse précédente, on constate que le contrôle commute chaque ΔT, suivant la même séquence.

Soit $\widehat{\xi}$ une géodésique correspondante à une extrémale avec $\boldsymbol{p}(0) = (1, \widehat{q}_0, \widehat{r})$, $|\widehat{q}_0| < 1$; en intégrant le système Hamiltonien, on constate que, pour $t \in [T_4, 8/\widehat{r}]$, l'expression de la géodésique est donnée par

$$\begin{cases} x(t) = t - 4\Delta T \\ y(t) = 0 \\ z(t) = \Delta T^2, \end{cases}$$

c'est-à-dire, sa valeur dépend uniquement de t et de r, mais pas de $\widehat{q}_0$. En particulier, au temps $t = T_4 = \frac{7-\widehat{q}_0}{\widehat{r}}$, $\widehat{\xi}$ rencontre toutes les géodésiques du type $\boldsymbol{\gamma_f}$ avec vecteur adjoint $(1, q_0, \widehat{r})$, $q_0 > \widehat{q}_0$, et coïncide avec elles jusqu'au temps $8/\widehat{r}$. De plus, pour tout $\epsilon > 0$, il est toujours possible de trouver un certain $q_0 > \widehat{q}_0$ tel que le graphe de la trajectoire correspondante est $\epsilon-$près dans la norme C^0 au graphe de $\widehat{\xi}$. Autrement dit, à $t = T_4$ la trajectoire $\widehat{\xi}$ perd son optimalité locale ; le quatrième temps de commutation est alors son *temps conjugué.*

Le *temps de coupure* est le temps où une trajectoire perd son optimalité *globale*, et il est évidemment inférieur ou égal au temps conjugué. Pour vérifier si ces deux temps sont différents, il est nécessaire de chercher des intersections de la trajectoire observée avec d'autres trajectoires dont les graphes n'appartiennent pas au voisinage de son graphe ; en particulier, ces trajectoires peuvent venir d'autres stratégies (c'est-à-dire, elles peuvent être les géodésiques $\boldsymbol{\gamma_{-f}}$ ou $\boldsymbol{\gamma_{\pm g}}$). Par des calculs, c'est possible de prouver que de telles intersections se produisent soit au temps conjugué, soit à $t = 8/r$. Donc, pour toute trajectoire, le quatrième temps de commutation T_4 est à la fois le temps conjugué est le temps de coupure.

En appliquant des rotations appropriées dans le plan xy et/ou une réflexion dans la coordonnée z, il est possible de récupérer toutes les géodésiques optimales du système de Heisenberg.

3 Forme normale de cas générique

Dans cette section, nous construisons la forme normale pour la paire des champs de vecteurs (f, g), en exploitant la symétrie du problème. Pour une construction alternative, voir [5].

Proposition 1. *Soient f, g deux champs de vecteurs dans M qui satisfont l'équation* (3) *pour tout $\boldsymbol{x}$ dans un voisinage U de $\boldsymbol{x}_0$. Alors, il existe une carte de coordonnées locales (x, y, z) centré en $\boldsymbol{x}_0$ telle que*

$$f = \begin{pmatrix} 1 \\ 0 \\ -\frac{y}{2} \end{pmatrix} + y \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 \end{pmatrix} \qquad g = \begin{pmatrix} 0 \\ 1 \\ \frac{x}{2} \end{pmatrix} - x \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 \end{pmatrix} \tag{6}$$

où, pour tout $i = 1, 2, 3$, $\boldsymbol{\beta}_i$ est une fonction lisse qui satisfait $\boldsymbol{\beta}_i(0, 0, z) = 0$.

Esquisse de la démonstration. Pour un certain $\boldsymbol{x}_0 \in M$ fixé, nous considérons la fonction $\boldsymbol{\Phi} : \mathbb{R}^3 \to M$ définie comme il suit :

$$\boldsymbol{\Phi}(x, y, z) = \exp(xf + yg) \circ \exp(z[f, g])(\boldsymbol{x}_0),$$

où $\exp(tf)$ représente le flot du champ de vecteurs f du temps 0 au temps t. Grâce à la condition (3), $\boldsymbol{\Phi}$ est un difféomorphisme local qui couvre un voisinage de $\boldsymbol{x}_0$ dans M, donc (x, y, z) sont des systèmes de coordonnées locales autour de $\boldsymbol{x}_0$.

Nous allons maintenant démontrer (6). Soit $\mathbf{x} \in \mathbb{R}^3$ fixé. En appliquant [2, equation (2.33)] nous obtenons que $f(\boldsymbol{\Phi}(\mathbf{x})) = \frac{\partial}{\partial x} - yW$ et $g(\boldsymbol{\Phi}(\mathbf{x})) = \frac{\partial}{\partial y} + xW$, où

$$W(\mathbf{x}) = \big([f, g] + \frac{1}{6}(x[f, [f, g]] + y[g, [f, g]]) + \cdots\big) \circ \boldsymbol{\Phi}(\mathbf{x}).$$

La thèse suit en observant que $[f, g](\boldsymbol{\Phi}(\mathbf{x})) = \frac{\partial}{\partial z} + xR_1(\mathbf{x}) + yR_2(\mathbf{x})$, avec $R_i(0, 0, 0) = 0$. □

Dans ce qui suit, pour tout $i = 1, 2, 3$ on écrit $\boldsymbol{\beta}_i = xL_{i1} + yL_{i2}$, pour certaines fonctions lisses

$$L_{ij}(x, y, z) = c_{ij} + ax_{ij}x + ay_{ij}y + az_{ij}z + \theta_{ij}(x, y, z),$$

où $c_{ij}, ax_{ij}, ay_{ij}, az_{ij}$ sont des constantes et θ_{ij} sont des fonctions lisses, telles que leurs coefficients d'ordre zéro du développement de Taylor sont nuls. Les constantes $c_{ij}, ax_{ij}, ay_{ij}, az_{ij}$ sont appelées *invariants* de la métrique. Comme nous le verrons plus loin, les fonctions suivantes des invariants jouent un rôle spécial dans l'emplacement et la forme des lieux conjugués et des lieux de coupure :

$$\begin{aligned} A &= 4ax_{32} + 4ay_{31} + c_{11} - c_{22} + 9c_{31}c_{31}, \\ D_1 &= 8ax_{31} - 2c_{21} + 9c_{31}^2, \\ D_2 &= 8ay_{32} + 2c_{12} + 9c_{32}^2. \end{aligned}$$

La forme normale (6) hérite certaines propriétés de symétrie du cas nilpotent. En effet, il est toujours vrai que les géodésiques du type $\boldsymbol{\gamma_g}, \boldsymbol{\gamma_{-g}}$ et $\boldsymbol{\gamma_{-f}}$ peuvent être récupérées des trajectoires du type $\boldsymbol{\gamma_f}$,en appliquant une rotation appropriée sur les trajectoires et une transformation appropriée sur les invariants, selon les énoncés des lemmes suivants.

Lemme 1 *Pour $\theta = \frac{k\pi}{2}$, $k \in \mathbb{Z}$, soient $R = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$ et $\widehat{R} = \begin{pmatrix} R & 0 \\ 0 & 1 \end{pmatrix}$. Considérons les champs de vecteurs*

$$\widetilde{f} = \begin{pmatrix} 1 + y(\widetilde{L}_{11}x + \widetilde{L}_{12}y) \\ y(\widetilde{L}_{21}x + \widetilde{L}_{22}y) \\ -\frac{y}{2} + y(\widetilde{L}_{31}x + \widetilde{L}_{32}y) \end{pmatrix} \quad \widetilde{g} = \begin{pmatrix} -x(\widetilde{L}_{11}x + \widetilde{L}_{12}y) \\ 1 - x(\widetilde{L}_{21}x + \widetilde{L}_{22}y) \\ \frac{x}{2} - x(\widetilde{L}_{31}x + \widetilde{L}_{32}y) \end{pmatrix},$$

où

$$\begin{pmatrix} \widetilde{L}_{11} & \widetilde{L}_{12} \\ \widetilde{L}_{21} & \widetilde{L}_{22} \\ \widetilde{L}_{31} & \widetilde{L}_{32} \end{pmatrix} \Bigg|_{\begin{pmatrix} x \\ y \\ z \end{pmatrix}} = \begin{pmatrix} R & 0 \\ 0 & \cos(2\theta) \end{pmatrix} \begin{pmatrix} L_{22} & -L_{21} \\ -L_{12} & L_{11} \\ L_{32} & -L_{31} \end{pmatrix} \Bigg|_{\widehat{R}\begin{pmatrix} x \\ y \\ z \end{pmatrix}} R^{-1}.$$

Soit $\widetilde{\boldsymbol{\xi}}(\cdot)$ la solution du problème de Cauchy

$$\dot{\widetilde{\boldsymbol{\xi}}} = \widetilde{u}_1\widetilde{f} + \widetilde{u}_2\widetilde{g}, \qquad \widetilde{\boldsymbol{\xi}}(0) = \widetilde{\boldsymbol{x}}_0 \in M. \tag{7}$$

Alors $\widehat{R}\widetilde{\boldsymbol{\xi}}(t)$ est la solution du problème de Cauchy (4b) *correspondant aux contrôles $\boldsymbol{u} = R\widetilde{\boldsymbol{u}}$ et à la condition initiale $\boldsymbol{x}_0 = \widehat{R}\widetilde{\boldsymbol{x}}_0$.*

Lemme 2 *Soit $(\widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{\xi}}, \widetilde{\boldsymbol{u}})$ une extrémale du problème en temps minimal* (4a)-(7)-(4c)-(4d)*. Alors*

$$\lambda = (\widehat{R}\widetilde{\boldsymbol{p}}, \widehat{R}\widetilde{\boldsymbol{\xi}}),$$

avec $\boldsymbol{u} = R\widetilde{\boldsymbol{u}}$, est une extrémale pour le problème en temps minimal (4a)-(4b)-(4c)-(4d)*.*

Remarque 1 *De même, les géodésiques associées aux extrémales avec $r(0) \leq 0$ peuvent être récupérées de celles avec $r(0) \geq 0$, grâce à une transformation adéquate.*

4 Structure locale des lieux conjugués et des lieux de coupure pour des petites sphères

Comme dit dans l'introduction, nous sommes intéressés par la structure locale de la métrique sous-Finslérienne associée à $\{f, g\}$. En effet, dans le cas global, de nombreux problèmes concernant l'intervalle de définition des solutions du problème de Cauchy (4b), l'existence du minimum, la validité de la forme normale, ainsi que des difficultés majeures dans l'intégration du système, peuvent survenir. D'autre part, comme fait dans [4,8] pour le cas sous-Riemannien, il est intéressant de voir comment une petite perturbation du cas nilpotent affecte la forme des lieux conjugués et de coupure.

Dans le cas générique comme dans le cas nilpotent, si la valeur du vecteur adjoint r est très large, les temps de commutations deviennent très petits (ils sont en effet proportionnels à $1/r$), et le cas générique peut être traité comme une petite perturbation du nilpotent. Comme dans ce dernier, si r est suffisamment grand, les seules trajectoires du type bang-bang qui satisfont le principe de Pontryagin sont des perturbations des géodésiques du type $\boldsymbol{\gamma_f}, \boldsymbol{\gamma_g}, \boldsymbol{\gamma_{-f}}$ et $\boldsymbol{\gamma_{-g}}$. Pour cela, dans la suite nous allons nous concentrer sur le comportement de ces extrémales.

4.1 Temps conjugués

Considérons une géodésique avec $r(0) \gg 1$. En appliquant les conditions d'optimalité de second ordre (voir Agrachev et Gamkrelidze, [1], et Poggiolini et Stefani,[12]), Sigalotti [13] a prouvé que, dans le cas générique, toutes les géodésiques qui sont formées des enchaînements bang-bang sont localement optimales au moins jusqu'au quatrième temps de commutation, et aucune géodésique est localement optimale après le cinquième temps de commutation. Plus précisément, une géodésique peut perdre son optimalité locale soit au quatrième soit au cinquième temps de commutation (en effet, il ne peut pas y avoir de perte d'optimalité locale le long d'un arc bang, voir [12]).

Dans le cas générique comme dans le cas nilpotent, les temps de coupure sont inversement proportionnels à r. Inspirés par [4,5,8], nous appliquons la reparamétrisation du temps suivante :

$$\tau(t) = \int_0^t r(s)\, dt,$$

et on définit $\rho = 1/r$. Dans la suite, τ_i dénote le $i-$ème temps de commutation, après reparamétrisation.

Dans le cas où $r(0) \gg 1$, ρ est approximativement constant. Pour cela, nous allons écrire $\rho(t) \equiv \rho_0$ et nous allons développer toutes nos variables en série de ρ_0.

En contrôle optimal, une méthode usuelle pour détecter la perte d'optimalité locale est de chercher les points où la mappe exponentielle n'est pas inversible ; pour le problème en cause, la mappe exponentielle est définie comme il suit :

$$\mathrm{Exp}(p_0, q_0, r_0, \tau) \mapsto \boldsymbol{\xi}(\tau),$$

où $(\boldsymbol{p}(\tau), \boldsymbol{\xi}(\tau))$ est la valeur au temps τ de l'extrémale avec une condition initiale égale à $(p_0, q_0, r_0, 0, 0, 0)$; nous remarquons que, comme nous avons toujours $|p_0| = 1$ ou $|q_0| = 1$, la mappe exponentielle est une fonction de trois variables.

Grâce à des calculs simples mais très longs, nous avons calculé le développement en série de puissances (de ρ_0) de la Jacobienne de la mappe exponentielle. En particulier, la Jacobienne se révèle d'être constante (par rapport au temps) le long de tout arc bang. Pour les extrémales correspondantes au géodésiques du type $\boldsymbol{\gamma_{\pm f}}$, on obtient

$$J\mathrm{Exp}|_{(q_0,r,\tau)} = \begin{cases} 0 & \tau \in [0, \tau_2) \\ 4\rho_0^3 & \tau \in (\tau_2, \tau_3) \\ 8\rho_0^3 & \tau \in (\tau_3, \tau_4) \\ 32 D_1 \rho_0^5 & \tau \in (\tau_4, \tau_5) \\ -8\rho_0^3 & \tau \in (\tau_5, \tau_6). \end{cases}$$

Pour les extrémales correspondantes au géodésiques $\boldsymbol{\gamma_{\pm g}}$, nous obtenons les mêmes valeurs de la Jacobienne, sauf pour $\tau \in (\tau_4, \tau_5)$, qui est égal à $32 D_2 \rho_0^5$. Nous déduisons le résultat suivant.

Théorème 1 *Si $D_1 < 0$, alors τ_4 est le temps conjugué pour toutes les extrémales correspondantes aux géodésiques du type $\boldsymbol{\gamma_{\pm f}}$; si $D_1 > 0$, le temps conjugué pour ces extrémales coïncide avec τ_5. De façon analogue, le temps conjugué pour toutes les extrémales correspondantes aux géodésiques du type $\boldsymbol{\gamma_{\pm g}}$ coïncide avec τ_4 si $D_2 < 0$ et avec τ_5 si $D_2 > 0$.*

Démonstration. Supposons que $D_1 \neq 0$. Alors, pour tout $q_0 \in [-1, 1)$, $\mathrm{Exp}(1, q_0, r, \tau)$ est inversible pour τ appartenant à chaque intervalle du type $(\tau_i, \tau_{i+1}), i = 0, \ldots, 5$ (où l'on pose $\tau_0 = 0$, et τ_i dépend de q_0). Aux points de commutation, Exp a des différentielles différentes pour $\tau \to \tau_k^{\pm}$. Par le théorème de la fonction inverse de Clarke [9], Exp est inversible à ces points si toutes les combinaisons convexes des différentielles à gauche et à droite sont inversibles. En particulier, ça se produit si et seulement si $J\mathrm{Exp}$ ne change pas de signe. Si $D_1 < 0$, alors $J\mathrm{Exp}$ change de signe à $\tau = \tau_4$, par conséquent Exp n'est pas inversible à τ_4, qui est donc le point conjugué de l'extrémale. Si $D_1 > 0$, $J\mathrm{Exp}$ change de signe à $\tau = \tau_5$, de telle sorte que Exp devient non inversible à $\tau = \tau_5$.

Le même argument s'applique pour toutes les autres extrémales correspondantes aux trajectoires bang-bang. □

4.2 Lieu de coupure

Dans ce problème, le temps conjugué est le premier temps où une géodésique rencontre d'autres géodésiques du même type (par exemple, une trajectoire avec $p_0 = 1$ rencontre d'autres trajectoires avec $p_0 = 1$, mais avec q_0 différent). Cependant, une géodésique peut perdre son optimalité globale avant le temps conjugué, si elle rencontre d'autres géodésiques avec des contrôles initiales différents (par

exemple, une trajectoire $\boldsymbol{\gamma_f}$ peut croiser une $\boldsymbol{\gamma_g}$). Ces points s'appellent *points de coupure*, et l'ensemble des points de coupure est le *lieu de coupure*.

Afin de localiser de telles intersections, nous nous inspirons du comportement des géodésiques du cas nilpotent : par exemple, dans ce cas une géodésique $\boldsymbol{\gamma_f}$ rencontre, à son quatrième temps de commutation, le quatrième arc d'une géodésique $\boldsymbol{\gamma_g}$ qui correspond à $p_0 = q_0 = 1$; au temps $t = 8/r$, elle rencontre toutes les autres géodésiques. Cela suggère que, dans le cas générique, une perte d'optimalité globale avant le point conjugué peut se produire de l'intersection du quatrième arc de la trajectoire $\boldsymbol{\gamma_f}$ avec le quatrième arc du $\boldsymbol{\gamma_g}$.

La procédure que nous utilisons pour déterminer ces intersections est décrite ci-dessous. Sans perte de généralité, nous décrivons le cas dans lequel nous voulons détecter si et où le quatrième arc de la trajectoire $\boldsymbol{\gamma_f}$ traverse le quatrième arc du $\boldsymbol{\gamma_g}$; les autres cas sont calculés de la même façon.

Procédure

— On fixe le vecteur adjoint initial de $\boldsymbol{\gamma_f}$ à $(1, q_0, 1/\rho_0)$.

— On fixe le vecteur adjoint initial de $\boldsymbol{\gamma_g}$ à $(\hat{p_0}, 1, 1/\hat{\rho}_0)$; en particulier, on suppose que $\hat{p}_0$ et $\hat{\rho}_0$ dépendent de ρ_0, avec développements $\hat{p}_0 = \beta_0 + \beta_1\rho_0 + \beta_2\rho_0^2 + \dots$ et $\hat{\rho}_0 = \rho_0 + \alpha_2\rho_0^2 + \dots$.

— Puisque probablement l'intersection a lieu à proximité du quatrième temps de commutation de $\boldsymbol{\gamma_f}$, on choisit un temps $\boldsymbol{t}$ proche de $(7 - q_0)\rho_0$, qui correspond au temps réparametrisé

$$\mathcal{T} = \int_0^{\boldsymbol{t}} r(s)ds = 7 - q_0 + \delta_1\rho_0 + \delta_2\rho_0^2 + \cdots$$

Comme en principe $\rho_0 \neq \hat{\rho}_0$, les temps réparamétrisés pour $\boldsymbol{\gamma_f}$ et $\boldsymbol{\gamma_g}$ peuvent être différents ; en particulier :

$$\hat{\mathcal{T}} = \int_0^{\boldsymbol{t}} \hat{r}(s)ds = 7 - q_0 + \hat{\delta}_1\rho_0 + \hat{\delta}_2\rho_0^2 + \cdots ,$$

avec

$$\hat{\delta}_1 = \delta_1 - \alpha_2(7 - q_0), \quad \hat{\delta}_2 = \delta_2 - \alpha_2\delta_1 + (\alpha_2^2 - \alpha_3)(7 - q_0).$$

— On calcule les développements de $\boldsymbol{\gamma_f}(\boldsymbol{t}(\mathcal{T}))$ et $\boldsymbol{\gamma_g}(\boldsymbol{t}(\hat{\mathcal{T}}))$, et on impose l'égalité pour chacune des trois coordonnées, pour x, y jusqu'au troisième ordre, et pour z jusqu'au quatrième. Grâce à ça, on peut trouver les valeurs des coefficients $\alpha_i, \beta_i, \delta_i$, en fonction des invariants et de q_0.

— En analysant ces expressions, nous obtenons des conditions pour l'existence d'intersection. Par exemple, dans le cas visé on obtient

$$\beta_0 = -1, \quad \beta_1 = 0, \quad \beta_2 = 2(q_0 - 1)D_1. \tag{8}$$

Comme $q_0 - 1 \leq 0$ et $|\hat{q}_0|$ doit être inférieur ou égal à 1, on constate que cette intersection peut se produire uniquement si $q_0 = 1$ (ce qui donne une trajectoire $\boldsymbol{\gamma_g}$, donc on rejette ce cas) ou si $D_1 \leq 0$.

Une fois trouvée l'expression de $\mathcal{T}$ en fonction des invariants, de q_0 et de ρ_0, nous calculons $\boldsymbol{\gamma_f}(\boldsymbol{t}(\mathcal{T}))$, en fonction de q_0 et ρ_0. Ensuite nous allons examiner la *suspension* de cette surface, autrement dit, son intersection avec le plan $\{z = 4\zeta^2\}$, pour certain petit $\zeta > 0$.

Ce programme général est appliqué à toutes les autres intersections qu'on doit détecter. Le lieu de coupure, en fonction des valeurs des trois invariants principaux A, D_1 et D_2, est décrit ci-après.

4.3 $D_1 < 0$ et $D_2 < 0$

Du Théorème 1, le temps conjugué pour chaque géodésique bang-bang coïncide avec le quatrième temps de commutation ; cela implique qu'il est significatif d'étudier seulement les intersections des géodésiques qui se produisent avant τ_4.

Considérons les intersections $\boldsymbol{\gamma_f} \cap \boldsymbol{\gamma_g}$ et $\boldsymbol{\gamma_{-f}} \cap \boldsymbol{\gamma_{-g}}$ (ces dernières peuvent être récupérées à partir des précédentes, en appliquant le Lemmes 1-2) ; nous rappelons que de telles intersections peuvent avoir lieu, puisque $D_1 < 0$. Suivant la méthode décrite dans la section précédente, on peut vérifier que le temps d'intersection vaut $\mathcal{T} = \tau_4 - 2(q_0 - 1)D_1\rho_0^2 + \mathcal{O}(\rho_0^3)$, donc il est plus petit que le temps conjugué.

Nous nous intéressons maintenant à l'intersection de $\boldsymbol{\gamma_{\pm f}}(t(\mathcal{T}))$ avec le plan $z = 4\zeta^2$. Comme $\boldsymbol{\gamma_{\pm f}}(t(\mathcal{T})) = (\mp(1 + q_0)\zeta, 0, 0) + \mathcal{O}(\zeta^2)$, au premier ordre en ζ l'intersection est une ligne horizontale. Quand $q_0 \to -1$, les deux coordonnées du point d'intersection sont

$$\begin{cases} x = -12c_{32}\zeta^2 \mp 4(A + D_2)\zeta^3 + \mathcal{O}(\zeta^4) \\ y = 12c_{31}\zeta^2 \pm 4A\zeta^3 + \mathcal{O}(\zeta^4). \end{cases}$$

Cela signifie que la ligne horizontale se sépare en deux demi-droites, divisées par $8(A + D_2)\zeta^3$.

En appliquant la même démarche pour $\boldsymbol{\gamma_{\pm g}} \cap \boldsymbol{\gamma_{\mp f}}$, on constate que le terme principal de la suspension est $(x, y, z) = (0, \mp(1-p_0)\zeta, 0)$ de sorte que, au premier ordre en ζ, l'intersection est une ligne verticale. La limite pour $p_0 \to 1$ donne

$$\begin{cases} x = -12c_{32}\zeta^2 \pm 4A\zeta^3 + \mathcal{O}(\zeta^4) \\ y = 12c_{31}\zeta^2 \pm 4(A - D_1)\zeta^3 + \mathcal{O}(\zeta^4), \end{cases} \tag{9}$$

c'est-à-dire, la ligne verticale se sépare en deux demi-droites, divisées par $8(A - D_1)\zeta^3$.

Si $A > 0$, alors les deux lignes $\boldsymbol{\gamma_{-g}} \cap \boldsymbol{\gamma_f}$ et $\boldsymbol{\gamma_f} \cap \boldsymbol{\gamma_g}$ s'intersectent à un certain point Q_f^+, à proximité du point $(-A\zeta^3, A\zeta^3)$ (voir Figure 1) ; d'une manière analogue, les lignes $\boldsymbol{\gamma_g} \cap \boldsymbol{\gamma_{-f}}$ et $\boldsymbol{\gamma_{-f}} \cap \boldsymbol{\gamma_{-g}}$ s'intersectent en un certain point Q_f^- à proximité du point $(A\zeta^3, -A\zeta^3)$. En particulier, toutes les trajectoires du type $\boldsymbol{\gamma_f}$ et $\boldsymbol{\gamma_{-f}}$ perdent leurs optimalités globales quand elles rencontrent certaines trajectoire du type $\boldsymbol{\gamma_g}$ ou $\boldsymbol{\gamma_{-g}}$.

D'autre part, la position relative des points d'intersections ci-dessus suggère que certaines trajectoires $\boldsymbol{\gamma_g}$ doivent rencontrer certaines trajectoires $\boldsymbol{\gamma_{-g}}$ avant

de croiser celles du type $\boldsymbol{\gamma_{\pm f}}$. On cherche ainsi ces intersections ; en appliquant la procédure de la section précédente, on trouve que les intersections du type $\boldsymbol{\gamma_g} \cap \boldsymbol{\gamma_{-g}}$ se produisent seulement si $A - D_1 \geq 0$, lequel est en effet le cas en cours d'étude.

On peut conclure que le lieu de coupure est composé de cinq branches : quatre sont constituées des courbes $\boldsymbol{\gamma_f} \cap \boldsymbol{\gamma_g}, \boldsymbol{\gamma_g} \cap \boldsymbol{\gamma_{-g}}, \boldsymbol{\gamma_{-f}} \cap \boldsymbol{\gamma_{-g}}$ et $\boldsymbol{\gamma_{-g}} \cap \boldsymbol{\gamma_f}$, avant leurs intersections, et la cinquième branche est le segment joignant Q_f^+ à Q_f^-. La forme est illustrée dans la Figure 1.

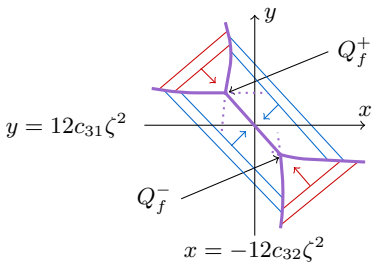


FIGURE 1: Lieu de coupure et fronts pour $D_1, D_2 < 0$ et $A > 0$. En bleu : les fronts pour $\boldsymbol{\gamma_{\pm g}}$; en rouge : les fronts pour $\boldsymbol{\gamma_{\pm f}}$; en violet continu, le lieu de coupure ; en pointillés violets, les intersections qui n'appartiennent pas au lieu de coupure.

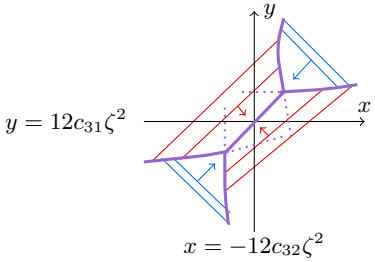


FIGURE 2: Lieu de coupure et fronts pour $D_1, D_2 < 0$ et $A < 0$.

Si $A < 0$, en répétant les mêmes arguments que ci-dessus, on obtient le lieu de coupure représenté dans la Figure 2.

4.4 $D_1 > 0$ et $D_2 < 0$

Si $D_2 < 0 < D_1$, alors le quatrième temps de commutation reste le temps conjugué pour les trajectoires du type $\boldsymbol{\gamma_{\pm g}}$, mais pas pour les trajectoires $\boldsymbol{\gamma_{\pm f}}$, donc leurs cinquièmes arcs bang peuvent participer au lieu de coupure ; d'autre part, grâce à (8) nous observons que les intersections entre le quatrième arc de $\boldsymbol{\gamma_f}$ ($\boldsymbol{\gamma_{-f}}$) et le quatrième arc de $\boldsymbol{\gamma_g}$ ($\boldsymbol{\gamma_{-g}}$) ne peuvent pas se produire ; il faut alors chercher aussi les intersections impliquant le cinquième arc de $\boldsymbol{\gamma_{\pm f}}$. Avant de le faire en détail, on remarque ce qui suit :

— en appliquant la procédure habituelle, il est possible de montrer que les intersections entre le cinquième arc de $\boldsymbol{\gamma_{\pm f}}$ avec le quatrième arc de $\boldsymbol{\gamma_{\pm g}}$ se produisent seulement si $q_0 = 1$ et $\hat{p}_0 = 1 + \mathcal{O}(\rho_0^3)$, c'est-à-dire, les deux trajectoires coïncident, puisqu'elles correspondent à l'extrémale avec $(p_0, q_0) = (1, 1)$.

— en gardant à l'esprit le cas nilpotent, on constate que les intersections du type $\boldsymbol{\gamma_f} \cap \boldsymbol{\gamma_{-f}}$ ou $\boldsymbol{\gamma_g} \cap \boldsymbol{\gamma_{-g}}$ (peu importe les arcs concernés) peuvent se produire que dans les alentours du point $\boldsymbol{O} = (-12c_{32}\zeta^2, 12c_{31}\zeta^2, 4\zeta^2)$, $\zeta > 0$, c'est-à-dire, pour $t \sim 8\rho_0$.

Ces observations suggèrent que la forme du lieu de coupure est déterminée par les positions relatives de ces courbes :

— les valeurs (en fonction de q_0) des courbes $\boldsymbol{\gamma_{\pm f}}$ au quatrième temps de commutation, notées par $\sigma_\pm = \boldsymbol{\gamma_{\pm f}}(t(\tau_{4\pm}))$; pour $q_0 \to \mp 1$, leurs suspensions aux alentours de $\boldsymbol{O}$ sont données respectivement par :

$$\begin{cases} x = -12c_{32}\zeta^2 \mp 4(A + D_2)\zeta^3 \\ y = 12c_{31}\zeta^2 \pm 4(A + D_1)\zeta^3. \end{cases}$$

— les suspensions du cinquième arc de $\boldsymbol{\gamma_f}$ correspondant à $q_0 = 1$ (respectivement, $\boldsymbol{\gamma_{-f}}$ pour $q_0 = -1$), qu'on appelle $\chi_\pm$; comme vu précédemment, elle coïncident avec les valeurs de $\boldsymbol{\gamma_{g\pm}}$ pour $\hat{p}_0 = 1$. Leurs suspensions sont données par :

$$\begin{cases} x = \pm(\tau - 8)\zeta - 12c_{32}\zeta^2 + (\tau - 8)\mathcal{O}(\zeta^2) + \mathcal{O}(\zeta^3) \\ y = 12c_{31}\zeta^2 \pm 4(A - D_1 + \frac{(9\tau - 72)}{2}c_{31}^2)\zeta^3 + \mathcal{O}(\zeta^4). \end{cases}$$

— les courbes $\kappa_\pm$ qui représentent, respectivement, $\boldsymbol{\gamma_{\pm f}} \cap \boldsymbol{\gamma_{\mp g}}$, dont les expressions près de l'origine sont (9). Ces courbes sont tracées en violet dans les Figures 3–6.

Les courbes $\kappa_\pm$ participent au lieu de coupure ; pour déterminer sa partie restante, il faut prendre en compte les positions relatives de $\sigma_\pm$ et $\chi_\pm$, qui dépendent des valeurs de D_1 et de A. Plus précisément, on peut distinguer quatre cas : $A > D_1$, $0 < A < D_1$, $-D_1 < A < 0$ et $A < -D_1$.

$\boldsymbol{A > D_1}$: Considérons les courbes $\sigma_\pm$ et $\chi_\pm$ dans un voisinage suffisamment petit $\boldsymbol{\mathcal{U}_O}$ de $\boldsymbol{O}$; pour les expressions données ci-dessus il s'ensuit que

$$\mathcal{Y}(\sigma_+) \succ \mathcal{Y}(\chi_+) \succ \mathcal{Y}(\chi_-) \succ \mathcal{Y}(\sigma_-),$$

où, avec un peu d'abus de notation, $\mathcal{Y}(\sigma_+) = \{y : \exists x : (x, y) \in \sigma_+ \cap \boldsymbol{\mathcal{U}_O}\}$, et $\mathcal{Y} \succ \mathcal{Y}'$ signifie que $\inf \mathcal{Y} > \sup \mathcal{Y}'$. Voir Figure 3 pour une représentation graphique.

En particulier, dans cette configuration $\boldsymbol{\gamma_f}$ et $\boldsymbol{\gamma_{-f}}$ ne s'intersectent pas, tandis que les intersections entre le cinquième arc de $\boldsymbol{\gamma_{\pm f}}$ et le quatrième arc de $\boldsymbol{\gamma_{\mp g}}$ sont possibles (suivant la procédure usuelle, on peut montrer qu'elles se produisent à chaque fois que $2A \geq (1 + q_0)D_1$, donc, pour tout $|q_0| \leq 1$). Les expressions de ces intersections sont

$$\begin{cases} x = -12c_{32}\zeta^2 \mp (4A - (1 + q_0)^2 D_1)\zeta^3 + \mathcal{O}(\zeta^4) \\ y = 12c_{31}\zeta^2 \pm 4(A - q_0 D_1)\zeta^3 + \mathcal{O}(\zeta^4), \end{cases}$$

donc elles décrivent deux arcs de paraboles joignant respectivement le point K_+ avec le point X_+, et le point K_- avec le point X_-, où

$$K_\pm = \boldsymbol{O} \pm (-4A\zeta^3, 4(A+D_1)\zeta^3, 0)$$
$$X_\pm = \boldsymbol{O} \pm (-4(A-D_1)\zeta^3, 4(A-D_1)\zeta^3, 0).$$

Le lieu de coupure est alors constitué de cinq branches lisses : les courbes $\kappa_\pm$, les deux paraboles (tracés en couleur lavande), et l'intersection des courbes $\boldsymbol{\gamma_g}$ et $\boldsymbol{\gamma_{-g}}$, qui résulte des segments joignant X_- à X_+ (tracé en rose). Puisque la tangente de chaque parabole au point $X_\pm$ a une pente égale à -1, et elle est verticale au point $K_\pm$, on peut conclure que le lieu de coupure est C^1 et lisse par morceaux.

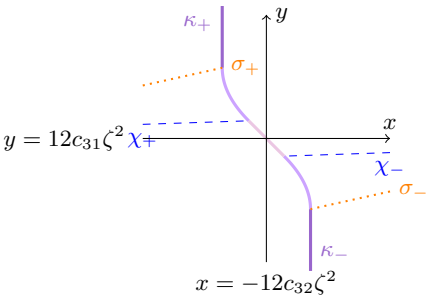


FIGURE 3: Lieu de coupure pour $A > D_1 > 0$, $D_2 < 0$.

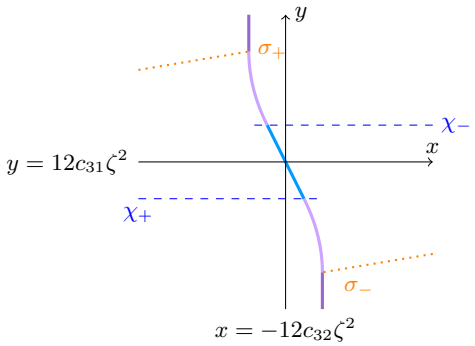


FIGURE 4: Lieu de coupure pour $A > D_1 > 0$, $D_2 < 0$.

$0 < A < D_1$: Dans ce cas

$$\mathcal{Y}(\sigma_+) \succ \mathcal{Y}(\chi_-) \succ \mathcal{Y}(\chi_+) \succ \mathcal{Y}(\sigma_-),$$

ce qui implique que des intersections entre $\boldsymbol{\gamma_f}$ et $\boldsymbol{\gamma_{-f}}$ peuvent se produire ; le cinquième arc de $\boldsymbol{\gamma_f}$ s'intersecte avec le quatrième arc de $\boldsymbol{\gamma_{-g}}$ uniquement pour q_0 tel que $2A \geq (1+q_0)D_1$.

Le lieu de coupure est constitué de cinq branches lisses : $\kappa_\pm$, les deux arcs de paraboles, et l'intersection $\boldsymbol{\gamma_f} \cap \boldsymbol{\gamma_{-f}}$ (en cyan dans la Figure 4).

Comme dans le cas précédent, en calculant les tangentes aux arcs des paraboles, il est possible de montrer que le lieu de coupure est C^1. Il est montré dans la Figure 4 .

$A < 0$: Dans cette partie, on regroupe les deux cas $-D_1 < A < 0$ et $A < -D_1$, comme ils sont similaires à ceux décrits précédemment.

• Si $-D_1 < A < 0$, on a

$$\mathcal{Y}(\chi_-) \succ \mathcal{Y}(\sigma_+) \succ \mathcal{Y}(\sigma_-) \succ \mathcal{Y}(\chi_+)$$

et, puisque $2A < 0 \leq (1+q_0)D_1$, le cinquième arc de $\boldsymbol{\gamma_{\pm f}}$ ne peut pas s'intersecter avec le quatrième arc de $\boldsymbol{\gamma_{\mp g}}$. Par contre, les intersections entre $\boldsymbol{\gamma_f}$ et $\boldsymbol{\gamma_{-f}}$ peuvent avoir lieu.

Nous répétons une analyse similaire à la précédente, et ainsi nous concluons que le lieu de coupure est C^1, lisse par morceaux, et composé de cinq branches :

— les deux lignes $\kappa_\pm$

— les intersections du cinquième arc de $\boldsymbol{\gamma_f}$ ($\boldsymbol{\gamma_{-f}}$) avec le quatrième arc de $\boldsymbol{\gamma_{-f}}$ ($\boldsymbol{\gamma_f}$), qui sont deux arcs de paraboles (bleu dans la Figure 5).

— les intersections entre les cinquièmes arcs de $\boldsymbol{\gamma_f}$ et $\boldsymbol{\gamma_{-f}}$ (cyan dans la Figure 5).

Ce cas est illustré dans la Figure 5.

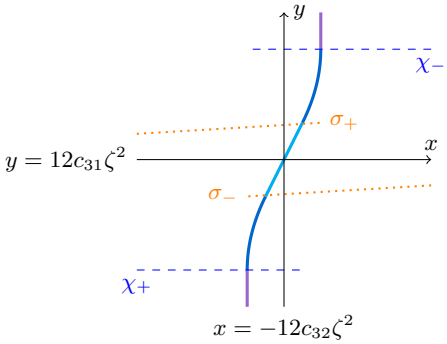


FIGURE 5: Lieu de coupure pour $-D_1 < A < 0$, $D_2 < 0$.

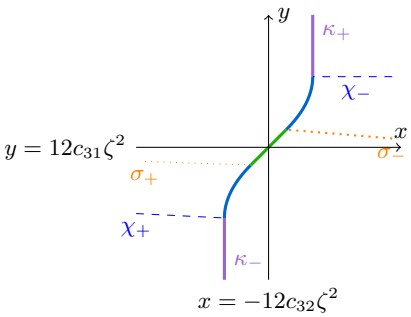


FIGURE 6: Lieu de coupure pour $A < -D_1$, $D_2 < 0$.

• Si $A < -D_1$, alors

$$\mathcal{Y}(\chi_-) \succ \mathcal{Y}(\sigma_-) \succ \mathcal{Y}(\sigma_+) \succ \mathcal{Y}(\chi_+).$$

Comme précédemment, le lieu de coupure est C^1, lisse par morceaux, et constitué de cinq branches :

— les deux lignes $\kappa_\pm$

— l'intersection du cinquième arc de $\boldsymbol{\gamma_f}$ ($\boldsymbol{\gamma_{-f}}$) avec le quatrième arc de $\boldsymbol{\gamma_{-f}}$ ($\boldsymbol{\gamma_f}$).

— les intersections entre les quatrièmes arcs de $\boldsymbol{\gamma_f}$ et $\boldsymbol{\gamma_{-f}}$ (vert dans la Figure 6).

Ce cas est illustré dans la Figure 6.

Remarque 2 *Comme les rôles entre D_1 et D_2 sont complètement échangeables, l'analyse de la partie 4.4 peut être adaptée au cas $D_1 < 0 < D_2$, en exploitant la symétrie et les Lemmes 1-2. Les représentations graphiques des lieux de coupure auront pour résultat la version horizontale des Figures 3–6.*

5 Conclusion

Cet article donne une analyse préliminaire de la géométrie sous-Finslérienne L^1 de dimension 3. Nous faisons les hypothèses génériques que la distribution de deux champs de vecteurs f, g ont un rang constant égal à 2, et nous nous concentrons sur les extrémales avec $r(0)$ assez grand, lesquelles correspondent aux extrémales bang-bang.

Nous avons donné une description des temps conjugués et de lieu de coupure dans le cas générique si les trois invariants principaux A, D_1 et D_2 sont non nuls, quand un parmi D_1 et D_2 est positif.

Le cas où D_1 et D_2 sont à la fois positifs, ainsi que les cas où un ou plusieurs des invariants principaux sont nuls, sont l'objet d'une étude courante des auteurs.

Références

1. A. A. Agrachev and R. V. Gamkrelidze, *Symplectic geometry for optimal control.* In Nonlinear controllability and optimal control, vol. 133 of Monogr. Textbooks Pure Appl. Math., Dekker, New York, 1990.
2. A. A. Agrachev and Yu. Sachkov, *Control Theory from the geometric viewpoint*, Springer-Verlag, 2004.
3. A. A. Agrachev, D. Barilari, U. Boscain, *A Comprehensive Introduction to sub-Riemannian Geometry*, en presse.
4. A. A. Agrachev, H. Chakir, J.P. Gauthier, *Subriemannian metrics on* $\mathbb{R}^3$. Geometric control and nonholonomic Mechanics, Proceedings of Canad. Math. Soc., vol. 25, 1998, pp. 29–78.
5. E. A.-L. Ali and G. Charlot, *Local contact sub-Finlserian geometry for maximum norms in dimension 3*, preprint `hal-02004281`
6. D. Barilari, U. Boscain, E. Le Donne and M. Sigalotti *Sub-Finsler geometry from the time-optimal control viewpoint for some nilpotent distributions*, J. Dyn. Control Syst., vol. 3, n. 3, 2017, 547–575.
7. E. Breuillard and E. Le Donne. *On the rate of convergence to the asymptotic cone for nilpotent groups and subfinsler geometry.* Proc. Natl. Acad. Sci. USA, 110(48) :19220–19226, 2013.
8. H. Chakir, J.P. Gauthier, I. Kupka, *Small Subriemannian Balls on* $\mathbb{R}^3$. J. Dyn. Control Syst., vol 2, n. 3, 1996, pp. 359–421.
9. F. Clarke, *On the inverse function theorem*, Pacific J. Math., vol. 64 n.1, 1976, pp. 97–102.
10. L.V. Lokutsievskiy, *Convex trigonometry with applications to sub-Finsler geometry*, preprint `arXiv:1807.08155` 2018

11. R. Montgomery, *A Tour of Subriemannian Geometries, Their Geodesics and Applications*, American Mathematical Society, 2006.
12. L. Poggiolini and G. Stefani *State-local optimality of a bang-bang trajectory : a Hamiltonian approach*,System Control Lett., vol 53, n. 3–4, 2004, pp. 269–279.
13. M. Sigalotti, *personal communication*

Multi-objective interval solid transportation problem with fuzzy equality under stochastic environment

Thiziri Sifaoui[1,2] and Méziane Aïder[2][0000−0001−7195−810X]

[1] LAROMAD, Fac. Sciences UMM-Tizi-Ouzou, Algeria
[2] LaROMaD, Fac. Maths, USTHB, Pb 32 El Alia, 16111 Algiers, Algeria
thiziri.sifaoui@gmail.com, m-aider@usthb.dz

Abstract. The paper deals with a solid transportation multi-objective problem (MOSISTPFE) in which the input data are expressed as stochastic intervals and the equality constraints are fuzzy. The formulated problem was transformed into a crisp equivalent problem, using the theories of interval, expected value and flexible index. An approach, combining goal programming, convex combination and fuzzy methods, is proposed for solving the crisp equivalent problem.

Keywords: Multi-objective solid transportation problem · fuzzy equality · stochastic environment.

1 Introduction

Basically, transportation problem (TP) is a particular class of linear programs that helps in solving problems on distribution and transportation of goods from a set of sources to a set of destinations to meet specific necessities. The goal is to satisfy the demand at destinations with with availability and supply constraints, at the minimum transportation cost possible.

Several extensions and generalizations of the classical transportation problem are possible and are studied in the literature, according to the specificities taken into account. The solid transportation problem (STP) is a special type of transportation problem that arises when heterogeneous conveyances are available for shipment of products. Indeed, in STP, instead of two items, namely source and destination, a third one is considered. This extra item is due to modes of transportation. The STP was stated by Shell [3] in 1955. Haley [4] developed a solution procedure of a STP and made a comparison between the STP and the classical TP.

On the other hand, many real world optimization problems require the minimization of multiple conflicting objectives. It is therefore a hard, and often an impossible, task to find a solution that simultaneously optimizes all objectives. Moreover, in general, the value of the parameters of these problems cannot be known in advance, hence they treated as random variables so the multi-objective

STP under uncertain environment has become a very popular research. Nagarajan and Jeyaraman [6] investigated in multi-objective solid transportation problems with interval parameters, and Chakraborty et al. [5] studied the multi-objective multi-item solid transportation with fuzzy inequalities constraints.

In this paper, we study a new model that consists to develop an expected model with fuzzy equality under stochastic environment and we develop a new approoach to resolve different kinds of STP.

Our paper is organized as follows. In section 2 we describe our problem, in section 3 we set a formulation of crisp equivalent, in section 4 three different compromise multi-objective methodologies are described, in section 5 we give a numerical experiment and discussion of our results. The main conclusion is presented at the end.

2 Multi-objective interval solid transportation Problem under stochastic environment with fuzzy equality constraints (MOISTPFE)

Multi-objective interval solid transportation problem with fuzzy equalities is a generalization of the multi-objective solid transportation problem in which the input data are expressed as stochastic intervals and the equality constraints are fuzzy.

It is well known that the aim of the classical STP is as follows: a homogenous product is to be transported from each of m sources to n destinations by k conveyances in which input data (objectives and constraints) are expressed as stochastic intervals and the equalities are assumed as fuzzy and may be uncertain due to some environmental impacts. Therefore, it is necessary to treat these coefficients as uncertain numbers.

We will use here the following notations.

- m : number of sources of the transportation problem;
- n : number of destinations;
- K : number of conveyances (modes of transportation);
- a_{L_i} : left limit of amount products in source i which can be transported to all the n destinations;
- a_{R_i} : right limit of amount products in source i which can be transported to n destinations;
- b_{L_j} : left limit of the demand of products at destination j;
- b_{R_j} : right limit of the demand of products at destination j;
- e_{L_k} : left limit of transportation ability of conveyance k;
- e_{R_k} : right limit of transportation ability of conveyance k;

- $C^p_{c_{ijk}}$: center of objective associated with transportation of a unit of product from source i to destination j by conveyance k, $p = 1, \ldots, P$;
- $C^p_{L_{ijk}}$: left limit of objective associated with transportation of a unit of product from source i to destination j by conveyance k, $p = 1, \ldots, P$;
- $C^p_{R_{ijk}}$: right limit of objective associated with transportation of a unit of product from source i to destination j by conveyance k, $p = 1, \ldots, P$;
- x_{ijk} : represents the unknown quantity to be transported from source i to destination j by conveyance k;
- $\cong$: means "essentially equals to";
- $\lesssim$: means "essentially smaller than";
- $\gtrsim$: means "essentially greater than".

Now, we can formulate this problem as:

$$\widetilde{P}b_{2.1}\begin{cases} \min \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K}[C^p_{Lijk}, C^p_{Rijk}]x_{ijk},\ p = 1, \ldots, P \\ \sum_{j=1}^{n}\sum_{k=1}^{K} x_{ijk} \cong [a_{L_i}, a_{R_i}] & i = 1, \ldots, m \\ \sum_{i=1}^{m}\sum_{k=1}^{K} x_{ijk} \cong [b_{L_j}, b_{R_j}] & j = 1, \ldots, n \\ \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ijk} \cong [e_{L_k}, e_{R_k}] & k = 1, \ldots, K \\ x_{ijk} \geq 0 & \begin{array}{l} i = 1, \ldots, m, \\ j = 1, \ldots, n, \\ k = 1, \ldots, K \end{array} \end{cases}$$

The constraints in the model mean that the source parameter lies between a left limit a_{L_i} and a right limit a_{R_i}, and similarly, the destination parameter lies between a left limit b_{L_i} and a right limit b_{R_i} and the mode transportation parameter lies between a left limit e_{L_i} and a right limit e_{R_i}. Moreover, the objective function means that the uncertain cost for the transportation lies between a left limit $C_{L^p_{ijk}}$ and a right limit $C_{R^p_{ijk}}$.

3 Formulation of Crisp equivalent

In this section we convert the original problem into a crisp equivalent using the theories of interval, expected value and flexible index.

3.1 Expected value model for multi-objective interval solid transportation problem with fuzzy equality

The expected value model ($\widetilde{P}b_E$) which optimizes expected objective function subject to expected constraints with fuzzy equalities in multi-objective problems may be expressed as follows :

$$\widetilde{Pb}_E \begin{cases} \min E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K}[C^p_{Lijk}, C^p_{Rijk}]x_{ijk}\Big] & p=1,\dots,P \\ E\Big[\sum_{j=1}^{n}\sum_{k=1}^{K} x_{ijk}\Big] \cong [a_{L_i}, a_{R_i}] & i=1,\dots,m \\ E\Big[\sum_{i=1}^{m}\sum_{k=1}^{K} x_{ijk}\Big] \cong [b_{L_j}, b_{R_j}] & j=1,\dots,n \\ E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ijk}\Big] \cong [e_{L_k}, e_{R_k}] & k=1,\dots,K \\ x_{ijk} \geq 0 & \begin{array}{l} i=1,\dots,m, \\ j=1,\dots,n, \\ k=1,\dots,K \end{array} \end{cases}$$

3.2 Formulation of the Crisp Constraint

Now, by using the theory of interval arithmetic, the problem is converted into its equivalent crisp form as follows:

$$\widetilde{Pb}_{3.2} \begin{cases} \min E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K}[C^p_{Lijk}, C^p_{Rijk}]x_{ijk}\Big] & p=1,\dots,P \\ E[a_{L_i}] \lesssim \sum_{j=1}^{n}\sum_{k=1}^{K} x_{ijk} \lesssim E[a_{R_i}] & i=1,\dots,m \\ E[b_{L_j}] \lesssim \sum_{i=1}^{m}\sum_{k=1}^{K} x_{ijk} \lesssim E[b_{R_j}] & j=1,\dots,n \\ E[e_{L_k}] \lesssim \sum_{i=1}^{m}\sum_{k=1}^{K} x_{ijk} \lesssim E[e_{R_k}] & k=1,\dots,k \\ x_{ijk} \geq 0 & \begin{array}{l} i=1,\dots,m, \\ j=1,\dots,n, \\ k=1,\dots,K \end{array} \end{cases}$$

with the following non-balanced conditions:

$$E\Big[\sum_{i=1}^{m} a_{L_i}\Big] \geq E\Big[\sum_{j=1}^{n} b_{L_j}\Big],\ E\Big[\sum_{i=1}^{m} a_{R_i}\Big] \geq E\Big[\sum_{j=1}^{n} b_{R_j}\Big],$$

$$E\Big[\sum_{k=1}^{K} e_{L_k}\Big] \geq E\Big[\sum_{j=1}^{n} b_{L_j}\Big],\ E\Big[\sum_{k=1}^{K} e_{R_k}\Big] \geq E\Big[\sum_{j=1}^{n} b_{R_j}\Big].$$

3.3 Formulation of the Crisp objective

The formulation of the original objective may be expressed as a Crisp equivalent.

Let the crisp equivalent of the right limit Z_R^P of the original problem $(\widetilde{Pb}_{2.1})$ be expressed as follows:

$$Z_R^p(x) = E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} C_{Cijk}^p\Big]x_{ijk} + E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} C_{Wijk}^p\Big]|x_{ijk}|,\ p = 1,\ldots,P$$

where $E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} C_{Cijk}^p\Big]$ is the expected value of the center and $E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} C_{Wijk}^p\Big]$ is the expected value of the half width of the coefficient of x_{ijk} in Z^P.

In the case where $x_{ijk} \geq 0,\ i = 1,\ldots,m, j = 1,\ldots,n, k = 1,\ldots,K,\ Z_R^p(x)$ is modified as follows:

$$Z_R^p(x) = E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} C_{Cijk}^p\Big]x_{ijk} + E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} C_{Wijk}^p\Big]x_{ijk}, p = 1,\ldots,P.$$

The expected value of center of the objective function $E[Z_C^p(x)]$ is expressed as:

$$Z_C^p(x) = E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} C_{Cijk}^p\Big]x_{ijk}, p = 1,\ldots,P$$

Definition 1 *[1] An interval number A is defined as*
$A = [a_L, a_R] = \{x : a_L \leq x \leq a_R, x \in R\}$
Here $a_L, a_R \in R$ are the lower and upper bounds of the interval A, respectively.
An interval number can also be expressed by its mean and width. in this form, an interval number $A = [a_L, a_R]$ is denoted by $< a_M, a_W >$, where $a_M = \frac{a_L + a_R}{2}$ and $a_W = \frac{a_R - a_L}{2}$ are known as the center and radius of the interval, respectively.

Definition 2 *A feasible solution $x^0 \in X$ is an optimal solution of $\widetilde{Pb}_{2.1}$ if and only if there is no solution $x \in X$ which satisfies $E(Z(x)) <_{LR} E(Z(x^0))$ or $E(Z(x)) <_{CW} E(Z(x^0))$,*
where the order relation $<_{LR}$ between $E(Z(x)) = [E(Z(x))_L, E(Z(x))_R]$ and $E(Z(x^0)) = [E(Z(x^0))_L, E(Z(x^0))_R]$ is defined as $E(Z(x)) <_{LR} E(Z(x^0))$ if and only if $E(Z(x))_L < E(Z(x^0))_L$ and $E(Z(x))_R < E(Z(x^0))_R$,
and the order relation $<_{CW}$ between $E(Z(x)) = [E(Z(x))_C, E(Z(x))_W]$ and $E(Z(x^0)) = [E(Z(x^0))_C, E(Z(x^0))_W]$ is defined as $E(Z(x)) <_{CW} E(Z(x^0))$ if and only if $E(Z(x))_C < E(Z(x^0))_C$ and $E(Z(x))_W < E(Z(x^0))_W$.
The orders $<_{LR}$ and $<_{CW}$ represent the preference of the decision maker with the lower and upper bounds and center and half width bounds.

Finally, we can formulate our problem as follows:

$$Pb^{\alpha}\begin{cases} \min Z^p = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} E[C^p_{Lijk}, C^p_{Rijk}]x_{ijk} & p = 1,\dots,P \\ (1-\alpha)d_{L_i} + E[a_{L_i}] \le \sum_{j=1}^{n}\sum_{k=1}^{K} x_{ijk} \le (1-\alpha)d_{R_i} + E[a_{R_i}] & i = 1,\dots,m \\ (1-\alpha)d_{L_j} + E[b_{L_j}] \le \sum_{i=1}^{m}\sum_{k=1}^{K} x_{ijk} \le (1-\alpha)d_{R_j} + E[b_{R_j}] & j = 1,\dots,n \\ (1-\alpha)d_{L_k} + E[e_{L_k}] \le \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ijk} \le (1-\alpha)d_{R_k} + E[e_{R_k}] & k = 1,\dots,K \end{cases}$$

with: $d_{L_i} \ge d_{R_i} \forall i \in 1,\dots,m$, $d_{L_j} \ge d_{R_j} \forall j \in 1,\dots,n$, $d_{L_k} \ge d_{R_k} \forall k \in 1,\dots,K$
and where $\alpha \in [0,1]$ and $d = [[d_{L_i}, d_{R_i}], [d_{L_j}, d_{R_j}], [d_{L_k}, d_{R_K}]] \ge 0$.
Let us denote x_α an optimal solution and B^α an optimal value of $\mathrm{l}(Pb^\alpha)$.

Definition 3 *Let B be an optimal basic matrix of (Pb^α). If there exists an interval $[\alpha_1, \alpha_2]$ such that B is an optimal base matrix of (Pb^α) for any $\alpha \in [\alpha_1, \alpha_2]$, while B is not an optimal matrix for any $\alpha \notin [\alpha_1, \alpha_2]$, we call that α_1 and α_2 critical values of (Pb^α) and $[\alpha_1, \alpha_2]$ a characteristic interval.*

Theorem 1. *[7] (Pb^α) has a finite characteristic interval on the interval $[0,1]$.*

Theorem 2. *[7] Let B be an optimal basic matrix of (Pb^α) on a characteristic interval $[\alpha_1, \alpha_2]$.*
If $(B^{-1}b)_i \neq 0$ $(1 \le i \le m)$, then

$$\begin{cases} \alpha_1 = \max\left[\frac{B^{-1}(b+d)_i}{B^{-1}d_i}, 0 \mid (B^{-1}d)_i < 0 (1 \le i \le m)\right] \\ \alpha_2 = \min\left[\frac{B^{-1}(b+d)_i}{B^{-1}d_i}, 0 \mid (B^{-1}d)_i > 0 (1 \le i \le m)\right] \end{cases}$$

is derived, where $(B^{-1}(b+d))_i$, and $(B^{-1}d)_i$ are the i-th *components of $B^{-1}(b+d)$ and $B^{-1}d$ respectively.*

Property 1 *[7] Let B be an optimal matrix of (Pb^α) on the characteristic interval $[\alpha_i, \alpha_j]$. Then $x_\alpha = B^{-1}(b + (1-\alpha)d), \alpha_i \le \alpha \le \alpha_j$, is a linear vector function about variable α. The optimal value function $z_\alpha = C_B B^{-1}(b + (1-\alpha)d)$ is a linear function about variable α and decreases with the increase of variable α.*

Property 2 *[7] The optimal value function z_α of (Pb^α) continues on the interval $[0,1]$.*

3.4 Algorithm to Fuzzy Linear Programming

Cao [7] presents a new algorithm to solve the fuzzy inequality constraints linear program. We develop and adapt this algorithm to solve our model.
Let z_1 be an optimal value of (Pb_1), and z_0 be an optimal value of (Pb_0), $d_0 = z_0 - z_1 > 0$.

Step 1: Solve linear programs (Pb_0) and (Pb_1) Let the optimal solutions be x_0, x_1, the optimal values be z_0, z_1 and the optimal matrix of (Pb_0) be B_0

Step 2: Solve $\left[B_0^{-1}(b+(1-\alpha)d)\right]_i = 0$. Assume the solutions as $\alpha_1, \ldots, \alpha_{n-1}$, $(0 < \alpha_1 < \ldots < \alpha_{n-1} < 1)$. Let $\alpha_0 = 0, \alpha_n = 1, \alpha = \alpha_1, k = 1$.

Step 3: Solve (Pb_α). Let the optimal value be z_α. If $z_\alpha \leq z_1 + d_0\alpha$, turn to Step 4, otherwise let $k = k+1$, $\alpha = \alpha_k$, turn to step 3.

Step 4: Solve the optimal decision

$$\alpha_* = \frac{z_1\alpha_k - z_1\alpha_{k-1} - z_{\alpha_{k-1}}\alpha_k + z_{\alpha_k}\alpha_{k-1}}{z_{\alpha_k} - z_{\alpha_{k-1}} - \alpha_k d_0 + \alpha_{k-1}d_0}.$$

Step 5: Solve linear program (Pb^*_α) and let x_{α^*} be an optimal solution and z_{α^*} an optimal value.

4 Approach to solve multi-objective solid transportation problem

4.1 Convex combination method

Let us consider the following multi-objective problem:

$$Pb_{4.1.1} \begin{cases} \min \quad [F_L^p(x), F_R^p(x)] \quad p = 1, \ldots, P \\ [g_L(x), g_R(x)] = 0 \\ x \in X \end{cases}$$

By the convex combination approach, we transform the above problem into:

$$Pb_{4.1.2} \begin{cases} \min \sum_{p=1}^{P} \mu_C^p F_C^p(x) + \sum_{p=1}^{P} \mu_R^p F_R^p(x) \\ g_L(x) \leq 0 \\ g_R(x) \geq 0 \\ x \in X \end{cases}$$

with $\sum_{p=1}^{P} \mu_C^p + \sum_{p=1}^{P} \mu_R^p = 1$, $0 < \mu_R^p < 1$, $0 < \mu_C^p < 1$.

4.2 Goal programming method

In order to solve (MOISTPFE), we have to solve each goal separately by specifying a goal that we will attempt to reach while minimizing the deviations. We can aggregate these deviations and solve the following problem.

$$Pb_{4.2.1} \begin{cases} \min \sum_{p=1}^{P} (\mu_R^{P^+} d_R^{P^+} + \mu_R^{P^-} d_R^{P^-} + \mu_C^{P^+} d_C^{P^+} + \mu_C^{P^-} d_C^{P^-}) \\ F_R^p(x) - d_R^{p^+} + d_R^{p^-} = L_R^p \qquad p = 1, \ldots, P \\ F_C^p(x) - d_C^{p^+} + d_C^{p^-} = L_C^p \qquad p = 1, \ldots, P \\ d_R^{P^+}, d_C^{P^+}, d_R^{P^-}, d_C^{P^-} \geq 0 \qquad p = 1, \ldots, P \\ x \in X \end{cases}$$

where

- L_R^p,L_C^p is the goal specified that we will attempt to reach,
- $d_R^{P^+}, d_C^{P^+}, d_R^{P^-}, d_C^{P^-}$ are the positives and negatives deviations from the goals.

4.3 Fuzzy Interactive Satisficing Method (FISM)

In this section we develop this approach to solve (MOISTPFE).
Let us consider the following multi-objective problem:

$$Pb_{4.3.1}\begin{cases}\min\ [F_L^p(x), F_R^p(x)]\ p=1,\ldots,P: \\ x\in X\end{cases}$$

Step 1: Introduce the membership function for each objective:

$$\zeta_p(F_R^p(x))\begin{cases}1 & \text{if } Z_R^p(x)\le L_R^p,\\ \frac{U_R^p-Z_R^p(x)}{U_R^p-L_R^p} & \text{if } L_R^p\le Z_R^p(x)\le U_R^p,\\ 0 & \text{if } Z_R^p(x)\ge U_R^p,\end{cases}$$

$$\zeta_p(F_C^p(x))\begin{cases}1 & \text{if } Z_C^p(x)\le L_C^p,\\ \frac{U_C^p-Z_C^p(x)}{U_C^p-L_C^p} & \text{if } L_C^p\le Z_C^p(x)\le U_C^p,\\ 0 & \text{if } Z_C^p(x)\ge U_C^p,\end{cases}$$

with:

$$\begin{cases}L_R^p=\min\limits_{x\in X} Z_R^p(x)\\ U_R^p=\max\limits_{x\in X} Z_R^p(x)\end{cases}$$

$$\begin{cases}L_C^p=\min\limits_{x\in X} Z_C^p(x)\\ U_C^p=\max\limits_{x\in X} Z_C^p(x)\end{cases}$$

Step 2: Solve the following problem:

$$Pb_{4.3.2}\begin{cases}\min\ \alpha\\ \bar{\mu}_R^p-\zeta_L^p(F_R^p(x))\le\alpha\ p=1,\ldots,P\\ \bar{\mu}_C^p-\zeta_C^p(F_C^p(x))\le\alpha\ p=1,\ldots,P\\ x\in X\end{cases}$$

where the decision maker updates reference membership levels $\bar{\mu_p}$, $p=1,\ldots,P$ through interaction.

4.4 Hybrid method

The stepwise algorithm can be written as follows:

Step 1: Solve all the p objective functions as single-objective problem ignoring all other objectives.

Step 2: Evaluate each objective function for its optimal solutions (max and min).

Step 3: Define the membership function as defined in step 1 fuzzy interactive satisficing approach.

Step 4: Define the weights of objectives as:

$$\sum_{p=1}^{P}\lambda_R^p+\sum_{p=1}^{P}\lambda_C^p=1,\quad \lambda_R^p\ge 0,\lambda_C^p\ge 0\ p=1,\ldots,P;$$

Step 5: Develop the proposed model as $\widehat{Pb}$ and solve it.

$$\widehat{Pb}\begin{cases} \min \phi^{'} = \sum_{p=1}^{P} \phi(1-\lambda_p) & p=1,\ldots,P \\ \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} E(C^p_{Rijk})x_{ijk} - \phi(1-\lambda^p_R) \le L^p_R & p=1,\ldots,P \\ \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} E(C^p_{Cijk})x_{ijk} - \phi(1-\lambda^p_C) \le L^p_C & p=1,\ldots,P \\ \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} E(C^p_{Rijk})x_{ijk} - \beta(U^p_R - L^p_R) \le L^p_R & p=1,\ldots,P \\ \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} E(C^p_{Cijk})x_{ijk} - \beta(U^p_C - L^p_C) \le L^p_C & p=1,\ldots,P \\ (1-\alpha)d_{L_i} + E[a_{L_i}] \le \sum_{j=1}^{n}\sum_{k=1}^{K} x_{ijk} \le (1-\alpha)d_{R_i} + E[a_{R_i}] & i=1,\ldots,m \\ (1-\alpha)d_{L_j} + E[b_{L_j}] \le \sum_{i=1}^{m}\sum_{k=1}^{K} x_{ijk} \le (1-\alpha)d_{R_j} + E[b_{R_j}] & j=1,\ldots,n \\ (1-\alpha)d_{L_k} + E[e_{L_k}] \le \sum_{i=1}^{m}\sum_{k=1}^{K} x_{ijk} \le (1-\alpha)d_{R_K} + E[e_{R_K}] & k=1,\ldots,K \\ \beta \ge 0 \end{cases}$$

Step 6: Evaluate each objective function at the solution obtained in step 5 .

Step 7: Present the solution obtained to the decision maker. If the DM is not satisfied go to Step 4. otherwise, go to Step 8.

Step 8: Stop.

5 A Numerical Example

The proposed method was tested on an example with 3 sources, 3 destinations and 3 conveyances. A comparative analysis of the results with three other methods is given.

Table 1. Interval Cost Matrix consisting of 3 sources, 3 destinations and 3 conveyance for first left limit objective C^1_{Lijk}.

	$j=1$			$j=2$			$j=3$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
$i=1$	$N(4,9)$	$N(7,6)$	$N(8,7)$	$N(3,2)$	$N(9,6)$	$N(7,4)$	$N(6,2)$	$N(7,2)$	$N(2,1)$
$i=2$	$N(4,2)$	$N(2,1)$	$N(6,1)$	$N(1,0)$	$N(3,1)$	$N(8,1)$	$N(8,3)$	$N(4,1)$	$N(5,2)$
$i=3$	$N(8,1)$	$N(1,1)$	$N(3,2)$	$N(4,1)$	$N(7,3)$	$N(3,2)$	$N(5,2)$	$N(6,3)$	$N(4,2)$

Table 2. Interval Cost Matrix consisting of 3 sources, 3 destinations and 3 conveyance for first Right limit objective C^1_{Rijk}.

	$j=1$			$j=2$			$j=3$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
$i=1$	$N(30,9)$	$N(20,8)$	$N(15,7)$	$N(20,4)$	$N(100,9)$	$N(60,8)$	$N(75,2)$	$N(50,2)$	$N(45,5)$
$i=2$	$N(5,4)$	$N(10,7)$	$N(25,8)$	$N(60,1)$	$N(45,1)$	$N(30,1)$	$N(15,3)$	$N(20,7)$	$N(100,5)$
$i=3$	$N(40,1)$	$N(15,1)$	$N(35,7)$	$N(10,5)$	$N(30,7)$	$N(5,2)$	$N(5,2)$	$N(100,5)$	$N(50,2)$

Table 3. Interval Cost Matrix consisting of 3 sources, 3 destinations and 3 conveyance for second left limit objective C^2_{Lijk}.

	$j=1$			$j=2$			$j=3$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
$i=1$	$N(9,1)$	$N(7,2)$	$N(8,2)$	$N(7,4)$	$N(9,3)$	$N(8,2)$	$N(6,2)$	$N(7,2)$	$N(2,1)$
$i=2$	$N(8,4)$	$N(4,2)$	$N(4,2)$	$N(7,1)$	$N(5,1)$	$N(8,1)$	$N(4,3)$	$N(4,2)$	$N(9,5)$
$i=3$	$N(8,1)$	$N(12,1)$	$N(34,7)$	$N(4,1)$	$N(7,2)$	$N(3,2)$	$N(5,2)$	$N(6,5)$	$N(4,2)$

Table 4. Interval Cost Matrix consisting of 3 sources, 3 destinations and 3 conveyance for second right limit objective C^2_{Rijk}.

	$j=1$			$j=2$			$j=3$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
$i=1$	$N(10,9)$	$N(20,8)$	$N(30,7)$	$N(25,4)$	$N(40,9)$	$N(60,8)$	$N(50,2)$	$N(15,2)$	$N(5,5)$
$i=2$	$N(25,4)$	$N(100,7)$	$N(50,8)$	$N(10,1)$	$N(10,1)$	$N(20,1)$	$N(5,3)$	$N(30,7)$	$N(55,5)$
$i=3$	$N(15,1)$	$N(20,1)$	$N(35,7)$	$N(20,5)$	$N(45,7)$	$N(60,2)$	$N(50,2)$	$N(10,5)$	$N(5,2)$

The proposed model can then be expressed as follows:

$$Pbm\begin{cases}
\min Z^1 = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}[C^1_{Lijk}, C^1_{Rijk}]x_{ijk} \\
\min Z^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}[C^2_{Lijk}, C^2_{Rijk}]x_{ijk} \\
N(5,5) \lesssim \sum_{j=1}^{3}\sum_{k=1}^{3} x_{1jk} \lesssim N(20,7) \\
N(10,5) \lesssim \sum_{j=1}^{3}\sum_{k=1}^{3} x_{2jk} \lesssim N(30,7) \\
N(5,2) \lesssim \sum_{j=1}^{3}\sum_{k=1}^{3} x_{3jk} \lesssim N(10,4) \\
EXP(15) \lesssim \sum_{i=1}^{3}\sum_{k=1}^{3} x_{i1k} \lesssim EXP(20) \\
EXP(10) \lesssim \sum_{i=1}^{3}\sum_{k=1}^{3} x_{i2k} \lesssim EXP(30) \\
EXP(20) \lesssim \sum_{i=1}^{3}\sum_{k=1}^{3} x_{i3k} \lesssim EXP(40) \\
N(10,6) \lesssim \sum_{i=1}^{3}\sum_{j=1}^{2} x_{ij1} \lesssim N(15,8) \\
N(20,5) \lesssim \sum_{i=1}^{3}\sum_{Wj=1}^{3} x_{ij2} \lesssim N(30,8) \\
N(10,5) \lesssim \sum_{i=1}^{3}\sum_{j=1}^{3} x_{ij3} \lesssim N(30,7) \\
x_{ijk} \geq 0,\ i = 1,2,3,\ j = 1,2,3,\ k = 1,2,3
\end{cases}$$

where the bounds in the constraints represent the probability distributions of the corresponding parameters ($N(\mu, \sigma^2)$ refers to the normal distribution with mean value μ and standard deviation σ and $EXP(\lambda)$ refers to the exponential distribution of parameter λ).

Let $d = ([0,0],[3,5],[0,0],[1,5],[0,0],[0,0],[0,0],[0,0],[0,0])^T$ be the flexible value.

By using the approach proposed by Cao, we convert Pbm into Pb_α.

We obtain $Z_0 = 570$, $Z_1 = 562.5$, $d_0 = 7.5 > 0$ by computing (Pb_0) and (Pb_1).

By calculating the equations

$\left[B_0^{-1}(5, 20, 13 - 3\alpha, 35 - 5\alpha, 5, 10, 16 - \alpha, 25 - 5\alpha, 10, 30, 20, 40, 10, 15, 20, 30, 10, 30)^T\right]_i = 0$, $(i = 1, \ldots, 18)$, respectively, we obtain $\alpha_0 = 0$, $\alpha_1 = 1$, $Z_{\alpha_0} = 570$ and $Z_{\alpha_1} = 562.5$, $570 > 562.5 + 0 * 7.5$ $Z_{\alpha_0} > Z_{\alpha_1} + d_0\alpha_0$.

We must continue to solve the linear program (Pb_{α_1}) $Z_{\alpha_1} = 562.5$, $Z_{\alpha_1} < Z_{\alpha_1} + d_0\alpha_1$.

So we stop here and calculate optimal decision α_*.

Now,

$$\alpha_* = \frac{Z_1\alpha_1 - Z_1\alpha_0 - Z_{\alpha_0}\alpha_1 + Z_{\alpha_1}\alpha_0}{Z_{\alpha_1} - Z_{\alpha_0} - \alpha_1 d_0 + \alpha_0 d_0} = 0.5$$

Hence, the optimal decision corresponds to $\alpha_* = 0.5$ and the optimal value $Z_{\alpha_*} = 566.25$. We must cover Pb_2^{α} by $\alpha_* = 0.5$

5.1 Solving by different kind of methods

Table 5. Compromise solution given by the different methods.

λ	$\begin{pmatrix} Z^1 \\ Z^2 \end{pmatrix}$			
$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	Hybrid approach	FISM	Convex Combination	Goal Programming
(0.2,0.2,0.2,0.4)	$\begin{pmatrix} 570.63,\ 878.35 \\ 531.1,\ 747 \end{pmatrix}$	$\begin{pmatrix} 569.05,\ 874.3 \\ 535.65,\ 904.65 \end{pmatrix}$	$\begin{pmatrix} 659.25,\ 1030 \\ 426.75,\ 560 \end{pmatrix}$	$\begin{pmatrix} 659.25,\ 1030 \\ 426.75,\ 560 \end{pmatrix}$
(0.4,0.2,0.2,0.2)	$\begin{pmatrix} 570.63,\ 878.35 \\ 531.1,\ 747 \end{pmatrix}$	$\begin{pmatrix} 592.75,\ 922.6 \\ 506.89,\ 849.6 \end{pmatrix}$	$\begin{pmatrix} 576.75,\ 885 \\ 526.75,\ 735 \end{pmatrix}$	$\begin{pmatrix} 576.75,\ 885 \\ 526.75,\ 735 \end{pmatrix}$
(0.2,0.4,0.2,0.2)	$\begin{pmatrix} 539.6,\ 957.4 \\ 489.49,\ 669.15 \end{pmatrix}$	$\begin{pmatrix} 578.7,\ 904.25 \\ 523.13,\ 745.4 \end{pmatrix}$	$\begin{pmatrix} 576.75,\ 885 \\ 526.75,\ 735 \end{pmatrix}$	$\begin{pmatrix} 576.75,\ 885 \\ 526.75,\ 735 \end{pmatrix}$
(0.2,0.2,0.4,0.2)	$\begin{pmatrix} 539.14,\ 805.95 \\ 727.9,\ 891.8 \end{pmatrix}$	$\begin{pmatrix} 659.25,\ 1030 \\ 426.75,\ 560 \end{pmatrix}$	$\begin{pmatrix} 659.25,\ 1030 \\ 426.75,\ 560 \end{pmatrix}$	$\begin{pmatrix} 659.25,\ 1030 \\ 426.75,\ 560 \end{pmatrix}$

Table 6. Execution time of the different methods.

λ	Cpu (in seconds)			
$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	Hybrid approach	FISM	Convex Combination	Goal Programming
(0.2,0.2,0.2,0.4)	0.05	0.09	0.12	0.07
(0.4,0.2,0.2,0.2)	0.05	0.10	0.06	0.06
(0.2,0.4,0.2,0.2)	0.05	0.10	0.06	0.06
(0.2,0.2,0.4,0.2)	0.05	0.07	0.06	0.06

6 Conclusion

In this paper, we explored the multi-objective interval Solid transportation Problem under stochastic environment with fuzzy equalities and developed an hybrid method to solve several kinds of multi-objective transportation problems.

In future we will hybridize the proposed method with a metaheuristic to solve more complex transportation problems with big instances.

References

1. Baidya, A.,Bera, U.K.: An interval valued solid transportation problem with budget constraint in different interval approaches, J.Transp Secur, 7(2), 147-155 (2014)
2. Hitchcock, F.: The distribution of a product from several sources to numerous localities. J. Math. Phys, 20(1-4), 224-230 (1941)
3. Schell, E.: Distribution of a product by several properties. In: Proceeding of 2nd Symposium in Linear Programming, DCS/comptroller, HQ US Air Force, Washington DC 615-642 (1955)
4. Haley, K.: New methods in mathematical programming-the solid transportation problem. Oper. Res. 10(4), 448-463 (1962)
5. Chakraborty, D., Jana, D.K., Roy, T.K.: Multi-objective multi-item solid transportation problem with fuzzy inequality constraints. J. Inequalities and Applications 338(1), 1-22 (2014)
6. Nagarajan, A., Jeyaraman, K.: Multi-Objective Solid Transportation Problem with Interval Cost in Source and Demand Parameters. I.J.C.O.T. 8(1), 33-41 (2014)
7. Cao, B.: Optimal Models and Methods with fuzzy quantities. Springer, Berlin (2010)
8. Alefeld, G., Herzberger, J.: Introduction to Interval Computations. Academic Press, New York (1983)
9. Moore, R. E.: Methods and Applications of Interval Analysis. SIAM Publications, Philadelphia, Pa. (1979)

MBO applied to the thermoforming process with convection and conduction considerations

Kahina Bachir Cherif[1], Djamal Rebaine[2], Fouad Erchiqui[3], and Issouf Fofana[1]

[1] Département des Sciences Appliquées, Université du Québec à Chicoutimi, Saguenay (Québec), Canada
Kahina.Bachir-Cherif1@uqac.ca, issouf_fofana@uqac.ca
[2] Département d'Informatique et de Mathématique, Université du Québec à Chicoutimi, Saguenay (Québec), Canada
Djamal_Rebaine@uqac.ca
[3] École de Génie, Université du Québec en Abitibi-Témiscamingue, Rouyn-Noranda (Québec), Canada
Fouad.Erchiqui@uqat.ca

Abstract. This paper addresses the problem of distributing uniformly infrared radiative energy intercepted by a thermoplastic sheet surface during the infrared radiation transmitted by an oven with convection and conduction considerations. After discretizing this problem, we proposed an objective function that captures the uniform distribution of the radiative energy. With this approximation scheme, the corresponding problem appears that it nothing else than a variant of a quadratic assignment problem. Then, we designed and applied a migrating bird optimization based algorithm (MBO for short), in order to minimize the corresponding objective function. To evaluate this approach we conducted a numerical experimental study.

Keywords: Thermoforming process, infrared radiative energy, convection and conduction, migrating bird optimization algorithm, quadratic assignment problem.

1 Introduction

Polymers are invading more and more all the branches of industry to the point that they became essential for the modern industry. As a consequence, the performance of the processes of transforming these materials into final products is a major step. The most used approach to form the final objects is the thermoforming process. This technique uses a heating phase to deform the thermoplastic materials according a predefined mold, and then a cooling phase to retrieve their initial physical properties while keeping the form obtained at the end of the previous phase. The quality of the formed object depends strongly on the thermal distribution imposed on the surface of the thermoforming sheet during the heating phase [6, 10]. Indeed, the distribution of the thicknesses of the object essentially depends on the distribution of the adjustment of the power of the elements of heating of the oven [2, 11].

When heated by infrared radiative energy (IR-energy), the plastic sheet is transformed from glassy state into a rubbery state. This hot state combined with the gravity creates a non-uniform thickness distribution in the plastic sheet. Adequate optimization of the heating stage can improve significantly the mass distribution in the finished part. One effective way to achieve better uniform thickness distribution is to reduce the differences of energy intercepted and absorbed by the different areas of the thermoplastic sheet. An illustration is pictured by Figure 1, in which the colors represent the intensity of energy received by the thermoforming sheet.

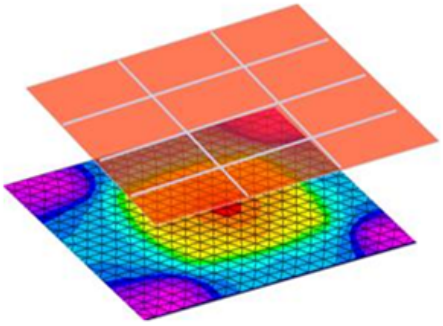


Fig. 1. Heating stage: thermoforming process

The main goal of this paper is to determine, from a given set of temperatures, the best distribution of the heating elements of the oven in order to minimize the difference of the temperature between the different areas the thermoplastic sheet.

The present paper is organized as follows. Section 2 introduces the thermal model we are addressing. In Section 3 we present the problem we are considering and the corresponding optimization formulation we derived. In section 4, we present the migrating bird optimization metaheuristic method [5] we used to solve the optimization model. Section 5 first presents details on the implementation issues of the MBO algorithm, and then the results of the experimental study we conducted on the quality of the solutions produced by our solution. Concluding remarks are presented in Section 6.

2 Thermal model of the oven and the thermoplastic sheet

The exchanges of energy between the oven and the thermoforming surface of the sheet are supposed to be of radiative type. The heating of the sheet is undertaken from its two sides through the lower and the upper heating elements of the oven. The oven has no system of ventilation, the cooling is made by natural convection between the sides of surfaces of the sheet and the ambient air. In the

thermoforming process, we consider the used plastic sheets of polymers mainly thin. Consequently, the transfer of heat by conduction is in the vertical direction of the thickness of the sheet [9]. The equation which governs the distribution of the temperature in the thickness of sheet is as follows:

$$\rho c_p \frac{\partial T}{\partial t} = k \frac{\partial^2 T}{\partial z^2}, \tag{1}$$

where c_p, ρ and k are the specific heat, density and thermal conductivity of the sheet plastic, respectively, T the temperature, z the coordinate of the sheet in the thickness direction and t the elapsed time. The upper and lower side of the thermoplastic sheet are exposed to exchanges of heat by convection and radiation coming from the heating elements of the upper and lower part of the oven. This is expressed by the boundary conditions of the differential equation of conduction.

$$\dot{q}_{tot} = \dot{q}_{conv} + \dot{q}_{rad}. \tag{2}$$

Parameter $\dot{q}_{conv}$ represents the heat transfer by convection from the surface of the thermoplastic sheet towards the ambient air. It expresses the Newton's condition:

$$\dot{q}_{conv} = h(T_\infty - T_s), \tag{3}$$

with T_s denoting the temperature of the thermoplastic sheet, T_∞ the temperature of ambient air, and h the convection coefficient. The convection is of natural type. The value of h is of order 2 to 10 $W.m^{-2}.K^{-1}$ [1].

The IR-energy $\dot{q}_{rad}$ received by the surface of the thermoplastic material governs the present study. In order to take into account the fluctuations of IR-energy on the surface of the thermoplastic sheet, we subdivide the surface of the sheet into small simple elementary surfaces. Thus, the surface of the thermoplastic sheet is discretized in n surface areas S_j. On the other hand, we assume that the oven has a set of m heating elements on the upper or lower surface of S_i (see Fig 2).

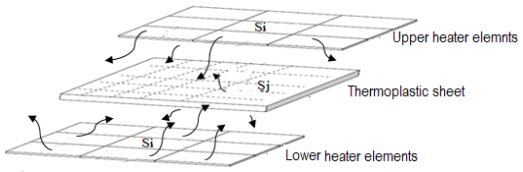


Fig. 2. Oven thermoforming

The IR-energy on cell j of the thermoplastic sheet of surface S_j is calculated according to the geometric orientation and temperature of oven heater cell i. View factors are required to obtain the energy for each cell of the thermoplastic

surface exposed to the infrared radiation. We assume that surface S_i of the heater is separated by a semi-transparent medium from surface S_j of the thermoplastic sheet. As a consequence, after some approximation manipulations, the amount of radiation Q_{ij}^k leaving surface S_i and intercepted by surface S_j for a given temperature τ_k is as follows [7]:

$$Q_{ij}^k = \frac{S_i}{S_j} F_{ij} \sigma \bar{\epsilon} \tau_k^4, \tag{4}$$

where F_{ij} denotes the view factor that calculates the fraction of energy emitted by cell S_i and received by cell S_j, $\bar{\epsilon}$ the average source emissivity of the material. Parameter σ is the Stefan-Boltzman constant of value 5.67×10^{-8} $W.m^{-2}.K^{-4}$, and τ_k the temperature assigned to cell j. The view factors are given by

$$F_{ij} = \frac{1}{S_i} \int_{S_i} \int_{S_j} \frac{\cos\theta_i \cos\theta_j}{\pi \times r^2} dS_j dS_i.$$

Parameter r is the distance separating cells i and j, and dS_i and dS_j are the elemental areas connected by line r, which forms the polar angles θ_i and θ_j with the surface normal n_i and n_j, respectively.

3 Statement of the optimization problem

The quality of the molded product depends heavily on the distribution of the temperatures generated by the energy flux received by the surface areas of the thermoplastic sheet exposed to the infrared radiations. More precisely, this energy spreading relies on the heating arrangement of the zones of the oven. Adequate modeling and optimization of the heating stage can improve significantly the final thickness distribution in the final product, thus improving its quality and reducing the number of rejections, which in turn improves productivity [8]. Let us recall that the energy received by each cell of the thermoplastic sheet from a single heating cell of the oven is expressed by Equation (4). The idea of minimizing the IR-energy gap received by the thermoplastic sheet areas may be constructed as follows. The difference in IR-energy between the areas of the medium and those of the edges of the thermoplastic sheet plastic sheet surface has an important influence on the quality of the final shape of the product. Obviously, the smaller are the gaps between these elements the better is the quality of the product. The goal is thus to make that the elements of the thermoplastic sheet receive approximately the same amount of infrared radiative energy from the heating elements. One way to capture this goal is to minimize the standard deviation of the energy received by the cells of the thermoplastic sheet. This process leads to the following optimization problem. Let us note that this problem is a variant of the quadratic assignment problem, known to be $\mathcal{NP}$-hard problem is the strong sense [4]. Therefore, the approximation approach is well justified. This is discussed in the next section

Given are m heating cells of an oven arranged as a (m_1, m_2)-matrix with $m = m_1 \times m_2$. A temperature from set $\tau = \{\tau_1, \ldots, \tau_p\}$, the temperature set,

is assigned to each heating cell of the oven in order to minimize the standard deviation of the set of fractions of the energy received by the n thermoplastic sheet cells, also arranged as a (n_1, n_2)-matrix with $n = n_1 \times n_2$.

To simplify our notation, we assume that the heating cells of the oven are lexicographically ordered and then numbered from 1 to m as follows: cell (i, j) precedes cell (k, t) if, and only if, we have either $i \leq k$ or $i = j$ and $j \leq t$.

Let x_{kj} be a decision variable such that $x_{kj} = 1$ if temperature $\tau_k \in \tau$ is assigned to cell j, and 0 otherwise. Then, the total energy received by cell i of the thermoplastic sheet is

$$q_i = \sum_{j=1}^{m} \sum_{k=1}^{p} Q_{ij}^{k} x_{kj}.$$

If we denote by $\overline{q}$ the average of the energy received by the n thermoplastic sheet cells, that is

$$\overline{q} = \frac{1}{n} \sum_{i=1}^{n} q_i.$$

The objective function we are addressing is therefore expressed as follows:

$$\min \frac{1}{\overline{q}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (q_i - \overline{q})^2}. \tag{5}$$

Let us note that if the standard deviation was not divided by $\overline{q}$, then the minimization process would tend to favor the small values of the temperature set as they produce small values for that objective function.

The above objective function is subject to the following constraints: a heating cell must receive exactly one temperature from set τ, and a temperature from set τ is used at most n times. These constraints are respectively expressed as below:

$$\sum_{k=1}^{p} x_{ki} = 1; \quad i = 1, \ldots, n,$$
$$\sum_{j=1}^{n} x_{ki} \leq n; \quad k = 1, \ldots, p.$$

In order to get the thermoplastic sheet ready for the thermoforming, the corresponding temperature, which corresponds to IR-energy, has to be within its thermoforming window. This constraint is expressed as follows:

$$q_{min} \leq q_i \leq q_{max}; \quad i = 1, \ldots, n.$$

4 Migrating bird optimization method

The Migration Birds Optimization (MBO) is inspired from the 'V'-shape of the flights of migrating birds. The property of bird flights lies in the energy conservation. Indeed, when a bird beats its wings, it generates a draft which will make the birds behind have to supply fewer efforts to rise. The organization of the flight of the birds is as follows: the bird in the head leads the group for a certain period of time, and spends more energy than the rest of its congeners. When it is tired, it moves behind the line of the group, and one of the birds currently behind takes the lead.

The parameters defining the above metaheuristic algorithm are p: the number of initial solutions, α: the number of neighbor solutions to consider, β: the number of neighbors to share with the next solution, γ: the number of iterations to perform before changing the leader solution, and L: the maximum number of iterations the algorithm executes. In [3], through their experimental study they conducted, the value of these parameters are: $p = 51$, $\alpha = 7$, $\beta = 3$, $\gamma = 10$ and $L = 10\ 000$. The MBO algorithm may be resumed as in Table 1. In order to

Table 1. Basic Migrating Bird Algorithm

1. Fix p, α, β, γ and L;
2. Generate at random p solutions, and place them in a V-shape;
3. for (i=0; $i < L$; i++)
   {
   For $(j = 0; j < \gamma; j++)$
   {
   - Improve the leader solution by generating and evaluating α of its neighbors and $i \longleftarrow i + \alpha$;
   - Except the solution leader, improve the solution in the V-shape by evaluating $(\alpha - \beta)$ neighbors with the β best solutions not used in the solution ahead and set $i \longleftarrow i + (\alpha - \beta)$;

   }
   - Move the leader to the back of the group, and move one of the next solutions that are behind it to the leader position;

   }

make the above algorithm operational, the neighborhood of a solution must be specified. The one we adopted is in Table 2.

5 Experimental study

To make our numerical simulation closer to real world applications, the dimensions of the thermoforming oven has an industrial scale. The oven is made of 120

Table 2. Procedure Neighborhood

Step 1. Generate randomly a number $r(0 \leq r \leq 1)$;
Step 2. If $(r \leq 0.5)$ then
  2.1. Select at random two temperature locations i and j from the actual solution;
  2.2. Exchange the two corresponding temperatures.
else
  2.3. Select at random a temperature location i.
  2.4. Choose at random a temperature from set τ and assign it to location i.
end-if

elements from above and below parts, each of is of dimension 0.06m by 0.245m. The lateral sides of the oven are open and the environment behaves as a black body. The power of every heating element is checked by a regulator; the maximum power available is 650W. The temperature on the surface of the heating elements is measured by a thermocouple of type K integrated to those elements. For our numerical simulation, the thermoplastic sheet (of type ABS) is heated until reaching the thermoforming window $140°C - 160°C$. The thermoplastic sheet, of dimension 1m×1,2m×0.012m, is placed at 0.20m from the upper and the lower parts of the oven. The heating time is fixed to 90 seconds. The thermal properties of the thermoplastic sheet are considered as independent from the temperatures. The average values that we considered are summarized as below [11]:

Thermal properties of ABS	values
Density $(kg.m^{-3})$	1050
Conduction $(W/m.k)$	0.174
Specific heat $(J/kg.k)$	2500

5.1 Analysis of the numerical results

The MBO metaheuristic was coded in C++ language, and debugged using Microsoft Visual Studio 2015 on an hp machine with an Intel(R)Core(TM) i5-6200U processor and a RAM of 16Go. The first stage of the simulation is dedicated to search the solution with MBO. The MBO algorithm was tested on 10 instances generated randomly. The solution that produces the best value of our objective function is chosen. The results are summarized in Table 3. Let us note that these values, generated by our meta-heuristic MBO algorithm, correspond to the percentage of the maximum power available for each of the (12×10) heating elements of the oven.

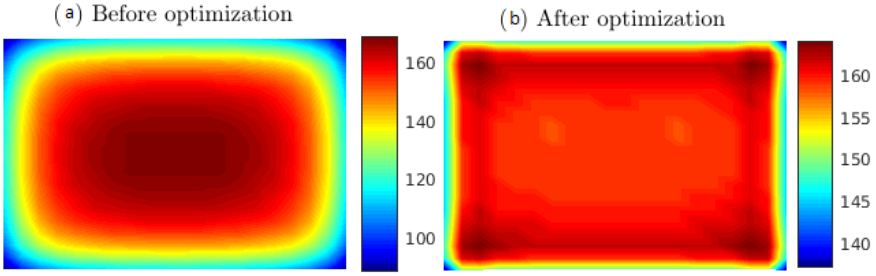
To show the quality of the solution produced by the MBO metaheuristic, we made a comparison between the temperature distribution obtained with the optimized solution of Table 3 and that calculated with a solution of which all the heating elements of the furnace are fixed to 68.57 % of the maximum power.

Table 3. Heater power distribution of the upper part of the oven

100	70.43	85	79	78.86	78.86	79	85	70.43	100
72.14	39.29	63.57	48.86	77.86	77.86	48.86	63.57	39.29	72.14
88.43	35	79.29	62.57	56.14	56.14	62.57	79.29	35	88.43
81.57	40.71	70.71	66	59.14	59.14	66	70.71	40.71	81.57
83.29	49.29	74.71	53.57	68.57	68.57	53.57	74.71	49.29	83.29
77.86	49.29	69.29	64.43	69.29	69.29	64.43	69.29	49.29	77.86
77.86	49.29	69.29	64.43	69.29	69.29	64.43	69.29	49.29	77.86
83.29	49.29	74.71	53.57	68.57	68.57	53.57	74.71	49.29	83.29
81.57	40.71	70.71	66	59.14	59.14	66	70.71	40.71	81.57
88.43	35	79.29	62.57	56.14	56.14	62.57	79.29	35	88.43
72.14	39.29	63.57	48.86	77.86	77.86	48.86	63.57	39.29	72.14
100	70.43	85	79	78.86	78.86	79	85	70.43	100

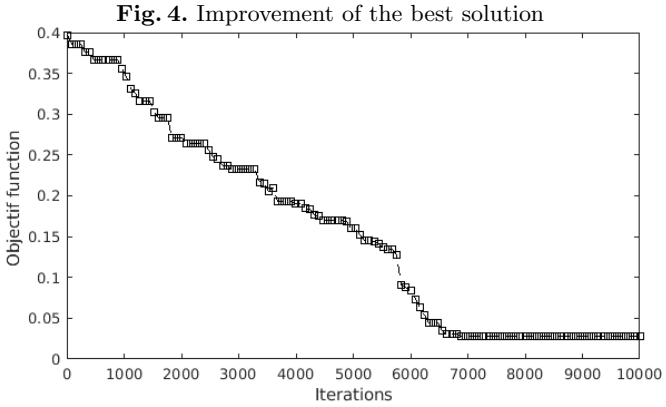
Figures 3-a and 3-b illustrate the distribution of the temperatures on the surface of the sheet before and after the optimization, respectively. The results show that the temperature differences between the center and the the borders of the sheet is greater before optimization. Indeed, for uniform heating of the temperatures of the elements of the oven, the temperature gap at the surface of the sheet is $60°C$ for a maximum of $160°C$ concentrated in the middle and a minimum of $100°C$ located at the borders of the surface of the thermoplastic sheet. With optimized heating, the gap temperatures between the borders and the center of the sheet is reduced to $20°C$. As we may see, the borders of the sheet are better heated with a temperature of $140°C$ and a central zone of $160°C$.

Fig. 3. Distribution of the temperature of the thermoplastic sheet



Figures 4 show the value improvement of the best solution versus the number of evaluated solutions. The results show, after 6500 iterations, the MBO method converged to the best solution. The execution time of the method does not exceed 20 minutes.

Fig. 4. Improvement of the best solution



6 Conclusion

We addressed in this paper the problem of distributing uniformly infrared radiative energy intercepted by a thermoplastic sheet surface during the infrared radiation transmitted by an oven with convection and conduction considerations. After discretizing this problem, we proposed an objective function that captures the uniform distribution of the radiative energy. With this approximation scheme, the corresponding problem appears that it nothing else than a variant of a quadratic assignment problem. Then, we designed and applied a migrating bird optimization based algorithm (MBO for short), in order to minimize the corresponding objective function.

We then conducted an experimental study to evaluate the quality of the solution produced by the MBO algorithm. This study reveals that the solutions generated by MBO improve the distribution of temperature, thus the infrared radiative energy, intercepted by the thermoplastic sheet surface during the infrared radiation stage.

References

1. Bejan A.: Convection Heat Transfer. 4th edition. Wiley, New York, USA (2013)
2. Bachir Cherif K., Rebaine D., Erchiqui F., Fofana I., Nabil N.: Numerically optimizing the distribution of the infrared radiative energy on a surface of a thermoplastic sheet surface. Heat transfer **140**(10), 1-7 (2018)
3. Bachir Cherif K., Rebaine D., Erchiqui F., Fofana I.: Metaheuristics as a solving approach for the infrared heating in the thermoforming process. In: GERAD-G-2015-139, Montreal, Canada (2015)
4. Burkard R., Dell'Amico M., Martello S.: Assignment Problems, revised reprint, SIAM (2012)

5. Duman E., Uysal M., Alkaya AF.: Migrating bird optimization: A new metaheuristic approach and its performance on quadratic assignment problem. Information Sciences **217**, 65-77 (2012)
6. Erchiqui F., Nahas N., Nourelfath M., Souli M.: A metaheuristic algorithms for optimization of infrared heating stage of the thermoforming process. International Journal of Metaheuristics **1**(3), 199-221, 2011
7. Monteix S., Schmidt Y., Le Maoult Y., Ben Yedder R., Diraddo R.W., Laroche D.: Experimental study and numerical simulation of perform or sheet exposed to infrared radiative heating. Journal of Materials Processing Technology **119**, 90-97 (2001)
8. Schmidt F., Le Maoult Y., Monteix S.: Modelling of infrared heating of thermoplastic sheet used in thermoforming process, Journal of Materials Processing Technology, 225-231 (2003)
9. Reddy J.: An Introduction to the Finite Element Method. 3rd edn. McGraw-Hill, New York (2006)
10. Throne JL.: Technology of Thermoforming. Hanser Publishers, Munich, Germany (1996)
11. Li ZZ., Heo KS., Seol SY.: Time-Dependent optimal heater control in thermoforming preheating using dual optimization steps. International Journal of Precision Engineering and Manufacturing **9**(4), 51-56 (2008)

Résolution d'un problème de Contrôle optimal en temps variant par la méthode des itérations variationnelles

Sarah GRIB[1], Abderrahmene AKKOUCHE[1,2] et Mohamed AIDENE[2]

[1] Laboratoire de Conception et Conduite des Systèmes de Production
Université MOULOUD MAMMERI de Tizi-Ouzou, 15 000 Tizi-Ouzou, Algérie
[2] Département de Mathématiques, Faculté des Sciences et des Sciences Appliquées,
Université AKLI MOHAND OULHADJ de Bouira, 10 000 Bouira, Algérie.
sarahgrib93@gmail.com, akkouche.abdo@yahoo.fr, aidene_2000@yahoo.fr

Résumé Dans cet article, une approche basée sur la méthode des itérations variationnelles (VIM) est proposée pour résoudre un problème de contrôle optimal de systèmes non autonomes, c'est-à-dire des problèmes de contrôle optimal en temps variant. L'idée consiste à déduire les conditions nécessaires d'optimalité en utilisant le principe du minimum de Pontryagin qui aboutit à un système d'équations différentielles ordinaires non autonomes soumises à des conditions aux limites, qui constituent un problème aux deux bouts (TPBVP). La loi de contrôle optimal et la trajectoire optimale sont déterminées en résolvant ce TPBVP en utilisant la méthode des itérations variationnelles.

Pour démontrer l'applicabilité de la méthode pour les problèmes du contrôle optimal en temps variant, trois exemples numériques sont traités et une comparaison avec la méthode de tir est effectuée.

Mots clés: Contrôle optimal en temps variant, Principe du minimum de Pontryagin, Méthode des itérations variationnelles, Équations différentielles ordinaires non autonomes.

1 Introduction

Le contrôle optimal est une branche de la théorie du contrôle qui s'applique dans plusieurs domaines tels que la physique, la chimie, l'économie, la robotique, etc...[16,18,7]. De point de vue mathématique, un système de contrôle est un système dynamique dépendant d'un paramètre appelé contrôle [23]. On peut le décrire soit par des équations différentielles ordinaires, des équations différentielles stochastiques, des équations intégrales, des équations aux dérivées partielles [15,13,22]. Les systèmes de contrôle gouvernés par des équations différentielles autonomes sont appelés systèmes à temps invariant, tandis que les systèmes décrits pas les équations différentielles non autonomes sont appelés systèmes à temps variant ou des systèmes non stationnaires.
Pour résoudre un problème de contrôle optimal, on trouve dans la littérature deux méthodes: la méthode directe et la méthode indirecte [21]. La méthode directe consiste à transformer le problème de contrôle optimal en problème

de programmation non linéaire, qu'on résout par la suite par des méthodes d'optimisation classiques [4].

La méthode indirecte, basée sur le principe optimiser ensuite discrétiser, consiste à dériver d'abord les conditions d'optimalité en utilisant soit la programmation dynamique basée sur le principe de Bellman [3], ou sur l'approche variationnelle basée sur le principe du minimum de Pontrayagin [17]. En utilisant le Principe du minimum de Pontryagin les conditions nécessaires d'optimalité sont données sous la forme d'un système d'équations différentielles avec des conditions aux limites, qui constituent un problème aux deux bouts.

Pour déterminer la solution du problème aux deux bouts ainsi obtenu, plusieurs méthodes que se soient numériques ou semi-analytiques ont été développées et utilisées par des chercheurs. Pour les méthodes numériques on peut citer la méthode de tir, la méthode de tir multiple, la méthode de collocation indirecte [19]. Pour les méthodes semi-analytiques on peut trouver la méthode de décomposition d'Adomian [2], Méthode d'analyse d'homotopie [12] , la méthode des itérations variationnelles [8,10], la méthode de perturbation d'homotopie [9], La méthode de transformation différentielle [25]. Ces méthodes semi-analytiques permettent de déterminer une solution exacte ou approchée d'une équation différentielle linéaire ou non linéaire en utilisant un schéma itératif. Ces méthodes sont aussi utilisées pour déterminer la solutions de problèmes de contrôle optimal à temps invariant en utilisant soit l'approche basée sur la programmation dynamique ou celle basée sur l'approche variationnelle [14,6,20,24,5].

Dans cet article on utilise la méthode des itérations variationnelles [8,10] pour approcher la solution d'un problème de contrôle à temps variant, c'est-à-dire les problèmes de contrôle optimal des sytèmes non stationnaires. L'idée consiste à utiliser l'approche indirecte basée sur le principe du minimum de Pontryagin pour dériver les conditions nécessaires d'optimalité qui résulte en un système d'équations différentielles ordinaires, appelées équations de Hamilton-Pontryagin, non autonomes. La méthode des itérations variationnelles a été développée par le mathématicien chinois J. H. He en 1997 pour résoudre les équations différentielles à retard [8], et depuis cette méthode est trés utilisée par des chercheurs pour déterminer la solution de différents type de problèmes linéaires ou non linéaires.

Le reste de cet article est structuré comme suit : dans la section (2) on donne la formulation générale du problème considéré dans cet article. Dans la section (3), on donne les conditions nécessaires d'optimalité obtenues en utilisant le principe du minimum de Pontryagin. Dans la section (4), on explique le principe général de la méthode des itérations variationnelles et une approche basée sur cette méthode pour déterminer la solution des conditions nécessaires d'optimalité est résumée à la section (5), et illustrée par des exemples d'application à la section (6), et on finira par une conclusion générale.

2 Position du Problème

Dans cet article on considère le problème de contrôle optimal suivant :

$$\min_{u(t)} J(u(t)) = \frac{1}{2}\int_{t_0}^{t_f} \Big(x(t)^T\, Q(t)\, x(t) + u(t)^T\, R(t)\, u(t)\Big)\, dt, \tag{1}$$

avec l'équation d'état :

$$\dot{x}(t) = A(t)\, x(t) + B(t)\, u(t), \tag{2}$$

et les conditions aux limites

$$\Xi(t_0,\, x(t_0)) = 0, \Upsilon(t_f, x(t_f)) = 0, \tag{3}$$

où $x(t) \in \mathbb{R}^n$ et $u(t) \in \mathbb{R}^m$ sont les vecteurs d'état et de contrôle, respectivement. Les matrices $Q(.) \in \mathcal{M}_n(\mathbb{R})$ est semi-définie positive, $R(.) \in \mathcal{M}_m(\mathbb{R})$ est définie positive, $A(.) \in \mathcal{M}_n(\mathbb{R})$, et $B(.) \in \mathcal{M}_{n,m}(\mathbb{R})$. De plus, on suppose que $x(t)$ est absolument continue sur $[t_0,\, t_f]$, Q, R et A sont des applications localement intégrables sur $[t_0,\, t_f]$. B et $u(t)$ sont des applications carrés intégrables sur $[t_0,\, t_f]$. $\Xi : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^{r\leq n}$, et $\Upsilon : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^{l\leq n}$ sont des fonctions vectorielles qui sont suposées différentiables par rapport à $x(t)$. L'objectif est de déterminer le meilleur contrôle $u(t)$ qui transfère le système de l'état initial vers l'état final en minimisant le cout $J(u(t))$.

3 Condition nécessaires d'optimalité

Pour dériver les conditions nécessaires d'optimalité, on introduit le principe du minimum de Pontraygin [17]. Comme le problème considéré est convexe alors le théorème suivant nous donne une condition nécessaire et suffisante pour l'optimalité de la solution.

Theorem 1. *Soit $(x(t), u(t))$, $t \in [t_0,\, t_f]$ une solution optimale du problème linèaire quadratique (1)-(??), avec $R(t) > 0$, pour tout $t \in [t_0,\, t_f]$. Alors il existe un vecteur de l'état adjoint différentiable $p(t) \in \mathbb{R}^n$, $t \in [t_0,\, t_f]$ telle que la fonction hamiltonienne définie par :*

$$\begin{aligned} H(x(t),\, u(t),\, p(t),\, t) = \frac{1}{2}&\left[x^T(t)\, Q(t)\, x(t) + u^T(t)\, R(t)\, u(t)\right] \\ &+ p^T(t)\left[A(t)\, x(t) + B(t)\, u(t)\right], \end{aligned} \tag{4}$$

satisfait les conditions suivantes :

$$\dot{x}(t) = \frac{\partial H}{\partial p(t)}(x(t),\, u(t),\, p(t),\, t) = A(t)\, x(t) + B(t)\, u(t), \tag{5}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial x(t)}(x(t),\, u(t),\, p(t),\, t) = -Q(t)\, p(t) - A^T(t)\, p(t), \tag{6}$$

$$\frac{\partial H(x(t),\, u(t),\, p(t),\, t)}{\partial u(t)} = R(t)\, u(t) + B^T(t)\, p(t) = 0. \tag{7}$$

avec les conditions aux limites:

$$\Xi(t_0,\, x(t_0)) = 0, \tag{8}$$

$$\Upsilon(t_f, x(t_f)) = 0, \tag{9}$$

$$p(0) = \mu^T \, \frac{\partial \Xi(t_0,\, x(t_0))}{\partial x}, \tag{10}$$

$$p(0) = -\eta^T \, \frac{\partial \Upsilon(t_0,\, x(t_0))}{\partial x}. \tag{11}$$

4 La méthode des itérations variationnelles

Pour illustrer le principe de la méthode des itérations variationnelles [8,10], on considère l'équation différentielle écrite sous la forme canonique suivante :

$$L\, y(t) + N\, y(t) = g(t), \tag{12}$$

où L et N sont respectivement l'opérateur linéaire et non linéaire, et $g(t)$ est une fonction analytique donnée.

Pour obtenir la solution de l'équation (12), on construit une fonctionnelle de correction de la forme :

$$y_{n+1}(t) = y_n(t) + \int_0^t \lambda(\tau) \, \left(L\, y_n(\tau) - N\, \tilde{y}_n(\tau) - g(\tau)\right) \, d\tau, \tag{13}$$

où λ est un multiplicateur de Lagrange [11], qui peut être identifié par la théorie des calcul des variations, $\tilde{y}_n$ est considérée comme la variation restreinte qui signifie que $\delta\, \tilde{y}_n = 0$.

L'étape principale de la méthode des itérations variationnelles est d'abord la détermination du multiplicateur de Lagrange λ de faÃ§on optimale par une intégration par partie. Une fois le multiplicateur de Lagrange λ est identifié. On choisit une approximation initiale $y_0(t)$ de la solution du problème (12), et les autres approximations successives $y_i(t)$, $i \geq 0$ seront obtenues en utilisant la fonctionnelle de correction (13). Par conséquent, la solution exacte est donnée par :

$$y(t) = \lim_{n \to \infty} y_n(t). \tag{14}$$

5 Procédure de l'approche proposée

L'idée générale de l'approche proposée consiste à utiliser la méthode indirecte, basée sur le principe du minimum de Pontryagin pour dériver les conditions nécessaires d'optimalité données sous la forme d'un système de $2n$ équations différrentielles non autonomes avec des conditions aux deux bouts bien détérminées, qui donne un problème aux limites. Ensuite on utilise la méthode des itérations variationnelles pour approcher la solution du problème aux deux bouts ainsi obtenu. Les différrentes étapes de l'approche proposée sont résumées comme suit :

1. Dériver les conditions nécessaires d'optimalité en utilisant le principe du minimum de Pontryagin,
2. Choisir un seuil de précision $\epsilon > 0$, et poser $k = 0$. Choisir $x^0(t) = x(0)$ et $p^0(t) = p(0)$ comme approximation initiales. Si les conditions initiales ne sont pas définies alors choisir $x^0(t) = \Sigma$ et $p^0(t) = \Pi$ où Σ et Π sont des vecteurs de paramètres inconnus à déterminer en imposant les conditions aux limites.
3. Déterminer les solutions approchées $x^{k+1}(t)$ et $p^{k+1}(t)$ en utilisant la fonctionnelle de correction (13).
4. Déterminer les paramètres inconnus en imposant les conditions aux limites,
5. Déduire le contôle optimal $u^k(t)$ et évaluer le coût $J(u^k(t))$,
6. Critère d'arrêt : si

$$\frac{\mid J(u^k(t)) - J(u^{k-1}(t)) \mid}{\mid J(u^k(t)) \mid} < \epsilon, \tag{15}$$

stop, sinon poser $k = k + 1$ et aller à l'étape 3.

6 Application

Pour illustrer l'efficacité de la méthode VIM pour la résolution des problèmes de contrôle optimal décrits par des équations différentielles non autonomes, on considère trois exemples d'applications. Dans le premier exemple on considère un système avec l'état initial fixe et l'état final libre. Dans le deuxième et le troisième exemples, on traitera des sytèmes avec des contraintes sur l'état initial et sur l'état final.

Exemple1 Considérons le problème de contrôle optimal suivant :

$$\min_{u(t)} J(u(t)) = \frac{1}{2} \int_0^1 u^2(t)dt, \tag{16}$$

$$\dot{x}_1(t) = t\, x_1(t) + x_2(t), \tag{17}$$

$$\dot{x}_2(t) = t^3 x_2(t) + u(t), \tag{18}$$

$$\dot{x}_3(t) = t^2 x_3(t) + u(t), \tag{19}$$

$$x_1(0) = -2, \quad x_2(0) = 0, \quad x_3(0) = 1, \tag{20}$$

et les conditions nécessaires d'optimalité obtenues en appliquant le principe du minimum de Pontryagin, sont données comme suit :

$$\dot{x}_1(t) = t\,x_1(t) + x_2(t), \tag{21}$$
$$\dot{x}_2(t) = t^3 x_2(t) - p_2(t) - p_3(t), \tag{22}$$
$$\dot{x}_3(t) = t^2 x_3(t) - p_2(t) - p_3(t), \tag{23}$$
$$\dot{p}_1(t) = -t\,p_1(t), \tag{24}$$
$$\dot{p}_2(t) = -p_1(t) - t^3\,p_2(t), \tag{25}$$
$$\dot{p}_3(t) = -t^2\,p_3(t), \tag{26}$$
$$x_1(0) = -2, \quad x_2(0) = 0, \quad x_3(0) = 1, \tag{27}$$
$$p_1(1) = 0, \quad p_2(1) = 0, \quad p_3(1) = 0, \tag{28}$$

et le contrôle optimal $u(t)$ est donné par :

$$u(t) = -p_2(t) - p_3(t). \tag{29}$$

Pour déterminer une solution approchée pour le système (21)–(28), on construit les formules itératives suivantes :

$$x_1^{k+1}(t) = x_1^k(t) - \int_0^t \left(\dot{x}_1^k(\tau) - \tau\,x_1^k(\tau) - x_2^k(\tau)\right)\,d\tau, \tag{30}$$
$$x_2^{k+1}(t) = x_2^k(t) - \int_0^t \left(\dot{x}_2^k(\tau) - \tau^3\,x_2^k(\tau) + p_2^k(\tau) + p_3^k(\tau)\right)\,d\tau, \tag{31}$$
$$x_3^{k+1}(\tau) = x_3^k(t) - \int_0^t \left(\dot{x}_3^k(\tau) - \tau^2\,x_3^k(\tau) + p_2^k(\tau) + p_3^k(\tau)\right)d\tau, \tag{32}$$
$$p_1^{k+1}(t) = p_1^k(t) - \int_0^t \left(\dot{p}_1^k(\tau) + \tau\,p_1^k(\tau)\right)d\tau, \tag{33}$$
$$p_2^{k+1}(t) = p_2^k(t) - \int_0^t \left(\dot{p}_2^k(\tau) + p_1^k(\tau) + \tau^3\,p_2^k(\tau)\right)d\tau, \tag{34}$$
$$p_3^{k+1}(t) = p_3^k(t) - \int_0^t \left(\dot{p}_3^k(\tau) + \tau^2\,p_3^k(\tau)\right)d\tau. \tag{35}$$

En choisissant $x_1^0(t) = x_1(0) = -2$, $x_2^0(t) = x_2(0) = 0$, $x_3^0(t) = x_3(0) = 1$, $p_1^0(t) = p_1(0) = a$, $p_2^0(t) = p_2(0) = b$, et $p_3^0(t) = p_3(0) = c$ comme approximations initiales, avec a, b et c sont des paramètres inconnus qui seront déterminés en imposant la conditions terminales $p_1(1) = p_2(1) = p_3(1) = 0$. Les résultats obtenus sont reportés dans la Table 1.

Table 1. Résultats de l'exemple 1

k	a	b	c	J^k	$\frac{\mid J^k - J^{k-1} \mid}{\mid J^k \mid}$
1	2	4	1.5	9.173413	–
2	1.6	2.577066	1.384615	4.836323	0.8967743
3	1.655172	2.781554	1.396552	5.384428	0.1017945
4	1.648069	2.754174	1.395549	5.309359	$0.1414e-1$
5	1.648776	2.756990	1.395616	5.317056	$0.1447e-2$
6	1.648717	2.756748	1.395612	5.316393	$0.1247e-3$
7	1.648722	2.756766	1.395612	5.316441	$0.9028e-5$
8	1.648721	2.756765	1.395612	5.316439	$0.3762e-6$

Fixons $\epsilon = 10^{-6}$, la méthode converge à la 8ieme itération, et les graphes de $x^8(t)$ et $u^8(t)$ tracés avec les résultats obtenus en appliquant la méthode de tir à la figure (**Fig.1**).

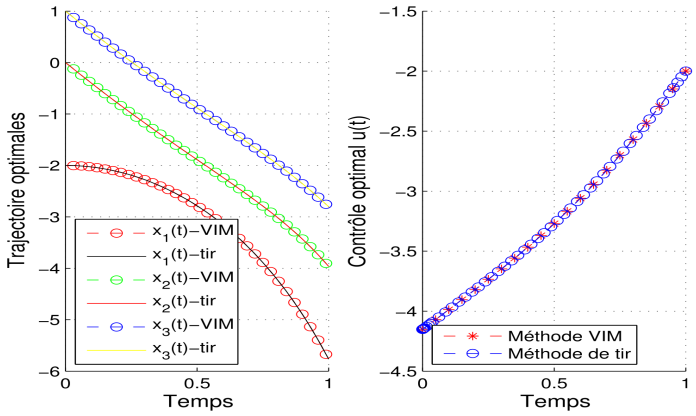


Fig. 1. Trajectoires optimales

Exemple 2 Considérons le problème de contrôle optimal suivant avec une entrée libre, c'est-à-dire une contrainte sur l'état initial:

$$\min_{u(t)} J(u(t)) = \frac{1}{4}\int_0^1 (x^2(t) + u^2(t))\,dt \tag{36}$$

$$\dot{x}(t) = \frac{1}{2}\,t\,x(t) + \frac{1}{2}\,u(t), \tag{37}$$

$$x(0) \quad \text{libre}, \quad x(1) = 1. \tag{38}$$

Les conditions nécessaires d'optimalité qui sont données comme suit :

$$\dot{x}(t) = \frac{1}{2}\,t\,x(t) - p(t), \tag{39}$$

$$\dot{p}(t) = -\frac{1}{2}\,x(t) - \frac{1}{2}\,t\,p(t), \tag{40}$$

$$x(0) = \text{libre}, \quad p(0) = a, \tag{41}$$

où a est un paramètre inconnu qui sera déterminé en imposant la condition terminale $x(1) = 1$. Les résultats obtenus par application de la mt́hode des itérations variationnelles sont reportés dans la Table 2.

Table 2. résultats de l'exemple 2

k	a	J^k	$\frac{\mid J^k - J^{k-1} \mid}{\mid J^k \mid}$
1	0.8	.202	–
2	0.653061	0.157337	0.283868
3	0.643215	0.154873	$0.159e-1$
4	0.637666	0.152894	$0.129e-1$
5	0.637393	0.152810	$0.549e-3$
6	0.637288	0.152769	$0.268e-3$
7	0.637284	0.152768	$0.655e-5$

En prenant $\epsilon = 10^{-5}$ et en imposant le critère d'arrêt (15), la méthode converge après sept itérations. L'approximation de la trajectoire $x(t)$ et la loi du contrôle optimal $u(t)$ est donnée à la figure (**Fig.2**).

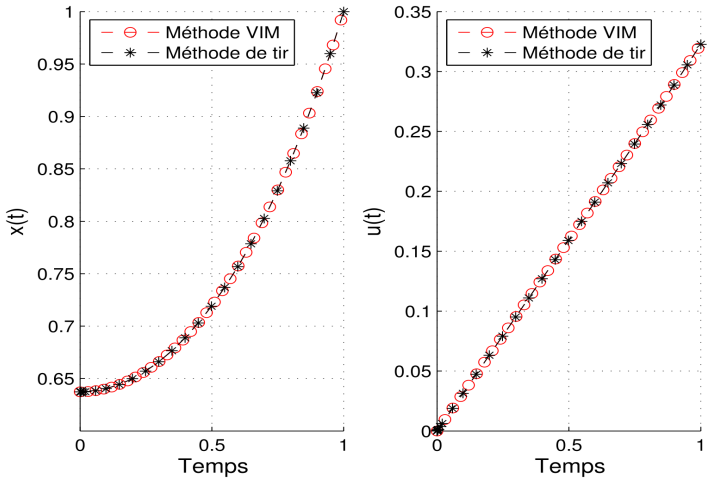


Fig. 2. Trajectoires optimales

Exemple 3 Considérons le problème suivant avec une contrainte sur l'état final:

$$\min_{u(t)} J(u(t)) = \frac{1}{2}\int_0^{\pi/2} \left(4\,x_2^2(t) + u^2(t)\right), \tag{42}$$

$$\dot{x}_1(t) = t\,u(t), \tag{43}$$

$$\dot{x}_2(t) = t^2\,x_1(t), \tag{44}$$

avec les conditions initiales

$$x_1(0) = 2, \quad x_2(0) = 1, \tag{45}$$

et une contrainte sur l'état final :

$$x_1(\frac{\pi}{2}) + x_2(\frac{\pi}{2}) = 1. \tag{46}$$

Les conditions nécessaires d'optimalité sont données comme suit:

$$\dot{x}_1(t) = -t^2\,p_1(t), \tag{47}$$

$$\dot{x}_2(t) = t^2\,x_1(t), \tag{48}$$

$$\dot{p}_1(t) = -t^2\,p_2(t), \tag{49}$$

$$\dot{p}_2(t) = -4\,x_2(t) \tag{50}$$

soumises aux conditions :

$$x_1(\frac{\pi}{2}) + x_2(\frac{\pi}{2}) = 1, \tag{51}$$

$$p_1(\frac{\pi}{2}) = -\eta, \tag{52}$$

$$p_2(\frac{\pi}{2}) = -\eta. \tag{53}$$

En choisissant $x_1^0(t) = x_1(0) = 2$, $x_2^0(t) = x_2(0) = 2$, $p_1^0(t) = p_1(0) = a$ et $p_2^0(t) = p_2(0) = b$ comme approximations initiales, et en choisissant $\epsilon = 10^{-4}$ comme un seuil de precision, la méthode VIM converge après 15 itérations, et les solutions $\epsilon-$optimales sont données comme suit:

$$\begin{aligned} x_1(t) &= 2 - 1.499\,t^3 + 0.4847\,t^6 - 0.1429\,t^7 - 0.9524e-2\,t^{10} + 0.0011\,t^{13} \\ &\quad - 0.1036e-3\,t^{16} + 0.2183e-4\,t^{17} + 0.6156e-6\,t^{20} - 0.3512e-7\,t^{23} \\ &\quad + 0.1823e-8\,t^{26} - 0.3208e-9\,t^{27} - 0.5508e-11\,t^{30} + 0.2021e-12\,t^{33} \\ &\quad - 0.7055e-14\,t^{36} + 0.1097e-14\,t^{37}, \end{aligned} \tag{54}$$

$$\begin{aligned} x_2(t) &= 1 + .6667\,t^3 - 0.2499\,t^6 + 0.0538\,t^9 - 0.0143\,t^{10} - 0.7326e-3\,t^{13} \\ &\quad + 0.6865e-4\,t^{16} - 0.5451e-5\,t^{19} + 0.1091e-5\,t^{20} + 0.2677e-7\,t^{23} \\ &\quad - .1351e-8\,t^{26} + 0.6285e-10\,t^{29} - 0.1069e-10\,t^{30} - 0.1669e-12\,t^{33} \\ &\quad + 0.5614e-14\,t^{36} - 0.1809e-15\,t^{39}, \end{aligned} \tag{55}$$

$$\begin{aligned} u(t) &= -t\,(4.498 - 2.908\,t^3 + t^4 + 0.09524\,t^7 - 0.01428\,t^{10} + 0.0016\,t^{13} \\ &\quad - 0.0037\,t^{14} - 0.1231e-4\,t^{17} + 0.8077e-6\,t^{20} - 0.4740e-7\,t^{23} \\ &\quad + 0.8661e-8\,t^{24} + 0.1652e-9\,t^{27} - 0.6670e-11\,t^{30} \\ &\quad + 0.2540e-12\,t^{33} - 0.4058e-13\,t^{34} - 0.5307e-15\,t^{37}). \end{aligned} \tag{56}$$

Dans la figure (**Fig.3**), on trace les graphes des trajectoires $x_1(t)$ et $x_2(t)$ ainsi que la trajectoire du contrôle optimal $u(t)$ avec celles obtenues par la méthode de tir, ce qui montre que les rsultats sont trés proches.

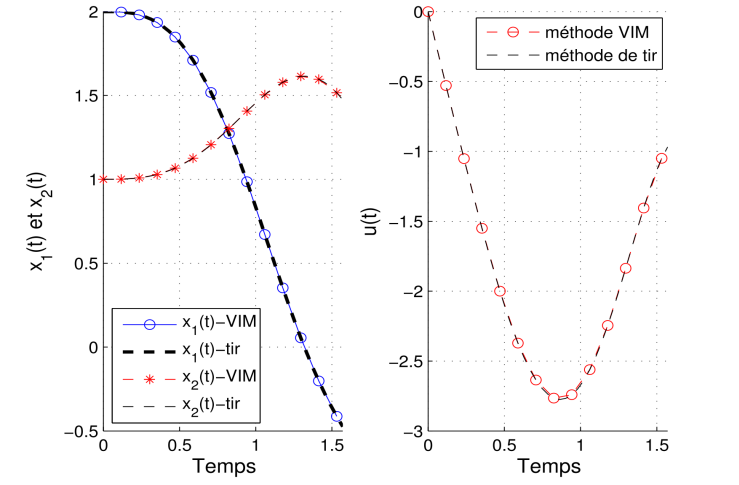


Fig. 3. Trajéctoire optimales.

7 Conclusion

Dans ce travail, on propose une approche basée sur la méthode des itérations variationnelles pour approcher la solution d'un problème de contrôle optimale à temps variant. l'idée principale consiste à résoudre les conditions nécessaires d'optimlaité, dérivées en utilisant le principe du minimum de Pontryagin, données par un système d'équations différentielles ordinaires non autonomes. La solution est obtenue d'une manière itérative en utilisant une fonctionnelle de correction. La méthode proposée donne une solution approchée avec une grande précision en un nombre fini d'itérations, et attaque le problème de manière directe sans discrétisation ou perturbation avec des petits paramètres. L'approche proposée est illustrée par trois exemples d'application, et les résultats obtenus sont très proches de ceux obtenus avec la méthode de tir ce qui confirme l'efficacité de l'approche developpé.

References

1. I.H. Abdel-Halim Hassan: Differential transformation technique for solving higher-order initial value problems: Applied Mathematics and Computation: 154 (2004) 299–311.
2. G. Adomian: Solving frontier problems of physics: the Decomposition Method: Kluwer Academic Publishers, Dordrecht, The Netherlands: (1994).
3. R. Bellman: Dynamic Programming: Princeton University Press: Princeton, New Jersey (1957).

4. J. T. Betts: Practical Methods for Optimal Control and Estimation Using Nonlinear Programming: Society for Industrial and Applied Mathematics (2009).
5. S. Effati and S. Nik: Solving a class of linear and non-linear optimal control problems by homotopy perturbation method: IMA Journal of Mathematical Control and Information: 28 (2011) 539–553.
6. M. A. El-Tawil and A. A. Bahnasawi and A. Abdel-Naby: Solving Riccati differential equation using Adomian's decomposition method: Applied Mathematics and Computation: 157 (2004) 503–514.
7. W. L. Garrard and J. M. Jordan: Design of nonlinear automatic flight control systems: Automatica: 13 (1997) 497–505.
8. J.H. He: Variational Iteration Method for Delay Differential Equations: Communications in Nonlinear Science and Numerical Simulation: 2 (1997) 235–236.
9. J. H. He: Homotopy perturbation technique: Computer Methods in Applied Mechanics and Engineering: 178, (1999) 257–262
10. J.H.He: Communications in Nonlinear Science and Numerical Simulation: International Journal of Non-Linear Mechanics: 4 (1999) 699-708.
11. M. Inokuti and H. Sekine and T. Mura: General use of the Lagrange multiplier in nonlinear mathematical physics, in : S. Namat-Nasser, Variational Method in the Mechanics of Solids: Pergamon Press, Oxford, (1978)
12. S. J. Liao: The proposed homotopy analysis technique for the solution of nonlinear problems: Ph.D. Thesis, Shanghai Jiao Tong University (1992).
13. J. L. Lions: Optimal Control of Systems Governed by Partial Differentiel Equations: Springer-Verlag, New York (1971).
14. A. Maidi and J.P. Corriou: Open-loop optimal controller design using variational iteration method: Applied Mathematics and Computation: 219 (2013) 8632–8645.
15. D. S. Naidu: Optimal Control Systems: CRC Press, Boca Raton, Florida: (2003).
16. T. Notsu, M. Konishi and J. Imai: Optimal water cooling control for plate rolling: International Journal of Innovative Computing, Information and Control: 4 (2008) 3169–3181.
17. L. S. Pontryagin and V. G. Boltyanskii and R. V. Gamkrelidze and E. F. Mishchenko: The Mathematical Theory of Optimal Processes: Pergamon Press, New York: (1964).
18. J.Flick, M.Ruggenthaler, H.Appel and A.Rubio: Atoms molecules in cavities, from week to strong coupling in quantum-electrodynamics(QED)chemistry: Proceedings of the National Academy of Sciences, 114(12)(2017) 3026-3034.
19. A. V. Rao: A survey of numerical methods for optimal control: Applied Mathematics and Computation: 345,(2007) 543–548.
20. H. Saberi Nik and S. Effati and A. Yildirim: Solution of linear optimal control systems by differential transform method: Neural Computing and Applications: 23 (2013) 1311–1317.
21. R. W. H. Sargent: Optimal Control: Journal of Computational and Applied Mathematics:124, (2000) 361–371.
22. J. L. Stein: Stochastic Optimal Control, International Finance, and Debt Crises: Oxford University Press, New York (2006)
23. E. Trélat: Contôle optimal : théorie et applications: Mathématiques concrètes, Vuibert, France (2011).
24. M.S. Zahedi and H.S. Nik: On homotopy analysis method applied to linear optimal control problems: Article in Press: (2013).
25. J. K. Zhou: Differential transformation and its applications for electrical circuits: Huazhong University Press, Wuhan (1986).

Répartition Economique Environnementale de l'Energie

avec l'Algorithme de Décoration Intérieure

Latifa DEKHICI*[1], Khaled GUERRAICHE[2] and Khaled BELKADI[1]

[1] LAMOSI, Faculté de Mathématiques et d'Informatique, Université des Sciences et de la Technologie d'Oran, Mohamed Boudiaf (USTO-MB), Algérie.

[2] École supérieure en génie électrique et énergétique, Oran, Algérie

latifa.dekhici@univ-usto.dz

Résumé. Ce papier a pour but de résoudre le problème de distribution économique et environnementale de l'énergie électrique. Ce problème bi-objectif vise la minimisation du coût du carburant et des polluants NOx lors de la distribution aux clients. L'énergie totale doit être égale à celle demandée en tenant compte de la puissance perdue. Pour cela, nous utilisons une métaheuristique relativement récente inspirée du comportement humain qui est la recherche de décoration intérieure (interior search algorithm). L'approche a été testée sur un système thermique à dix nœuds.

Mots clés: Répartition de l'énergie, environnement, coût, l'Algorithme de Décoration Intérieure, métaheuristiques.

1 Introduction

Actuellement, un ensemble de métaheuristiques bio-inspirées basées sur le comportement naturel des essaims, des abeilles, des oiseaux, et des loups avaient émergé comme une alternative pour surmonter les difficultés présentées par les méthodes classiques dans l'optimisation. Un des problèmes d'optimisation est le problème de répartition des puissances. Sa résolution informatique par des méthodes approchées fait éviter des pertes pécuniaires considérables. Ce papier a pour but de résoudre le problème de distribution économique et environnementale de l'énergie électrique qui a comme objectif de minimiser le coût du carburant et l'émission des polluants NOx lors de la répartition des charges. L'énergie totale doit être égale à celle demandée en tenant compte de la puissance perdue. Pour cela, nous utilisons une métaheuristique relativement récente inspirée du comportement humain qui est la recherche de décoration intérieure (interior search). Dans la deuxième section, nous décrivons le problème de la répartition des charges et sa formulation. Dans la troisième section, nous présentons la métaheuristique, son origine et ses paramètres. Dans la dernière section, nous discuterons les résultats de la métaheuristique sur un

exemple de réseau électrique. Enfin, nous donnons une conclusion et des perspectives.

2 Problème de distribution économique et environnementale de L'énergie

2.1 Description

De nos jours, avec l'augmentation des besoins en énergie électrique dans tous les domaines de la vie, l'optimisation de la planification des systèmes d'énergie suscite un intérêt considérable en termes de coûts d'énergie pour l'acheminement de l'énergie électrique. La Répartition Economique consiste à répartir la demande électrique totale du réseau entre ses différentes unités de production de manière à avoir le coût de la production le plus réduit possible. Ce coût diffère d'une unité à une autre et il est, entre autre, fonction du carburant utilisé pour la production électrique (charbon, fuel, gaz naturel, uranium, eau…). En outre, la sensibilisation croissante du public à la protection de l'environnement fait de la minimisation des émissions atmosphériques de polluants des centrales thermiques un autre enjeu important de répartition de l'énergie. Dans la littérature, le problème de la répartition de la puissance économique désigne généralement le contrôle de la production du générateur engagé afin de minimiser le coût total du carburant tout en satisfaisant la demande de puissance et d'autres contraintes. Alors que la répartition environnementale de la puissance vise à minimiser l'émission de polluants. Donc, le problème de répartition de la puissance peut être traité comme un problème d'optimisation multi-objectif avec des objectifs non commensurables et contradictoires.

Le lecteur peut se référer aux revues récentes de [11] [4] au sujet des algorithmes utilisés. En [13], Qu et al. donnent un aperçu des algorithmes évolutionnaires multi-objectifs dans le domaine. La plupart ont utilisé des méthodes conventionnelles . Peu de chercheurs ont utilisé des métaheuristiques.. Nous citons par exemple [6] où les auteurs combinent la répartition économique / environnementale en utilisant l'algorithme des lucioles (firefly algorithm) et l'algorithme de chauve-souris (bat algorithm). L'optimisation de coucou (cuckoo algorithm) a été utilisée dans [10] pour la répartition économique. Des chercheurs ont proposé un algorithme de recherche d'organismes symbiotiques pour une répartition économique / d'émission multizone à grande échelle en tenant compte de l'effet Valve-point et des pertes de transmission [3]. Jayabarathi et al. en [7] ont inclus le croisement et la mutation à l'optimisation inspirée du loup gris (Grey Wolf optimization) pour résoudre les problèmes de répartition économique avec des zones d'opération interdites, l'effet de chargement de la valve et les limites de vitesse de rampe. Narang et autres [14] ont décrit une approche intégrée qui intègre l'optimisation par essaims civilisés et la recherche de motifs de Powell. En [12], un algorithme génétique amélioré et une méthode basée sur la programmation linéaire mixte améliorée ont été proposé pour la distribution

économique des unités de micro-réseaux. Les techniques de résolution des problèmes de répartition de la puissance sont soumises à des contraintes de temps et sont irréalisables avec l'émergence de nouvelles contraintes liées à l'évolution de la technologie. L'adoption de méthodes conventionnelles ou évolutives classiques reste insuffisante. Cependant, avec la croissance des métaheuristiques inspirées de la nature, l'optimisation de la répartition de la puissance avec ces méthodes intelligentes devra inévitablement intéresser la plupart des chercheurs.

2.2 Formulation

Problème économique

Etant donné un réseau de N nœuds générateurs où : : Puissance du nœud générateur k et : Puissance perdue. Dans la version la plus simple du problème avec prise en compte des pertes, l'objectif est de réduire au minimum la fonction objectif de la répartition optimale des puissances actives définie par (1).

(1)

Tel que : : Coût total et : Coût du nœud k

En utilisant en considération la contrainte d'égalité (2) à la puissance demandée Pd, et la contrainte d'inégalité (3).

(2)

(3)

La fonction du coût peut être formulée par (4)

(4)

Tel que : : Coefficients du coût de production pour le $k^{ième}$ nœud par heure

Avec des effets thermiques, le coût des générateurspeut être formulé par (5).

(5)

Avec , : coefficients du coût de la production thermique

Problème Environnemental. Des polluants atmosphériques peuvent être émis tels que les oxydes de soufre (SOx) et les oxydes d'azote (NOx) provoqués par des unités thermiques classiques. Cependant l'émission totale de ces polluants peut être exprimée en(6):

(6)

Si l'objectif est économique, on peut ajouter la contrainte environnementale(7).

(7)

Tel que : ME est le maximum de polluants permis et sont des coefficients d'émission. Sinon (6) est considérée comme fonction objectif.

Problème Eco./Environnemental. L'optimisation choisie utilise une fonction Pareto objectif avec pénalité. Pour évaluer la fonction globale du problème de répartition de puissance multi-objectif qui minimise le coût et l'émission, des fonctions de Pareto et de pénalité sont utilisées. La fonction de pénalité est formée en ajoutant à la contrainte dure comme la demande un grand nombre positif C (999999). Le poids pour l'émission de NOx E et le coût du combustible FCost sont respectivement λ_1 et λ_2. Ces coefficients sont choisis de manière à ce que l'émission et le coût du combustible soient dans la même échelle.

(8)

La somme des abus des contraintes dures est estimée par le non respect de la puissance demandée et de la charge minimale et maximale de chaque générateur.

3 Algorithme de Recherche ou de Décoration Intérieure

La méthode de décoration intérieure dont le nom en anglais « Interior Search algorithm» n'est pas expressif de son origine est une métaheuristique récente inspirée de la décoration persienne [5]. Elle est aussi basée sur des formules en fonction des bornes du domaine de la solution. Idéale pour l'optimisation avec contraintes, elle a été utilisée dans plusieurs problèmes tels que d'écoulement de puissance optimal [1], la conception de différentiateurs numériques [8], l'identification adaptative du système de réponse impulsionnelle infinie [9], l'ordonnancement des ateliers de types flow shop hybride [2].

3.1 Métaphore de la Décoration Intérieure

Dans la décoration Intérieure deux principes d'optimisation peuvent être vus :

La Composition du décor. La décoration est un projet qui utilise des ressources et doit satisfaire le client et les contraintes d'espace et atteindre l'objectif. Pendant ce processus qui commence souvent des murs (bornes) au centre, le décorateur peut changer l'emplacement des éléments existants pour plus d'esthétique.

La Miroiterie. Les perses utilisent les miroirs pour reproduire un décor dans une pièce. Ils placent ces miroirs devant le plus beau décor pour renforcer son effet.

3.2 Paramètres de l'algorithme (ISA : Interior Search Algorithm)

Inspiré des deux principes cités, l'algorithme ISA génère aléatoirement et sans dépasser les bornes X_{min} et X_{max} de l'espace des localisations X_i pour les éléments de décor. Il évalue leurs intensités et détecte le plus bel élément. Il divise les éléments en 2 groupes selon un paramètre d'éventualité α :

– Un groupe de composition où le déplacement à l'itération t est régénéré selon les bornes du domaine (9) :

= (9)

– Un groupe de miroiterie où le déplacement à l'itération t (11) est un reflet de miroir (10) autour du meilleur élément X^*.

(10)

= (11)

Alors que le meilleur élément est déplacé de sa position à l'itération précédente t-1 d'un pas aléatoire rand() multiplié par λ.

= (12)

Algorithme ISA

```
Initialisation
Pour(t=1 à nombre Itérations)
Pour (i=1 à nombre Eléments Décor)
si  f()==f()
=   // déplacer le meilleur d'un pas
Sinon
   si rand()<=α // deviser en 2 groupes
        // placer un miroir
      = // reflet du miroir
   sinon // groupe composition
=//regénérer position
   Finsi
Finsi
Corriger selon bornes et contraintes
Evaluer f()
Si f() <f() accepter
   Sinon
Finsi
Fin Pour
Trouver
Fin Pour
```

4 Résultats Expérimentaux et Discussion

Dans cet exemple, nous considérons un système électrique thermique à 10 nœuds. La puissance totale perdue PL a été estimée en tant que constant égale à 14.6982. Les données relatives aux dix générateurs sont dans la Table 1. La demande estimée PD est 1036 MW. Aucune émission de NOx n'est tolérée au-delà de 48 (ton/h). Après

plusieurs simulations, le nombre des éléments de décors a été fixé à 15 quant au nombre d'itération est 10000.

Données de La Station Thermique

generateur	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A $	786.79	451.3251	1049.9977	1243.5311	1658.5696	1356.6592	1450.7045	1450.7045	1455.6056	1469.4026
b($/MW	38.5397	46.1591	40.3965	38.3055	36.3278	38.2704	36.5104	36.5104	39.5804	40.5407
c($/MW²)	0.1524	0.1058	0.028	0.0354	0.0211	0.0179	0.0121	0.0121	0.109	0.1295
D	450	600	320	260	280	310	300	340	270	380
E	0.041	0.036	0.028	0.052	0.063	0.048	0.086	0.082	0.098	0.094
	103.3908	103.3908	300.391	300.391	320.0006	320.0006	330.0056	330.0056	350.0056	360.0012
	-2.444	-2.444	-4.0695	-4.0695	-3.8132	-3.8132	-3.9023	-3.9023	-3.9524	-3.9864
	0.0312	0.0312	0.0509	0.0509	0.0344	0.0344	0.0465	0.0465	0.0465	0.047
	0.5035	0.5035	0.4968	0.4968	0.4972	0.4972	0.5163	0.5163	0.5475	0.5475
	0.0207	0.0207	0.0202	0.0202	0.02	0.02	0.0214	0.0214	0.0234	0.0234
Pmin (MW)	150	135	73	60	73	57	20	47	20	10
Pmax (MW)	470	135	340	300	243	160	130	120	80	55

La deuxième table montre que le résultat de la minimisation du coût sous contrainte d'émission Nox de l'algorithme ISA est meilleur que celui de l'algorithme de Lucioles : Firefly Algorithm(FF) ou l'optimisation par essaims de Particules: Particle Swarm Optimization (PSO).

VALEURS MINIMALES, MOYENNES ET MAXIMALES DU COÛT POUR L'OPTIMISATION DU COÛT

	FF	**PSO**	**ISA**
Max	122000	124000	122000
Min.	60955.1	61619.8	**60656.3**
Moy.coût	61347.7	62247.4	61084.9
Emission	42.4775	46.477	47.1363
Pg1	150	150	150
Pg2	135	135.686	135
Pg3	73	74.6194	73
Pg4	60	120.377	60.0454
Pg5	221.892	221.278	222.638
Pg6	118.122	116.105	130.295
Pg7	129.566	76.4705	129.599
Pg8	120	118.419	120
Pg9	30.9671	22.0813	20.1001
Pg10	12.1535	15.6645	10.0226
P. totale	1050.7	1050.7	1050.7

Dans une deuxième phase, nous avons lancé 20 simulations de l'optimisation de l'émission sans contrainte de seuil de Polluant. Les résultats d'ISA sont satisfaisants. (Table 3) L'algorithme a pu minimiser les polluants jusqu'à 37.12 (ton/h) en gardant un coût d'environ 62751($/h). face à PSO avec une émission élevée 39.1841(ton/h) et un coût élevé 64030.3($/h).

VALEURS MINIMALES, MOYENNES ET MAXIMALES POUR L'OPTIMISATION DE L'ÉMISSION

	FF	**PSO**	**ISA**
Max	74.8048	78.0854	74.2578
Moy.	37.3937	38.9013	37.1289
Min. émission	37.4111	39.1841	**37.392**
coût	62779.2	64030.3	62751.2
Pg1	150	158.993	150
Pg2	135	135.408	135
Pg3	90.1068	113.29	73
Pg4	90.2424	100.967	94.8286
Pg5	128.656	105.549	121.5
Pg6	128.403	154.509	141.371
Pg7	96.643	90.1427	104
Pg8	96.6486	58.7255	96
Pg9	80	80.8715	80
Pg10	55	52.2441	55
P. totale	1050.7	1050.7	1050.7

Dans la troisième phase, nous appliquons les métaheuristiques sur le problème bi-objectif qui vise à minimiser le coût et l'émission (table 4). Dans cet exemple, l'émission est pondérée avec 100 pour qu'elle soit mise à la même échelle dans la fonction Pareto objectif. Le résultat étonnant est que l'algorithme choisi ISA arrive à obtenir la meilleur fonction Pareto 434031 et la meilleure émission 37.13 et avec un coût très satisfaisant 62733 qui ne diffère pas beaucoup du meilleur coût 62262.

VALEURS MINIMALES, MOYENNES ET MAXIMALES POUR L'OPTIMISATION BI-OBJECTIF

	FF	**PSO**	**ISA**
Maxf	437330	451700	**434032**
Min. f	436680	444926	**434031**
Moy. f	438672	456534	**434033**
Emission	37.641	38.2278	**37.13**
Coût	**62262**	74256	62733
Pg1	150	154.832	150
Pg2	135	135.338	135
Pg3	73	84.5405	90.5312
Pg4	93.8999	92.5016	91.2945
Pg5	145.8	123.485	128.224
Pg6	144	144.185	128.432
Pg7	78	91.2572	95.895
Pg8	96	114.814	96.3235
Pg9	80	55.6949	80
Pg10	55	54.0515	55
P. totale	1050.7	1050.7	1050.7

5 Conclusion

Dans cette article, une métaheuristique inspirée de la décoration a été adapté au problème bi-objectif de répartition de charge aussi appelé distribution éco. environnementale de l'énergie. L'expérimentation sur un système thermique à dix

générateurs a montré que l'algorithme de décoration intérieure donne de meilleurs résultats en termes de coût et/ou d'émission de polluant par rapport à l'algorithme des lucioles et l'algorithme par essaims particulaires. D'autres tests sont en cours pour valider l'efficacité de la méthode sur d'autres systèmes standards à 40 nœuds et à éoliennes.

Références

6 Bentouati B., Saliha Chettih, Lakhdar Chaib Victor Sreeram (2017). Interior search algorithm for optimal power flow with non-smooth cost functions. Cogent Engineering

7 Dekhici L., Belkadi K., Interior Search Algorithm for Hybrid Flow Shop Scheduling, First International Conference on Business Intelligence and Applications, ICBIA'16, Blida, 2016.

8 DinuCalinSecui, Large-scale multi-area economic/emission dispatch based on a new symbiotic organisms search algorithm, Energy Conversion and Management, Vol. 154, 15 December 2017, Pp. 203-223

9 Fahad Parvez M, PandianVasant, Vish Kallimani, Junzo Watada, Patrick Yeoh Siew Fai, M. Abdullah-Al-Wadud, A holistic review on optimization strategies for combined economic emission dispatch problem, Renewable and Sustainable Energy Reviews, Vol. 81, Part 2, January 2018, Pp. 3006-3020, ISSN 1364-0321

10 Gandomi Amir H., Interior search algorithm (ISA): A novel approach for global optimization, ISA Transactions, Vol. 53, Issue 4, July 2014, Pp. 1168-1183.

11 Gherbi Y. A., Hamid Bouzeboudja, Fatima Zohra Gherbi, The combined economic environmental dispatch using new hybrid metaheuristic, Energy, Vol. 115, Part 1, 15 November 2016, Pp. 468-477

12 Jayabarathi T., T. Raghunathan, B.R. Adarsh, Ponnuthurai Nagaratnam Suganthan, Economic dispatch using hybrid grey wolf optimizer, Energy, Vol. 111, 15 September 2016, Pp. 630-641.

13 Kumar Manjeet, Tarun Kumar Rawat, Aman Jain, Atul Anshuman Singh, Aviral Mittal, Design of Digital Differentiators Using Interior Search Algorithm, Procedia Computer Science, Vol. 57, 2015, Pp. 368-376

14 Kumar Manjeet, Tarun Kumar Rawat, Apoorva Aggarwal, Adaptive infinite impulse response system identification using modified-interior search algorithm with Lèvy flight, ISA Transactions, Vol. 67, March 2017, Pp. 266-279, ISSN 0019-0578,

15 Mellal Mohamed Arezki, Edward J. Williams, Cuckoo optimization algorithm with penalty function for combined heat and power economic dispatch problem, Energy, Vol. 93, Part 2, 15 December 2015, Pp. 1711-1718

16 Nazari-Heris M., B. Mohammadi-Ivatloo, G.B. Ghareh petian, A comprehensive reviewof heuristic optimization algorithms for optimal combined heat and power dispatch from economic and environmental perspectives, Renewable and Sustainable Energy Reviews, Vol. 81, Part 2, January 2018, Pp. 2128-2143

17 Nemati Mohsen, Martin Braun, Stefan Tenbohlen, Optimization of unit commitment and economic dispatch in micro grids based on genetic algorithm and mixed integer linear programming, Applied Energy, Vol. 210, 15 January 2018, Pp. 944-963

18 Qu B.Y., Y.S. Zhu, Y.C. Jiao, M.Y. Wu, P.N. Suganthan, J.J. Liang, A survey on multi-objective evolutionary algorithms for the solution of the environmental/economic dispatch problems, Swarm and Evolutionary Computation, Vol. 38, February 2018, Pp. 1-11

19 Narang N, E Sharma, JS Dhillon, Combined heat and power economic dispatch using integrated civilized swarm optimization and Powell's pattern search method, Applied Soft Computing, 2017, 52, 190-202

Improving Twitter Sentiment Analysis using Preprocessing

Tolba Marwa , Ouadfel Salima , Meshoul Souham ,Sofiane chemaa

Computer Science Department, Faculty of NTIC,
University Constantine 2 - Abdelhamid Mehri
Constantine, Algeria

Abstract. In recent years, the strong rise in the use of social network platforms such as Twitter has resulted in millions of users sharing their thoughts and opinions about different aspects and events on the micro-blogging platform. Exploiting these opinions by extracting useful information from it has become a great challenge in data mining and knowledge discovery. Twitter sentiment analysis (TSA) tackles the problem of analyzing the tweets in terms of the opinion they express. This analysis offers organizations a fast and effective way to monitor the publics feelings about their brand, business and directors among others. One of the main challenges of TSA is the data sparsity due to the extensive use of incorrect English and misspellings. Moreover tweets contain a lot of textual peculiarities such as emphatic lengthening, slang and abbreviations. In this paper, we describe a framework for effective TSA. It suggests using an improved preprocessing phase prior to sentiment classification to deal efficiently with textual peculiarities of tweets. Very promising and even competitive results have been obtained using state of the art data sets such as STS-Gold, SemEval2017, and Sanders.

Keywords: Twitter sentiment analysis, Data Sparsity, Microsoft Cognitive Services

1 Introduction

Twitter is one of the largest micro-blogging services on the internet. It allows users to publish short messages that are visible to other users and most commonly known as tweets. The number of Tweets sent per day is very huge as it is approximately five hundred millions[1]. These tweets can be a very good source to mine opinions and valuable knowledge useful for a variety of applications that require understanding the public opinion about a concept. For example in business intelligence field, public opinions help enterprises to capture the views of customers about their products. Therefore mining such opinions is very useful to managers in decision making. Politics is another example where there is a need to analyze trends, to identify ideological bias, to evaluate public opinions

[1] `https://www.blogdumoderateur.com/chires-twitter/`

and to gauge reactions. Mining opinions is also very important in sociology because adoption of new ideas is generally a consequence of idea propagation and reaction to opinions and ideas through groups. Therefore, the challenge is how to mine opinionated information even within a huge amount of data in twitter. This is most commonly known as twitter sentiment analysis (TSA). TSA aims to automatically detect tweets polarity that is, the extent to which the sentiment is either positive or negative with regard to a given aspect. Generally, TSA involves tools from natural language processing (NLP), machine learning (ML) and statistics in order to analyze tweets and characterize the sentiment content they convey. It is related to other tasks such as information extraction, question answering and summarization. Compared to other sentiment analysis (SA) tasks, TSA is even more difficult because of the short length of tweets as they are up to 280 characters. Moreover, the language used in Twitter is very different from the language used in other text genres (web, blogs, news...) as it contains a lot of textual peculiarities such as emphatic lengthening, abbreviations, slang. This leads to the data sparsity problem that causes misclassification of tweets and leads to incorrectly classified them as neutral tweets [7].
In this paper we describe a general framework to build a fully automatic system to process tweets and determine their polarity. The main features lie on the use of improved preprocessing to deal with the data sparsity problem and a text analytics API from the Microsoft Cognitive Services to handle the sentiment classification task.
The rest of the paper is organized as follows. In section 2, an overview of related work in the literature is given. Section 3 is devoted to the description of the proposed system for TSA. Section 4 presents the experimental study and shows obtained results. Finally the paper ends with a conclusion

2 Related work

SA has been applied mainly at three different levels namely document level, sentence level, and aspect (or entity) level. In the first level, SA aims to identify the sentiment polarity in the whole document expressing a single entity (e.g., a single product) whereas the sentence level determines whether each sentence expresses a positive, negative, or neutral opinion. However at the document-level and the sentence-level, analyses do not discover what exactly people like and dislike. Thats why, aspect level has been introduced to perform finer-grained analysis in order to detect the sentiment polarity of a specific entity/target of a particular opinion [3, 8] .The majority of tweets contain a single sentence due to the length limitation of twitter microbloging (up to 280 characters). Consequently, TSA is applied on two levels only: Sentence and Aspect (Entity) levels. TSA has been performed in the literature using machine learning approaches and/or Lexicon-Based approaches [3]. Different classifiers have been used in the first type of TSA approaches such as Naive Bayes classifiers (NB), Maximum Entropy (MaxEnt), Support Vector Machines (SVM) and number of different features to detect tweets that express positive or negative sentiment. Among

recent and related studies within this context, Hamdan et al. [5] make use of features that include con-cepts from DBPedia, verb groups and similar adjectives from WordNet, senti-features from Senti-WordNet and also employed a dictionary of emotions and abbreviations. Their method has been assessed on the dataset SemEval-2013 evaluation campaign [6] and its improved F-measure accuracy by 2% and 4% by considering these features compared to the SVM trained on unigrams and NB classifier, respectively. In the same context, Aston et al. [2] proposes a tweet representation in terms of character n-grams, This led to increase exponentially the number of possible grams by a factor of 95n as the gram size n increases; this large number of features was reduced by selecting the top N features of a gram using 6 different evaluation algorithms: Chi Squared, Filtered Feature, Gain Ratio, Info Gain, One R, and Relief. They reported an F-measure of 85% for subjectivity and 78% for sentiment classification.
Lexicon-based approach uses a manually or automatically built list of positive and negative words to calculate the polarity according to the positive and negative words in the text. Khan et al [7] presented a framework based on hybrid Classifier: Enhanced Emoticon Classifier (EEC) used list of emoticon (70 positive and 75 negative) proposed in [9], Improved Polarity Classifier (IPC) used a list of positive and negative words created from the Bing Liu list [2] and the Bill McDonald list [3] SentiWordNet Classifier (SWNC) based on classification of tweets using SentiWordNet dictionary. Whereas Saif et al. [12] defined an approach that allows detecting sentiments at both entity-level and tweet-level. This approach called SentiCircles builds a dynamic representation of words that captures their contextual semantics (i.e., semantics inferred from the co-occurrence patterns of words in text) in order to tune their pre-assigned sentiment strength and polarity in a given sentiment lexicon [12]. Their approach was evaluated on three different datasets: OMD [13], HCR [14], and STS-Gold [11] and proved the effectiveness for both entity- and tweet-level sentiment detection against SentiStrength, MPQA, SentiWordNet Methods.
Moreover, Alnashwan et al [1] propose an ensemble learning approach based on the meta-level features of seven existing lexicon resources (SentiWordNet, Bing Liu, AFINN, NRC-hashtag, Sentiment140 lexicon, Sentiment140 method, SentiStrength) for automated polarity sentiment classification.

3 The Proposed System Including Improved Preprocessing For TSA

To perform a TSA task, one needs to follow the following steps namely tweet acquisition, tweet preprocessing and tweet classification. As a consequence and as above mentioned, TSA requires tools from different fields such as NLP, ML and statistics and may be others. To the best of our knowledge, no fully automatic system for TSA that integrates all these functionalities is available. In this

[2] http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar
[3] http://www3.nd.edu/~mcdonald/Word_Lists.html

work, we propose an architecture for fully automatic TSA which combines NLP tools for tweets pre-processing that includes removal of URLs, hash-tags, username and special characters; spelling correction; substitution of abbreviations and slangs with expansions and ML tools for tweet classification using Microsoft cognitive APIs: Bing Speech API and Text Analytics API .The ultimate goal of our system is to increase the accuracy of tweet sentiment classification and resolve the data sparsity issues.
The overall architecture of the proposed system is depicted in Figure 1 where the two components system can be easily seen. In what follows, we briefly describe each of the modules involved in this architecture while giving examples when necessary. Examples are extracted from the datasets we used in our experiments.

3.1 Preprocessing component

As shown on the proposed architecture, this component includes several modules that help to prepare tweets for classification with the aim to characterize the sentiment content in the text unit.

URL / Hashtags / Username Removal module : Tweet may contain URL, hash tags and user name. After tokenization, URLs and Usernames are identified and removed since their impact on the classification is not significant. In the case of hash tags, the sign # is kept because sometimes the hash tags helps to detect polarity of tweet.

Detect and Replace slang / abbreviation module: Among the challenges of tweets classification is data sparsity problem as it leads to incorrectly classify most of the tweets. The main reason behind these problems is use of slangs and abbreviations (or shorthand grammars) due to the limit of tweet message (280 characters). For this reason, the Abbreviations and/or shorthand notations will be replaced by expansions. In our approach we are use **Netlingo**[4] and **Dictionary of Text Messaging and Online Chat Abbreviations**[5] for this purpose. The following example illustrates how slangs and abbreviations are handled:

This Friday as in 5th June Friday?! Argh. **OMG**. Which website?! I' m so excited!

⇓

This Friday as in 5*th June Friday*?! *Argh* .**Oh my god** .*Which website*?! *I'm so excited*!.

[4] `http://www.netlingo.com/acronyms.php`
[5] `http://www.webopedia.com/quick_ref/textmessageabbreviations.asp`

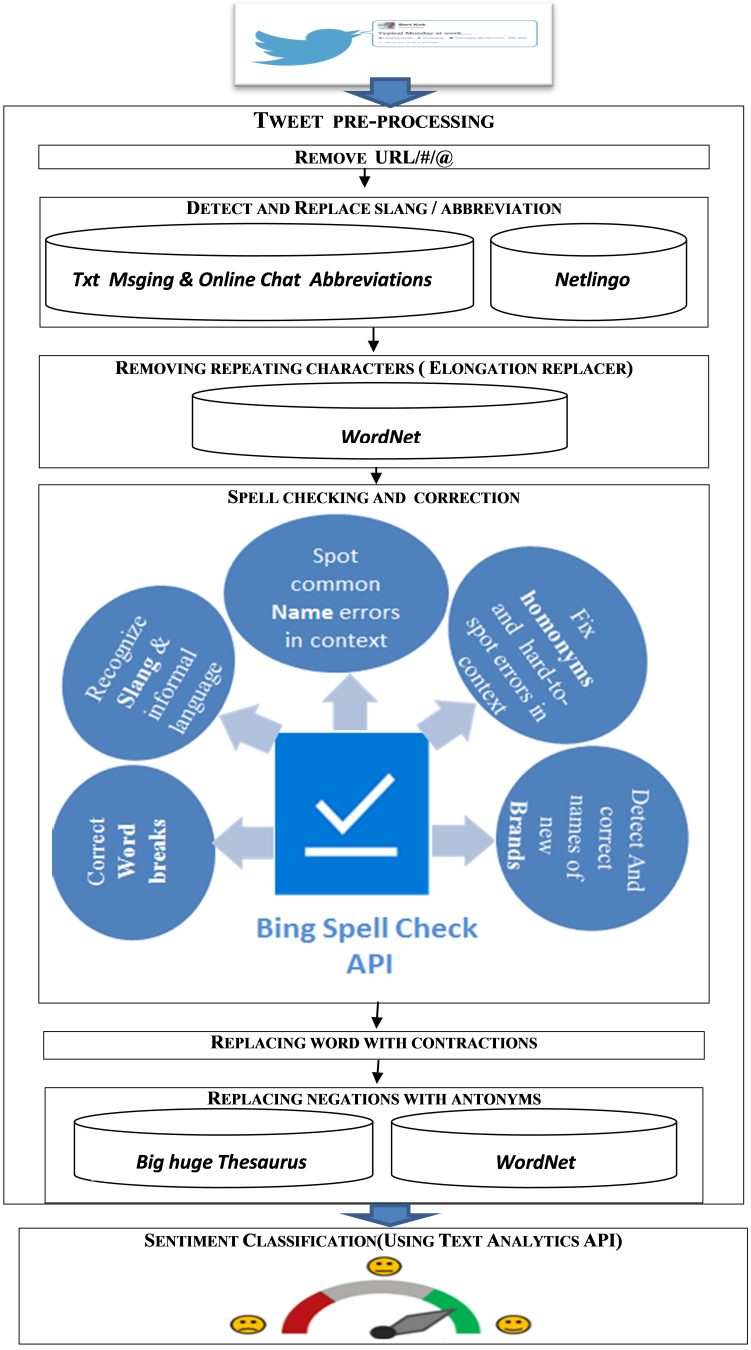


Fig. 1. Twitter Sentiment classification integrating the improved pre-processing using Microsoft Cognitive Services APIs for language .

Removing repeating characters (or Elongation replacer) module: Elongation replacer recursively removes repeating characters until no more characters are removed or recognized by WordNet. A WordNet lookup before removing repeating characters is done to ensure unnecessary removal of characters. In the example, removing repeating characters from cooool without WordNet lookup gives col but the word cool gives.

fkell that's freakin **cooool** i love twitter .haha
⇓
fkell that's freakin **cool** i love twitter .haha

Spell checking and correction module: Due to its informal type of communication and the length limitation, Tweets contain a lot of noise resulting in the extensive use of incorrect English and misspellings. To address this problem, we used **Bing Spell Check API**[6] which is based on machine learning and statistical machine translation to dynamically train a constantly evolving and highly contextual algorithm further a massive corpus of web searches and documents. The reason behind the use of such an API is that it allows [7]:

- Recognizing slang and informal language.
- Recognizing common name errors in context.
- Correcting word breaking issues with a single flag.
- Being able to correct homophones in context, and other difficult to spot errors.
- Supporting new brands, digital entertainment, and popular expressions as they emerge.
- Words that sound alike but differ in meaning and spelling, for example see and sea.

ugh now I feel retarted, and jayz is **gona** be at tao and diddy is **gona** be **vegas** too fight nights are so fun oh well.
⇓
ugh now I feel retarted , and jayz is **gonna***be at tao and diddy is* **gonna** *be* **Vegas** *too fight nights are so fun oh well .*

Replacing word with contractions (Replacing words matching regular expressions) module : Contractions such as didnt, dont, couldnt are common in tweets. These types of bi-grams often determine the polarity of tweets. These contractions are replaced with their expanded forms.

Doesn't look good for the cavs ⇒ **Does not***look good for the cavs .*

[6] `https://azure.microsoft.com/fr-fr/services/cognitive-services/spell-check/`

[7] https://docs.microsoft.com/fr-fr/azure/cognitive-services/bing-spell-check/proof-text

Explicit negation handling (Replacing negations with antonyms) module : We used an antonym replacer using WordNet and Big huge Thesaurus to replace words preceded by not, never, etc. a word is replaced if an unambiguous antonym is present in the two thesaurus. We used an antonym replacer using WordNet and Big huge Thesaurus[8] to replace words proceeded by not, never, etc. a word is replaced if an unambiguous antonym is present in the two thesauruses.

I'm **not happy** that I can't watch community channel's videos on my iPod.
$\Downarrow$
I'm **unhappy** *that I cannot watch community channel's videos on my iPod .*

3.2 Sentiment Classification component

In this step, we used The Text Analytics API to get the polarity score to classify tweets that express positive or negative sentiment., The input features of the classifier include n-grams, features generated from part-of-speech tags and word embed-dings. Moreover, it detects a sentiment in text written in English, Spanish, French, Portuguese and 11 additional languages are available in preview[9]

4 Experimental Setup

In order to assess the performance of the proposed framework for TSA, three of the most commonly used datasets in the literature have been considered and two performance metrics namely accuracy and F-measure have been used for this purpose. In this section we give a brief description of the used datasets then we discuss the obtained results.

4.1 Datasets

The following is a description of the three datasets: STS-Gold [11] and SemEval-2017 task 4 [10] and used to evaluate the performance of the proposed system.

STS-Gold Dataset The STS-Gold dataset is constructed by [11] from the Stanford Twitter Sentiment Corpus (STS) [4] for the evaluation of sentiment classification models at both the entity and tweet levels. It contains 2034 tweets and 58 entities manually and independently annotated by three different human evaluators [11, 12].

SemEval-2017 task4 (subtasks B and D) This dataset is provided by the Semantic Evaluation of Systems (SemEval-2017) challenge for Twitter Sentiment Analysis task. It consists of five subtasks (A, B, C, D and E) and includes sentiment analysis on: highly-positive, positive, neutral, negative, highly-negative

[8] `https://words.bighugelabs.com/`

[9] `https://westus.dev.cognitive.microsoft.com/docs/services/TextAnalytics.V2.0/operations/56f30ceeeda5650db055a3c9`

points scales .each subtask offered for both Arabic and English languages. In this paper we used testing datasets of subtasks B and D annotated for on a 2-point scale (positive and negative) with 6185 tweets [10].

Sanders dataset[10] It is composed of 5513 manually classified tweets (positive, negative, neutral, or irrelevant) with respect to four different topics: @apple, #google, #microsoft, #twitter. Resulting in 570 positive, 654 negative, 2,505 neutral, and 1,786 irrelevant tweets. In our study we used also positive and negative tweets. Table 1 summarizes information about these datasets.

Table 1. STS-Gold, SemEval-2017 and Sanders Tweets description.

Dataset	Tweets	Positive	Negative
STS-Gold	**1081**	**393**	**688**
SemEval-2017Task4 (subtask B and D)	**6185**	**2423**	**3722**
Sanders	**1224**	**570**	**654**

4.2 Results and Discussion

To implement the proposed system, we make use of the aforementioned Text Analytics API for sentiment classification. This API does not take into account the particularity of tweets such as: the presence of slang, abbreviations elongated words, Incorrect English. Table 2 shows sentiment scores of few sentences before and after preprocessing, knowing that the score generated by the API lies within the range [0,1] : Scores close to 1 indicate positive sentiment and scores close to 0 indicate negative sentiment. To eliminate these imperfections we applied improved tweet prepreprocessing that begins by replacing slang and abbreviation using the on-line dictionary called Netlingo which contains a largest list of text messages shorthand and Internet Acronyms. However, this list is insufficient to capture the wide range of slogan words employed in Twitter. Therefore we added in parallel lists of more than 1,460 text messages and online chat abbreviations used in Facebook, Twitter, instant messaging, email, Internet, online gaming services, chat rooms, discussion boards and mobile phone text messaging (SMS)[11] The next step, based on the correction of spelling mistakes that appear in tweets, one of the most important sources of these mistakes is elongated words which means the number of words with one character repeated more than 2 times like looooove and it is used to emphasize words. Emphatic lengthening is very frequent in Twitter; occurring in approximately one of every six tweets [3]. This type of mistakes is not handled by Bing spell check.

[10] `https://westus.dev.cognitive.microsoft.com/docs/services/TextAnalytics.V2.0/operations/56f30ceeeda5650db055a3c9`

[11] `http://www.webopedia.com/quick_ref/textmessageabbreviations.asp`

On the other hand, the presence of negated bi-gram phrases plays an important role in detecting the sentiment polarity of a tweet. The detection and the proper handling of negations are not trivial and remain a challenge [3]. This is not handled by Text Analytics API from Microsoft cognitive service APIs. The third example in Table 2 illustrates this situation. For this reason we used an antonym replacer using WordNet and Big Huge Thesaurus which contains over 145,000 words in English language. After this integrated treatment of tweets we used Text Analytics API to generate the sentiment score and to deduce tweet polarity.

Table 2. Examples of few tweets with sentiment score given by Text Analytics API before and after correction to show the importance of preprocessing component.

Tweet	**score**	**Correct tweet**	**score**
looooooooove!!!!!!!!!!!!	**0.5**	love!!!!!!!!!!!!	**0.78**
Facebook is **BOREEEEEEEEEEEEEEEEEENG**	**0.73**	Facebook is **boring**	**0.03**
I'm **not happy** that I can't watch community channel's videos on my iPod	**0.78**	I'm **unhappy** that I cannot watch community channel's videos on my iPod	**0.06**
This Friday as in 5th June Friday?! Argh. **OMG**. Which website?! I m so excited!	**0.29**	This Friday as in 5th June Friday?! Argh.**Oh My God** Which website?! Im so excited!	**0.72**

Our system aims at efficiently detecting sentiment polarity in tweets based on Text Analytics API by integrating preprocessing to fix the aforementioned issues. As can be seen on the examples shown on table 2, correct scores or sentiment polarity have been found after preprocessing. In order to corroborate this finding, the proposed system has been tested using the three data sets. Table 3 shows results in terms of accuracy and F-measure before and after preprocessing. As can be seen very promising results have been obtained.

Moreover the proposed system has been compared to the best three related

Table 3. Result of sentiment classification before and after using preprocessing with Text Analytics API

Datasets	Text Analytics API		Preprocessing+Text Analytics API	
	F-measure	Accuracy	F-measure	Accuracy
SemEval 2017	**75.73**	**83.16**	**77.70**	**84.19**
Sanders	**79.54**	**80.21**	**82.24**	**84.26**
STS-GOLD	**75.18**	**80.89**	**81.83**	**87.36**

methods described in [12] using STS-Gold dataset because only results for this dataset are available. Table 4 shows the obtained results where we can see that

the proposed framework outperforms these methods and confirm the importance of preprocessing.

Table 4. Result given by our system compared with the best three method for tweet-level sentiment detection cited in [12] using STS-Gold datasets

Metodes	SentiStrength		SentiCircle with Median method		SentiCircle with Pivot-Hybrid		Preprocessing+ Text Analytics API	
Dataset	F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy
STS-Gold	78.56	81.32	76.15	79.74	77.52	80.33	**81.83**	**87.36**

4.3 Conclusion

In this paper we tackled TSA problem which addresses the task of analyzing the messages posted in twitter in terms of the sentiment they express. This task is non-trivial and very challenging compared to detecting sentiment in conventional text such as blogs and forums. This can be explained by the length limitation according to which tweets can be up to 280 characters and this leads to use informal language and abbreviations to edit messages.

In our proposed system, we took advantage of two APIs namely Bing Spell Check API and Text Analytics API to analyze sentiments in tweets. The first API has been used to fix the common spelling mistakes that appear in tweets while the second API enables to analyze a sentiment of text units. In spite of the benefits granted by Text Analytics API, it is necessary to normalize the tweets before their use because are often represented in cryptic and informal language, systematic preprocessing of tweets is required to enhance the accuracy and F-measure of sentiment analyzer. The use of two API cited above with improved tweet preprocessing is described in Section (4.b). An improved preprocessing component has been integrated in order to further improve the classification task. Obtained preliminary results are very encouraging. As ongoing work, we plan to design intelligent systems to handle multilingual and multimodal tweets

References

1. Alnashwan, R., O'Riordan, A.P., Sorensen, H., Hoare, C.: Improving sentiment analysis through ensemble learning of meta-level features. In: KDWEB 2016: 2nd International Workshop on Knowledge Discovery on the Web. Sun SITE Central Europe (CEUR)/RWTH Aachen University (2016)
2. Aston, N., Liddle, J., Hu, W.: Twitter sentiment in data streams with perceptron. Journal of Computer and Communications 2(03), 11 (2014)

3. Giachanou, A., Crestani, F.: Like it or not: A survey of twitter sentiment analysis methods. ACM Computing Surveys (CSUR) 49(2), 28 (2016)
4. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1(12) (2009)
5. Hamdan, H., Béchet, F., Bellot, P.: Experiments with dbpedia, wordnet and sentiwordnet as resources for sentiment analysis in micro-blogging. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). vol. 2, pp. 455–459 (2013)
6. Hltcoe, J.: Semeval-2013 task 2: Sentiment analysis in twitter. Atlanta, Georgia, USA 312 (2013)
7. Khan, F.H., Bashir, S., Qamar, U.: Tom: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems 57, 245–257 (2014)
8. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5(1), 1–167 (2012)
9. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL student research workshop. pp. 43–48. Association for Computational Linguistics (2005)
10. Rosenthal, S., Farra, N., Nakov, P.: Semeval-2017 task 4: Sentiment analysis in twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 502–518 (2017)
11. Saif, H., Fernandez, M., He, Y., Alani, H.: Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold (2013)
12. Saif, H., He, Y., Fernandez, M., Alani, H.: Contextual semantics for sentiment analysis of twitter. Information Processing & Management 52(1), 5–19 (2016)
13. Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of the first SIGMM workshop on Social media. pp. 3–10. ACM (2009)
14. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the First workshop on Unsupervised Learning in NLP. pp. 53–63. Association for Computational Linguistics (2011)

A new Method for Facial Expression Recognition

Kahina Amara[1,3], Naeem Ramzan[2], Nouara Achour[1], Mahmoud Belhocine[3], Nadia Zenati[3], and Cherif Larbes[4]

[1] LRPE Laboratory, USTHB University,
B.P 32 El Alia 16111, Bab Ezzouar, Algiers, Algeria
[2] School of Engineering and Computing, University of the West of Scotland,
Paisley, Scotland, United Kingdom
[3] CDTA Center of developpement of advanced technologies, ALgiers
[4] ENP Ecole Nationale politechnique Hassen Badi Avenue, Algierskamara@cdta.dz

Abstract. Virtual reality (VR) technology, in particular, has the potential to simulate real-world social and communication interactions and hence could be used as an interactive platform in several research fields. Emotional facial recognition is considered among the core building blocks of social communication. Studies have shown that facial emotional processing and understanding is impaired in social human interaction. In this paper, we propose a new method for facial expression recognition based on new geometrical features. We collected a novel dataset of 17 subjects performance of six facial expressions (anger, fear, happiness, surprise, sadness, and neutral) using Kinect (v1) and Kinect (v2) and RGB HD camera. To assess the performance of the proposed system we used leave-one-out-subject cross-validation. A comparison between RGB and RGB-D data for facial expression recognition is provided. The obtained results show the superior performance of the RGB-D features provided by Kinect (v2). Based on our experiment, we observed that the 2D images are not robust enough for facial expression recognition.

Keywords: Virtual reality · Interaction · Emotion recognition · Facial Expression · RGB · RGB-D · Classification · Geometrical Features· RGB and RGB-D database.

1 Introduction

Facial expressions are common, effective and nonverbal way of expressing feeling. In the literature, many studies addressed the facial expression [1–3]. The facial expression recognition has found fertile ground for applications in many research areas of security system (surveillance video), in interaction for virtual reality [4, 5], healthcare with virtual reality [6], in humans-robots interaction, action tendency, health-care [7, 8], serious games [9], and recently in Social Interaction Assistance [2]. Virtual reality based facial emotion recognition has been widely studied with Schizophrenia for symptom assessment [10], autism spectrum [11], training of medication management skills, hallucinations training, and social perception (figure 1).

Fig. 1. Facial expression recognition's application field.

Recent approaches use 3D facial points, such techniques have gained more attention lately due to the proliferation of affordable commodity depth sensing devices, such as the Kinect. According to the data used in this work, different approaches have been proposed for feature extraction. Positional and temporal features have been investigated in [12] using the facial data collected by Kinect. They defined a feature vector composed of the coordinates of tracked points and Euclidean distance between the tracked points and the angle between those points for the positional features. Many studies are based on 2D images for facial emotion recognition. In [1], the authors proposed a software for the analysis of facial behaviour. Furthermore, several approaches have demonstrated state-of-the-art performance on RGBD input, or only depth input. We can cite the work presented in [3, 13]. In [13], the authors proposed the skeleton based approach to extract facial features for facial emotion recognition by using a depth camera. Billy *et al.* [14] used Kinect sensor to recognize emotions under different conditions. The authors used a publicly available database which contains facial images (RGB-D) captured by Kinect sensor with different poses, expressions, illumination and disguise. Their results demonstrated that using RGB-D information could improve the performance of facial emotion recognition compared with the methods using 2D information. The conventional approaches for facial emotion recognition still suffer from some constraints and limitations which directly affect the system performance [15]. We can cite among these problems, the lack of publicly available database, the environmental changes including illumination changes, the different personnel style for emotion expression.
The existing approaches for facial expression recognition suffer from some constraints and limitations. Firstly, the lack of publicly available database. Furthermore, the selection of non-significant features for depicting different expressions can cause model failure. To deal with these problems, we propose in this work

a system for mono-modal facial expression recognition based on facial mouvements. The main contributions of the presented work are:

- **RGB and RGB-D facial expressions dataset:** We created a dataset including the performance of 17 participants (9 males and 8 females). The participants are from more than 10 different nationality and have different color skin. Unique features collected from Kinect sensors (version 1 and 2) and RGB HD camera were used to collect the data. The dataset was captured in controlled conditions of varying face appearance and illuminations.
- **Significant facial key points features selection for facial expression recognition:** Significant facial points features were selected for depicting the different defined facial expressions (anger, fear, happiness, surprise, sadness, and neutral). We choose significant facial points and used geometrical features between each selected points for the data provided by Kinect 1 and Kinect 2 captors and HD RGB camera. Kinect-like data (i.e. RGB-D videos and joint sequences) can be used for extracting significant features
- **Kinect 1 and Kinect 2 data comparison:** In this study, we carried out a facial expression recognition comparison using data provided by Kinect 1 and data provided by Kinect 2.
- **RGB and RGB-D data comparison:** The 2D RGB data provided by RGB HD camera were tested against Kinect sensors RGB-D data. The use of the depth data may improve the emotion recognition.
- **Demonstration of the proposed method performance against other state-of-the-art methods:** The performance of the method is tested against other state-of-the-art methods.

In this paper, we present a novel facial expression recognition method using 3D angle and 3D distance features for the RGB-D data and 2D angle and 2D distance for the RGB data provided by HD RGB camera. A comparison between RGB and RGB-D data is provided and a new RGB-D and RGB facial database is presented. This paper consists of four sections. In the second section, we describe the proposed approach, the dataset collection; the feature extraction will be presented in detail. In the third section, we discuss the experimental results. The conclusions and future works will be drawn in the last section.

2 Proposed method

Emotion can be expressed in different ways and plays important role in daily life. The facial expression is a common, nonverbal and effective way of expressing emotion. The presented work is included in this area; the process for facial expression recognition is depicted in figure 2, we aim to distinguish the expressions as accurate as possible by establishing computational models. We carried out experiments on synthetic RGB-D sequences captured by Kinect (v1) and Kinect (v2) sensors and RGB sequences recorded using RGB HD camera. In this work, we target six basic emotions (anger, fear, happiness, surprise, sadness, and neutral).

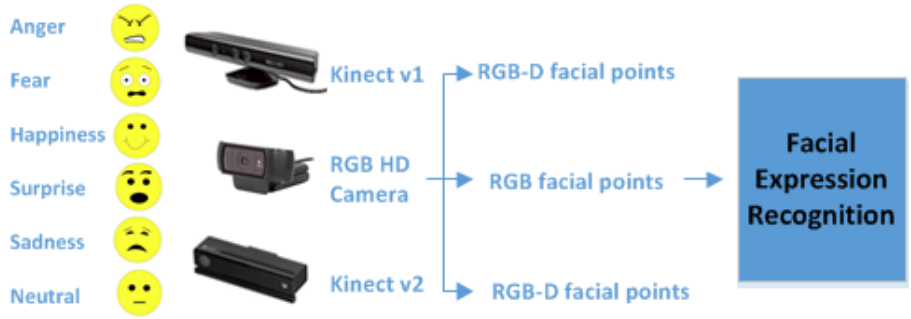


Fig. 2. The proposed framework for facial expression recognition.

We collected our own database including the performance of 17 students (9 male and 8 female) recruited from the School of computing and engineering at the University of West of Scotland. The participants are from different cultures with different skin colour. In order to obtain actual facial expression data, we conducted an emotion priming experiment using different emotional videos. The subjects were first asked to perform emotional states depicted on projected images on screen. In the second part of the experiment, we used 20 emotional videos used in [16] collected from Youtube. We used many types of emotional videos including neutral, happy, surprise, anger, fear and sad video which could induce corresponding emotional state. The participants were asked to perform their feeling according to their personal style. They have to repeat the performance for three times. The face-recorded videos were segmented and stored in a database. The collected dataset contains 1581 RGB videos recorded by RGB HD camera and more than 3000 synthetic RGBD sequences captured by Kinect (v1) and Kinect (v2). Figure 3 displays the participants facial expression.

In this work, the facial expressions were tracked using the face and skeleton tracking API available in the Microsoft Kinect Software Development Toolkit. The synthetic RGB-D sequences captured by Kinect sensors provided 3D facial points. We choose representative points, which represent significant movement in order to describe the subtle changes of facial expression. The face tracking Kinect toolkit provides 121 facial points for Kinect 1 and 1347 facial points for Kinect 2. However, not all of these points are significant to facial expressions. In [20], the authors proposed that the main areas englobing the eyes, eyebrows, and the mouth are involved in facial expression displays. Out of the available points, facial points around eyebrows, eyes, mouth, nose, chin and cheeks were tracked and some other key positions were finally selected to improve the recognition accuracy. For the RGB video recorded by RGB HD camera, we used the open source tool OpenFace [1]. It provides facial landmark using the Conditional Local Neural Fields (CLNF) [18]. The CLNF performs the detection of 68 facial landmark. The CLNF is an instance of a Constrained Local Model (CLM) [19]. The CLM is composed of two main components: Point Distribution Model

Fig. 3. participant Acted facial expressions.

(PDM) which captures landmark shape variations; patch experts, which capture local appearance variations of each facial landmark [1]. Finally, we choose 26 facial points. Figure 4 illustrates the selected facial key points.

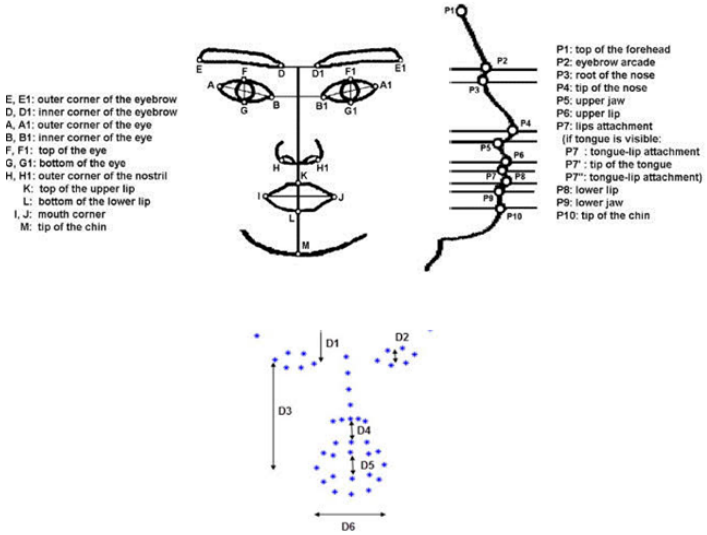


Fig. 4. Some of the selected facial key points.

The new face feature vector (FV_{Face}) is defined using geometric features: distance and angle with the horizontal axis and the coordinates of tracked points from one frame.
Given two facial points $P_n^{face}(t)$ and $P_{n-1}^{face}(t)$ with coordinates $(x_n(t), y_n(t), z_n(t))$ and $(x_{n-1}(t), y_{n-1}(t), z_{n-1}(t))$ respectively at frame t,

$$D_n^{face}(P_{n-1}^{face}(t), P_n^{face}(t) = \begin{cases} (x_{n-1}(t) - x_n(t)) \\ (y_{n-1}(t) - y_n(t)) \\ (z_{n-1}(t) - z_n(t)) \end{cases} \tag{1}$$

$$\theta_n^{face}(P_{n-1}^{face}(t), P_n^{face}(t))) = \begin{cases} \theta(x_{n-1}(t), x_n(t)) \\ \theta(y_{n-1}(t), y_n(t)) \\ \theta(z_{n-1}(t), z_n(t)) \end{cases} \tag{2}$$

$$FV_{Face} = \begin{cases} D_1^{face}(P_0^{face}(t), P_1^{face}(t)), ..., D_n^{face}(P_{n-1}^{face}(t), \\ P_n^{face}(t)), \theta_1^{face}(P_0^{face}(t), P_1^{face}(t)), \\ ..., \theta_n^{face}(P_{n-1}^{face}(t), P_n^{face}(t)) \end{cases} \tag{3}$$

The feature vector is based on position of the tracked points from one frame. The face feature vector is defined as follows (Equation 3). It is a set of distance difference $D(t)$ and $\theta(t)$ which is the angle between each selected facial points which are depicted in Equation 1 and Equation 2. We calculated 36 distance and 36 angle between each tracked points. We selected key facial points representing significant changes based on psychological studies [20].
Experiments in this study were conducted on a computer with an Intel Xeon CPU E3-1245 v3 3.40 Ghz and 8 GB RAM. All the experiments have been run in Matlab 2016b environment, using Matlab's own implementation of classification algorithms (Bagged Trees, k-NN, Linear SVM).
Support vector machine [21, 22] proposed by Vapnik and Chervonenk is a powerful statistical learning method, it models the situation by creating a feature space. The goal is to train a model that assigns new unseen objects into a particular category. Linear SVM is one method used in statistics and machine learning to find a linear combination of features which characterize or separate two or more classes or events. Since emotion recognition may not be linearly separable, we also considered non-linear classification algorithms. Bagging is a method for improving results of machine learning classification algorithms. This method was formulated by Leo Breiman and its name was deduced from the phrase bootstrap aggregating [17]. Bagged Trees can be used to reduce the variance associated with prediction and improve the prediction process. Many bagging samples are drawn from the available data, some prediction method is applied to each bagging sample, and then the results are combined, by simple voting process for classification, to obtain the overall prediction, with the variance being reduced due to the averaging. The bagging method generates additional data for training from the original dataset using combinations with repetitions to produce multi-sets of the same cardinality/size as the original data. By increasing the size of the training set, it cannot improve the model predictive force, but just decrease the variance, narrowly tuning the prediction to expected outcome.

The k-NN algorithm as non-parametric lazy learning algorithm is one of the simplest classification algorithm [23]. Even with such simplicity, it can give highly competitive results. Non-parametric means that it does not make any assumptions on the underlying data distribution. This is pretty useful , as in the real world , most of the practical data does not obey the typical theoretical assumptions made (gaussian mixtures, linearly separable etc). It is also a lazy algorithm which means is that k-NN does not use the training data points to do any generalization. There is no explicit training phase or it is very minimal and fast . All the training data is needed during the testing phase, the k-NN algorithm keeps all the training data. This is in contrast to other techniques like SVM where it is possible to discard all non support vectors without any problem. Most of the lazy algorithms especially k-NN makes decision based on the entire training data set. Predictions are made for a new instance by searching through the entire training set for the k most similar instances (the neighbors) and summarizing the output variable for those k instances. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance [24] which is calculated as the square root of the sum of the squared differences between a new point and an existing point.

3 Results and discussion

Six models have been defined to recognise the six target emotional states (anger, fear, happiness, sadness, surprise, and neutral) to solve our multi-class problem. As different classifiers may yield to different classification performance for the same dataset, we used for the training linear and non-linear classifiers including Bagged Trees, Fine k-NN and Linear SVM.

For models performance evaluation, we used leave-one-subject-out validation. The data of 16 participants were used for training and we left the performance of one participant for testing. Comparing the results of different training algorithms, we notice that Bagged Trees algorithm outperforms the remaining classifiers with accuracy rate of 98.46%, 97.51%, and 73.83% for Kinect (v2), Kinect (v1) and HD RGB camera respectively.

Comparing the devices used to collect the data, the Kinect captors achieved better results than OpenFace, which gives the lowest performance with 73.83%, 67.97%, and 71.49% of accuracy for Bagged Trees, Fine k-NN and Linear SVM respectively. The obtained results showed that Kinect (v2) performs better than Kinect (v1). The figure 5 showcases the obtained results using the three devices. The histogram presents the accuracy rates obtained by each classifier. Based on our experiment, the data collected by the HD RGB camera showed the lowest performance, which can be explained by the sensitivity to the surroundings, especially to illumination conditions [25]. The RGB-D images can capture essential geometrical features, and enable higher precision and preservation of facial details insensitive to different conditions. Table 1 shows the performance comparison between the proposed work in this paper and state-of-the-art works.

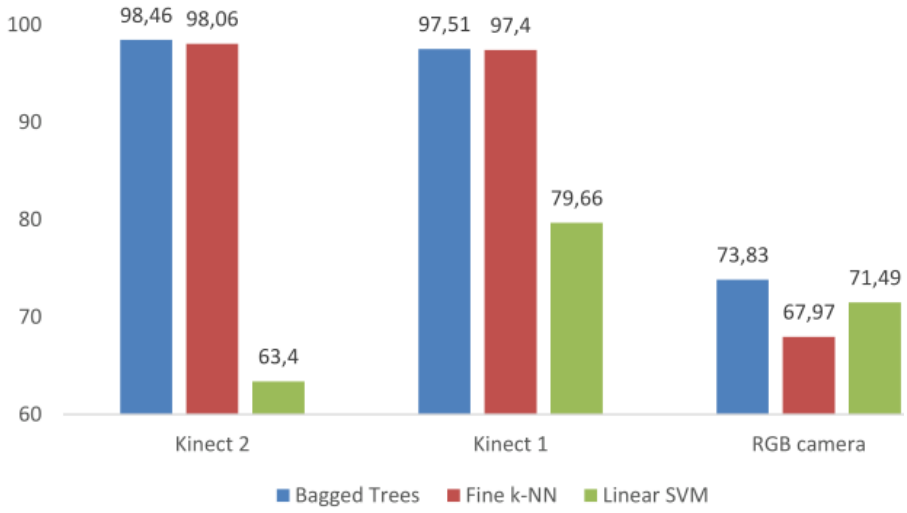


Fig. 5. The classifiers accuracy performance.

Table 1. Accuracy performance comparison with state of the art works. *Kinect 1, **Kinect 2, [1] k-NN, [2] Bagged-Trees.

Works	Number of classes	Accuracy %
[15] **[1]	6	89.44
	8	90.33
[26] *	6	80.75
	7	80.57
[3] **[1]	6	96.74
	8	96.92
Proposed approach **[2]	6	**98.46**
Proposed approach *[2]	6	**97.51**

Based on the comparison depicted on table 1, we believe that the proposed approach for facial expression recognition outperform the state-of-the-art works.

4 Conclusion

In this paper, we present a new facial expression recognition method by using a combination of angle and distance between facial key points as features. A comparison between RGB images and RGB-D images is depicted. The non-linear algorithms presented consistent results due the data nature, the Bagged Trees and k-NN consistently outperformed all the tested classification algorithms on our new dataset. Based on our experiments, we observed that the 2D images are not robust enough for facial emotion recognition which are 3D objects. Fur-

thermore, the RGB-D images can capture essential geometrical features, and enable higher precision and preservation of facial details insensitive to different conditions. Future works will concentrate on building real time system for facial expression recognition.

References

1. T. Baltruaitis, P. Robinson and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, 2016, pp. 1-10. doi: 10.1109/WACV.2016.7477553
2. P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion Recognition From EEG Using Higher Order Crossings," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 2, pp. 186-197, March 2010. doi: 10.1109/TITB.2009.2034649
3. N. Chanthaphan, K. Uchimura, T. Satonaka and T. Makioka, "Facial Emotion Recognition Based on Facial Motion Stream Generated by Kinect," 2015 11th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS), Bangkok, 2015, pp. 117-124. doi: 10.1109/SITIS.2015.31
4. Bekele E., Zheng Z., Swanson A., Davidson J., Warren Z., Sarkar N. (2013) Virtual Reality-Based Facial Expressions Understanding for Teenagers with Autism. In: Stephanidis C., Antona M. (eds) Universal Access in Human-Computer Interaction. User and Context Diversity. UAHCI 2013. Lecture Notes in Computer Science, vol 8010. Springer, Berlin, Heidelberg
5. Bekele E., Bian D., Zheng Z., Peterman J., Park S., Sarkar N. (2014) Responses during Facial Emotional Expression Recognition Tasks Using Virtual Reality and Static IAPS Pictures for Adults with Schizophrenia. In: Shumaker R., Lackey S. (eds) Virtual, Augmented and Mixed Reality. Applications of Virtual and Augmented Reality. VAMR 2014. Lecture Notes in Computer Science, vol 8526. Springer, Cham
6. Marcos-Pablos S, Gonzlez-Pablos E, Martn-Lorenzo C, Flores LA, Gmez-Garca-Bermejo J, Zalama E. Virtual Avatar for Emotion Recognition in Patients with Schizophrenia: A Pilot Study. Frontiers in Human Neuroscience. 2016;10:421. doi:10.3389/fnhum.2016.00421.
7. A. Psaltis et al., "Multimodal affective state recognition in serious games applications," 2016 IEEE International Conference on Imaging Systems and Techniques (IST), Chania, 2016, pp. 435-439.doi: 10.1109/IST.2016.7738265
8. C. A. Frantzidis et al., "On the Classification of Emotional Biosignals Evoked While Viewing Affective Pictures: An Integrated Data-Mining-Based Approach for Healthcare Applications," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 2, pp. 309-318, March 2010. doi: 10.1109/TITB.2009.2038481
9. B. Fasel, Juergen Luettin, Automatic facial expression analysis: a survey, Pattern Recognition, Volume 36, Issue 1, 2003, Pages 259-275, ISSN 0031-3203, http://dx.doi.org/10.1016/S0031-3203(02)00052-3.
10. Bekele E., Bian D., Zheng Z., Peterman J., Park S., Sarkar N. (2014) Responses during Facial Emotional Expression Recognition Tasks Using Virtual Reality and Static IAPS Pictures for Adults with Schizophrenia. In: Shumaker R., Lackey S. (eds) Virtual, Augmented and Mixed Reality. Applications of Virtual and Augmented Reality. VAMR 2014. Lecture Notes in Computer Science, vol 8526. Springer, Cham.

11. Bekele E., Zheng Z., Swanson A., Davidson J., Warren Z., Sarkar N. (2013) Virtual Reality-Based Facial Expressions Understanding for Teenagers with Autism. In: Stephanidis C., Antona M. (eds) Universal Access in Human-Computer Interaction. User and Context Diversity. UAHCI 2013. Lecture Notes in Computer Science, vol 8010. Springer, Berlin, Heidelberg.
12. Z. Zhang, L. Cui, X. Liu and T. Zhu, "Emotion Detection Using Kinect 3D Facial Points," 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, 2016, pp. 407-410.doi: 10.1109/WI.2016.0063
13. X. Zhao, J. Zou, H. Li, E. Dellandra, I. A. Kakadiaris and L. Chen, "Automatic 2.5-D Facial Landmarking and Emotion Annotation for Social Interaction Assistance," in IEEE Transactions on Cybernetics, vol. 46, no. 9, pp. 2042-2055, Sept. 2016. doi: 10.1109/TCYB.2015.2461131
14. B. Y. L. Li, A. S. Mian, W. Liu and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," 2013 IEEE Workshop on Applications of Computer Vision (WACV), Tampa, FL, 2013, pp. 186-192. doi: 10.1109/WACV.2013.6475017
15. N. Chanthaphan, K. Uchimura, T. Satonaka, T. Makioka, "Novel facial feature extraction technique for facial emotion recognition system by using depth sensor", (2016), International Journal of Innovative Computing, Information and Control ,12 20672087.
16. C . A. Gabert-Quillen, E. E. Bartolini, B. T. Abravanel, C. A. Sanislow, Ratings for emotion film clips, Behavior Research Methods 47, 2015, 773787.
17. Breiman, L. Machine Learning (1996) 24: 123. https://doi.org/10.1023/A:1018054314350.
18. T. Baltrusaitis, P. Robinson and L. P. Morency, "Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild," 2013 IEEE International Conference on Computer Vision Workshops, Sydney, NSW, 2013, pp. 354-361. doi: 10.1109/ICCVW.2013.54
19. Cristinacce D, Cootes TF. Feature detection and tracking with constrained local models. British Machine Vision Conference. 2006:929938.
20. B. Fasel, J. Luettin, Automatic Facial Expression Analysis: A Survey, (2003),Idiap-RR Idiap-RR-19-1999, IDIAP, 1999. Published in Pattern Recognition, 36(1):259-275.
21. Vladimir N. Vapnik. 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., New York, NY, USA.
22. Cortes, C. and Vapnik, V. Machine Learning (1995) 20: 273. https://doi.org/10.1023/A:1022627411411.
23. S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," in IEEE Transactions on Neural Networks and Learning Systems, vol. PP, no. 99, pp. 1-12. doi: 10.1109/TNNLS.2017.2673241.
24. Hui Wang, "Nearest neighbors by neighborhood counting," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 6, pp. 942-953, June 2006.doi: 10.1109/TPAMI.2006.126
25. S. M. Lajevardi and H. R. Wu, "Facial Expression Recognition in Perceptual Color Space," in IEEE Transactions on Image Processing, vol. 21, no. 8, pp. 3721-3733, Aug. 2012. doi: 10.1109/TIP.2012.2197628
26. Mao, Qi-rong,Pan, Xin-yu, Zhan, Yong-zhao, and Shen, Xiang-jun,"Using Kinect for real-time emotion recognition via facial expressions", Frontiers of Information Technology & Electronic Engineering,2015,vol.16, no.4, pp.272-282,issn="2095-9230".

An Improved Clustering for CpG Islands Identification based on Parallel Generalized Island Models

Abdelbasset Boukelia[1,2,3], Mohamed Batouche[1], and Brahim Matougui[1,2]

[1] Computer Science Departement, Faculty NTIC, University Abdelhamid Mehri - Constantine 2, Constantine, Algeria.
[2] National Center for Biotechnology Research, Constantine, Algeria.
{abdelbasset.boukelia,mohamed.batouche,brahim.matougui}
@univ-constantine2.dz

Abstract. DNA methylation is an important epigenetic alteration that regulates gene expression in several cellular processes. It is about including a group methyl to a cytosine in the C-5 position using enzymes called DNMTs. CpG islands are special regions where most of them are methylated and located in the transcription start of genes (TSS). Many genetic diseases are related to an aberrant number of CpG island methylation. In this paper, we present a novel method called iCpGIM for CpG islands identification in DNA sequences based on an improved clustering with a parallel generalized island model which comprises a genetic algorithm (GA), a particle swarm optimization (PSO) and a differential evolution algorithm (DE). This cooperative combination can effectively conquer the advantages of every technique while keeping up their points of interest. Using six DNA sequences downloaded from NCBI, iCpGIM shows a high performance in CpG island identification compared to the existing approaches in the literature.

Keywords: CpG Islands, Generalized Island Model, Clustering, Cooperative Optimization, Epigenetic, DNA Methylation

1 Introduction

Our genetic heritage conditions the color of our eyes, our hair, our skin, our size, and physical appearance. Each of our cells contains 24 chromosomes inherited from our parents on which we count about 25,000 genes. But if all our cells contain the same information, they obviously do not all have the same functions. As an example, a skin cell does not function as a cell of the liver. Similarly, two twins sharing the same genome are never perfectly identical [1].
However, each cell line has its own morphological and functional characteristics.These latter are linked directly to specific transcriptional pathways where in each different cell type, genes are expressed while others are transiently suppressed or permed. This differential gene expression can be modulated by epigenetic mechanisms that include changes in the methylation of DNA, in the histone

modifications as well as in the chromatin structure[1–3].
DNA methylation is an epigenetic alteration that includes the methyl gathering to the C-5 position of the cytosine dinucleotides of DNA by enzymes called DNA methyltransferases (DNMTs).In mammals, most CG sites are methylated on cytosine residues, whereas CG dinucleotides within promoters tend to be protected from methylation. Defects in DNA methylation are closely associated with cancer, although no mutation or deficiency in any DNMT has been identified as causally linked to tumor development, most likely because of their critical role during embryogenesis. Epigenetic hallmarks of cancer include global DNA hypomethylation and locus-specific hypermethylation of CpG islands (CGIs) [4].
Most CGIs are destinations of translation start (TSS).The methylation of a gene promoter is connected to the inactivation or restraint of the translation of this gene. However, nonmethylation of these promoter CG sites can instigates its transcription. This marks of CGI methylation can be used as biomarkers for several human diseases like Cancer and Alzheimer. Thus, the identification of CGIs can lead to an early diagnostic and prognostic for a majority of genetic diseases [5, 6].
The most utilized strategy to find the potential CGIs in the genome are based on clustering technique and classical optimization algorithms where GGF [7] criteria is taking as solution of the problem where a gander at districts of the DNA sequence with no less than 200 nucleotides long, a GC rate is no less than 50% and a Observed to-expected CpG proportion is over 60% [7, 8].
The amount of generated data become ubiquitous due to the evolution of genome sequencing technologies. Thus, the classical optimization algorithms are not designed to scale to the instances of this size. Many parallel optimization approaches are proposed to deal with this type of problems with large-scale of data [9, 10].The cooperative island model methods [11–16] are one of the most used techniques for solving big data optimization problems due to its effectiveness and the availability of high performance computing (HPC).
In this paper, a novel hybrid strategy for CGI identification iCpGIM is proposed to deal with the large scale of DNA reads. It consists essentially of three phases: A clustering stage which allows separating the genome so as to select the potential CGI candidates, and a parallel optimization strategy based on parallel generalized island models (GIM) which refines the cluster candidates by finding the best CGIs, then a filtering stage using a binomial distribution.
The remainder of the paper is organized as follows. In section 2, we discuss recent works in literature related to CGI identification . Section 3 is devoted to the proposed approach. Section 4 describes the experimental results. Finally, conclusions and future work are drawn.

2 Related Work

Most systems for identifying CGIs in DNA sequences are based on clustering and optimization algorithms.The CpGCluster [17] is one of the standards in literature. This technique is utilized to define statistically huge bunches of CG

dinucleotides by ascertaining the distance between CG sites. A threshold is determined to build CGI clusters. A *pvalue* is calculated using binomial distribution for each cluster, as a measure for taking a decision on behalf a defined *pvalue_limit*.
The CPSORL strategy [18] joins complementary particles swarms optimization (CPSO) with a reinforcement learning policy (RL) to identify CpG islands in the human genome. This technique uses GGF criteria [7] as metrics to recognize CpG islands.
It is important to highlight the most recent works in the literature like The ClusterPSO [19], 3CPSO [20] and CpGTBLO [21] that use two steps: The clustering technique of CpGCluster is used to define the CGI candidates that effectively used to reduce the huge volume of unnecessary DNA fragments and an optimization algorithm to refine these clusters with respecting the GGF criteria.The ClusterPSO technique joins CpGCluster strategy and a PSO calculation, where 3CPSO joins the Chaotic Complementary PSO to the clustering technique and the CpGCluster-TLBO approach also based on these two steps but it use TLBO algorithm to accurately predict CpG islands among the promising CpG island candidates.

In order to alleviate the limitations of existing methods, we have proposed a novel approach named iCpGIM for CGIs identification based on clustering and parallel generalized island models [11, 22], in the next section our method is drawn.

3 The Proposed Approach

The proposed method iCpGIM for CGIs identification in genome combines an improved clustering algorithm, a parallel optimization method based on GIM, and a filtering step **Figure.1**.The GIMs consists three optimization metaheuristics based population namely Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Differential Evolution algorithm (DE) in form of islands. iCpGIM is explained bellow.

- **Step One : Build CGI Clusters.** Two important properties to build clusters :
  - The start and the end of each cluster with the positions of GC sites.
  - The distance between GC sites.

  In this step the Clustering is described as follow :

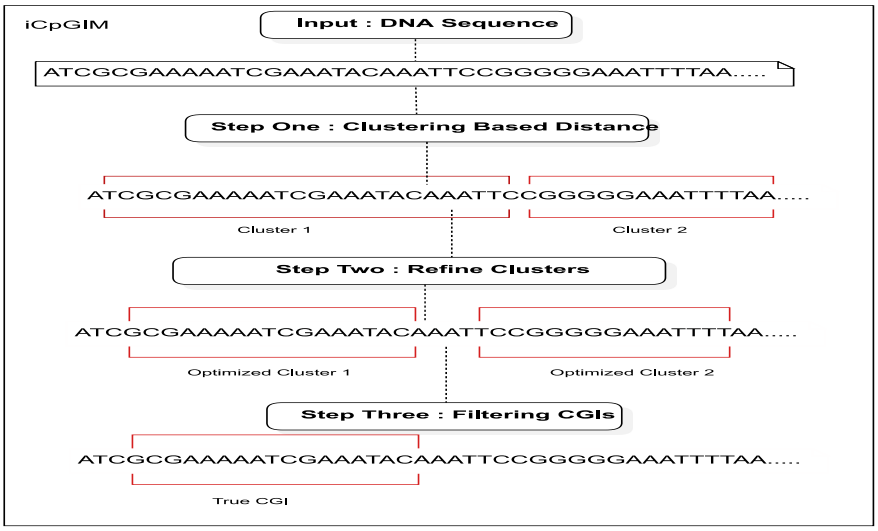


Fig. 1. The general framework of iCpGIM.

1. Save the position 'C' of GC sites by reading the DNA sequence in the direction 3' to 5' to gather the its positions $C=(c1,c2,...cn)$ where 'n' is the number of GC sites.
2. Calculate the distance between adjacent GC dinucleotides (by counting dinucleotides between the GC sites) is estimated using the following measure:

$$D_i = c_{i+1} - c_i - 1$$

   the value taken of the minimum distance between sites is equal to 1 *'CGCG'*.
3. Sort the distance list without eliminate any repeat distance detected to express the threshold 'd_f' in position 65^{th} [17] of the list.
4. Collect the positions using the threshold to define clusters. When a distance between neighbors GC sites is smaller than threshold d_f , so the two sites belong to the same cluster otherwise if the distance overrides the threshold d_f thats mean the end of the previous cluster and the start of the next one. repeat to find every cluster.

This clustering step is depicted on the **Figure.2**

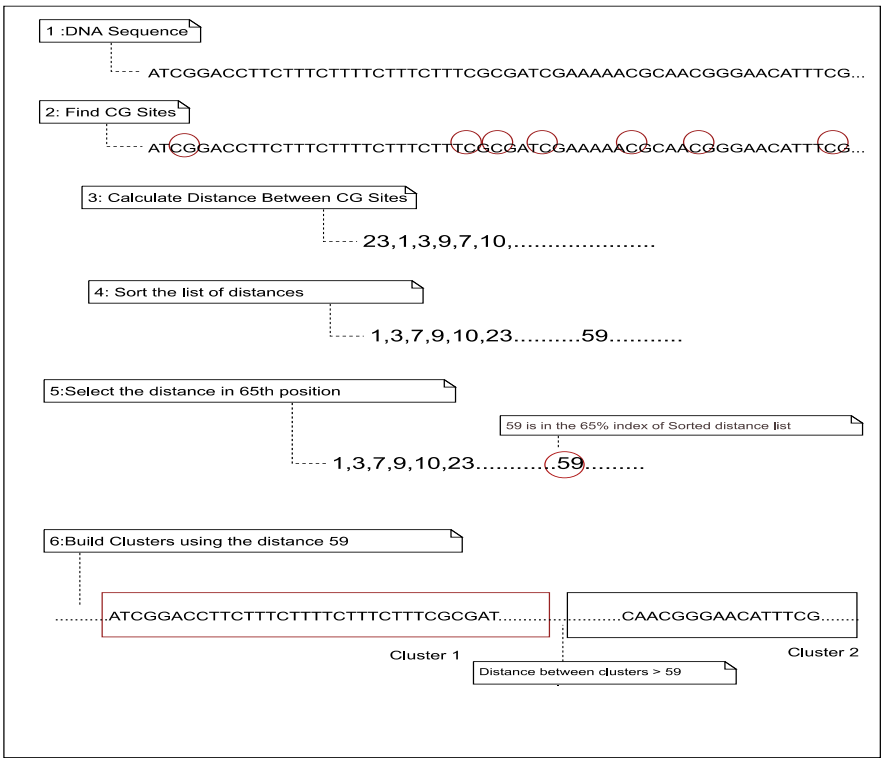


Fig. 2. The Clustering phase used in iCpGIM.

– **Step 2 : Refine the Clusters using GIM model.** After building the clusters, a GIM comprising a PSO, a GA, and a DE is assigned to each cluster. In every GIM, metaheuristics work in parallel and cooperate to find the best solution using a migration operator. it should be noted that there is no communication between GIMs. This step can be described as follows.

- Extend the cluster candidates.
- Create GIMs and assign each GIM to a cluster candidate.
  For each GIM:

1. **Initialization** of the populations in each island randomly.
   Set randomly the position and velocity for each island particles.
   the encode of position is used as follow :

$$P_i = (F_{si}, F_{li})$$

Where Fsi is the predicted start position of the particle Pi and Fli is her predicted length in the cluster. The Fsi and Fli are initialized by using **Equation 1** and **Equation 2** and the length of each candidate cluster

$$F_{si}^{initial} = rand * (end_{seq} - 200 - start_{seq}) + start_{seq} \quad (1)$$

$$F_{li}^{initial} = rand * (end_{seq} - 200 - F_{si}^{intial}) + 200 \quad (2)$$

2. **The fitness function** used was inspired by the GGF criteria (GC content 50, O/E ratio0.6, length 200 bp). The fitness functions of the length, the GC content and the O/E ratio are defined in **Equations (3-5)**, respectively. In addition, **Equation (6)** is used for the calculation of each particle fitness value. Note that the fitness values must be between 0 and 1 to adjust the function result:

$$CpG_{length}(P_i) = \frac{F_{li}}{L} \quad (3)$$

$$GC(P_i) = \frac{\#C + \#G}{F_{li}} \quad (4)$$

$$Obs/_{Exp}(P_i) = \frac{\#CpG \times F_{li}}{\#C \times \#G} \quad (5)$$

$$Fitness(P_i) = CpG_{length}(P_i) \times GC(P_i) \times Obs/_{Exp}(P_i) \quad (6)$$

Where, #C, #G, and #CpG are respectively the numbers of dinucleotides cytosine (C), guanine (G) and the number of CpG sites in the predicted CGI region at Pi. L is the number of nucleotide in the considered cluster (extended CGI candidate).

3. **Migration**. After each i^{th} iteration, the best solution of each island is chosen to undergo the migration, and it is transmitted to all neighboring islands. Once all selected solution are received, a recombination policy is used to form the new population. This process is repeated until a stopping condition (Optimal solution or maximum of iterations) is reached.

The Distrubed Optimization step is shown in the **figure 3**.

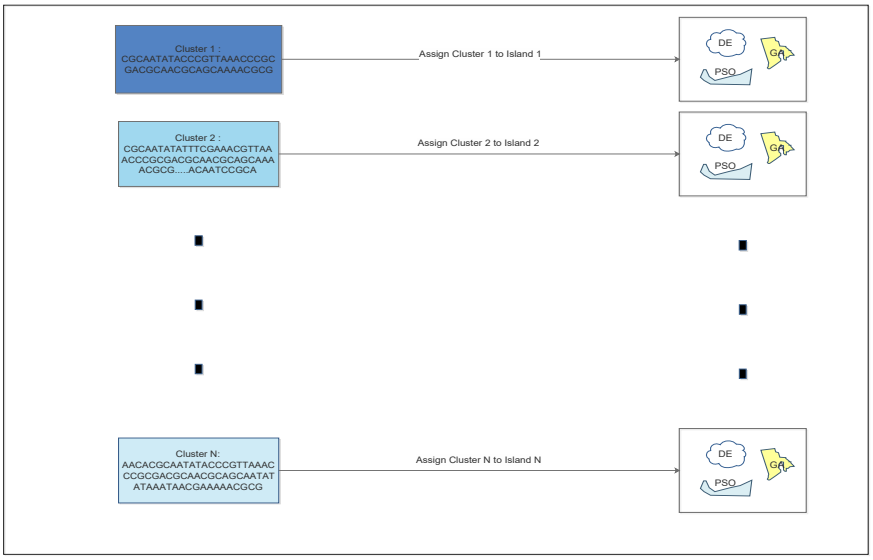


Fig. 3. The Optimization step using distributed GIM.

– **Step Three : Filtering the CGIs.** After the optimization step of all the clusters, the ***p-value*** of each optimized CGI is calculated for defining the probability to discover a CGI in a random sequence. The negative binomial distribution evaluate the probability to moderate the requirements of the CGI. If the number of successes is fixed in the progress, the distribution fails. Therefore, only the successes describe the truly CGI. The negative binomial distribution is calculated by the cumulative density function at point n_f of the CGI, and is taken as the *p-value*.

$$P^{Cum}_{(N,p)}(x <= n_f) = \sum_{x=0}^{n_f} \binom{(x-(N+1)-1)}{((N-1)-1)} * p^{N-1} * (1-p)^x \tag{7}$$

$$n_f = L - 2 * N \tag{8}$$

$$p = N_s / N_{is} \tag{9}$$

N is the number of CG sites in the CGI. n_f is the number of independent non-CG sites in the CGI. L is the length of the CGI. p is the probability of success discovering a CG site. and N_s is the number of CG sites and N_{is} the number of independent dinucleotides in the DNA sequence. This phase examines statistically significant CGIs and assumes that all CG sites are included in these CGI. If the $p - value$ of a cluster is bigger than the threshold value, the clusters are accepted; else, the cluster is rejected.

More formally, the whole algorithm is shown on **figure 4**.

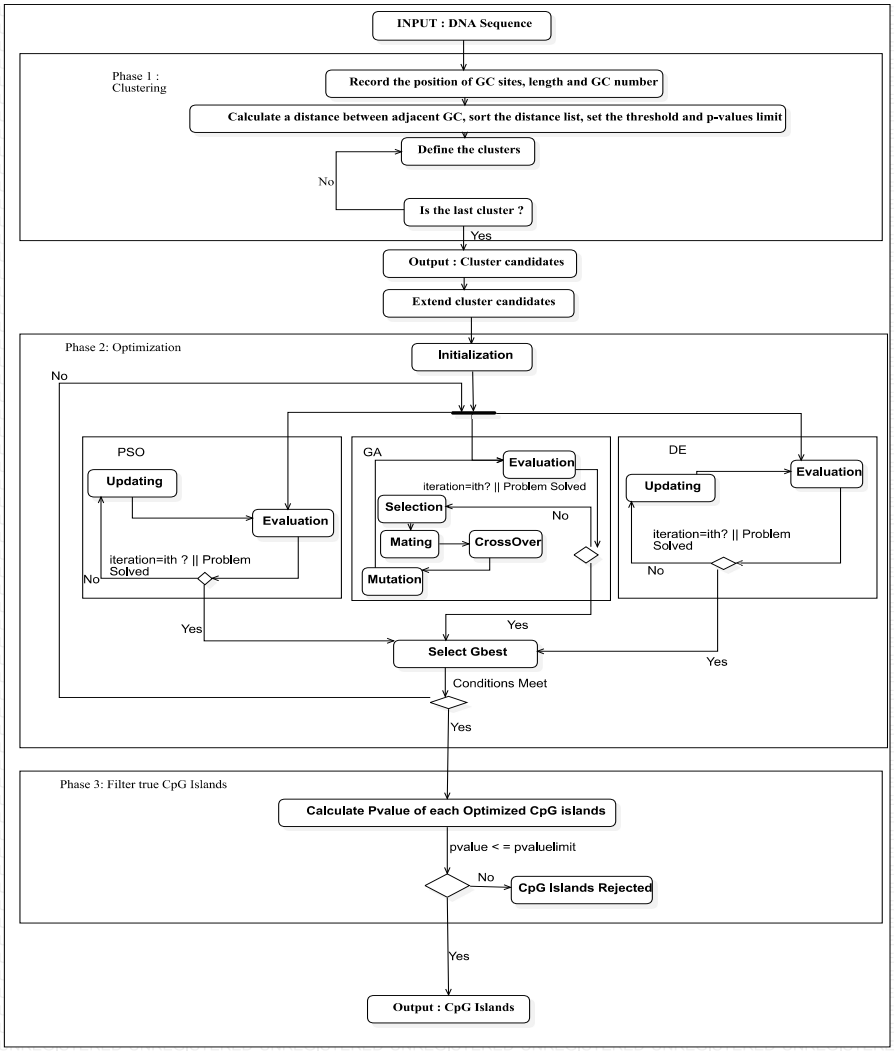


Fig. 4. The general framework of the iCpGIM.

4 Results & Discussion

iCpGIM was implemented using Pagmo2 integrated with Apache Cuda for GPU calculating and Apache Anaconda for the disturbed and parallel computing. In the clustering technique, a distance threshold parameter was set to 65th position

and the p-value to 10-2. In GIM Models, the population size was set to 100 particles in each island. The number of iterations is set to 200 and c1 = c2 = 2. All the mentioned settings are set arbitrary. Five common criteria are used to determine the prediction accuracy, namely the sensitivity (SN), specificity (SP), accuracy (ACC), performance coefficient (PC) and correlation coefficient (CC). In this study, we calculated the five prediction performances which are defined as follows

$$SN = \frac{TP}{TP + FN} \tag{10}$$

$$SP = \frac{TN}{TN + FP} \tag{11}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$CC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN) * (TP + FP) * (TN + FP) * (TN * FN)}} \tag{13}$$

$$PC = \frac{TP}{TP + FN + FP} \tag{14}$$

Where TP is true positive, TN is true negative, FN is false negative, and FP is false positive.
In this study, we compared our method to the best tools so far in literature using six sequences downloaded from NCBI (https://ncbi.nlm.nih.gov/).**Table .1** shows that iCpGIM achieved the highest SP using contig.*113958.2* **(100%)**,contig.*113953.1* **(100%)**, contig.*113954.1* **(100%)** and contig.*028395.3* **(100%)**.

Table 1. Comparaison of best methods of CGI identification so far

Contig.	Performance	Algorithms			
		ClusterPSO	3C-PSO	CpGTLBO	iCpGIM
NT 113952.1	SN	95.98	92.46	87.94	**98.20**
	SP	99.47	**100**	99.64	99.89
	ACC	99.32	99.51	99.12	**99.57**
	PC	86.16	74.94	81.48	**88.84**
	CC	92.28	85.67	89.36	**95.62**
NT 113955.2	SN	94.67	84.78	91.86	**95.65**
	SP	99.51	**100**	99.58	99.15
	ACC	99.33	**99.98**	99.30	99.71
	PC	83.81	84.78	82.69	**90.02**
	CC	90.92	93.94	90.17	**98.36**
NT 113958.2	SN	88.56	88.84	80.11	**90.16**
	SP	99.1	100	99.22	**100**
	ACC	98.43	99.9	98.00	**99.99**
	PC	78.2	88.41	71.82	**92.61**
	CC	86.93	93.98	82.63	**95.65**
NT 113953.1	SN	82.74	**96.72**	80.91	96.02
	SP	99.47	**99.98**	99.56	**100**
	ACC	98.99	**99.97**	99.02	99.51
	PC	70.39	93.09	70.63	**98.10**
	CC	82.09	96.37	82.31	**97.65**
NT 113954.1	SN	78.02	84.65	78.85	**90.65**
	SP	98.23	**100**	98.49	**100**
	ACC	97.48	99.32	97.76	**99.56**
	PC	53.34	50.81	56.56	**80.10**
	CC	68.72	71.18	71.37	**86.25**
NT 028395.3	SN	81.53	72.19	81.80	**86.25**
	SP	81.53	72.19	99.40	**100**
	ACC	99.24	100	98.78	**93.00**
	PC	98.6	**99.94**	70.43	99.85
	CC	67.53	72.2	82.02	**90.06**

Table 2. Number of CpG Islands & true CpG Islands detected by iCpGIM — 3CPSO

Chr	Contig	Number CGIs detected		Number of true islands	
		iCpGIM	3C-PSO	iCpGIM	3C-PSO
21	NT 113952.1	**20**	15	**18**	12
21	NT 113955.2	**22**	15	**16**	11
21	NT 113958.2	**60**	31	**45**	23
21	NT 113953.1	**19**	10	8	8
21	NT 113954.1	12	**15**	**10**	10
22	NT 028395.3	**51**	46	**45**	29

Table .2 depicts the number of CpG islands and the true ones identified by iCpGIM and 3CPSO, where the obtained results from iCpGIM are the highest so far in the literature. This demonstrates that iCpGIM outperforms existing methods for CpG island identification in the literature.
The complexity of iCpGIM is significantly reduced compared to other methods based solely on optimization algorithms. Indeed, the proposed method for CGIs detection combines a clustering technique with an optimization technique and filtering step. The clustering is used to find out potential CGIs. Thus, reducing the search space. The optimization technique based on cooperative generalized island model is used to refine the search and yields the best CpG islands included inside the CGI cluster candidates by migrating the optimal solutions of every island of the model.Thus, allowing to reach the optimal solution quickly comparing to the existing methods.

5 Conclusion

In this paper, an improved method iCpGIM for CpG islands identification is presented. This method consists of three steps : a clustering to define candidates and remove unnecessary fragments from DNA sequence, an optimizing step based on GIM to determine the true CpG islands from these candidates and a filtering step to using statistically measure called binomial distribution. The experimental results shows a high performance and sensitivity of iCpGIM comparing to the others tools in literature. As perspective, a new approach based on deep learning for CpG islands will be proposed to improve and identify new features and signatures for CpG islands.

References

1. Makoto Tachibana. Epigenetic regulation of mammalian sex determination. *The Journal of Medical Investigation*, 62:19–23, 2015.
2. Cotton Allison M et al. Chromosome-wide dna methylation analysis predicts human tissue-specific x inactivation. *Human genetics*, 130(2):187–201, Aug 2011.
3. Xiaoyu Pan, Desheng Gong, Duc Ninh Nguyen, Xinxin Zhang, Qi Hu, Hanlin Lu, Merete Fredholm, Per T Sangild, and Fei Gao. Early microbial colonization affects dna methylation of genes related to intestinal immunity and metabolism in preterm pigs. *DNA Research*, 2018.
4. R. Massicotte, E. Whitelaw, and B. Angers. Dna methylation: A source of random variation in natural populations. *Epigenetics: official journal of the DNA Methylation Society*, 6(4):421, 2011.
5. Yi Cai, Hsing-Chen Tsai, Ray-Whay Chiu Yen, Yang W Zhang, Xiangqian Kong, Wei Wang, Limin Xia, and Stephen B Baylin. Critical threshold levels of dna methyltransferase 1 are required to maintain dna methylation across the genome in human cancer cells. *Genome research*, 27(4):533–544, 2017.
6. Lasse Sommer Kristensen, Marianne Bach Treppendahl, and Kirsten Gronbaek. Analysis of epigenetic modifications of dna in human cells. *Current protocols in human genetics*, pages 20–2, 2013.

7. Gardiner-Garden M and Frommer M. Cpg islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–82, Jul 1987.
8. D Takai and P A Jones. Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A*, 99:3740–5, 2002.
9. El-Ghazali Talbi and Celso Ribeiro. Special issue on "optimization and machine learning". *ITOR*, 25(4):1407, 2018.
10. AJ Umbarkar and MS Joshi. Review of parallel genetic algorithm based on computing paradigm and diversity in search space. *ICTACT Journal on Soft Computing*, 3(4):615–622, 2013.
11. Izzo Dario, Ruciński Marek, and Biscani Francesco. The generalized island model. In *Parallel Architectures and Bioinspired Algorithms*, pages 151–169. Springer, 2012.
12. Kai Wang, Zhen Shen, et al. A gpu-based parallel genetic algorithm for generating daily activity plans. *IEEE Trans. Intelligent Transportation Systems*, 13(3):1474–1480, 2012.
13. Houda Abadlia, Nadia Smairi, and Khaled Ghedira. Particle swarm optimization based on dynamic island model. In *Tools with Artificial Intelligence (ICTAI), 2017 IEEE 29th International Conference on*, pages 709–716. IEEE, 2017.
14. Mohamed Kurdi. An effective new island model genetic algorithm for job shop scheduling problem. *Computers & operations research*, 67:132–142, 2016.
15. Meryem Ammi and Salim Chikhi. Cooperative parallel metaheuristics based penguin optimization search for solving the vehicle routing problem. *International Journal of Applied Metaheuristic Computing (IJAMC)*, 7(1):1–18, 2016.
16. Caner Candan, Adrien Goeffon, Frédéric Lardeux, and Frédéric Saubion. A dynamic island model for adaptive operator selection. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*, pages 1253–1260. ACM, 2012.
17. Hackenberg Michael et al. Cpgcluster: a distance-based algorithm for cpg-island detection. *BMC bioinformatics*, 7:446, 2006.
18. Chuang Li-Yeh et al. Particle swarm optimization with reinforcement learning for the prediction of cpg islands in the human genome. *PloS one*, 6(6):e21036, 2011.
19. Cheng-Hong Yang, Yu-Da Lin, Yi-Cheng Chiang, and Li-Yeh Chuang. A hybrid approach for cpg island detection in the human genome. *PloS one*, 11(1):e0144748, 2016.
20. Boukelia Abdelbasset and Benmounah Zakaria et al. A novel algorithm for cpg island detection in human genome based on clustering and chaotic particle swarm optimization. In *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 70–81. Springer, 2016.
21. Yang Cheng-Hong et al. A cpgcluster-teaching–learning-based optimization for prediction of cpg islands in the human genome. *Journal of Computational Biology*, 25(2):158–169, 2018.
22. Zakaria Benmounah and Mohamed Batouche. A parallel distributed system for gene expression profiling based on clustering ensemble and distributed optimization. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 176–185. Springer, 2013.

Support Global Algorithm for Nonconvex Quadratic Minimization with One Negative Eigenvalue

Amar ANDJOUH[1] and Mohand Ouamer BIBI[1]

[1] Research Unit LaMOS, University of Bejaia, 06000 Bejaia, Algeria
omarandjouh@yahoo.fr, mobibi.dz@gmail.com

Abstract. This paper provides a new support method of global optimization to solve the quadratic minimization problem with one negative eigenvalue, subject to box constraints. We investigate the support of the objective function and exploit properties of the indefinite associated matrix for finding global optimality criterion (necessary and sufficient conditions). Furthermore, using these conditions and computational techniques, we apply the support method that can effectively solve a quadratic minimization problem with an indefinite associated matrix, having one negative eigenvalue. Particularly, we study the case where the associated matrix is positive subdefinite, and we use the suggested support algorithm in order to find the optimal solution. We present numerical applications to solve some box-constrained nonconvex problems with one negative eigenvalue.

Keywords: Quadratic Minimization with One Negative Eigenvalue, Support Method, Support Feasible Solution (SFS), Global Optimality Criterion.

1 Introduction

The resolution of a quadratic problem with linear constraints is very difficult in the nonconvex case, clearly the nonconvex quadratic problems are NP-Complete. In particular, global quadratic minimization problem with one negative eigenvalue is NP-hard [10]. So the global research of the solutions is a very difficult and very complicated application, and several efforts have been made to find efficient methods in order to simplify the resolution of this type of problems[9].

Our contribution in this paper is the development of a new method for solving the nonconvex quadratic problem, where the associated matrix is indefinite and contains precisely one negative eigenvalue. In particular, the problem with an associated positive subdefinite matrix is often not NP-hard [12]. By Pardalos in [1], the latter can have two local minima, the one that verifies the sufficient global optimality criterion will be a global minimum. On the other hand, we can meet problems with one negative eigenvalue that admit a great number of local minima, and these latters are NP-hard [10].

Many researches deal with the nonconvex quadratic problems. So Jeyakumar et al.[3] have developed sufficient global optimality conditions for nonconvex quadratic minimization problems while exploiting subdifferentiability techniques for global optimization. Also several applications of nonconvex quadratic programming problems with box constraints are mentioned [11], and the nonconvex quadratic programming is the topic of actuality [4].

This work is essentially inspired by those of Gabasov et al.[5] and others which have developed support methods for convex quadratic programming problems (see [7] and [8]). In this paper we characterize the global minimum with sufficient global optimality conditions while using the property of the associated matrix D of the objective function F. Hence we develop a new direct support algorithm for nonconvex quadratic minimization with one negative eigenvalue. The iteration of the algorithm is based on the following principle: From an initial Support Feasible Solution $\{x, J_S\}$, we move to a new SFS $\{\overline{x}, \overline{J_S}\}$ such that $F(\overline{x}) \leq F(x)$.

2 Model description

We consider the nonconvex quadratic minimization problem with box constraints:

$$(QP) \qquad \min F(x) = \frac{1}{2}x^t Dx + c^t x, \tag{1}$$

$$s.t \quad \ell_i \leq x_i \leq u_i, \qquad i = \overline{1,n}, \tag{2}$$

where $D^t = D = (d_{ij}, 1 \leq i, j \leq n)$ is a symmetric matrix of order n, supposed indefinite with one negative eigenvalue (in particular, D is positive subdefinite); ℓ_i and u_i, with $\ell_i < u_i$, are finite real numbers for all $i \in J = \{1, 2, ..., n\}$; $\ell = \ell(J)$, $u = u(J)$, $c = c(J) \in \mathbb{R}^n$ and $x = x(J) \in S \subseteq \mathbb{R}^n$, where $S = \{x \in \mathbb{R}^n : \ \ell \leq x \leq u\}$ is the set of the feasible solutions of the problem (QP), and the symbol $(^t)$ represents the transposition operation.

2.1 Properties of the positive subdefinite matrices

Definition 1. *[12]The symmetric matrix $D = (d_{ij}, 1 \leq i, j \leq n)$ is said to be positive subdefinite* (PSubD)*, if for all $x \in \mathbb{R}^n$ we have*

$$x^t Dx < 0 \ \ implies \ \ Dx \neq 0 \ \ and \ \ Dx \geq 0 \ \ or \ \ Dx \leq 0, \tag{3}$$

and is said to be positive semidefinite (PSD) *if*

$$x^t Dx \leq 0 \ \ implies \ \ Dx = 0. \tag{4}$$

The class of PSubD *matrices is a natural generalization of the class of positive semidefinite matrices (*PSD*) and it is useful in the study of quadratic programming problems [2].*

Theorem 1. *[6]The real symmetric matrix D is called Merely Positive SubDefinite matrices (* MPSubD *: matrices that are not* PSD*), if and only if*

1. $\eta(\mathrm{D}) = 1$,
2. $D \leq 0 \Leftrightarrow D = (d_{ij} \leq 0, 1 \leq i, j \leq n)$ *and* $D \neq 0$,

where $\eta(\mathrm{D})$ *is the number of the negative eigenvalues of* D.

Let $\rho(D)$ be the spectral radius of the matrix D :

$$\mid \lambda_1 \mid = \rho(D) = \max\{\mid \lambda_i \mid, \quad i = \overline{1,n}\},$$

and the eigenvalue λ_1 is said to be dominant, because its absolute value is equal or greater than the absolute value of any other eigenvalue $\lambda_i,\ i = \overline{2,n}$.

Theorem 2. *The negative eigenvalue of an* MPSubD *matrix is dominant. Moreover its absolute value is not less than the sum of the other nonnegative eigenvalues.*

Proof: If D is MPSubD, then $D \leq 0$, and $tr(D) = \sum_{i=1}^{n} d_{ii} = \sum_{i=1}^{n} \lambda_i \leq 0$. Hence, for the negative eigenvalue λ_1 we must have $\mid \lambda_1 \mid = -\lambda_1 \geq \sum_{i=2}^{n} \lambda_i$. □

3 Local optimality conditions

3.1 First Order Necessary Optimality Conditions

Let x be a global (local) minimum of (QP). Then the following conditions must be satisfied[11]:

$$\begin{aligned} E_i(x) \geq 0, \quad & \forall\ i \in J_L = \{i \in J : \quad x_i = \ell_i\}, \\ E_i(x) \leq 0, \quad & \forall\ i \in J_U = \{i \in J : \quad x_i = u_i\}, \\ E_i(x) = 0, \quad & \forall\ i \in J_F = \{i \in J : \quad \ell_i < x_i < u_i\}, \end{aligned} \tag{5}$$

where $E = Dx + c$ is the gradient of the objective function F at x. The point x satisfying the conditions (5) is called a stationary point of the problem (QP).

Remark 1 *If* D *is positive semidefinite matrix* ($D \succcurlyeq 0$)*, then the relations (5) are also sufficient for the global optimality of the vector* x.

3.2 Second Order Necessary Optimality Conditions

Let x be a stationary point of the problem (QP). Then the following condition

$$D_F = D(J_F, J_F) \succcurlyeq 0 \ \ (J_F\ is\ defined\ in\ (5)\ and\ verifies\ E(J_F) = 0) \tag{6}$$

is necessary for the global (local) optimality of the vector x.

3.3 Second Order Sufficient Optimality Conditions

Let x be a stationary point verifying the conditions (5) and we consider the set

$$J_0 = \{i \in J : \quad E_i = 0\}. \tag{7}$$

If $D(J_0, J_0) \succ 0$, then x is a local minimum of the problem (QP).

4 Global optimality criterion

Let x be a feasible solution of the problem (QP) and let's consider another arbitrary feasible solution $\overline{x} = x + \Delta x$. So the increment formula of the objective function is given by:

$$F(\overline{x}) - F(x) = E^t(x)\Delta x + \frac{1}{2}\Delta x^t D \Delta x. \tag{8}$$

So we can write :

$$F(\overline{x}) - F(x) = E^t(x)\Delta x + \frac{1}{2}\Delta x^t Q \Delta x + \frac{1}{2}\Delta x^t (D - Q)\Delta x,$$

where the matrix $Q = diag(\alpha_1, ..., \alpha_n)$, $\alpha_i \in \mathbb{R}$, is constructed such that $D - Q \succcurlyeq 0$, with D supposed MPSubD or indefinite having one negative eigenvalue.
There are several techniques to generate the matrix Q that verifies $D - Q \succcurlyeq 0$. Therefore we give two examples [3]:

a) $Q_1 = \overline{D}$, where $\overline{D} = diag(\overline{d}_1, ..., \overline{d}_n)$, $\overline{d}_i \in \mathbb{R}$, is constructed such that $D - \overline{D} \succcurlyeq 0$. So we define $\overline{d}_i$ as follows:

$$\overline{d}_i = d_{ii} - \sum_{j=1, j\neq i}^{n} |d_{ij}|, \quad \forall i = 1, ..., n.$$

The matrix $(D - \overline{D})$ will be diagonally dominant with nonnegative diagonal elements. Hence we deduce that $D - \overline{D} \succcurlyeq 0$.

b) $Q_2 = \lambda_1 I_n$, where λ_1 is the negative eigenvalue of the matrix D, and I_n is an identity matrix of order n. Consequently, we get $D - \lambda_1 I_n \succcurlyeq 0$.

For testing the sufficient global optimality conditions, it is preferable to construct another matrix Q combining the matrices Q_1 and Q_2 [3]. So, in order to satisfy the global optimality criterion, we chose an arbitrary real number $\rho \in [0, 1]$ and we determine Q as follows: $Q = \rho Q_1 + (1 - \rho)Q_2 = diag(\alpha_1, ..., \alpha_n)$, where $\alpha_i = \rho\overline{d}_i + (1 - \rho)\lambda_1,\ i = 1, ..., n$. Clearly $D - Q_1 \succcurlyeq 0$ and $D - Q_2 \succcurlyeq 0$, hence with $\rho \in [0, 1]$ we have

$$\rho(D - Q_1) + (1 - \rho)(D - Q_2) \succcurlyeq 0 \Leftrightarrow D - \rho Q_1 - (1 - \rho)Q_2 \succcurlyeq 0 \Leftrightarrow D - Q \succcurlyeq 0.$$

For deriving the sufficient conditions of global optimality we define the matrix $\widehat{Q} = diag(\widehat{\alpha}_1, ..., \widehat{\alpha}_n)$, where the numbers $\widehat{\alpha}_i,\ i = 1, ..., n$ are defined as follows:

$$\widehat{\alpha}_i = \min\{0, \alpha_i\} = \begin{cases} \alpha_i, \, if \quad \alpha_i < 0, \\ 0, \quad if \quad \alpha_i \geq 0, \ i \in J. \end{cases} \tag{9}$$

So the condition $\widehat{\alpha}_i \leq 0$ holds,for all$i = 1, ..., n$. Furthermore, we have

$$\widehat{Q} \preccurlyeq Q \Rightarrow D - \widehat{Q} \succcurlyeq D - Q \succcurlyeq 0.$$

4.1 Sufficient optimality conditions

We consider the increment formula (8) of the objective function in the following form:

$$F(\overline{x}) - F(x) = (E(x) + \frac{1}{2}\widehat{Q}\Delta x)^t \Delta x + \frac{1}{2}\Delta x^t (D - \widehat{Q})\Delta x, \tag{10}$$

where $D - \widehat{Q} \succcurlyeq 0$. Since $\frac{1}{2}\Delta x^t (D - \widehat{Q})\Delta x \geq 0$, $\forall x,\ \overline{x} \in S$, then we deduce

$$F(\overline{x}) - F(x) \geq [E(x) + \frac{1}{2}\widehat{Q}\Delta x]^t \Delta x. \tag{11}$$

If we have $[E(x) + \frac{1}{2}\widehat{Q}\Delta x]^t \Delta x \geq 0,\ \forall \overline{x} \in S$, then x is a global minimum of (QP).

Theorem 3. *Let x be a feasible solution of the problem (QP) and we note by $\widehat{E}$ the vector of estimations such that*

$$\widehat{E}_i(x) = \begin{cases} E_i(x) + \frac{1}{2}\widehat{\alpha}_i(u_i - \ell_i),\ if \quad x_i = \ell_i, \\ E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\ell_i - u_i),\ if \quad x_i = u_i, \\ E_i^2(x) - \frac{1}{2}\widehat{\alpha}_i(u_i - \ell_i), if \quad \ell_i < x_i < u_i,\ i \in J. \end{cases} \tag{12}$$

Then the following conditions:

$$\begin{cases} \widehat{E}_i(x) \geq 0, if \quad x_i = \ell_i, \\ \widehat{E}_i(x) \leq 0, if \quad x_i = u_i, \\ \widehat{E}_i(x) = 0, if \quad \ell_i < x_i < u_i,\ i \in J, \end{cases} \tag{13}$$

are sufficient for the global optimality of the vector x.

Proof: Let $A(x,\ \overline{x}) = \sum_{i=1}^{n}[E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - x_i)](\overline{x}_i - x_i)$. If we can prove that $A(x,\ \overline{x}) \geq 0,\ \forall\ \overline{x} \in S$, then

$$F(\overline{x}) - F(x) \geq 0 \Rightarrow x\ is\ a\ global\ mimimum\ of\ (QP).$$

Indeed, we have $x_i \geq \ell_i \Rightarrow \overline{x}_i - x_i \leq \overline{x}_i - \ell_i \Rightarrow \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - x_i) \geq \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - \ell_i)$. So

$$E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - x_i) \geq E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - \ell_i) \geq E_i(x) + \frac{1}{2}\widehat{\alpha}_i(u_i - \ell_i).$$

On the other hand, we have also

$$x_i \leq u_i \Rightarrow \overline{x}_i - x_i \geq \overline{x}_i - u_i \Rightarrow \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - x_i) \leq \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - u_i).$$

$$\Rightarrow E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - x_i) \leq E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - u_i) \leq E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\ell_i - u_i).$$

We now consider the following three cases:
Case 1: if $x_i = \ell_i$ and $\widehat{E}_i(x) \geq 0$, then $\overline{x}_i - x_i = \overline{x}_i - \ell_i \geq 0$ and we obtain

$$[E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - x_i)](\overline{x}_i - x_i) \geq [E_i(x) + \frac{1}{2}\widehat{\alpha}_i(u_i - \ell_i)](\overline{x}_i - \ell_i) = \widehat{E}_i(x)(\overline{x}_i - \ell_i) \geq 0.$$

Case 2: if $x_i = u_i$ and $\widehat{E}_i(x) \leq 0$, then $\overline{x}_i - x_i = \overline{x}_i - u_i \leq 0$ and we obtain

$$[E_i(x)+\frac{1}{2}\widehat{\alpha}_i(\overline{x}_i-x_i)](\overline{x}_i-x_i) \geq [E_i(x)+\frac{1}{2}\widehat{\alpha}_i(\ell_i-u_i)](\overline{x}_i-u_i) = \widehat{E}_i(x)(\overline{x}_i-u_i) \geq 0.$$

Case 3: if $\ell_i < x_i < u_i$ and $\widehat{E}_i(x) = 0$, then we have

$$E_i^2(x) - \frac{1}{2}\widehat{\alpha}_i(u_i - \ell_i) = 0 \Rightarrow \frac{1}{2}\widehat{\alpha}_i(u_i - \ell_i) = E_i^2(x) \geq 0.$$

On the other hand, we have $\widehat{\alpha}_i \leq 0$ and $(u_i - \ell_i) > 0$. So

$$\frac{1}{2}\widehat{\alpha}_i(u_i - \ell_i) \leq 0 \Rightarrow \widehat{\alpha}_i = 0 \; and \; E_i = 0.$$

Then $E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - x_i) = 0$, $\forall \overline{x} \in S$. We deduce

$$[E_i(x) + \frac{1}{2}\widehat{\alpha}_i(\overline{x}_i - x_i)](\overline{x}_i - x_i) = 0.$$

Hence, $\forall \overline{x} \in S$, $F(\overline{x})-F(x) = \sum_{i=1}^{n}[E_i(x)+\frac{1}{2}\widehat{\alpha}_i(\overline{x}_i-x_i)](\overline{x}_i-x_i) = A(x,\, \overline{x}) \geq 0$. Then x is a global minimum of the problem (QP). □

5 Support global optimality criterion

Let J_S and J_N be the subsets of J such that $J_S \cup J_N = J$ and $J_S \cap J_N = \emptyset$.

Definition 2. *The subset $J_S \subset \overline{J}_0$ is called a support of the objective function if*

$$D_S = D(J_S, J_S) \succ 0, \tag{14}$$

where $\overline{J}_0$ is such that

$$\overline{J}_0 = \{i \in J : \quad \widehat{\alpha}_i = 0\}, \quad |\overline{J}_0| \leq n - 1. \tag{15}$$

The subset J_S is then said to be the support of the problem (QP). The couple $\{x, J_S\}$ formed by a feasible solution x and a support J_S is called a support feasible solution (SFS).

Definition 3. *The support feasible solution $\{x, J_S\}$ is said to be consistent if $E(J_S) = 0$.*

5.1 Support necessary optimality conditions

Consider a consistent support feasible solution $\{x, J_S\}$ of the problem (QP) and $E(x) = \nabla F(x) = Dx + c$ is the gradient of the function F, with $E(J_S) = 0$. Then the following conditions:

$$\begin{cases} E_i \geq 0, if \quad x_i = \ell_i, \\ E_i \leq 0, if \quad x_i = u_i, \\ E_i = 0, if \quad \ell_i < x_i < u_i, \; i \in J_N = J \setminus J_S, \end{cases} \tag{16}$$

are necessary for the global (local) optimality of the SFS $\{x, J_S\}$.

5.2 Support sufficient optimality conditions

Let $\{x, J_S\}$ be an SFS of the problem (QP), with $E(J_S) = 0$. We set $V(x) = V(J) = E(x) + \frac{1}{2}\widehat{Q}\Delta x$. So

$$V(J_S) = E(J_S)+\frac{1}{2}\widehat{Q}(J_S, J)\Delta x = E(J_S)+\frac{1}{2}\widehat{Q}(J_S, J_S)\Delta x(J_S)+\frac{1}{2}\widehat{Q}(J_S, J_N)\Delta x(J_N).$$

Since $\widehat{Q}$ is diagonal and $\widehat{\alpha}_i = 0,\ i \in J_S$, we deduce $V(J_S) = E(J_S) = 0$. Hence the relations (11) yields:

$$\begin{aligned} F(\overline{x}) - F(x) &\geq V^t(x)\Delta x = V^t(J_S)\Delta x(J_S) + V^t(J_N)\Delta x(J_N) \\ &\geq V^t(J_N)\Delta x(J_N). \end{aligned}$$

So, according for the theorem 3, the following conditions

$$\begin{cases} \widehat{E}_i(x) \geq 0, if \quad x_i = \ell_i, \\ \widehat{E}_i(x) \leq 0, if \quad x_i = u_i, \\ \widehat{E}_i(x) = 0, if \quad \ell_i < x_i < u_i,\ i \in J_N = J \setminus J_S, \end{cases} \tag{17}$$

are sufficient for the global optimality of the SFS $\{x, J_S\}$.

Remark 2 *If the relations (17) of the global optimality are not verified, then there exists at least an index $i_0 \in J_N$ such that:*

$$\widehat{E}_{i_0}(x) < 0, \quad and\ x_{i_0} < u_{i_0} \quad or \quad \widehat{E}_{i_0}(x) > 0, \quad and\ x_{i_0} > \ell_{i_0}. \tag{18}$$

Theorem 4. *Let x^* be a global (local) minimum for the problem (QP). If D is $MPSubD$, then $J_S = \emptyset$, and we have*

$$x_i^* = \ell_i \ or\ x_i^* = u_i, \qquad \forall\ i \in J,$$

i.e., the global (local) minimum x^* is a vertex.

Proof: According to the second order necessary optimality conditions, the global (local) minimum must verify $D_F \succcurlyeq 0$. Since the matrix D is MPSubD, then all its diagonal elements are nonpositive. Consequently, every principal square submatrix of the matrix D has the nonpositive diagonal elements. What implies for all J_F, the matrix D_F is not positive semidefinite. Consequently, at the optimum, $J_F = \emptyset$, i.e., the global (local) minimum must be an extreme point (vertex). Hence $J_S = \emptyset$. □

6 Support Method of resolution

Given $\{x, J_S\}$ an initial SFS. If x is not an optimal solution, then we consider one of the nonoptimal indices. This allows to construct an improvement direction of the objective function and the step along this direction and find a new FS

$\{\overline{x}, \overline{J_S}\}$. Construct a new FS $\overline{x} = x + \theta d$, where $d = (d_i, i \in J)$ is an improving feasible direction at a point x, and $\theta \geq 0$ is the step along this direction.
For this, let J_1 be the subset of J_N defined as follows:

$$J_1 = \left\{ i \in J_N : \ [\widehat{E}_i < 0, \ and \ x_i < u_i] \ \ or \ \ [\widehat{E}_i > 0, \ and \ x_i > \ell_i] \right\}. \tag{19}$$

In order to obtain a maximal increment, we must choose the subscript i_0 such that

$$|\widehat{E}_{i_0}| = \max |\widehat{E}_i|, \ i \in J_1. \tag{20}$$

For the computation of the direction d, we must take into account the following conditions:

(i) $E_i(\overline{x}) = \widehat{E}_i(\overline{x}) = 0, \ i \in J_S$, must be verified for the new FS $\overline{x} = x + \theta d$,
(ii) the objective function must decrease while passing from x to $\bar{x}$.

So we set

$$d_{i_0} = -sign\widehat{E}_{i_0}, \ d_i = 0 \ if \ i \neq i_0, \forall i \in J_N. \tag{21}$$

The relation $E_i(x+\theta d) = \widehat{E}_i(x+\theta d) = 0$, $i \in J_S$ must be verified, it is equivalent that the equality

$$D(J_S, J_S)d(J_S) + D(J_S, J_N)d(J_N) = 0,$$

i.e.,

$$d(J_S) = -D_S^{-1}D(J_S, J_N)d(J_N) = D_S^{-1}D(J_S, i_0)sign\widehat{E}_{i_0}. \tag{22}$$

Therefore, whatever is the choice of the component $d(J_N)$ of the vector d, the relations $\widehat{E}_i(x + \theta d) = 0$, $i \in J_S$, remain always verified.

While choosing the step θ along the direction d, the following conditions must be satisfied :

(a) the constraint (2) of program (QP) must be verified for $\overline{x} = x + \theta d$,
(b) the passage from x to $\overline{x}$ ensures a maximal decreasing of the objective function.

For (a), we must have

$$\ell_i \leq x_i + \theta d_i \leq u_i, \quad i \in J_N, \quad \ell_i \leq x_i + \theta d_i \leq u_i, \quad i \in J_S.$$

If we consider $\theta \in [0; 1]$, we give

$$d_{i_0} = \begin{cases} \ell_{i_0} - x_{i_0}, \ if \quad \widehat{E}_{i_0}(x) > 0, \\ u_{i_0} - x_{i_0}, \ if \quad \widehat{E}_{i_0}(x) < 0, \end{cases}$$

where $d(J_N \backslash i_0) = 0$, $d(J_S) = -D_S^{-1}D(J_S, i_0)d_{i_0}$ while excluding $d_{i_0} = -sign\widehat{E}_{i_0}$.
Then

$$\theta_{i_0} = \theta_N = \begin{cases} \frac{\ell_{i_0} - x_{i_0}}{d_{i_0}} = 1, \ if \quad \widehat{E}_{i_0}(x) > 0, \\ \frac{u_{i_0} - x_{i_0}}{d_{i_0}} = 1, \ if \quad \widehat{E}_{i_0}(x) < 0, \end{cases} \tag{23}$$

and $\theta_S = \theta_{i_1} = \min \theta_i,\ i \in J_S$, where

$$\theta_i = \begin{cases} \frac{u_i - x_i}{d_i},\ if\ d_i > 0, \\ \frac{\ell_i - x_i}{d_i},\ if\ d_i < 0, \\ \infty, \quad else. \end{cases} \tag{24}$$

Now, we define the step θ_F which must verify the condition (b). So θ_F is defined as the global solution of the following one-dimensional minimization problem

$$\min_{\theta \in [0;1]} \phi(\theta) = \min_{\theta \in [0;1]} F(x + \theta d) \tag{25}$$

The objective function ϕ is quadratic, the solution θ can easily be computed by

$$\phi'(\theta) = (Dx + c)^t d + \theta d^t D d = 0 \Leftrightarrow E_N^t d_N + \theta d^t D d = 0.$$

Then

$$\theta = -\frac{E_{i_0} d_{i_0}}{d^t D d}.$$

We will take θ_F as follows

$$\theta_F = \begin{cases} min\{1, -\frac{E_{i_0} d_{i_0}}{d^t D d}\},\ if \quad d^t D d > 0, \\ 1, \qquad\qquad\qquad\quad if \quad d^t D d \leq 0. \end{cases} \tag{26}$$

So, we will select $\theta = min(\theta_{i_0}, \theta_S, \theta_F)$ and consider the following three cases:
Case 1: if $\theta = \theta_{i_0} = \theta_N$, we have

$$\overline{x}_{i_0} = \begin{cases} \ell_{i_0}\ if \quad \widehat{E}_{i_0}(x) > 0, \\ u_{i_0}\ if \quad \widehat{E}_{i_0}(x) < 0. \end{cases} \tag{27}$$

Case 2: if $\theta = \theta_S = \theta_{i_1}$, we have

$$\overline{x}_{i_1} = \begin{cases} \ell_{i_1}\ if \quad d_{i_1} < 0, \\ u_{i_1}\ if \quad d_{i_1} > 0, i_1 \in J_S. \end{cases} \tag{28}$$

We exclude the index i_1 from J_S and we set $\overline{J_S} = J_S \setminus i_1$.
Case 3: if $\theta = \theta_F$:
We set $\overline{J_S} = J_S \cup i_0$. The matrix $\overline{D}_S = D(\overline{J_S}, \overline{J_S})$ is obtained while adding to the matrix D_S the line $D(i_0, \overline{J_S})$ and the column $D(\overline{J_S}, i_0)$, and $\overline{D}_S \succ 0$ must hold.

6.1 Algorithm

Begin

(0.) Generate x the initial FS of (QP), such that $\ell \leq x \leq u$, then calculate $E(x)$.
* Calculate $\widehat{\alpha}(J)$ and set $\widehat{\alpha}_i = 0$ if $E_i = 0,\ i \in J$ (optimality criterion).
* Set $\overline{J}_0 = \{i \in J :\ \widehat{\alpha}_i = 0\}$ then select $J_S \subset \overline{J}_0$ verifying $D_S = D(J_S, J_S) \succ 0$ and $E(J_S) = 0$ ($\{x, J_S\}$ is consistent). Or set $J_S \subseteq J_0 \cap \overline{J}_0$ with $D_S \succ 0$.

(1.) Let $\{x, J_S\}$ be the initial SFS of (QP), with $E(J_S) = 0$ and set $J_N = J \setminus J_S$.
* Set $\widehat{E}$ the vector of estimations given by the relations (12),with $\widehat{E}(J_S) = 0$.
* If $\{x, J_S\}$ is optimal, then go to **End**.
* Else determine the index $i_0 \in J_N$ by the relation (20) for which the support optimality criterion is not verified.
* Set $D_S = D(J_S, J_S)$, and determinate the subscript i_0 then calculate

$$d_{i_0} = \begin{cases} \ell_{i_0} - x_{i_0}, \ if \ \ \widehat{E}_{i_0}(x) > 0, \\ u_{i_0} - x_{i_0}, \ if \ \ \widehat{E}_{i_0}(x) < 0, \end{cases}$$

with $d(J_N \setminus i_0) = 0$, $d(J_S) = -D_S^{-1} D(J_S, i_0) d_{i_0}$.
* Calculate:

$$\theta_i = \begin{cases} \frac{u_i - x_i}{d_i}, \ if \ d_i > 0, \\ \frac{\ell_i - x_i}{d_i}, \ if \ d_i < 0, \\ \infty, \quad else, \ i \in J_S. \end{cases}$$

$$\theta_S = \theta_{i_1} = \min \theta_i, \ i \in J_S.$$

$$\theta_F = \begin{cases} min\{1, -\frac{E_{i_0} d_{i_0}}{d^t D d}\}, \ if \ \ d^t D d > 0, \\ 1, \qquad\qquad\qquad if \ \ d^t D d \leq 0. \end{cases}$$

$$\theta_{i_0} = \theta_N = 1 (\theta_{i_0} = 1 \ is fixed).$$

$$\theta = min(1, \theta_S, \theta_F), \quad \overline{x} = x + \theta d.$$

(2.) **If** $\theta = \theta_{i_0}$ then $\overline{J_S} = J_S$, **end if.**
(3.) **If** $\theta = \theta_S = \theta_{i_1}$ then $\overline{J_S} = J_S \setminus i_1$, **end if.**
(4.) **If** $\theta = \theta_{\overline{F}}$ then $\overline{J_S} = J_S \cup i_0$, **end if.**
(5.) $\{x, J_S\} = \{\overline{x}, \overline{J_S}\}$, **go to (1.)**.

End.

7 Numerical Examples

7.1 Example 1

Consider the following problem :

$$(EP1) \qquad \min F(x) = \frac{1}{2} x^t D x + c^t x, \tag{29}$$

$$s.t \quad \ell_i \leq x_i \leq u_i, \qquad i = \overline{1,3}, \tag{30}$$

where : $D = \begin{pmatrix} -1 & -1 & -2 \\ -1 & -1 & -2 \\ -2 & -2 & -1 \end{pmatrix}$, $c = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$, $\ell = \begin{pmatrix} -1 \\ 0 \\ -1 \end{pmatrix}$, $u = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$.
The matrix D is an indefinite matrix having one negative eigenvalue $\lambda_1 = -4.3723$, such that D is MPSubD, so the global minimum is a vertex.
Given: $\overline{D} = diag(-4, -4, -5)$. In this example, $J_S = \emptyset$, so $J_N = J = \{1, 2, 3\}$. For θ, we calculate only $\theta_{i_0} = \theta_N = 1$. So for $\rho = 1 \in [0, 1]$, then $Q = diag(-4, -4, -5)$ and $\widehat{Q} = diag(-4, -4, -5)$ is fixed during the resolution, because $J_S = \emptyset$. So, the iterations of the algorithm are given as follows:

Step (0): $k = 0$, $x^0 = (-1, 0, -1)^t$, $E(x^0) = (2, 3, 4)^t$. So $\widehat{E}(x^0) = (-2, -1, -1)^t$, so x^0 doesn't verify the global optimality criterion but it verifies the local optimality conditions, then x^0 is a local minimum. In this case, our algorithm choses a feasible direction d. So $i_0 = 1 \in J_N$, $d^0(J_N) = (+2, 0, 0)^t$, $\theta_0 = \theta_{i_0} = 1$, $k = k + 1$, $x^1 = x^0 + \theta_0 d^0$.

Step (1): $k = 1$, $x^1 = (1, 0, -1)^t$, $E(x^1) = (0, 1, 0)^t$, so $\widehat{E}(x^1) = (-4, 5, 5)^t$, then x^1 doesn't verify the global optimality criterion. After computation given: $i_0 = 2 \in J_N$, $d^1 = d^1(J_N) = (0, +2, 0)^t$, $\theta_1 = \theta_{i_0} = 1$, $k = k + 1$, $x^2 = x^1 + \theta_1 d^1$.

Step (2): $k = 2$, $x^2 = (1, 2, -1)^t$, $E(x^2) = (-2, -1, -4)^t$, so $\widehat{E}(x^2) = (2, 3, -9)^t$, then x^2 doesn't verify the global optimality criterion. After computation given: $i_0 = 3 \in J_N$, $d^2 = d^2(J_N) = (0, 0, +2)^t$, $\theta_2 = \theta_{i_0} = 1$, $k = k + 1$, $x^3 = x^2 + \theta_2 d^2$.

Step (3): $k = 3$, $x^3 = (1, 2, 1)^t$, $E(x^3) = (-6, -5, -6)^t$, so $\widehat{E}(x^3) = (-2, -1, -1)^t$. Since x^3 verify the global optimality criterion, then it is an optimal solution.

The global optimal solution founded for the problem $(EP1)$ is $x^* = (1, 2, 1)^t$, the minimal value of the objective function is $F(x^*) = -22$.

7.2 Example 2

Consider the following problem :

$$(EP2) \qquad \min F(x) = \frac{1}{2} x^t D x + c^t x, \tag{31}$$

$$s.t \quad \ell_i \leq x_i \leq u_i, \qquad i = \overline{1,4}, \tag{32}$$

where : $D = \begin{pmatrix} 2 & -2 & -1 & 0 \\ -2 & 2 & -1 & 1 \\ -1 & -1 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{pmatrix}$, $c = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, $\ell = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}$, $u = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$.

The matrix D is indefinite, having one negative eigenvalue: $\lambda_1 = -0.8758$. So, given $\overline{D} = diag(-1, -2, 0, 1)$. For $\rho = \frac{1}{2} \in [0, 1]$, then
$Q = diag(-0.9379, -1.4379, -0.4379, 0.0621)$ and $\widehat{Q} = diag(-0.9379, -1.4379, -0.4379, 0)$.
So, the iterations of the algorithm are given as follows:

Step (0): $k = 0$, $x^0 = (1, 0, 0, 0)^t$, $E(x^0) = (1, -2, -1, 0)^t$, then $J_S = \{4\}$, $J_N = \{1, 2, 3\}$, $\widehat{\alpha} = (-0.9379, -1.4379, -0.4379, 0)^t$, $\widehat{E}(x^0) = (1.89379, 6.0676, 1.1918, 0)^t$.
So x^0 doesn't verify the global optimality criterion.
After computation given: $i_0 = 2 \in J_N$, $d^0(J_N) = (0, +1, 0)^t$, $d^0(J_S) = (-0.5)$, $d^0(J) = (0, +1, 0, -0.5)^t$, $\theta_{i_0} = 1$, $\theta_S = 1$, $\theta_{\overline{F}} = 1$ then $\theta_0 = min\{1; 1; 1\} = 1$, $k = k + 1$, $x^1 = x^0 + \theta_0 d^0$.

Step (1): $k = 1$, $x^1 = (1, 1, 0, -0.5)^t$, $E(x^1) = (-1, -0.5, -2, 0)^t$, then $J_S = \{4\}$, $J_N = \{1, 2, 3\}$, $\widehat{\alpha} = (-0.9379, -1.4379, -0.4379, 0)^t$, $\widehat{E}(x^1) = (-0.0621, 0.9379, 4.1918, 0)^t$.
So x^1 doesn't verify the global optimality criterion. After computation given: $i_0 = 3 \in J_N$, $d^1(J_N) = (0, 0, +1)^t$, $d^1(J_S) = (0)$, $d^1(J) = (0, 0, +1, 0)^t$, $\theta_{i_0} = 1$, $\theta_S = +\infty$, $\theta_{\overline{F}} = 1$ then $\theta_1 = min\{1; +\infty; 1\} = 1$, $k = k + 1$, $x^2 = x^1 + \theta_1 d^1$.

Step (2): $k = 2, x^2 = (1, 1, 1, -0.5)^t$, $E(x^2) = (-2, -1.5, 0, 0)^t$, then $J_S = \{3, 4\}$, $J_N = \{1, 2\}$, $\widehat{\alpha} = (-0.9379, -1.4379, 0, 0)^t$, $\widehat{E}(x^2) = (-1.0621, -0.0621, 0, 0)^t$. So the global optimality criterion is verified, then x^2 is an optimal solution.

The global optimal solution founded for a problem $(EP2)$ is $x^* = (1, 1, 1, -0.5)^t$, the minimal value of the objective function is $F(x^*) = -2.2500$.

8 Conclusion

We have considered an indefinite quadratic problem with box constraints, where the corresponding matrix has one negative eigenvalue. In particular, when the matrix D is merely positive subdefinite, we have proved that the global minimum is an extreme point. We have developed a new support method for solving the nonconvex problems while investigating the support of the objective function. We have presented the algorithm which can find a global minimum, while starting by an initial SFS. So, if the global optimality criterion is verified, then the SFS is optimal, else we generate an other SFS.

References

1. P. M. Pardalos : Global optimization algorithms for linearly constrained indefinite quadratic problems. *Computers Math Applic.* 21, 87–97 (1991)
2. S. R. Mohan, S. K. Neogy and A. K. Das : More on positive subdefinite matrices and the linear complementarity problem. *Linear Algebra and its Applications.* 338, 275-285 (2001)
3. V. Jeyakumar, A. M. Rubinov and Z. Y. Wu : Sufficient global optimality conditions for nonconvex quadratic minimization problems with box constraints. *Journal of Global Optimization.* 36, 471-481 (2006)
4. Z. Y. Wu and A. M. Rubinov : Global Optimality Conditions for Some Classes of Optimization Problems. *Journal of Optimization Theory and Applications.* 145, 164-185, (2010)
5. R. Gabasov, F. M. Kirillova, O. I. Kostyukova and V. M. Raketsky : Constructive methods of optimization. *Part 4 : Convex Problems. University Press, Minsk.* (1987)
6. J. A. Ferland : Positive Subdefinite Matrices. *Linear Algebra and its Applications.* 31, 233-244 (1980)
7. B. Brahmi and M. O. Bibi: Dual Support method for Solving convex quadratic programs. *Optimization.* 59, 851-872 (2010)
8. E. A. Kostina, and O. I. Kostyukova: An algorithm for solving quadratic programming problems with linear equality and inequality constraints. *Computational Mathematics and Mathematical Physics.* 41, 960-973 (2001)
9. Vandenbussche, D., and Nemhauser, G. L.: A branch-and-cut algorithm for nonconvex quadratic programs with box constraints. *Math Program.* 102, 559-575 (2005)
10. P. M. Pardalos and S. A. Vavasis: Quadratic programming with one negative eigenvalue is NP-hard. *Journal of Global optimization.* 1, 15-22 (1991)
11. L. Fernandes, A. Fischer, J. Judice, C. Requejo and J. Soares : A block active set algorithm for large-scale quadratic programing with box contraints, *Annals of Operations Research.* 18, 75-95 (1998)
12. B. Martos: Subdefinite matrices and quadratic forms, *SIAM J. Appl. Math.* 17, 1215-1223 (1969)

Optimal control strategy of an SIR epidemic model

Abderrahmene AKKOUCHE[1,2], Sarah Grib[2], Lydia Dehbi[2], and Mohamed AIDENE[2]

[1] Département de Mathématiques, Faculté des Sciences et des Sciences Appliquées, Université AKLI MOHAND OULHADJ de Bouira, 10 000 Bouira, Algérie.
akkouche.abdo@yahoo.fr

[2] Laboratoire de Conception et Conduite des Systèmes de Production Université MOULOUD MAMMERI de Tizi-Ouzou, 15 000 Tizi-Ouzou, Algérie
sarahgrib93@gmail.com, lydiadehbi@gmail.com, aidene_2000@yahoo.fr

Abstract. In this paper, we consider the spread of a non fatal disease in a population given by a nonlinear susceptible-infected-recovered (SIR) epidemic model, which describes the interaction between susceptible and infected individuals in a community, for which we seek to determine the best optimal strategy to reduce the number of susceptible and infected individuals and to increase the number of recovered individuals. In order to do this, we use the optimal control theory to introduce two optimal control problems. In the first optimal control problem, we introduce one control function, vaccination of susceptible individuals. In the second one, we introduce two optimal control functions, vaccination of susceptible individuals and treatment of infected individuals. In order to obtain the solution of the optimal control problems, we utilise the minimum principle of Pontryagin to derive the optimality conditions to be solved afterwards by the shooting method.
Keywords: SIR epidemic model, optimal control, Minimum principle of Pontryagin, ordinary differential equations.

1 Introduction

Several mathematical epidemic models are used to describe and study the spread of an infectious disease through a community, such as acquired immunodeficiency syndrome (AIDS) [9], Cholera [6], Pest [13], Lyme disease [8], Hantavirus [1], Hepatitis C virus [3],...

These mathematical models provide several significant advantages. First, a mathematical model can help us to understand how a transmission of the disease can occur and how a single infective may spread the disease by contact with others. Second, formulating epidemic models that adequately describe communicable disease data is that the model provides a convenient summary of the data, and an adequate model helps to isolate the more important features of diseases spread. Another more important reason, is that such models can help to provide insight into the biological and sociological mechanisms underlying the process of

disease spread [2]. And a very interesting issue in the study of epidemic models is to identify therapeutic strategies that minimize the relevant negative features of the disease. This naturally leads to the formulation and investigation of optimal control problems [7].

Mathematical models of epidemic consist of a system of differential equations governing the dynamics of the relevant state variables (susceptible, infective, recovered, etc.)[7]. In a susceptible-infected-recovered epidemic model, the population is divided into a set of three distinct states. These states are labelled as S, I and R. Where 'S' denotes the susceptible individuals who are never been infected, but can catch the disease as a result of an adequately close contact with an infect. Once they have it, they move into the infected state. 'I' denotes infected individuals who are infected and can spread the disease to susceptible individuals. The time that individuals spend in the infected state is the infectious period; after, they enter the recovered state. 'R' denotes the recovered individuals who are the infected individuals passed through a stage during which they are latent and then through a stage during which he is infectious, before being resolved by isolation, by death or by naturally losing his infectiousness and becoming immune for the remaining duration of the epidemic [11,12].

In this paper, we consider the classical SIR model for the spread of an infectious disease introduced by Kermack and Mckendrick [5], for which we seek to adopt the best strategy, formulated as optimal control problem, to eradicate the infection successfully. For this purpose, two control strategies are proposed: the first one is the vaccination strategy of the susceptible individuals, and the second one is the vaccination of the susceptible and the treatment of infected individuals. To do this, for the first control strategy, we introduce a control variable that represents the fraction of susceptible to be vaccinated. For the second strategy we introduce two control variables, one represents the percentage of the susceptible individuals to be vaccinated and the other control variable represents the fraction of the infected individual to be treated. Hence, the optimal control strategy is to minimize the infected and susceptible individuals, and to maximize the total number of recovered individuals.

The rest of this paper is organised as follows. in Section 2, we present the SIR epidemic model to be studied. In Section 3, we introduce the optimal control techniques, vaccination and treatment. Then we solve numerically, using the shooting method, the optimality system obtained by the application of the minimum principle of Pontryagin. Finally, we conclude by discussing the obtained results of the numerical simulation for the proposed control strategies of the studied epidemic model.

2 SIR epidemic model

Consider the spread of a non-fatal disease in a population which is assumed to have constant size over a period of the epidemic [5]. Suppose that at time t the population consists of $x_1(t)$ susceptible, $x_2(t)$ infected and $x_3(t)$ recovered individuals. Assume that there is a steady contact rate between susceptible and

infective and that a constant proportion of these contacts result in a transmission. Then in time δt, $\delta x_1(t)$ of the susceptible become infective, where

$$\delta x_1(t) = -\beta\, x_1(t)\, x_2(t)\, \delta t, \tag{1}$$

and β is a positive constant. If $\gamma > 0$ is the rate at which current infectives become recovered, then

$$\delta x_2(t) = [\beta\, x_1(t)\, x_2(t) - \gamma\, x_2(t)]\ \delta t. \tag{2}$$

The number of new recovered $\delta x_3(t)$ is given by

$$\delta x_3(t) = \gamma\, x_2(t)\, \delta t. \tag{3}$$

Now let $\delta t \to 0$, then the SIR system is given as :

$$\begin{cases} \dot{x}_1(t) = -\beta\, x_1(t)\, x_2(t), \\ \dot{x}_2(t) = \beta\, x_1(t)\, x_2(t) - \gamma\, x_2(t), \\ \dot{x}_3(t) = \gamma\, x_2(t), \end{cases} \tag{4}$$

with initial conditions

$$x_1(0) = N_1 \geq 0,\ x_2(0) = N_2 \geq 0,\ x_3(0) = N_3 \geq 0. \tag{5}$$

For numerical results of the system (4), we use the following initial values, and the model parameters value that are given in the Table 1, and the obtained results are depicted in Fig. 1, which show that there is a few individuals that are recovered, and the numbers of susceptible decrease but the infected individuals increase.

$N_1 = 20$	Initial population of $x_1(t)$, who are susceptible
$N_2 = 15$	Initial population of $x_2(t)$, who are infective
$N_3 = 10$	Initial population of $x_3(t)$, who are recovered
$\beta = 0.01$	Rate of change of susceptible to infective population
$\gamma = 0.02$	Rate of change of infective to immune population

Table 1: Initial conditions and model parameter values

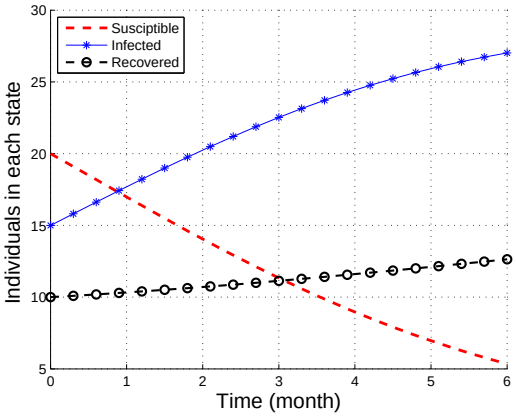


Fig. 1: Plot of SIR epedimic model without control

in the next section, we introduce the optimal control problem in order to try eradicate the epidemic in the population.

3 Optimal control problem

In this section, we suggest to develop an effective strategy to control the spread of the non-fatal infectious disease in a population considered in this paper. We seek to reduce the number of susceptible and infected individuals and increase the number of recovered individuals. First, we control the susceptible individuals by developing an optimal vaccination strategy, and the second strategy is to control both susceptible and infected individuals using vaccination and treatment.

3.1 The Optimal Vaccination Strategy

To formulate the optimal control problem that describe the vaccination strategy, we introduce into the model (4) a control variable $u(t) \in U_{ad}$ to be the fraction of susceptible individuals to be vaccinated to protect against possible infection per unit of time. The objective is to minimize the number of susceptible and infected individuals and the cost of applying the control $u(t)$. Hence the optimal control problem can be stated as :

$$\min_{u(t)} J(u(t)) = \int_0^{t_f} (x_1(t) + x_2(t) + u^2(t))dt, \tag{6}$$

$$\dot{x}_1(t) = -\beta\, x_1(t)\, x_2(t) - x_1(t)\, u(t), \tag{7}$$

$$\dot{x}_2(t) = \beta\, x_1(t)\, x_2(t) - \gamma\, x_2(t), \tag{8}$$

$$\dot{x}_3(t) = \gamma\, x_2(t) + x_1(t)\, u(t), \tag{9}$$

where $U_{ad} = \{u \text{ such that } u(t) \text{ is measurable, } 0 \leq u(t) \leq 1,\ t \in [0,\ t_f]\}$ is the set of admissible controls, and Let $\mathcal{X}$ be the set of reachable state. The optimal control problem consists of determining the vector function $x = (x_1^*,\ x_2^*,\ x_3^*) \in \mathcal{X}$ associated with the admissible control $u^*(t) \in U_{ad}$ on the time interval $[0,\ t_f]$, minimizing the cost functional (6), i.e.,

$$J(x^*(t), u^*(t)) = \min_{(x(t),\, u(t)) \in \mathcal{X} \times U_{ad}} J(x(t),\, u(t)). \tag{10}$$

Existence of optimal control According to [4], we give the conditions that ensure the existence of the solution of the optimal control problem (6)–(9). Let make the following assumptions:

- **H1.** The set of controls and corresponding state variables is non empty.
- **H2.** The admissible control set U_{ad} is closed and convex.
- **H3.** Each right hand side of equations (7)-(9) is continuously differentiable.
- **H4.** There exist a constant $\rho > 1$, $w1 > 0$ and $w_2 > 0$ such that the objective functional is convex on control u and satisfies :

$$J(u(t)) \geq w_2 + w_1 \left(\mid u(t) \mid^2 \right)^{\rho/2}. \tag{11}$$

Under these assumptions, we have the following theorem :

Theorem 1. *If the hypotheses **H1**- **H4** are satisfied, then there exists an optimal control $u^*(t)$ such that*

$$J(u^*(t)) = \min_{u(t) \in U} J(u(t)), \tag{12}$$

subject to the control system (6)-(9) with the initial conditions (5).

Proof. [11]

Necessary optimality conditions To derive the necessary optimality conditions for the optimal control problem $(6) - (9)$, we apply the minimum principle of Pontruagin [10]. Let $\mathcal{H}(x(t),\, p(t),\, u(t))$ be the Hamilton function defined as :

$$\begin{aligned} \mathcal{H}(x_1(t), x_2(t), x_3(t), p_1(t), p_2(t), p_3(t), u(t)) &= x_1(t) + x_2(t) + u^2(t) \\ &+ p_1(t)\ (-\beta\, x_1(t)\, x_2(t) - x_1(t)\, u(t)), \\ &+ p_2(t)\ (\beta\, x_1(t)\, x_2(t) - \gamma\, x_2(t)) \\ &+ p_3(t)\ (\gamma\, x_2(t) + x_1(t)\, u(t)) \end{aligned} \tag{13}$$

where $p(t) = (p_1(t), p_2(t), p_3(t))$ is called the adjoint vector, and the necessary optimality conditions are :

$$\dot{x}_1(t) = -\beta\, x_1(t)\, x_2(t) - 0.5\, x_1(t)^2\, (p_1(t) - p_3(t)), \tag{14}$$
$$\dot{x}_2(t) = \beta\, x_1(t)\, x_2(t) - \gamma\, x_2(t), \tag{15}$$
$$\dot{x}_3(t) = \gamma\, x_2(t) + 0.5\, x_1(t)^2\, (p_1(t) - p_3(t)), \tag{16}$$
$$\dot{p}_1(t) = -1 + \beta\, x_2(t)\, p_1(t) + 0.5\, x_1(t)\, (p_1(t) - p_3(t))\, p_1(t) - \beta\, x_2(t)\, p_2(t) - 0.5\, x_1(t)\, (p_1(t) - p_3(t))\, p_3(t), \tag{17}$$
$$\dot{p}_2(t) = -1 + \beta\, x_1(t)\, p_1(t) - \beta\, x_1(t)\, p_2(t) + \gamma\, p_2(t) - \gamma\, p_3(t), \tag{18}$$
$$\dot{p}_3(t) = 0. \tag{19}$$

with the boundary conditions :

$$x(0) = x_0, \quad p(t_f) = 0, \tag{20}$$

and the optimal control $u(t)$ is given as :

$$u(t) = \max\left\{0, \min\left\{1, \frac{x_1(t)\, (p_1(t) - p_3(t))}{2}\right\}\right\}, \tag{21}$$

Numerical simulation To solve the optimality system (14)–(19) obtained by the application of the minimum principle of Pontryagin, which results in a six coupled ordinary differential equations, we use the shooting method and the Runge-Kutta forth order procedure. for the shooting method is used to determine the initial values for the adjoint system, and the Runge-Kutta forth order is used to determine an approximate solution for the differential system.

In Fig.2–Fig.4, we have plotted the solution curves for the susceptible, infected and recovered individuals of the two systems (4) and (14)-(19), respectively without control and with control cases.

Using the vaccination of susceptible individuals strategy, we see in Fig.2 that the number of susceptible individual decrease significantly from the first time of application of vaccine. In Fig.4 we observe that the population of the recovered increases. while in Fig. 3 the infected individuals decline shortly. In Fig.5, we have plotted the optimal control that represents the vaccination of susceptible individuals.

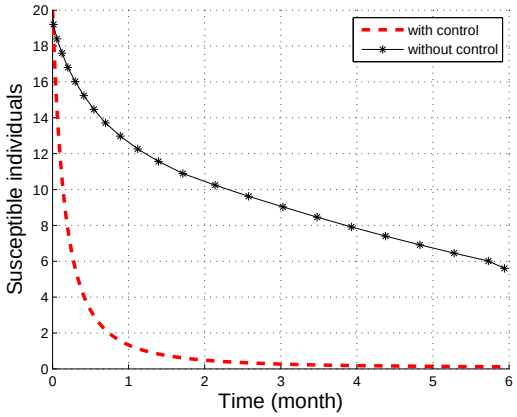


Fig. 2: Susceptible individuals without and with control

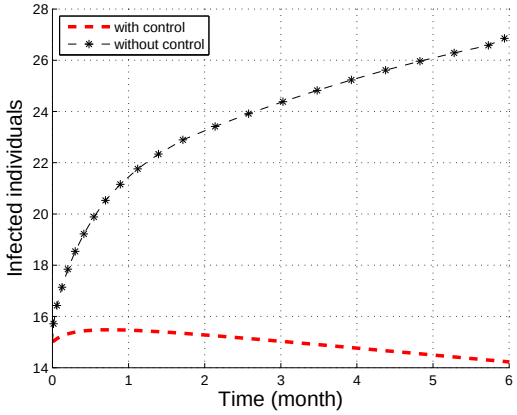


Fig. 3: Infected individuals without and with control

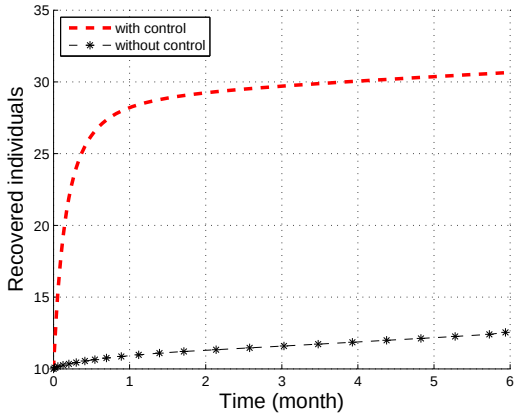


Fig. 4: Recovered individuals without and with control

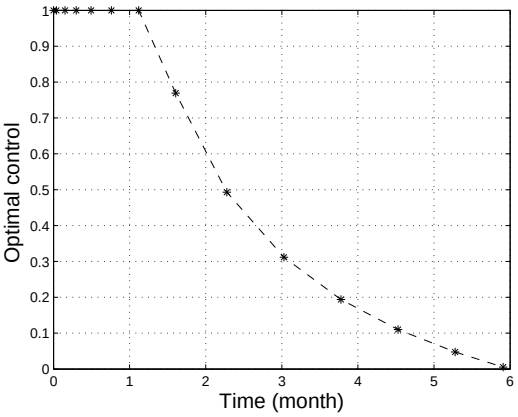


Fig. 5: Optimal control (Vaccination)

3.2 Optimal control vaccination and treatment strategy

In the first strategy, we have used only one control variable, which represents the vaccination of susceptible individuals. However, using only the vaccination of susceptible individuals may be difficult to eradicate the infection successfully, so we adopt a new strategy by taking into account an other control variable which is the treatment of the infected individuals, and the optimal control problem can

be formulated as follows :

$$\min_{u(t)} J(u(t)) = \int_0^{t_f} (x_1(t) + x_2(t) + u_1^2(t) + u_2^2(t))dt, \tag{22}$$

$$\dot{x}_1(t) = -\beta\, x_1(t)\, x_2(t) - x_1(t)\, u_1(t), \tag{23}$$

$$\dot{x}_2(t) = \beta\, x_1(t)\, x_2(t) - \gamma\, x_2(t) - x_2(t)\, u_2(t), \tag{24}$$

$$\dot{x}_3(t) = \gamma\, x_2(t) + x_1(t)\, u_1(t) + x_2(t)\, u_2(t), \tag{25}$$

Here $U_{ad} = \{ (u_1(t), u_2(t))$, such that $u_1(t)$ and $u_2(t)$ are measurables with $0 \leq u_1(t) \leq 1$ and $0 \leq u_2(t) \leq 1$ for $t \in [0,\ t_f] \}$ is the control set.

Existence of optimal control Let us give the following hypotheses to be satisfied.

- **H1.** The set of controls and corresponding state variables is non empty.
- **H2.** The admissible control set U_{ad} is closed and convex.
- **H3.** Each right hand side of equations (23)-(25) is continuously differentiable.
- **H4.** There exist constants $\rho > 1$, $w1 > 0$ and $w_2 > 0$ such that the objective functional is convex on control u and satisfies :

$$J(u(t)) \geq w_2 + w_1 \left(\mid u_1(t) \mid^2 + \mid u_2(t) \mid^2 \right)^{\rho/2}. \tag{26}$$

Under these hypotheses, we have the following theorem.

Theorem 2. *If the hypotheses **H1-H4** are satisfied, then there exists an optimal control $u^*(t)$ such that*

$$J(u_1^*(t), u_2^*(t)) = \min_{(u_1(t),\, u_2(t)) \in U} J(u_1(t),\, u_2(t)),, \tag{27}$$

subject to the control system (22)-(25) with the initial conditions (5).

Proof. [12]

Numerical simulation To obtain the optimal control strategy for the optimal control problem (22)–(25), we apply the minimum principle of Pontryagin, and the obtained results are depicted in Fig.6–Fig.8. In Fig.6 we compare the solution curves for the susceptible individuals for different strategies, no control, with one control and with two controls cases. Likewise, in Fig.7 we intend to compare the evolution of the infected population in the case with no control, with one control and in the presence of two controls.

We see that when the vaccination and treatment is applied, the number of susceptible and infected decrease sharply from the first time. So, we see that the later strategy gives more number of recovered than the other strategy as shown in Fig.8.

The optimal controls, vaccination and treatment, are represented in Fig.9, which show that the vaccination and treatment are given to everyone in the first 40 days to eradicate the infectious disease.

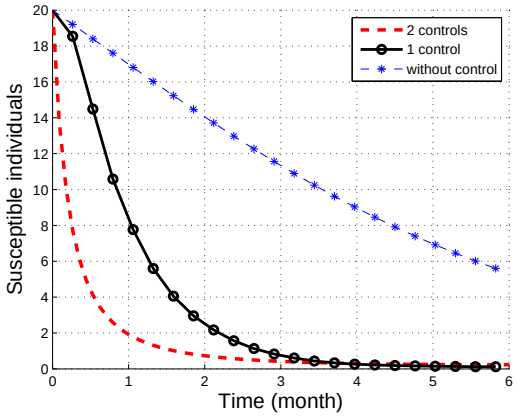


Fig. 6: Susceptible individuals without and with control

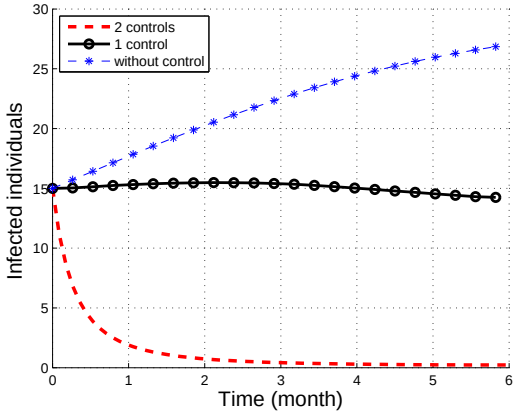


Fig. 7: Infected individuals without and with control

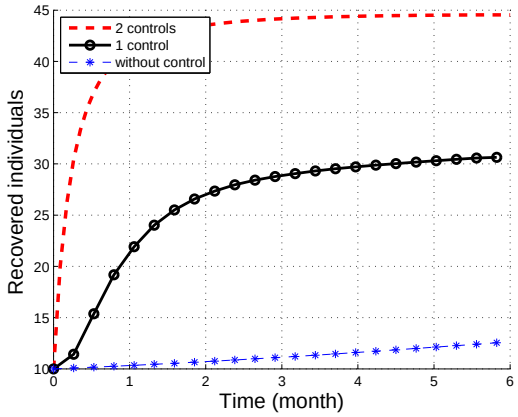


Fig. 8: Recovered individuals without and with control

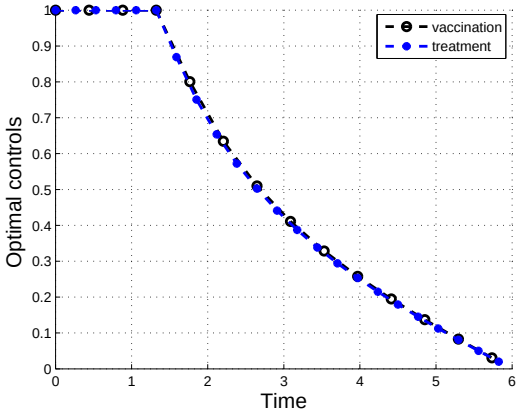


Fig. 9: Optimal controls (vaccination and treatment)

4 Conclusion

In this work, we have used the optimal control theory to determine the best strategy that minimizes the susceptible and infected individuals and maximizes the number of recovered individuals of an infectious disease described by an SIR epidemic model. To do this, we have adopted two strategies : the first one is the vaccination of the susceptible individuals, and the second one is the vaccination and treatment of both susceptible and infected individuals. The optimal control

strategy is obtained by solving a two-point-boundary value problem constituted by nonlinear coupled ordinary differential equations derived from the minimum principle of Pontryagin.

The numerical results reveal that using the optimal control vaccination strategy permit to decrease the number of susceptible individuals and increase the number of the recovered individuals. While, the second strategy, vaccination and treatment, is enable to decrease sharply the size of the susceptible and infected individuals and gives more number of recovered individuals than the first strategy. Thus, the second strategy is more efficient than the first one to eradicate the infectious disease of the population.

References

[1] L. J. S. Allen, R. K. McCormack and C. B. Jonsson: Mathematical Models for Hantavirus Infection in Rodents: Bulletin of Mathematical Biology: 68 (2006) 511-524.

[2] N. Becker: The Uses of Epidemic Models: Biometrics: 35 (1979) 295–305.

[3] A. Chatterjee, J. Guedj and A.S. Perelson: Mathematical modelling of HCV infection: what can it teach us in the era of direct-acting antiviral agents?: Antiviral Therapy: 17 (2012) 1171–1182.

[4] W.H. Fleming and R. W. Rishel: Deterministic and Stochastic Optimal Control: (1975) Springer, New York.

[5] W. Kermack and A. McKendrick : Contributions to the mathematical theory of epidemics: Proceeding of the Royal Society of London, Series A, 115 (1927) 700–721.

[6] A. P. Lemos-Paião, Cristiana J. Silva, Delfim F.M. Torres: An epidemic model for cholera with optimal control treatment: Journal of Computational and Applied Mathematics: 318 (2017) 168–180.

[7] A. D. Liddo: Optimal Control and Treatment of Infectious Diseases. The Case of Huge Treatment Costs: Mathematics: 21 (2016) 1–27.

[8] Y. Lou and J. Wu Modeling Lyme disease transmission: Infectious Disease Modelling: 2 (2017) 229–243.

[9] K.O. Okosun, O.D. Makinde and I. Takaidza: Impact of optimal control on the treatment of HIV/AIDS and screening of unaware infectives: Applied Mathematical Modelling: 37 (2013) 3802–3820.

[10] L.S. Pontryagin and V. G. Boltyanskii and R. V. Gamkrelidze and E. F. Mishchenko: The Mathematical Theory of Optimal Processes: Pergamon Press, (1964), New York.

[11] G. Zaman, Y. H. Kang, G. Cho and I. H. Jung: Stability analysis and optimal vaccination of an SIR epidemic model: BioSystems: 93 (2008) 240–249.

[12] G. Zaman, Y. H. Kang, G. Cho and I. H. Jung: Optimal strategy of vaccination and treatment in an SIR epidemic model: Mathematics and Computers in Simulation: 136 (2017) 63–77.

[13] H. Zhang, J. Jiao and L. Chen: Pest management through continuous and impulsive control strategies: BioSystems: 90 (2007) 350–361.

Une nouvelle méta-heuristique basée sur la recherche locale et le croisement pour le problème de positionnement d'antennes dans les réseaux cellulaires*

Benmezal Larbi[1,2], Benhamou Belaid[2] et Boughaci Dalila[1]

[1] LRIA, USTHB BP32 El-Alia, Bab-Ezzoaur, 16111, Algérie
larbi07@hotmail.fr, dboughaci@usthb.dz
[2] LIS, AMU Domaine universitaire de Saint Jérôme Avenue Escadrille Normandie Niemen 13397 Marseille cedex 20, France
belaid.benhamou@univ-amu.fr

Résumé. Le problème de positionnement d'antennes dans les réseaux cellulaires est un problème connu dans le domaine de la télécommunication. Il consiste à choisir parmi un ensemble de sites candidats, les meilleurs emplacements pour installer les stations de bases, de façon à maximiser la couverture réseau, tout en minimisant le nombre de stations employées. En théorie, le problème est NP-difficile. Pour le résoudre dans la pratique, nous proposons une nouvelle méthode basée sur la recherche locale et tire profit de quelques opérations utilisées dans les métaheuristiques évolutionnaires. Pour valider notre approche, nous avons testé notre algorithme sur une instance réelle du problème. Les résultats expérimentaux obtenus montrent que la méthode proposée améliore les performances de beaucoup de méthodes testées dans la littérature sur la même instance.

Mots-clé: : Réseaux cellulaires, Optimisation combinatoire, Métaheuristique, Recherche Locale, Croisement, Mutation.

1 Introduction

De nos jours, les moyens de télécommunication sont de plus en plus utilisés et s'impliquent presque dans tous les domaines. En ce qui concerne les réseaux cellulaire par exemple, beaucoup de nos données doivent passer par ces derniers avant d'arriver à leur destination. En effet, que se soit pour passer un simple appel ou bien pour se connecter à internet, avec la 3G ou bien la 4G, on doit toujours passer par ces réseaux. D'où il est primordiale que ces réseaux supportent cette demande et répondent à ce besoin incessant. Pour répondre à ces besoins, le réseau doit être performant, et pour qu'il soit ainsi, il doit être bien conçu. La planification des réseaux cellulaires est une étape clé dans leur conception. En effet, elle permet d'une part, d'assurer un réseau performant, et

* Ce travail est supporté par le projet PHC-Tassili

d'autre part, d'optimiser l'utilisation des ressources, généralement limitées. La planification des réseaux comprend deux principaux problèmes: le problème de positionnement d'antennes ou APP (Antenna Positionning Problem) [5][11] et le problème d'allocation de fréquences ou FAP (Frequency Assignment Problem) [6][9].

On s'intéresse dans ce papier au problème de positionnement d'antennes. Ce dernier englobe un ensemble de décisions concernant le déploiement des ressources physiques dans les réseaux cellulaires. Pour optimiser le réseau, il faudrait minimiser le nombre d'antennes utilisées tout en veillant aux performances du réseau. C'est à dire, installer un nombre minimal d'antennes à des emplacements judicieusement choisis de façon à garantir une couverture réseau maximale et un bruit minimal. Le problème peut s'étendre aussi à la spécification du type d'antennes et leur paramétrage (Tilt, Puissance, Azimut).

Le problème est de nature multi-objectif, c'est-à-dire, on cherche une, ou des solutions qui optimisent plusieurs objectifs en même temps. Par contre, il peut aussi être réduit à un problème mono-objectif. Dans ce travail nous traiterons seulement deux objectifs: maximiser la zone couverte par le réseau, et minimiser le nombre d'antennes employées. Ces deux objectifs représentent le centre d'intérêt du problème de positionnement d'antennes dans les réseaux cellulaires. La difficulté réside dans le fait que ces deux objectifs sont difficiles à satisfaire en même temps. En plus, le problème est caractérisé aussi par le nombre important de solutions possibles. Le problème à résoudre est un problème d'optimisation combinatoire NP-Difficile [17].

Plusieurs travaux de recherche ont abordé le problème et plusieurs méthodes ont été proposées pour le résoudre dans la pratique. Ces dernières sont généralement des méthodes approchées; c'est les plus adaptées à ce type de problème. Les méthodes exactes ne peuvent résoudre que des instances de petites tailles. On ne cherche pas à trouver une solution exacte, mais on cherche plutôt une solution de bonne qualité dans un temps raisonnable.

Dans ce travail, nous présentons une nouvelle méta-heuristiques pour résoudre le problème de positionnement d'antennes dans les réseaux cellulaires. Cette dernière combine les propriétés des méthodes de voisinage avec certaines propriétés des méthodes évolutionnaires. En effet, cette nouvelle méthode emploie la recherche locale et le croisement comme moyens d'intensification de la recherche. Son idée de base est de faire évoluer par le croisement plusieurs optima locaux générés par la recherche locale. Cela lui permet de bénéficier à la fois des caractéristiques des méthodes de voisinage et à la fois des méthodes évolutionnaires.

Ce papier est organisé comme suit : Dans la section suivante, nous présentons les différents travaux qui ont été réalisés sur le problème. Ensuite, nous présentons dans la section 3, une modélisation du problème basée sur les hypergraphes, proposée par Mendes et al. Dans la section 4, nous décrirons la nouvelle méthode que nous proposons. Nous réaliserons dans la section 5 une étude expérimentale où nous présenterons les résultats obtenus lorsque notre algorithme est appliqué sur trois instances du problème dont une qui est réelle. Les conclusions et perspectives sont données dans la dernière section.

2 Travaux connexes

Plusieurs travaux ont été réalisés dans le domaine. On peut citer comme exemple les travaux de Reininger [14] et Reininger et Caminada [15] qui ont proposée une modélisation au problème APP. Dans cette dernière, plusieurs fonctions qui représentent divers objectifs et contraintes du problème sont définies, comme la couverture réseau, le bruit, le trafic pris en charge, le nombre d'antennes utilisées et d'autres. Ces fonctions doivent être optimisées. Cette modélisation reflète bien la nature multi-objectif du problème. Plusieurs autres travaux ont utilisé cette modélisation pour traiter le problème, comme dans Meunier et al.[12], où ils ont utilisé un algorithme génétique multi-objectif lors de la phase de résolution, et dans Vasquez et al. [13] où ils ont proposé une approche en trois phases, basée sur la recherche taboue. Une première phase de prétraitement, vise à éliminer les solutions qui utilisent un nombre d'antennes, soit très petit, soit très grand. La deuxième phase est une phase de recherche, et la troisième a pour but le paramétrage des antennes. El-Ghazali et al. [7] se sont aussi basés sur cette modélisation pour proposer des modèles parallèles hiérarchiques. Ils ont montré que l'utilisation de ces derniers peut accélérer la recherche et permet aussi de trouver des solutions de très bonne qualité. Un algorithme génétique a été utilisé pour la résolution du problème.

La modélisation de Reininger est une modélisation qui reflète bien la nature complexe du problème traité, dans le sens où elle modélise plusieurs aspects. Néanmoins, la résolution d'instances basées sur cette modélisation peut être très difficile.

Une autre modélisation a été définie par Calegari et al. [5]. Contrairement à la modélisation de Reininger et al., cette modélisation fait abstraction à plusieurs aspects techniques du problème, comme les paramètres d'antennes et les stations mobiles. La zone est modélisée par un graphe biparti, où l'on cherche à trouver le plus petit ensemble dominant. Deux objectifs seulement sont pris en compte par cette modélisation : la maximisation de la couverture et la minimisation du nombre d'antennes employées. Ces deux objectifs sont représentés par une seule fonction objectif.

Plusieurs méthodes ont été appliquées sur des instances basées sur cette modélisation. On cite par exemple, l'algorithme génétique et l'algorithme CHC dans Alba et al. [1,2], et NSGA-II et MOCHC dans Nebro et al. [3]. Ces deux derniers sont une adaptation multi-objectif des algorithmes GA et CHC afin de leur permettre de traiter plusieurs fonctions objectif en même temps. Pour ça, Ils ont défini leurs propres fonctions à optimiser.

Mendes et al. ont proposé une nouvelle modélisation dans [11]. Cette dernière améliore celle de Calégari en proposant de modéliser la surface géographique par un hyper-graphe au lieu d'un graphe. Une large panoplie d'algorithmes a été testée dans ce travail sur une instance réelle du problème.

Dans ce papier on comparera les performances des algorithmes étudiés avec ceux testés dans le travail de Mendes et al., en les exécutant sur la même instance.

3 Modélisation du problème

Pour représenter le problème, nous utilisons la modélisation de Mendes, décrite dans [11]. Cette modélisation propose un modèle basé sur la théorie des graphes pour définir les différentes composantes du réseau.

3.1 Zone géographique

La zone géographique à couvrir est discrétisée pour avoir en résultat un ensemble fini d'emplacements. Chaque emplacement est représenté par un point caractérisé par ses coordonnés x et y. Parmi ces points, quelques uns sont choisis pour être des sites candidats à héberger des stations de bases. La zone est composée de deux types de points:

- Points de type L : ce sont les points qui doivent être couverts;
- Points de type M : les points qui représentent les sites candidats pour héberger les stations de base.

La zone est modélisée par un hypergraphe $H(V, E)$ tel que, V est l'ensemble des sommets de l'hypergraphe H, et E est une partie de $P(V)$ (l'ensemble des parties de l'ensemble V). On appelle les éléments de E les hyper-arêtes. Une hyper-arête généralise la notion d'arête, dans le sens où elle permet de relier plusieurs sommets au lieu de deux seulement. Dans notre cas, l'ensemble V contient tous les points de la zone géographique ; $V = M \cup L$, et E l'ensemble qui contient toutes les hyper-arêtes qui lient chaque point de type M avec les points couverts par ce dernier. Notre hypergraphe peut être vu comme un ensemble de graphes orientés. Chaque graphe possède un sommet central. Ce dernier est un point de type M (représente une station de base) associé avec des points de l'ensemble V représentant les points qui ce trouvent dans la cellule formée par cette station de base.

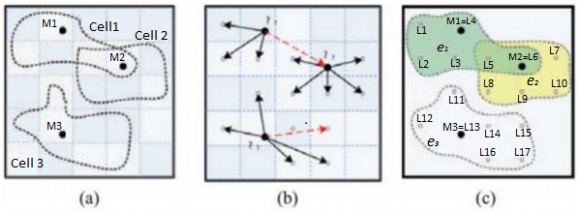


Fig. 1. Exemple de trois transmetteurs représentés par un hypergraphe

La figure 1 illustre un exemple de trois transmetteurs représentés avec leurs cellules associées (Fig.1.a). Ils ont été modélisés par un ensemble de graphes

qui associent les transmetteurs avec les points qu'ils couvrent (Fig.1.b). La représentation de l'hypergraphe associée est donnée dans la figure (Fig.1.c). On voit bien la ressemblance entre les cellules formées par les transmetteurs et leur modélisation par l'hypergraphe.

3.2 Antenne

Des antennes dites Isotropes sont utilisées dans cette modélisation ; ce sont des modèles théoriques qui rayonnent uniformément dans toutes les directions. Une seule antenne seulement de ce type est installée dans une station de base.

3.3 Fonction objectif

Les deux objectifs à optimiser sont combinés dans une seule fonction objectif notée $f(x)$. Cette dernière a été définie dans [5].

$$f(x) = \frac{Taux\_de\_couverture(x)^2}{Nombre\_de\_transmetteurs\_utilises(x)} \tag{1}$$

Avec:

$$Taux\_de\_couverture(x) = \frac{|Voisin(M', E')|}{|Voisin(M, E)|}, \tag{2}$$

tel que

$$voisins(M, E) = \{v \in V | \exists u \in M, (u, v) \in E\} \tag{3}$$

et M' est l'ensemble des transmetteurs utilisés par la solution x. E' est l'ensemble des hyper-arêtes qui lient les points de l'ensemble M' avec leurs points associés

Le problème à résoudre rappel le problème de couverture par ensembles (Set Covering Problem dans la littérature originale). Il consiste à trouver un sous ensemble de sommets de taille minimum qui couvre tout le graphe. Une seule différence réside dans le but des deux problèmes, là on ne cherche pas à assurer une couverture totale de la zone, mais on cherche plutôt à la maximiser. Le problème est connu pour être NP-Difficile

3.4 Solution

Une solution au problème indique pour chaque emplacement candidat, si une station de base est installée ou non. Cela peut être modélisé par un vecteur binaire, où chaque bit représente un emplacement candidat. Si une station de base est installée dans un emplacement, son bit associé est alors mis à 1, sinon il est mis à 0. De ce fait, La taille du vecteur solution est égale au nombre d'emplacements candidats.

4 Méthode proposée

Pour résoudre le problème de positionnement d'antennes décrit dans la section précédente, nous proposons une nouvelle méthode qui combine la recherche locale et le croisement spécifiques aux algorithmes évolutionnaires pour intensifier la recherche. Pour diversifier, une opération de mutation, qui est aussi spécifique aux algorithmes évolutionnaires, est utilisée.

Pour la recherche locale, nous définissons le voisinage d'une solution comme suit:

Définition: Un voisinage N est une fonction $N : S \rightarrow P(S)$, tel que S représente l'espace de recherche et $P(S)$ l'ensemble des parties de S. Le voisinage d'une solution $s \in S$ est l'ensemble des solutions $N(s)$ de $P(S)$ qui peuvent être obtenues en effectuant un seul mouvement à partir de s.

Nous avons défini deux mouvements guidés par deux propriétés des réseaux cellulaire. Le premier mouvement est guidé par une propriété qui dit qu'une bonne solution ne possède généralement pas beaucoup d'antennes installées dans un petit périmètre. Le deuxième mouvement est guidé par une autre propriété, celle-ci dit que chaque point dans la zone à couvrir doit être couvert par au moins une station de base.

- **Mouvement** $M1$**:** Un mouvement $M1$ à partir d'une solution $s \in S$ sélectionne une station de base non active pour l'activer, et désactive les autres stations localisées dans un périmètre de rayon r.
- **Mouvement** $M2$**:** Un mouvement $M2$ à partir d'une solution $s \in S$ sélectionne une station de base active pour la désactiver, et active une station choisie aléatoirement localisée dans un périmètre de rayon r.

4.1 Description de la nouvelle méthode

L'idée de base de notre méthode est d'effectuer des croisements sur des optima locaux résultants de plusieurs recherches locales. Ceci nous permettra de tirer profit du croisement pour avoir de nouvelles solutions de bonne qualité en permettant l'échange d'information entre les optima locaux pour avoir de nouvelles générations. D'une autre part, le fait que les individus soient issues de la recherche locale, ceci garanti en quelque sort la bonne qualité des individus de la population. En tout, la population est alimentée à la fois par la recherche locale et par le croisement.

Notre méthode effectue plusieurs fois des recherches locales pour parcourir plusieurs optima locaux. A chaque fois qu'un optimum local est généré, il est inséré dans une population d'individus. Ensuite, des opérations de croisements sont réalisées entre ses individus. Les fils générés par le croisement sont insérés dans la population. Une fois le croisement terminé, le processus réitère pour refaire une nouvelle recherche locale à partir d'une nouvelle solution choisie parmi les individus de la population. Le choix est guidée par une stratégie que nous décrivons dans la section 4.2.

Juste avant de ré-entamer le processus, la solution choisie subie une mutation pour apporter une diversification à la recherche et pour prévenir la convergence prématurée, c'est une façon aussi de s'échapper de l'attraction d'un optimum local, dans le cas où celui-ci est choisi comme nouveau point de départ pour la prochaine recherche locale.

Même si notre méthode rappel l'algorithme mémétique dans le sens où les deux combinent le croisement et la recherche locale dans leurs processus, notre méthode se distingue par son schéma et par sa façon d'utiliser ces opérations. En effet, contrairement à un algorithme mémétique qui effectue la recherche locale après le croisement, notre méthode effectue tout d'abord plusieurs recherches locales pour générer plusieurs optima locaux, ensuite les croisements sont effectués sur ces optima locaux.

Algorithm 1 Algorithme de la méthode proposée

Require: Une instance du problème
Ensure: La meilleure solution trouvée
1: Générer les individus(P)
2: $S0$=Choix (P)
3: **while** Condition d'arrêt non satisfaite **do**
4: S=RechercheLocale($S0$)
5: Insérer(P, S)/*insérer l'optimum local dans la population P*/
6: Croisement(P)
7: $S0$=Choix (P) /*choisir une nouvelle solution de départ parmi les solutions de P comme expliqué dans la section 4.2*/
8: Mutation($S0$)
9: **end while**

La population est de taille N à chaque instant t. A chaque fois qu'un nouveau individu entre dans la population, un autre sort. Dans ce travail, on a adopté une stratégie qui fait sortir à chaque fois l'individu qui a la plus mauvaise qualité. Toute fois, d'autre stratégies peuvent être adoptées à la place de cette dernière.

4.2 Stratégie de choix de la nouvelle solution de départ

Les individus S de P sont classés par ordre croissant par rapport à leur fonction objectif $f(S)$. Donc on a la propriété suivante : $f(S_{i+1}) > f(S_i)$. La nouvelle solution de départ est une solution qui appartient à P. Chaque individu S_i de P peut être choisi comme solution de départ suivant une probabilité $p(S_i)$. La stratégie de choix adoptée est de telle sorte à favoriser au début de la recherche les solutions qui sont au milieu du tableau, c'est à dire qui ont une fonction objectif moyenne. Ensuite, plus la recherche avance, plus on favorise les solutions qui sont en haut du tableau, c'est à dire les meilleurs solutions. Ceci est pour éviter la convergence prématurée de la recherche et pour l'intensifier au fur et à mesure qu'elle avance.

Formellement, cette stratégie est guidée par une variable aléatoire X qui suit une loi binomiale $X \to B(n,p)$, où n est le nombre de tentatives (la taille de la population dans notre cas) et p est la probabilité de succès (ça sera une variable qui va gérer le choix de la solution). La valeur de la variable X représente le nombre de succès. Dans notre cas, elle va représenter l'indice de la solution retenue. Par exemple, si $X = 0$ alors on choisi la solution qui a l'indice 0, et de façon général, si $X = i$, la solution S_i est choisie.

La probabilité de succès p: Lorsqu'il s'agit d'une variable aléatoire qui suit une loi binomiale de paramètre n et p, on sait que plus la probabilité de succès p est grande, plus la probabilité d'avoir un nombre de succès sur n tentatives devient grand. C'est cette propriété qui va nous permettre d'implémenter la stratégie décrite en haut. Au début de la recherche, on prend $p = p0$ ($p0$=0.5) de tel façon à ce que les probabilités $P(X \approx N/2)$ soient les plus grandes, c'est à dire favoriser le choix des solutions qui sont classées au milieu du tableau. Plus la recherche avance, plus le paramètre p est augmenté. Donc les probabilités les plus grandes sont pour X plus grand, c'est à dire favoriser le choix des solutions qui sont en haut du tableau.

la probabilité de choisir la solution $S_i = P(X = i)$.

$$P(X = i) = C_n^k p^k (1-p)^{n-k} \tag{4}$$

Où:

$$p = p_0 + (Niveau\_Recherche/(2 * (Niveau\_Recherche + t))) \tag{5}$$

et: *niveaure_cherche* est une valeur en pourcentage qui exprime le niveau de la recherche. t est un paramètre qui contrôle la rapidité de convergence de p.

Si p_0 est égal à 0.5 (comme dans notre cas), la formule de p tend vers 1 lorsque $Niveau\_Recherche$ tend vers l'infini. Ceci nous garanti que la probabilité p augmente au fur et à mesure que la recherche avance, et nous garanti aussi que p n'atteindra jamais la valeur 1.

Avant de réitérer et juste après le choix de la nouvelle solution de départ, une mutation est réalisée sur les éléments de la solution choisie avec une probabilité p_m. Dans cette nouvelle méthode la population P évolue grâce à deux concepts: premièrement, le croisement, deuxièmement, la recherche locale qui permet d'insérer à chaque fois un optimum local comme nouvel individu.

4.3 Réinitialisation de la recherche

Après un certain nombre d'itérations sans amélioration de la meilleure solution trouvée, la recherche est réinitialisée. Le processus redémarre à partir d'une nouvelle solution de départ générée aléatoirement et à partir d'une nouvelle population P.

Il est à souligner que la réinitialisation est déclenchée par rapport à la non amélioration de la meilleure solution trouvée depuis la dernière réinitialisation et non pas par rapport à la meilleure solution trouvée depuis le début de la

recherche. Ceci car le contraire risque d'introduire notre algorithme dans une série de réinitialisations sans amélioration, car la meilleure solution trouvée deviens de plus en plus bonne, donc rarement améliorée au fil du temps, ce qui cause ces réinitialisations infructueuses. Et vu que la réinitialisation est en quelque sort un moyen de diversification de la recherche, cela crée un déséquilibre entre l'intensification et la diversification. Donc, c'est pour ça qu'on se base plutôt sur la meilleure solution trouvée depuis la dernière réinitialisation.

5 Résultats expérimentaux

Afin d'évaluer la performance de notre nouvelle méthode, nous l'avons testé sur trois instances du problème dont une issue de données réelles. Cette dernière est une représentation de la zone géographique de la ville de Malaga (Fig.2). Cette zone à une surface de 27,2 km^2 discrétisée par une grille de taille 450 X 300. Ce qui donne au total 135000 points. Chaque point représente une surface de 15 X 15 m^2. Parmi ces points, 1000 sites candidats sont répartis sur l'ensemble de la grille. Chaque site candidat est défini par sa position (x, y) dans la grille. Des antennes isotropes de rayon de couverture égal à 30 points ont été utilisées[11]. La zone étudiée possède des endroits dont il est impossible d'installer des antennes. Ces endroits peuvent être, la mer, les montagnes ...etc., ce qui implique que seulement un taux de 95,522% de couverture peut être atteint. Cette instance est fournie dans [8].

Fig. 2. Image correspondant à la zone géographique à couvrir (ville de Malaga)

Les deux autres instances sont des instances générées aléatoirement. Elles sont fournies par l'université de Constantine dans [18]. Le tableau 1 illustre plus de détails sur ces instances.

Pour tester notre méthode, nous avons effectué plusieurs séries d'exécutions pour différents nombres d'évaluations. Pour chaque série de tests, 10 exécutions ont été réalisées. Les résultats obtenus sont représentés dans la Figure 3. L'axe

Tableau 1. caractéristiques des instances utilisées

Instance	Dimension	Emplacements candidats	Type d'antenne	Couverture
Malaga Instance	450 X 300	1000	Circulaire	30 points
Instance 749	300 X 300	749	Circulaire	26 points
Instance 549	300 X 300	549	Circulaire	26 points

x représente le nombre d'évaluations de la fonction objectif et l'axe y représente la moyenne des valeurs de la fonction objectif obtenues dans les 10 exécutions.

5.1 Paramètres

Plusieurs séries de tests ont été réalisées pour fixer les meilleurs paramètres pour notre algorithme. La valeur de chaque paramètre est donnée comme suit: $p0 = 0.5; t = 15; p_m = 0.005$.

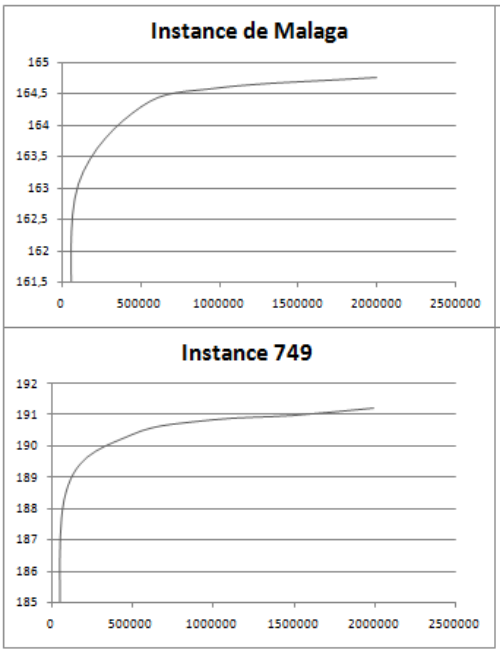


Tableau 2. Tableau récapitulatif des résultats obtenus sur les trois instances

Instance	Fonction objectif	Couverture (%)	Antennes utilisées
Malaga	**164.76**	89.66	46
749	**191,222**	90.678	43
549	**188,418**	90.010	43

Fig. 3. Valeur moyenne de la fonction objectif obtenue en fonction du nombre d'évaluations.

Les graphes de la figure 3 représentent les valeurs de la fonction objectif en fonction du nombre d'évaluations obtenues sur les trois instances.

Le tableau 2 récapitule les valeurs moyennes maximales de la fonction objectif. Pour l'instance de Malaga, la valeur moyenne maximale atteinte est égale à 164.76. En ce qui concerne les instances 549 et 749, une valeur égale à 188,418 a été obtenue pour la première et une valeur égale à 191,222 a été obtenue pour la seconde. En général, l'algorithme converge avant d'atteindre la barre des 2 000 000 évaluations. Le tableau montre aussi le taux de couverture moyen assuré par les solutions trouvées ainsi que le nombre d'antennes moyen employé.

Pour situer ces résultats obtenus, nous les avons comparé contre ceux obtenus dans les travaux de Mendes et al[11] et de Dahi et al [19], qui ont aussi utilisé les mêmes instances dans leurs tests. De plus, dans ces travaux, la même fonction objectif (donnée par la formule 1) est utilisée pour mesurer la qualité d'une solution, ce qui implique que la comparaison peut se faire sur le même critère.

Le tableau 3 classe, selon la valeur de la fonction objectif obtenue, les algorithmes proposés et testés par Mendes et al sur l'instance de Malaga ainsi que notre algorithme.

Tableau 3. Comparaison avec les algorithmes testés par Mendes et al dans [11] sur l'instance de Malaga

Algorithme	Fonction objectif
Notre méthode proposée	**164.760**
MS_GEPVNS	164.701
BLS	164.608
ILS	163.911
CHC	163.278
PBIL	162.651
HYBRID_RUFNS	162.411
AGC	162.134
MS_VNS	162.120
MS_FNS	161.884
CAPMC	161.727
GRASP_EVNS	161.352
SA	156.478
DE	148.802
GRASP_SRCL	148.196

Notre méthode produit la valeur moyenne de la fonction objectif la plus élevée, comparée à beaucoup d'autres algorithmes testés par Mendes et al. En effet, elle dépasse des algorithmes connus comme l'algorithme CHC, PBIL, le recuit simulé (SA), l'algorithme génétique canonique AGC et d'autres variantes.

En plus de l'instance de Malaga, Dahi et al [19] ont testé leur nouvelle méthode appelée QIGA sur les instances 549 et l'instance 749. Le tableau suivant compare nos résultats obtenus avec ceux de Dahi et al.

Tableau 4. Comparaison entre la nouvelle méthode proposée et l'algorithme QIGA présenté par Dahi et al dans [19]

Instance / Algorithme	Méthode proposée	QIGA
Malaga	**164.76**	137.121
749	**191.222**	163.651
549	**188.418**	162.436

Le tableau 4 montre clairement que la nouvelle méthode proposée surclasse l'algorithme QIGA proposé par dahi et al et ce, pour toutes les instances.

Les résultats obtenus sur les trois instances montrent que notre méthode fournit la meilleure valeur de la fonction objectif, c'est à dire qu'elle arrive à trouver les meilleurs solutions par rapport aux méthodes utilisées dans la littérature pour résoudre le problème.

6 Conclusion

Nous avons présenté dans ce papier une nouvelle méthode pour résoudre le problème de positionnement d'antennes dans les réseaux cellulaires. Une méthode qui combine deux principaux aspects des algorithmes de voisinages et des algorithmes évolutionnaires, pour profiter des avantages des deux types.

Cette méthode a été implémentée et testée sur un benchmark réel qui représente la ville de Malaga, ainsi que sur deux autres instances aléatoire. Les résultats ont été comparés avec d'autres travaux dans la littérature. Notre méthode fournit des résultats très prometteurs et elle arrive même à surclasser les autres méthodes testées dans la littérature sur les mêmes instances. Toute fois, notre étude expérimentale doit être étendue sur d'autres instances du problème pour prouver son efficacité. C'est le but de notre travail en ce moment.

En perspective, nous essayerons de tester d'autres stratégies de remplacement des individus de la population. Aussi, nous essayerons de faire de même avec la stratégie du choix de la solution de départ pour la recherche locale. Enfin, aborder le problème sous ça nature réelle, c'est à dire le résoudre sous une formalisation multiobjectif.

References

1. E. Alba, G. Molina, and F. Chicano. Optimal placement of antennae using metaheuristics. In Numerical Methods and Applications (NM&A-2006), Borovets, Bulgaria, August 2006.

2. Alba, E. and Chicano, F., 2005. On the Behavior of Parallel Genetic Algorithms for Optimal Placement of Antennae in Telecommunications. International Journal of Foundations of Computer Science, 16 (2), 343 – 359.
3. Nebro, A. J., Alba, E., Molina, G., Chicano, F., Luna, F., & Durillo, J. J. (2007, July). Optimal antenna placement using a new multi-objective CHC algorithm. In Proceedings of the 9th annual conference on Genetic and evolutionary computation (pp. 876-883). ACM.
4. Benlic U., Hao J.K.: Breakout local search for maximum clique problems. Accepted to Computers & Operations Research, DOI:10.1016/j.cor.2012.06.002 (2012)
5. Calégari, P., Guidec, F., Kuonen, P., & Wagner, D. A. W. D. (1997, May). Genetic approach to radio network optimization for mobile systems. In Vehicular Technology Conference, 1997, IEEE 47th (Vol. 2, pp. 755-759). IEEE.
6. D. Castelino , S. Hurley , And N. Stephens, "A Tabu Search Algorithm for frequency assignment", in Annals of Operations Research 63, pp : 301-319, 1996.]
7. El-Ghazali Talbi, Hervé Meunier, "Hierarchical parallel approach for GSM mobile network design", In J.Parallel Distrib. Comput. 66 (2006) 274 – 290
8. J. Gómez-Pulido. (2008). Web Site of Net-Centric Optimization [Online]. Available: http://oplink.unex.es/rnd
9. Yasmine LAHSINAT – Belaïd BENHAMOU – Dalila BOUGHACI "Trois hyperheuristiques pour le problème d'affectation de fréquence dans un réseau cellulaire" , Dans les proceedings de JFPC 2015, LABRI, Bordeaux, 22-24, 2015, jui 2015
10. H. R. Lourenco, O. C. Martin, and T. Stützle, "Iterated local search," in Handbook of Metaheuristics, Boston, MA: Kluwer, 2002, pp. 321–353
11. Mendes, S.P., Molina, G., Vega-Rodrguez, M.A., Gomez-Pulido, J.A., Sez, Y., Miranda, G., Segura, C., Alba, E., Isasi, P., Len, C., Snchez-Prez, J.M.: Benchmarking a Wide Spectrum of Meta-Heuristic Techniques for the Radio Network Design Problem. IEEE Transactions on Evolutionary Computation, 1133–1150 (2009)
12. Meunier, H., Talbi, E. G., & Reininger, P. (2000). A multiobjective genetic algorithm for radio network optimization. In Evolutionary Computation, 2000. Proceedings of the 2000 Congress on (Vol. 1, pp. 317-324). IEEE.
13. Michel Vasquez , Jin-Kao Hao A Heuristic Approach for Antenna Positioning in Cellular Networks. Journal of Heuristics, 7: 443–472, 2001 Kluwer Academic
14. Reininger, P. (1997). "ARNO Radio Network Optimisation Problem Modelling," ARNO Deliverable N 1-A 1-Part 1. FT. CNET, July 15, 1997.
15. Reininger, P. and A. Caminada. (1998a). "Model for GSM Radio Network Optimisation." In 2nd Intl. ACM/IEEE Mobicom Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications (DIALM), Dallas, December 16, 1998.
16. Larbi Benmezal, Belaid Benhamou and Dalila Boughaci "Some neighbourhood approaches for the Antenna Positioning Problem" , Proceedings of the International Conference on Tools with Artificial Intelligence (ICTAI), nov 2017
17. M.A. Vega-Rodriguez, J.A.G. Pulido, E. Alba, D. Vega-Perez, S. Priem-Mendes, G. Molina, Evaluation of different metaheuristics solving the RND problem, in: Proceedings of the EvoWorkshops on EvoCoMnet, EvoFIN, EvoIASP, EvoINTERACTION, EvoMUSART, EvoSTOC and EvoTransLog: Applications of Evolutionary Computing, Springer, 2007, pp. 101–110.
18. Random Instances: http://www.fichier-rar.fr/2014/10/03/random-instance/.
19. Z. A. E. M. DAHI, C. MEZIOUD, and A. DRAA, "A quantum-inspired genetic algorithm for solving the antenna positioning problem," Swarm and Evolutionary Computation, vol. 31, no. 2, pp. 24–63, 2016.

A K-mer based Multi Convolutional Neural Network Classifier of Low-Ranking Taxonomic Bins from Metagenome

Brahim Matougui[1,2] , Mohamed Batouche[1] and Abdelbasset Boukelia[1,2]

[1] Computer Science Department, Faculty NTIC, University Abdelhamid Mehri - Constantine 2, Constantine, Algeria.
[2] National Center for Biotechnology Research, Constantine, Algeria.
{brahim.matougui,mohamed.batouche,abdelbasset.boukelia}
@univ-constantine2.dz

Abstract. Metagenomics is the study of genomic content gotten in mass from an environment of interests such as the human gut or soil. Among the most important field of metagenomics, we can find the taxonomy which is the science of defining and naming groups of microbial organisms that share the same characteristics. The taxonomy classification is the science of identification and quantification of microbial species sampled by high throughput sequencing. Although many methods exist to deal with the classification problem, assignment to low-ranking taxonomic units remains an important challenge for binning methods as is scalability to Gb-sized datasets generated with deep sequencing techniques. In this paper, we introduce a novel composition based taxonomic assignment method, which relies on the use of a mixture of Convolutional Neural Networks trained by short oligonucleotides (k-mers). We compared our results with the existing taxonomic classification methods. The experimental results have shown that our method outperforms the other methods especially for the low-ranking taxonomic class such as species.

Keywords: Metagenomics, Taxonomic, Classification, Convolutional Neural Network, Composition based Binning Method.

1 Introduction

A handful of soil is rich in microbial life, but the number of microorganisms inside remains obscure. The human body contains many thousands of microbiome such as species of bacteria, fungi, and archaea. In addition to their species variety, the composition and the function of microbial communities is always changing depending on their environment [1]. Metagenomic offers an essential way to analyse microbial diversity inside environmental communities without culturing them in an artificial environment (Laboratory). Such analysis allows the functional and taxonomic characterization of the microbial community. These are often achieved by a combination of sequence assembly and binning methods [2].

The term binning is used to describe the problem of separation of sequence fragments of a metagenome according to their microbial population origins [3, 4]. This definition also includes bins which represent all sequences that have the same higher level clade when it is not possible to rank fragments to individual population (low-level clade) [5].
The taxonomy is the science that studies organisms to define, describe and classify them, including the connection between taxa and the standards underlying such a classification. Taxonomic hierarchy is the arrangement of several categories in hierarchical levels of the biological classification (see **Fig. 1**). Each level represents a taxonomic unit or rank.

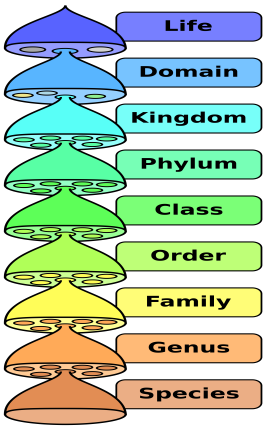


Fig. 1. Taxonomic Hierarchy

We can divide the existing classification (binning) methods into two main categories: Alignment based and composition based methods. In the first category, the DNA read is assigned to a taxonomic label by the alignment of that DNA read to a reference genome. The limitation of these methods is that the assignment is less accurate due to the unavailability or the incompleteness of the reference genome in public databases. This limitation is alleviated by the use of the second methods where DNA reads with the same signature are assigned to the same group. However, the second methods are suffering from binning the low taxonomic ranks. Furthermore, these methods (based on the traditional machine learning algorithm such as SVM, ANN ...) as described in **Fig. 2**, are stuffing from performance issues with data scalability. To cope with these limitations, in this paper, we introduce a novel composition based approach to classify metagenomics DNA fragments, based on the use of a combination of deep learning CNN models.

Therefore, the rest of the paper is organized as follows. In section 2, we discuss related works. Section 3 is dedicated to the description of the proposed approach. In section 4, we present the experimental results and discussions. Finally, conclusions and future work are drawn.

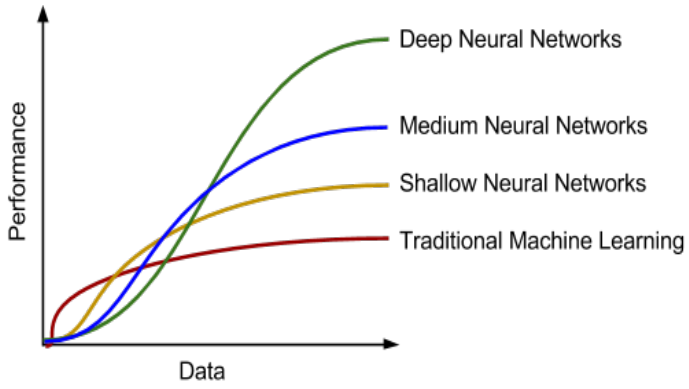


Fig. 2. Performance comparison of deep learning vs other machine-learning algorithms

2 Related Work

Various taxonomic assignments methods exist. We can split them into two main categories namely alignment based and composition based methods.

2.1 Alignment Based Methods

Methods such as MEGAN [6], PhymmBL [7], and NBC [8] are based on the local similarity of a query sequence to the known taxonomic origin ones, where the standard alignment tool BLAST [9, 10] is used to assign each DNA read to the taxon that affords the best score of alignment with its reference genomes. Machine learning algorithms are used in the classification to increase the accuracy.

2.2 Composition Based Methods

In this approach, the sequences are grouped into the same taxonomic group based on their G/C content, oligonucleotide sequence frequency (k-mers), and codon usage [5], because there is a hypothesis, which says that sequences from closely related species are more similar to each other in the features listed above

than non-related or distant species [1]. Some binning methods that use this approach are PhyloPythia [11], PhyloPythiaS [12], PhyloPythiaS+ [2], TETRA [13], TACOA [14], Phymm [7] , S-GSOM [15], and PCAHIER [16, 1] .
In the alignment based methods, each DNA read is assigned to the taxon, which gives the best alignment score with its reference genome [10].The main limitation of these methods that they depend on the availability of reference sequences, which means if there are no complete genome sequences of related organisms, the assignment is less accurate [2]. Whereas, in the composition based methods, the DNA fragment with the same signature such as GC content, codon usage or short oligomers (k-mers), are grouped into the same taxonomic group. Although the composition based methods do not need the complete genome sequences of related organisms to do the assignment, they still have an accuracy problem especially for the low ranking taxa (Family, Genus, Species). Therefore, we propose a k-mer multi convolutional neural network-based approach to classify metagenomic reads of living organisms. The highest classification accuracy is achieved using our method compared to the other taxonomic classification methods existing in the literature. It is an efficient way to classify reads that belong to low taxonomic rank such as species.

3 The Proposed Approach for Taxonomic Classification

The big emergence of Convolution neural networks (CNNs) in the field of DNA barcode analysis has led to integrating them with a lot of genome platforms and analysis tools [17]. Furthermore, due to its capacity to process raw data CNN is often used as a classifier [18]. We propose a new approach for taxonomic classification based on CNN named K-mer mCNN. It consists mainly of two steps: Data Preparation and Model Construction for taxonomic classification.

3.1 Dataset Extraction, Transformation and Load

First, a marker gene analysis is applied to the metagenome sample to define the list of taxa to incorporate in the composition-based- taxonomic metagenome classifier. Then, this list (a list of labeled DNA reads from the metagenome sample) is extended by adding the NCBI reference sequences which have a taxonomy rank or a parent in common with the list above in order to build the reference sequences (see **Fig. 3**) [2]. The reference sequences are split into six different ranks according to their reads length (1000 bp, 3000 bp, 5000 bp, 10000 bp, 15000 bp, 50000 bp). For each rank, a k-mer file is built [2]. Then, we adapted the k-mer files to serve as input of the CNN model (see **Fig. 3**). The used length of k-mers is set to 4-6 dinucleotide.

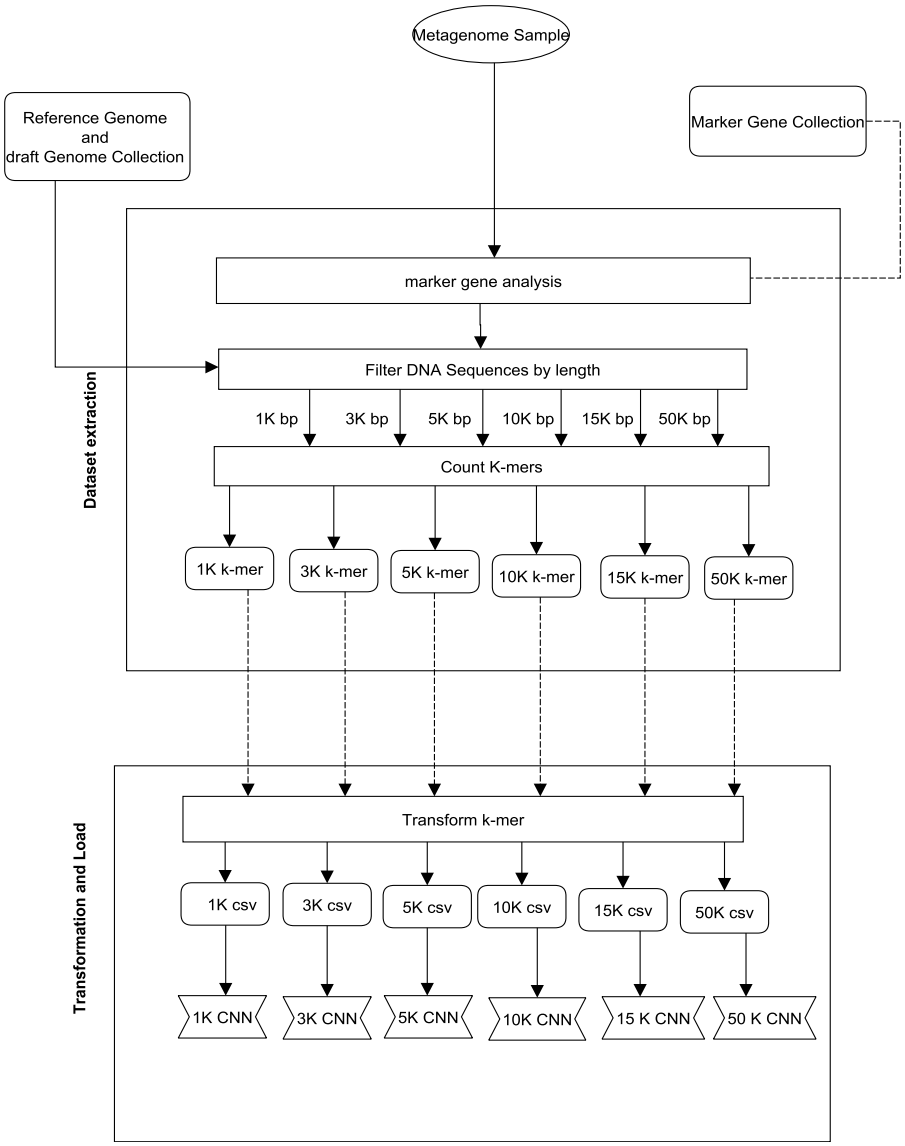


Fig. 3. Dataset extraction, transformation and load

3.2 Classification Model

The Network Structure of our Convolution Neural Network (CNN) model consists of five layers: a convolutional layer, a max-pooling layer and FCL (Fully Connected Layers) (see **Fig. 4**).

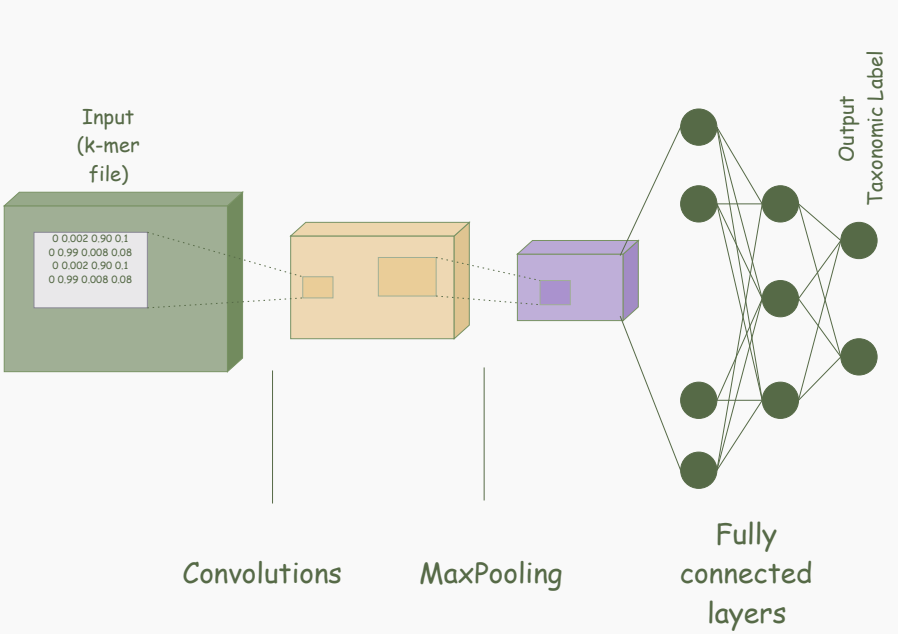


Fig. 4. Illustration of the convolutional neural network (CNN) model used

The **table 1** below describes the parameters, architecture of the proposed CNN model.

Table 1. The architecture used for the classification model

Layer	Name	Type	Parameters
Input	data	Input adapted k-mer	Vector[2828] of k-mers , reshaped to a Matrix [371][8].
1	Conv relu	Convolution ReLU	Features minimization using Kernel[16][8] and ReLU for nonlinear rectified.
2	pool	max-pooling	Take the max of each value from kernel[4][4].
3	dense1	HardTanH	Output equal to 512 for classification using non linear equation.
4	dense2	HardTanH	Output equal to 256 for classification using non linear equation.
5	output	SOFTMAX MCXENT	The output layer contains the number of classes, and Loss Function MCXENT for multiclass classification And SOFTMAX for the classification activity.

Activation Functions

Nodes in a neural network have activation functions that transform signals sent by the neurons of the previous layer by using a mathematical function. This can significantly affect network performance. The following functions are used in our CNN:

– In the Convolutional layer, we used the ReLU as activation function defined as next

$$R(z) = MAX(0, z)$$

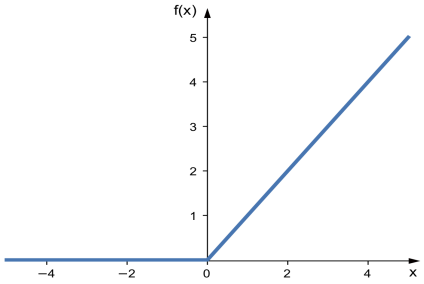


Fig. 5. The plot of the ReLU function

– In the fully connected layers, we used the hard hyperbolic Tangent ($HARDTANH$) function defined as next:

$$HardTanH(x) = \begin{cases} -1 : x < -1 \\ x : -1 < x < 1 \\ 1 : x > 1 \end{cases}$$

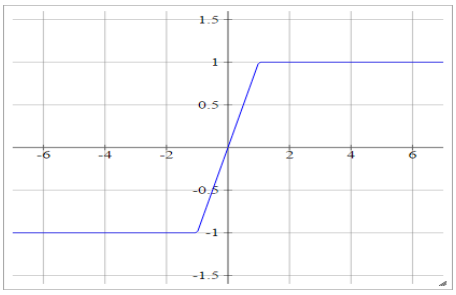


Fig. 6. The plot of the HardTanH function

– In the output layer, we used the normalized exponential function (SOFTMAX):

$$SoftMax(x_i) = \frac{e^{x_i}}{\sum e^{x_i}}$$

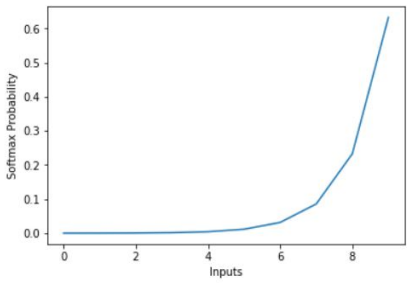


Fig. 7. The plot of the SoftMax function

3.3 Taxonomic Classification

Every rank includes a multiclass CNN classifier which is trained with k-mer files generated from fragments of a different length. The 1K CNN classifier is trained with fragments with length equal to 1K bp. The 3K CNN classifier is trained with fragments with length equal to 3K bp etc. (see **Fig.3**).
The output of each CNN model is a taxonomic class. So, if we want to know the taxonomic origins of a fragment having three suggestions, the assignment of the fragment to a specific class is done with a voting mechanism (see **Fig. 8**).

Workflow

If we want to know the taxonomic class of a sequence, several operations need to be processed. First, we start by counting its length and its k-mers content. Then, we have to adapt the k-mer file to fit into the model. Not all models are involved in the prediction process, only three models with rank close to the sequence length are considered (see **Fig. 8**).

Voting Scheme

As mentioned in the previous step, three models are selected to predict the taxonomic label. Each CNN model has a taxonomic label as output (see **Fig. 8**). To select the taxonomic label among the three candidates the following strategy is applied. The low ranking taxonomic label must be selected first. If all the ranks are equal, a majority vote is applied to select the appropriate output.

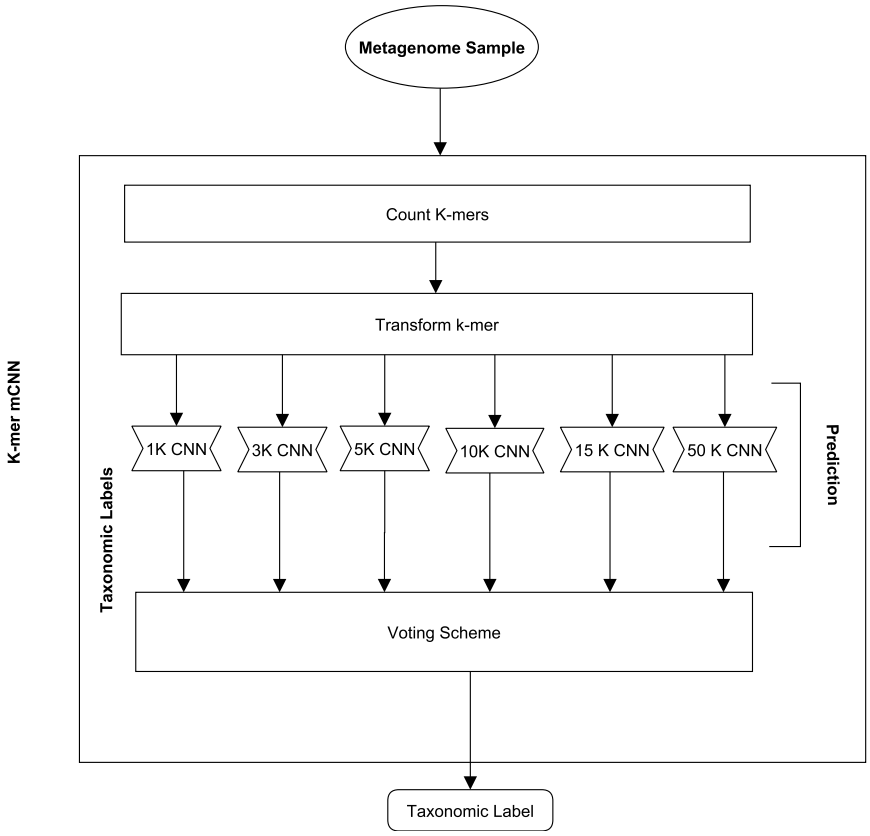


Fig. 8. Taxonomic classification

4 Experimental Results

To develop our CNN model, we used Deeplearning4j which is a Java-based toolkit for building, training and deploying deep neural networks. We used simulated datasets [2] which contains 47 strains from 45 different species including all major taxonomic ranks, at superkingdom, phylum, class, order, family, genus and species rank. The simulated datasets follow a uniform distribution (µ=1). The learning rate used in the training step was set to (5 x 10-3). To validate our classifier, we used k-fold cross validation policy where we set k to 5. For the assessment of classifiers, we used three metrics namely F1-score, Precision, Recall which are defined as follows:

$$Recall = \frac{truepositive}{truepositive + falsenegative}$$

$$Precision = \frac{truepositve}{truepositve + falsepositive}$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

We compared our proposed method to the best-known binning tools so far such as Kraken, Taxator-Tk, Megan, PhyloPhythia, PhyloPhythia +. **Table 2** shows the results of evaluation metrics.

Table 2. Comparison table to assess the performances

Methods	**Ranks**	**F1-score (%)**	**Precision (%)**	**Recall (%)**
taxator-tk	Family	66.6	98.2	50.4
PPS	Family	60.4	72.6	51.7
MEGAN	Family	78.8	88.9	70.7
Kraken	Family	74.7	79.6	70.4
PPS+	Family	88.4	96.4	81.6
k-mer mCNN	**Family**	**99.68**	**100**	**99.38**
taxator-tk	Genus	46.1	93.2	30.6
PPS	Genus	45.8	68.2	34.5
MEGAN	Genus	63.1	75.7	54.1
Kraken	Genus	59.3	63.4	55.7
PPS+	Genus	77.4	91.8	66.9
k-mer mCNN	**Genus**	**97.64**	**96.06**	**99.28**
taxator-tk	Species	16.7	87.8	9.2
PPS	Species	N/A	N/A	N/A
MEGAN	Species	34.2	49.6	26.1
Kraken	Species	32.8	35.7	30.3
PPS+	Species	51.5	71.4	40.3
k-mer mCNN	**Species**	**98.94**	**98.81**	**99.08**

The next plot (Fig.9) summarizes **table 2**.

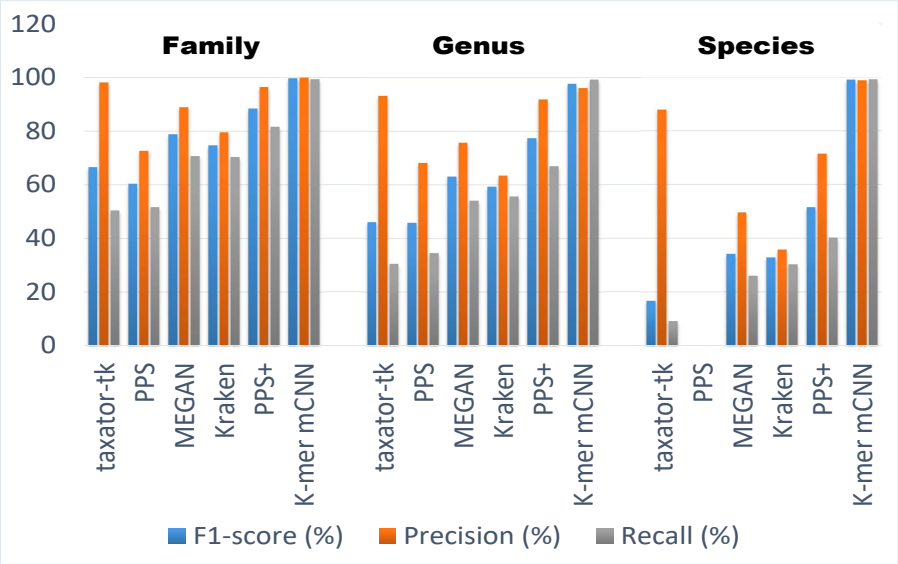


Fig. 9. Performance comparison Histogram

As described in the **table 2** and **Fig. 9**, our proposed method shows the best results compared to other tools where the average F1 Score equals to 98.75, the average Recall equals to 99.24 and the average Precision equals to 98.29 and shows a high performance in the classification of low ranking taxonomic labels.

5 Conclusion

In this paper, we have proposed a deep learning approach for taxonomic low ranks classification based on sequence composition. The proposed approach consists of two steps: a dataset preparation step, which relies on the length and the composition of DNA reads, and a classification step based on convolutional neural network which consists of several classifiers for each dataset length with a voting mechanism for the prediction of a taxonomic label. The experimental results are promising and show that the proposed method has achieved the best performance compared to the existing methods in the literature. As future work, we expect using other deep learning platforms.

References

1. Xinkun Wang. *Next-generation sequencing data analysis*. CRC Press, 2016.
2. Ivan Gregor, Johannes Dröge, Melanie Schirmer, Christopher Quince, and Alice C McHardy. Phylopythias+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ*, 4:e1603, 2016.

3. Tanja Woyke, Hanno Teeling, Natalia N Ivanova, Marcel Huntemann, Michael Richter, Frank Oliver Gloeckner, Dario Boffelli, Iain J Anderson, Kerrie W Barry, Harris J Shapiro, et al. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, 443(7114):950, 2006.
4. Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37, 2004.
5. Johannes Dröge and Alice C McHardy. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in bioinformatics*, 13(6):646–655, 2012.
6. Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. Megan analysis of metagenomic data. *Genome research*, 17(3):000–000, 2007.
7. Arthur Brady and Steven L Salzberg. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, 6(9):673, 2009.
8. Gail L Rosen, Erin R Reichenberger, and Aaron M Rosenfeld. Nbc: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, 27(1):127–129, 2010.
9. Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
10. Xinan Liu, Ye Yu, Jinpeng Liu, Corrine F Elliott, Chen Qian, and Jinze Liu. A novel data structure to support ultra-fast taxonomic classification of metagenomic sequences with k-mer signatures. *Bioinformatics*, 34(1):171–178, 2017.
11. Alice Carolyn McHardy, Héctor García Martín, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos. Accurate phylogenetic classification of variable-length dna fragments. *Nature methods*, 4(1):63, 2007.
12. Kaustubh Raosaheb Patil, Linus Roune, and Alice Carolyn McHardy. The phylopythias web server for taxonomic assignment of metagenome sequences. *PloS one*, 7(6):e38581, 2012.
13. Hanno Teeling, Jost Waldmann, Thierry Lombardot, Margarete Bauer, and Frank Oliver Glöckner. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC bioinformatics*, 5(1):163, 2004.
14. Naryttza N Diaz, Lutz Krause, Alexander Goesmann, Karsten Niehaus, and Tim W Nattkemper. Tacoa–taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC bioinformatics*, 10(1):56, 2009.
15. Chon-Kit Kenneth Chan, Arthur L Hsu, Saman K Halgamuge, and Sen-Lin Tang. Binning sequences using very sparse labels within a metagenome. *BMC bioinformatics*, 9(1):215, 2008.
16. Hao Zheng and Hongwei Wu. Short prokaryotic dna fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. *Journal of bioinformatics and computational biology*, 8(06):995–1011, 2010.
17. Saed Khawaldeh, Usama Pervaiz, Mohammed Elsharnoby, Alaa Eddin Alchalabi, and Nayel Al-Zubi. Taxonomic classification for living organisms using convolutional neural networks. *Genes*, 8(11):326, 2017.
18. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Application du Modèle Arc-Flot pour la Résolution des Problèmes de Bin Covering et Open-End Bin Packing

TOUATI Sofiane[1], RADJEF Mohammed Said[1], BOUAROURI Ouahib[1], and BENMATOUK Hamou[2]

[1] Unité de recherche LaMOS, Département de Recherche Opérationnelle, Faculté des Sciences Exactes, Université de Bejaia, Algérie
[2] Département de Recherche Opérationnelle, Faculté des Sciences Exactes, Université de Bejaia, Algérie

Résumé Dans ce papier, nous traitons deux variantes du problème de Bin Packing, à savoir le Bin Covering et le Bin Packing à boites ouvertes (Open-end Bin Packing). Ces deux problèmes appartenant à la classe des problèmes NP-difficiles, ils ont été traités par des heuristiques, ainsi que par la programmation linéaire en nombres entiers. L'approche, que nous proposons, s'inscrit dans cette dernière classe. Elle consiste en une formulation arc-flot, où on représente les rangements possibles par un graphe. Le programme linéaire obtenu est de taille pseudo polynomiale, qui peut être résolu par des solvers de programmation linéaire. L'avantage de la méthode est qu'elle fournit une bonne borne inférieure, ce qui permet de résoudre des instances tests connues de la littérature.
mots-clés : Bin Packing, Bin Covering, Open-End Bin Packing, Arc-Flot.

1 Etat de l'art

Le modèle arc-flot a été développé par Schapiro [15] en 1968 pour résoudre le problème de sac-à-dos, puis adapté par De Carvalho [4] pour le problème classique du Bin Packing. Le modèle d'arc-flot fait partie des méthodes de programmation linéaire en nombres entiers [5], et a été adapté pour d'autres variantes du Bin Packing [2]. On peut trouver dans [6] une étude comparative entre les différents modèles de Bin Packing, qui montre que le modèle arc-flot permet de résoudre un grand nombre d'instances. Plus récemment, Delorme et Iori [7] ont amélioré le modèle arc-flot, en réduisant le nombre de variables, permettant ainsi de résoudre des benchmarks plus rapidement que le modèle classique.

Avant de présenter les problèmes traités dans ce travail, nous commençons par rappeler le modèle d'arc-flot de De Carvalho [4] pour le bin packing classique. Le problème de Bin Packing classique (BPP) est défini par les données suivantes :

- des boites de capacité C.
- $I = \{1, 2, ..., n\}$, un ensemble d'objets de tailles $0 < a_i \leq C$, avec une demande $b_i \in \mathbb{N}$, $\forall i \in I$.

L'objectif est de ranger tous les objets dans un nombre minimal de boites, sans fragmenter les objets et sans dépasser la capacité des boites.

On se restreint au cas où la taille des objets ainsi que la capacité des boites sont des entiers, les objets sont considérés triés par ordre décroissant de taille. Le modèle d'arc-flot [4] est un modèle graphique, dans lequel un rangement dans une boite est représenté par un chemin d'un sommet source à un sommet puits.

Soit $G = (S, A)$ le graphe défini par l'ensemble des sommets $S = \{0, 1, 2, \ldots, C\}$ et l'ensemble A des arcs. Un objet est représenté dans le graphe par un arc, désigné par un triplet (i, j, l), signifiant que l'objet $l \in I$, de taille $a_l = j - i$ est rangé de la position $i \in S$ à la position $j \in S$.

Nous construisons l'ensemble des arcs comme suit :

$$\begin{cases} A_0 = \{(0, a_l, l) : l \in I\}; \\ A_i = \{(i, j, h) \in S \times S \times I : \exists (e, i, l) \in \cup_{l=\overline{0,i-1}} A_l, l \leq h\} \ ,\forall i = \overline{1, C-1}. \end{cases} \quad (1)$$

Dans ce graphe, l'ordre dans lequel sont rangés les objets est important, néanmoins le (BPP) ne spécifie aucune contrainte sur l'ordre. Ainsi, tout arc sortant du sommet i est nécessairement d'une longueur inférieure ou égale à celle de l'arc de plus grande taille entrant vers i.

On complète le graphe en joignant les sommets sans successeur au sommet final C par des arcs $(i, C, n + 1)$. On remarque qu'un chemin du sommet 0 au sommet C représente une boite remplie de manière maximale (on ne peut plus ajouter un autre objet), et le nombre de chemins est égal au nombre de boites.

On peut résoudre le (BPP) en determinant le nombre minimum de chemins contenant tous les objets. Pour celà, considérons les variables de décisions suivantes :

- $x_{(i,j,l)}$ = nombre de fois où l'on a rangé l'objet l de la position i à la position j.
- z = nombre de boites utilisées.

On peut modéliser le problème par le programme linéaire en nombres entiers suivants :

$$\min z \quad (2)$$

$$\sum_{(0,j,l)\in A} x_{(0,j,l)} = z, \quad (3)$$

$$\sum_{(i,C,l)\in A} x_{(i,C,l)} = z, \quad (4)$$

$$\sum_{(e,i,l)\in A} x_{(e,i,l)} = \sum_{(i,j,l)\in A} x_{(i,j,l)}, \forall i = 1, \ldots, C-1, \quad (5)$$

$$\sum_{(i,j,l)\in A} x_{(i,j,l)} \geq b_l, \forall l = 1, \ldots, n, \quad (6)$$

$$z, x_{(i,j,l)} \in \mathbb{N}, \forall (i,j,l) \in A. \tag{7}$$

On peut remarquer que les équations (2), (3), (4), (5) sont celles du flot de coût minimum, la contrainte de demande (6) impose de ranger tous les objets, et la contrainte d'intégrité (7) impose de ne pas fragmenter les objets.

Exemple 1 *Considérons l'instance suivante :* $C = 7$, $a^T = (5,3,2)$ *et* $b^T = (3,2,2)$. *Le modèle graphique ainsi que la solution sont donnés dans la figure suivante. Seules les variables non nulles sont représentées dans la solution (graphe du bas), avec leur valeurs inscrites sur l'arc.*

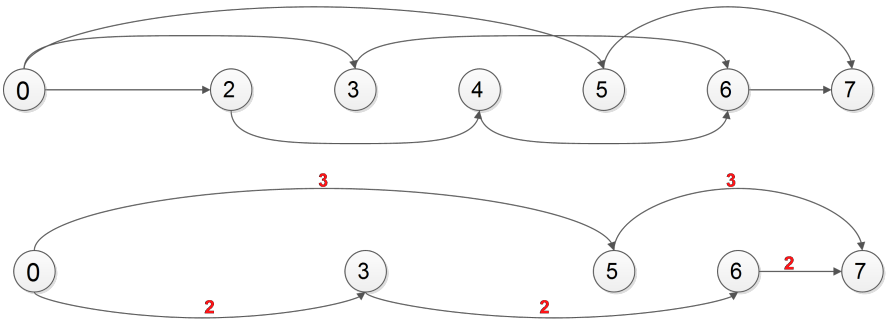


Figure 1. Modèle graphique (graphe du haut), solution du problème (graphe du bas)

Le graphe du bas représente la solution optimale du problème. Le chemin $((0,5),(5,7))$ *représente un rangement constitué d'un objet de taille 5 puis d'un autre de taille 2. Le chemin* $((0,3),(3,6),(6,7))$ *représente une boite remplie avec deux objets de taille égale à 3. Le premier rangement est effectué 3 fois, et le deuxième 2 fois, on a donc une solution à 5 boites.*

2 Modélisation du Bin Covering et de l'Open-end Bin Packing

2.1 Problème de Bin Covering

Le problème de Bin Covering (BCP) [3,10] (aussi appelé Problème de Bin Packing Dual [14]) est posé avec les mêmes données que le (BPP). L'objectif est de ranger tous les objets dans un nombre maximal de boites, en dépassant la capacité des boites. Le (BCP) a plusieurs applications, par exemple dans le milieu hospitalier [16] et dans le transport [11]. Le (BCP), étant NP-difficile [11], a été traité par des heuristiques [11,3], ainsi que par la programmation linéaire [14]. Le modèle obtenu dans [14] est une adaptation du modèle de Gilmore et

Gomory [9], dans lequel le nombre de variables est exponentiel par rapport à la taille du problème, et donc impossible à résoudre directement par un solver. Pour celà, il a été nécessaire de développer un algorithme de Branch and Price spécifique.

Le problème de Bin Covering peut être modélisé par les variables d'affectations suivantes :

$$x_{ij} = \text{nombre de fois où l'objet i est rangé dans la boite j.}$$

$$y_j = \begin{cases} 1, \text{ si la boite j est utilisée}; \\ 0, \text{ sinon.} \end{cases}$$

Le modèle sous forme d'un programme linéaire est le suivant :

$$\max \sum_{j=\overline{1,n}} y_j \tag{8}$$

$$\sum_{i=\overline{1,n}} a_i x_{ij} \geq C y_j, \forall j = \overline{1,n}, \tag{9}$$

$$\sum_{j=\overline{1,n}} x_{ij} = 1, \forall i = \overline{1,n}, \tag{10}$$

$$y_j \in \{0,1\}, x_{ij} \in \mathbb{N}, \forall i,j = \overline{1,n}. \tag{11}$$

Nous proposons une nouvelle modélisation, inspirée de [4], sous forme d'un programme linéaire en nombres entiers, de taille pseudo polynomiale, que nous testons sur des benchmarks connus de la littérature.

On modélise le problème par un graphe $G = (S, A)$, où $S = \{0, 1, \ldots, C + a_1\}$ est l'ensemble des sommets, l'ensemble A des arcs est constitué des chemins possibles du sommet 0 au sommet $C + a_1$:

$$\begin{cases} A_0 = \{(0, a_l, l) : l \in I\}, \\ A_i = \{(i, j, h) \in S \times S \times I : \exists (e, i, l) \in \cup_{l=\overline{0,i-1}} A_l, l \leq h\} \ , \forall i = \overline{1, C + a_1 - 1}. \end{cases} \tag{12}$$

On peut remarquer que le graphe est construit en considérant les objets rangés par ordre décroissant des tailles. On rajoute un arc à chaque fois qu'il est de taille inférieure ou égale à l'un de ses prédécesseurs, on s'arrête dès que l'on dépasse le sommet C (c'est-à-dire la capacité de la boite). Puis, on complète le graphe en joignant les sommets $s \in \{C, C + 1, \ldots, C + a_1 - 1\} : \exists (i, s, l) \in A$ au sommet puits $C + a_1$, en notant ces arcs $(i, C + a_1, n + 1)$.

La solution du problème sera obtenue en résolvant le programme linéaire en nombres entiers suivant :

$$\max z \tag{13}$$

$$\sum_{(0,j,l) \in A} x_{(0,j,l)} = z, \tag{14}$$

$$\sum_{(i,C+a_1,l)\in A} x_{(i,C+a_1,l)} = z, \tag{15}$$

$$\sum_{(e,i,l)\in A} x_{(e,i,l)} = \sum_{(i,j,l)\in A} x_{(i,j,l)}, \forall i = 1,\ldots,C+a_1-1, \tag{16}$$

$$\sum_{(i,j,l)\in A} x_{(i,j,l)} \leq b_l, \forall l = 1,\ldots,n, \tag{17}$$

$$z, x_{(i,j,l)} \in \mathbb{N}, \forall (i,j,l) \in A. \tag{18}$$

On remarque que la contrainte (17) du modèle (13)-(18) n'impose pas de ranger tous les objets, néanmoins il suffira de rajouter les objets manquants à n'importe quelle boite, sans compromettre l'optimalité de la solution.

Exemple 2 *Considérons l'instance suivante :* $C = 100$, $a^T = (98, 73, 45, 20)$, $b^T = (1, 1, 1, 1)$. *La figure 2 représente le modèle graphique ainsi que la solution optimale à deux boites. Dans cet exemple, le nombre optimal de boites est 2.*

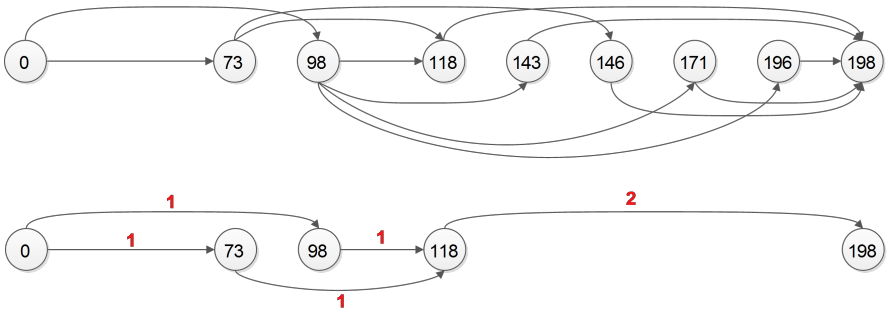


Figure 2. Modèle graphique (graphe d'en haut), solution du problème (graphe du bas)

2.2 Problème de Bin Packing avec boites ouvertes

Le problème de Bin Packing avec boites ouvertes (OEBP) [12] est une variante du (BPP), de complexité NP-difficile [12], dans laquelle on doit ranger tous les objets dans un minimum de boites, à ceci près qu'il est permis qu'un objet dépasse la capacité de la boite, cet objet est appelé objet de débordement.

L'(OEBP) a été introduit par Leung et al. [12], où il a modélisé un problème de billetterie à la station de metro de Honk-Kong. Ce problème a surtout été résolu par des heuristiques [12],[8],[13], même si un modèle de programmation linéaire a été fourni dans [13], qui n'a pas été cependant résolu.

Dans [13], les données du problème (OEBP) se présentent sans le vecteur b, représentant l'occurrence des objets. Le modèle présenté dans [13] est basé sur les variables d'affectations suivantes :

$$x_{ij} = \begin{cases} 1, \text{ si l'objet i est affecté à la boite j ;} \\ 0, \text{ sinon.} \end{cases}$$

$$y_j = \begin{cases} 1, \text{ si la boite j est utilisée ;} \\ 0, \text{ sinon.} \end{cases}$$

Les auteurs se basent sur le fait qu'il existe forcément une solution optimale dans laquelle les objets de débordement sont les plus gros en taille, ainsi on peut se limiter au cas où l'objet candidat pour déborder la boite j est forcément le j^{eme} objet. Le modèle d'affectation de l'(OEBP) est donné ci-dessous :

$$\min \sum_{j=\overline{1,n}} y_j \tag{19}$$

$$\sum_{i\in\{1,\ldots,n\}\setminus\{j\}} a_i x_{ij} \leq (C-1)y_j, \forall j \in \{1,\ldots,n\}; \tag{20}$$

$$y_i + \sum_{j\in\{1,\ldots,n\}\setminus\{i\}} x_{ij} = 1, \forall i \in \{1,\ldots,n\}; \tag{21}$$

$$x_{ij}, y_j \in \{0,1\}, \forall i,j \in \{1,\ldots,n\}. \tag{22}$$

La formule (19) indique que l'objectif est de minimiser le nombre de boites. Dans la formule (20), le j^{eme} objet est un objet de débordement, occuppant seulement un espace de taille 1, et les objets qui ne sont pas de débordement occupent un espace de taille inférieure ou égale à $(C-1)$. La formule (21) indique que l'on affecte tous les objets. La formule (22) est la contrainte d'intégrité, qui spécifie que l'on ne fragmente pas les objets.

Nous proposons une autre modélisation sous forme d'un programme linéaire. On considère que les objets sont ordonnés par ordre croissant de tailles $a_1 \leq a_2 \leq \cdots \leq a_n$, afin de ranger un maximum d'objets dans une boite, et pour que l'objet de débordement soit laissé en dernier.

Notons par $G = (S, A)$ le modèle graphique, où $S = \{0, 1, \ldots, C + a_1\}$ est l'ensemble des sommets, l'ensemble A des arcs est construit de manière similaire au Bin Covering, en considérant les objets par ordre croissant des tailles.

On modélise le problème par le programme linéaire en nombres entiers suivant :

$$\min z \tag{23}$$

$$\sum_{(0,j,l)\in A} x_{(0,j,l)} = z, \tag{24}$$

$$\sum_{(i,C+a_1,l)\in A} x_{(i,C+a_1,l)} = z, \tag{25}$$

$$\sum_{(e,i,l)\in A} x_{(e,i,l)} = \sum_{(i,j,l)\in A} x_{(i,j,l)}, \forall i = 1, \ldots, C + a_1 - 1, \tag{26}$$

$$\sum_{(i,j,l)\in A} x_{(i,j,l)} \geq b_l, \forall l = 1, \ldots, n, \tag{27}$$

$$z, x_{(i,j,l)} \in \mathbb{N}, \forall (i,j,l) \in A. \tag{28}$$

Exemple 3 *Considérons l'instance suivante :* $C = 60$, $a^T = (98, 46, 23)$. *La figure suivante représente le modèle graphique ainsi que la solution optimale à deux boites.*

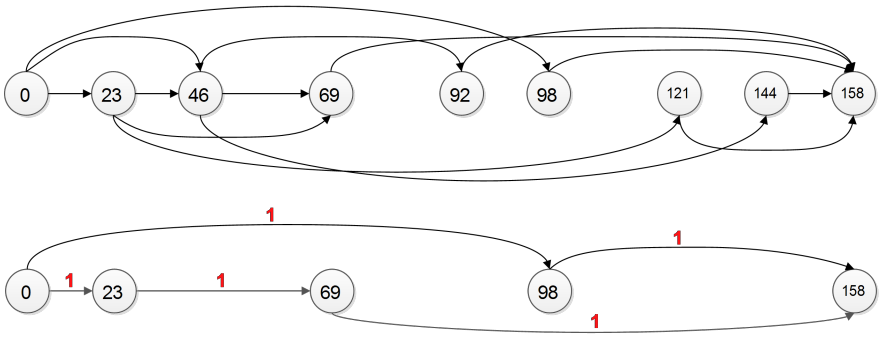


Figure 3. Modèle graphique (graphe d'en haut), solution du problème (graphe du bas)

3 Résultats numériques

Nous avons testé les modèles (13)-(18) du (BCP) et (23)-(28) de l'(OEBP) sur quelques instances issues des benchmarks proposés par Beasley, disponibles sur le web [1]. Ces Benchmarks ont été résolus avec exactitude pour le (BPP). Dans [13], ces benchmarks ont servi à tester une heuristique, que les auteurs ont développée pour la résolution du problème (OEBP). A notre connaissance, aucune méthode exacte de résolution des problèmes (BCP) et (OEBP) n'a été testée sur ces benchmarks pour pouvoir faire une analyse comparative des méthodes.

Dans ce travail, nous avons choisi de valider les résultats obtenus en comparant le modèle arc-flot avec le modèle d'affectation (le modèle (8)-(11) pour le (BCP), le modèle (19)-(22) pour le l'(OEBP)). Les PLNE (8)-(11), (13)-(18), (19)-(22) et (23)-(28) ont été résolus par le solver GLPK 4.61, sur un pc portable i5 avec 6 Go Ram. Le temps de résolution est limité à une minute, la génération des coupes de gomory est activée, l'exploration de l'arbre se fait en

largeur d'abord et une stratégie de séparation basée sur la variable dont la partie fractionnaire est la plus proche de 0.5. Les instances tests sont de tailles 120 et 250, chacune contient vingt instances numérotées de 0 à 19.

Nous avons adopté la notation suivante :

- nvar : nombre de variables.
- ncont : nombre de contraintes.
- nbins : le nombre minimal de boites.
- temps : temps (en secondes) nécessaire pour construire le graphe et résoudre le programme linéaire.

Table 1. Résultats numériques U_{120}, Bin Covering

	modèle arc-flot				modèle affectation		
N°	nvar	ncont	nbins	temps	nvar	ncont	nbins
00	3559	255	47	4.50	2773	106	34
01	3704	257	48	7.06	2880	108	29
02	4028	262	45	4.88	2790	107	32
03	4340	267	48	8.11	3312	117	-
04	3759	263	48	4.89	3087	112	36
05	3790	256	47	5.81	2914	109	38
06	4222	261	47	15.16	3102	113	32
07	3911	257	48	3.98	3120	113	-
08	4310	266	49	4.34	3332	117	32
09	4173	267	45	5.20	2925	110	-
10	3850	257	51	4.19	3315	116	33
11	3571	255	48	3.33	2928	109	36
12	4137	260	47	5.08	3008	111	31
13	3900	261	48	3.76	3024	111	34
14	3678	257	49	3.12	3038	111	36
15	3896	263	47	5.31	3008	111	33
16	3815	255	50	5.27	3315	116	31
17	3842	258	51	2.71	3315	116	37
18	4173	265	48	4.08	3216	115	35
19	4481	270	48	6.17	3312	117	35

Table 2. Résultats numériques U_{250}, Bin Covering

	modèle arc-flot				modèle affectation		
N°	nvar	ncont	nbins	temps	nvar	ncont	nbins
00	4530	272	98	6.99	7056	170	-
01	5348	282	99	20.08	7821	178	-
02	5134	280	101	14.41	7777	178	-
03	5387	282	99	6.98	7920	179	-
04	5085	281	100	9.41	7700	177	-
05	5269	280	100	10.96	7900	179	-
06	5389	282	101	7.31	7979	180	-
07	5178	279	101	11.27	7956	180	-
08	5277	281	104	11.05	8216	183	-
09	5463	285	100	7.61	8100	181	-
10	5156	280	104	7.67	8112	182	-
11	5456	283	100	8.25	8100	181	-
12	5013	277	104	16.12	8008	181	-
13	5049	279	101	10.05	7676	177	-
14	5398	285	99	5.60	7920	179	-
15	5240	280	104	10.58	8216	183	-
16	5208	281	96	10.61	7392	173	-
17	5175	281	99	3.91	7623	176	-
18	5145	281	99	6.58	7623	176	-
19	5587	285	101	7.28	8282	183	-

Nous remarquons des résultats numériques (table 1 à 4) les points suivants :

- Le nombre de variables du modèle arc-flot est dans la majorité des cas inférieur à celui du modèle d'affectation.
- La solution obtenue par le modèle arc-flot est toujours meilleure que celle du modèle d'affectation.
- Le modèle affectation ne permet pas d'avoir toujours une solution réalisable dans le temps imparti.

Au vu des résultats numériques, le modèle arc-flot nous apparaît meilleur que le modèle d'affectation.

Table 3. Résultats numériques U_{120}, Open-End Bin Packing

	modèle arc-flot				modèle affectation		
N°	nvar	ncont	nbins	temps	nvar	ncont	nbins
00	7047	285	30	8.02	14400	241	32
01	7140	287	31	8.99	14400	241	32
02	7538	293	29	19.31	14400	241	30
03	8378	298	31	14.15	14400	241	32
04	7542	289	31	14.53	14400	241	32
05	7411	289	30	12.66	14400	241	31
06	7991	295	30	18.87	14400	241	31
07	7572	288	31	11.24	14400	241	32
08	8239	297	31	22.74	14400	241	32
09	7853	294	29	18.58	14400	241	30
10	7720	291	32	10.65	14400	241	33
11	7162	287	31	12.43	14400	241	32
12	7693	291	31	18.82	14400	241	32
13	7488	290	30	14.86	14400	241	31
14	7186	286	31	10.33	14400	241	32
15	7677	292	30	14.27	14400	241	31
16	7537	288	32	17.17	14400	241	33
17	7480	288	33	15.22	14400	241	34
18	7911	293	31	12.57	14400	241	32
19	8341	299	31	16.34	14400	241	32

Table 4. Résultats numériques U_{250}, Open-End Bin Packing

	modèle arc-flot				modèle affectation		
N°	nvar	ncont	nbins	temps	nvar	ncont	nbins
00	8750	301	62	18.68	62500	501	64
01	9807	311	63	37.20	62500	501	65
02	9480	307	64	24.75	62500	501	66
03	9878	311	63	28.48	62500	501	65
04	9506	308	64	26.87	62500	501	66
05	9759	310	64	29.45	62500	501	66
06	9807	311	64	35.86	62500	501	66
07	9566	308	64	21.26	62500	501	67
08	9651	309	66	18.25	62500	501	69
09	9948	312	63	34.05	62500	501	66
10	9563	308	66	25.78	62500	501	68
11	10010	312	63	26.00	62500	501	66
12	9344	304	66	34.25	62500	501	68
13	9358	307	65	23.27	62500	501	67
14	9938	312	63	31.85	62500	501	65
15	9738	309	66	27.29	62500	501	68
16	9545	309	62	30.26	62500	501	64
17	9488	308	63	47.31	62500	501	65
18	9545	309	63	36.06	62500	501	65
19	10200	314	64	61.40	62500	501	67

4 Conclusion et Perspectives

Dans ce papier, nous avons traité deux variantes du problème de bin packing, à savoir le Bin Covering et l'Open-End Bin Packing. Nous avons commencé par fournir une nouvelle modélisation basée sur le modèle arc-flot. Dans ce modèle, on représente un rangement dans une boite par un graphe, puis nous déterminons la solution optimale en résolvant un programme linéaire à variables entières, construit grâce au graphe. Le modèle d'arc-flot a montré la possibilité de résoudre des Benchmarks références en un temps acceptable.

En perspective, nous prévoyons d'adapter ce modèle à la classe de problèmes d'ordonnancement NP-difficiles, notamment l'ordonnancement sur machine parallèle ainsi que certains problèmes de sac à dos.

Références

1. http ://people.brunel.ac.uk/ mastjjb/jeb/orlib/binpackinfo.html.
2. Brandao, F., & Pedroso, J. P. (2016). Bin packing and related problems : general arc-flow formulation with graph compression. Computers & Operations Research, 69, 56-67.
3. Csirik, J., Johnson, D. S., & Kenyon, C. (2001, January). Better approximation algorithms for bin covering. In Proceedings of the twelfth annual

ACM-SIAM symposium on Discrete algorithms (pp. 557-566). Society for Industrial and Applied Mathematics.

4. De Carvalho, J. V. (1999). Exact solution of bin-packing problems using column generation and branch-and-bound. Annals of Operations Research, 86, 629-659.

5. De Carvalho, J. V. (2002). LP models for bin packing and cutting stock problems. European Journal of Operational Research, 141(2), 253-273.

6. Delorme, M., Iori, M., & Martello, S. (2016). Bin packing and cutting stock problems : Mathematical models and exact algorithms. European Journal of Operational Research, 255(1), 1-20.

7. Delorme, M., & Iori, M. (2017). Enhanced Pseudo-Polynomial Formulations for Bin Packing and Cutting Stock Problems. Technical report, DEI "Guglielmo Marconi", University of Bologna, Italy.

8. Gent, I. P. (1998). Heuristic solution of open bin packing problems. Journal of Heuristics, 3(4), 299-304.

9. Gilmore, P. C., & Gomory, R. E. (1961). A linear programming approach to the cutting-stock problem. Operations research, 9(6), 849-859.

10. Jansen, K., & Solis-Oba, R. (2003). An asymptotic fully polynomial time approximation scheme for bin covering. Theoretical Computer Science, 306(1-3), 543-551.

11. Labbé, M., Laporte, G., & Martello, S. (1995). An exact algorithm for the dual bin packing problem. Operations Research Letters, 17(1), 9-18.

12. Leung, J. Y. T., Dror, M., & Young, G. H. (2001). A note on an open-end bin packing problem. Journal of Scheduling, 4(4), 201-207.

13. Mohamed, M., Mohamed, T., & Billal, R. (2016). Modeling and solving the open-end bin packing problem. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 7(12), 399-404.

14. Peeters, M., & Degraeve, Z. (2006). Branch-and-price algorithms for the dual bin packing and maximum cardinality bin packing problem. European Journal of Operational Research, 170(2), 416-439.

15. Shapiro, J. F. (1968). Dynamic programming algorithms for the integer programming problem-I : The integer programming problem viewed as a knapsack type problem. Operations Research, 16(1), 103-121.

16. Vijayakumar, B., Parikh, P. J., Scott, R., Barnes, A., & Gallimore, J. (2013). A dual bin-packing approach to scheduling surgical cases at a publicly-funded hospital. European Journal of Operational Research, 224(3), 583-591.

17. Yang, J., & Leung, J. Y. T. (2003). The ordered open-end bin-packing problem. Operations Research, 51(5), 759-770.

GEO : jeu sérieux adaptatif basé sur le profil de l'apprenant

Ahmed yassine Benanane[1], Zoulikha Mekkakia Maaza[2]

[1,2] Université des sciences et de la technologie "Mohamed Boudiaf" (USTO-MB) Oran, Algerie
{yacineose1@gmail.com, zoulikha.mekkakia@univ-usto.dz}

Résumé Plusieurs recherches dans le domaine des jeux sérieux adaptatifs ont amené à mettre en valeur le profil de l'apprenant caractérisé par ses connaissances, ses capacités (émotifs et mémorisation visuelles ou sonores) ou ses actions constatées et enregistrées pendant le jeu.
Les jeux (Learning games) destinés à la formation dans le but de développement des compétences des apprenants doivent être adaptés à leurs profils, nous avons besoin de la compréhension de ce qui s'est passé durant le jeu, identifier les comportements et les performances de l'apprenant pour formuler son profil en se basant sur l'analyse de la trace et la description des connaissances acquises. Notre objectif est d'adapter le jeu sérieux (GEO destiné pour l'apprentissage de la géographie) au niveau de l'apprenant-joueur par l'apprentissage automatique et les colonies d'abeille.

Keywords: Jeux sérieux, Jeux d'apprentissage, Modélisation de l'apprenant, IA, Jeux adaptatifs.

I. Introduction

Les enfants qui grandissent dans notre société ces dernières années ont une manière différente de vivre au quotidien, ou` le temps consacré à l'utilisation de l'électronique et de l'informatique est très important. Les smart phones et les ordinateurs sont devenus indispensables dans la vie quotidienne de cette jeune génération.

Les jeux vidéo dominent le quotidien des enfants et des adolescents qui prennent du plaisir à y jouer, ces jeux vidéo ont pris même une place dans l'économie mondiale. Au lieu que nos enfants passent leurs temps libre avec des jeux « inutiles » ils peuvent se consacré aux jeux sérieux pour consolider ce qu'ils aient appris en classe tout en s'amusant.

Le jeu sérieux (serious games) est considéré comme un jeu vidéo conçu avec un objectif autre que simple divertissement, il utilise les nouvelles technologies dans le but de passer un message de manière attractive. Aujourd'hui les serious games font une entrée remarquable dans l'apprentissage individualisé qui est beaucoup efficace que l'apprentissage dans une salle de classe.

Ces outils pédagogiques permettent à l'enseignant de devenir un guide dans l'apprentissage de l'élève, quant à l'apprenant il devient le responsable de la construction de sa propre connaissance dans les différentes matières étudiées. Il est indispensable de faire comprendre la différence entre les joueurs dans la conception des jeux, l'adaptabilité des jeux en temps réel nécessite une modélisation précise du joueur. Une étude pour l'évaluation des approches d'apprentissage basées sur le jeu par le biais de jeux sérieux a été faite par **[1**], les auteurs confirment que les jeux sérieux peuvent contribuer à l'amélioration de la motivation.
Dans la section suivante, nous décrivons les motivations de notre recherche, ensuite nous présentons les antécédents et l'état de l'art, en suite nous présentons notre jeu adaptatif GEO, dans la section **V** une proposition pour l'adaptation est présentée. Dans la section **VI** des expériences avec 2 classes de (34 élèves et 23 élèves respectivement) ont été réalisées, la section **VII** donne notre analyse des résultats et la discussion enfin, la section **VIII** conclut le document.

II. Les motivations

L'un des principaux objectifs des développeurs des jeux sérieux est d'encourager les apprenants– joueurs à réutiliser le jeu. Cela permet de soutenir un apprentissage efficace, puisque les enfants sont toujours désireux de jouer et de s'éloigné des devoirs et les exercices en classe. Le succès des jeux sérieux fait face à des problèmes qui résultent d'une gamme des apprenants qui n'ont pas de patience même pour jouer parce qu'ils sont moins motivés, par exemple un jeu sérieux difficile entraine la frustration rapide, par contre un jeu sérieux avec des défis facile entraine l'ennui chez l'apprenant–joueur, ainsi la classification des compétences cognitives fournit un niveau idéal de difficulté du jeu.

Selon [**2**] la méthode d'apprentissage dans le jeu sérieux adaptatifs est meilleur que la traditionnelle car les animations du jeu sérieux d'apprentissage déclenche des souvenir à long terme chez les apprenants, le processus de l'adaptation peut fournir des informations quand le joueur a besoin d'aider et équilibrer ses émotions.

Ces données renforcent l'utilisation des jeux sérieux adaptatifs par les enfants pour acquérir une meilleure expérience efficace d'apprentissage en passant par des situations pratiques dans le jeu.

III. Travaux connexes

L'idée d'utilisation des jeux sérieux adaptatifs dans différents domaines n'est pas récente, et l'adaptabilité des jeux sérieux fait l'objet de plusieurs recherches .Il s'agit notamment des approches basées sur la maintenance de l'intérêt élevé du joueur pour un bon apprentissage de la maitrise par exemple

l'article [**3**] discute comment tirer les caractéristiques des joueurs dans un jeu sérieux. Les auteurs proposent le cognitive skill game (CGS) (Jeu d'habileté cognitive) qui est basé sur un agent intelligent artificiel, il peut prévoir le caractère cognitif du joueur. La méthode Apprentissage de Quantification Vectorielle (LVQ : Learning Vector Quantization) est utilisée pour classer le niveau cognitif des joueurs, utilisée dans (CGS). LVQ est un réseau de neurones, lorsque certaines entrées ont des vecteurs de distance très proches, ces vecteurs seront regroupés selon trois classes (1- qui font des essais et erreurs, 2- les prudents et 3- les experts).

D'autres recherches ont étudié comment détecter la frustration, le stress, la motivation et les états affectifs de l'élève. Dans [**4]** les auteurs proposent des techniques bayésiennes pour élaborer des modèles de prédiction de l'affect d'élèves. Les joueurs ont 55 minutes pour résoudre le mystère du jeu CRYSTAL ISLAND, ils sont invités toutes les 7 minutes à déclarer leur état d'esprit et de choisir une émotion (frustration, confusion, anxiété, ennui….).
Les auteurs ont rassemblé les suggestions comment les émotions sont produites, et ils ont suggéré d'utiliser un moteur pour traiter des informations émotionnelles dynamiquement.

Les auteurs de [**5**] proposent un environnement d'apprentissage de renforcement pour modéliser et évaluer les compétences de l'apprenant-joueur à des applications qui intègrent la rééducation robotique des patients. Cette application fait face au défi d'adapter la difficulté des jeux par rapport aux compétences des joueurs (patients), en gardant la motivation et l'engagement des joueurs. La rééducation traditionnelle est basée sur des exercices répétitifs, les robots et les jeux sérieux ont une opportunité et des nouveaux moyens pour améliorer le processus de traitement. Le niveau de difficulté des jeux doit être ajusté à chaque niveau de compétence du joueur. Des expériences avec une durée de trente (30) minutes sont présentées impliquant quatre (04) joueurs volontaires, afin d'ajuster le niveau de difficulté individuel, l'intégration de l'algorithme Q-learning dans ce contexte est fait pour permettre de modifier les paramètres des jeux et évaluer les compétences de l'utilisateur.

Dans [**6**] les auteurs proposent le développement d'une plate-forme générique et évolutive permettant la génération de scénarios adaptés aux caractéristiques et besoins des utilisateurs. Ils ont proposé une architecture permettant d'organiser les connaissances du domaine en trois couches : concepts du domaine, ressources pédagogiques et ressources du jeu. Les traces d'interaction sont utilisées pour faire évoluer le profil de l'utilisateur à partir de ses performances.

Des recherches récentes [**7**] se concentrent sur la reconnaissance des émotions des joueurs dans les jeux sérieux et comment comprendre la réponse émotionnelle. Ils proposent une combinaison des modalités faciales, corporels, et paroles (méthode multimodale) de reconnaissance des émotions. Ils ont créé une liste d'actions corporelles et d'expression faciales couramment rencontrées dans un jeu typique, sur la base de cette liste, une base de données bimodale a été créée à l'aide du capteur KINECT MICROSOFT. Pour évaluer l'approche, les auteurs ont créé un ensemble de données avec enregistrement

MICROSOFT KINECT de personne qui exécute 5 émotions de base (colère, peur, bonheur, tristesse, surprise) qui se rencontre couramment dans un jeu.

Les auteurs [**8**] décrivent leurs adaptations de la rétroaction formative basée sur les états émotionnels des élèves dans (l'environnement iTalk2Learn). L'approche utilisée est l'utilisation de feedback formative: comment doit être fournis? A quel moment? De quelle façon doit être présenté? Le feedback vise à améliorer les états affectifs des apprenants: de l'état négatif vers l'état positif ou bien garder l'apprenant à l'état positif. A cet effet ils utilisent deux réseaux bayésiens, le premier détermine le type de feedback et le second pour détecter la rétroaction formative.

IV. Le jeu sérieux adaptatif « GEO »

On peut distinguer cinq types de jeux sérieux : Jeux publicitaires (Adver gaming), Ludo- Éducatifs (Edutainment), Jeux de marché (Edumarket game), les jeux engagés (ou détournés), jeux de simulation et jeux expérimentaux.

Dans le cadre de notre présent travail de recherche, nous présentons l'adaptation du jeu de formations pour les enfants Ludo- Éducatifs (learning game).

Notre jeu sérieux a été réalisé au sein de notre laboratoire (SIMPA de l'USTO-MB). Il est dédié à la géographie qui a trop souvent était présenté à l'élève du collège d'enseignement moyen « CEM » comme une description et cartographie de la surface de la terre, mais il considère ses cours comme des cours ennuyants et il trouve que c'est très difficile pour acquérir des connaissances dans ce domaine.

Fig.1 Jeu sérieux pour l'apprentissage de la géographie (GEO)

Pour que le parcours obtenu dans notre jeu sérieux GEO soit optimal et adapté à l'objectif pédagogique et au profil de l'apprenant, nous avons besoin de formuler son profil en se basant sur la description de ces connaissances acquises. Nous distinguons trois types de systèmes intégrant des dispositifs d'adaptation : le système adapté dans le quel les techniques d'adaptation sont appliquées durant la phase de conception du système , le deuxième est le

système adaptable dans le quel les utilisateurs qui saisissent leurs préférences, qui les enregistrent dans un modèle qui, par la suite, n'est modifié que sur nouvelle demande explicite de l'utilisateur , dans notre cas on s'intéresse au troisième système adaptatif ou la mise à jour du modèle utilisateur est réalisée par le système lui-même, par observation des actions et des réactions de l'utilisateur (interaction avec le système).

Notre jeux sérieux GEO, dédié aux collégiens de l'enseignement moyens, selon le programme de l'éducation national algérienne, est composé initialement de trois cours indépendants ; le premier cours est (le globe terrestre) il est présenté sous forme de narration, l'élève doit répondre au quiz global qui est noté pour qu'à la fin il ait son bonus qui est « la découverte des planètes ». Le second cours est (les mers et les océans), il est présenté sous forme d'animation et de texte, le troisièmes cours est (les montagnes et les collines), il est présenté sous forme de narration, animations et textes.

Chaque partie est définie par un environnement et un scénario différent du précédent, les cours débutent par la recherche d'un objet référence du thème en suite la progression dans le cours se fait par la recherche d'éléments qui dévoilent son contenu.

Seulement deux agents peuvent contrôler l'adaptation : l'apprenant et le système.

Exemple: auto adaptation Contrôlée par l'utilisateur

	Système	Utilisateur
Initiative	X	
Proposition	X	
Décision		X
Exécution	X	

Tab.1 Auto adaptation contrôlée par l'utilisateur

V. L'approche proposée et les progrès

Notre contribution se situe sur l'aspect d'apporter, structurer et organiser les connaissances de l'apprenant.

L'idée est de permettre à l'apprenant de comprendre ses erreurs, combler ses lacunes et d'identifier les éléments qu'il ne maîtrise pas. Si l'apprenant n'arrive pas à saisir une notion, il sera obligé de revoir la notion précédente, c'est à dire il faut intervenir avec un feed-back immédiat.

L'adaptabilité nécessite de disposer de mécanismes permettant de choisir les modifications nécessaires sur l'interface à partir d'observations effectuées sur le triplet : apprenant-joueur, scénario du jeu et l'interaction de l'utilisateur.

Pour cela nous présentons notre architecture (description) basée sur des techniques de l'intelligence artificielle qui vont nous permettre d'analyser la trace.

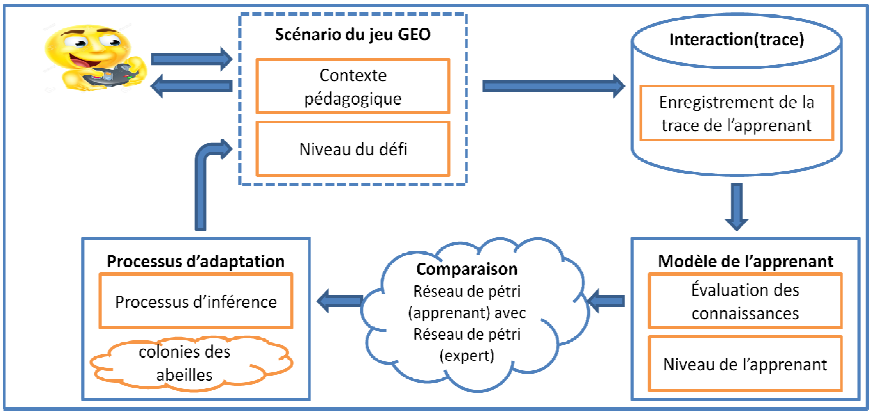


Fig.2 ***Architecture d'un jeu adaptatif***

Afin de décrire le comportement des joueurs nous avons créé une liste des traits de comportement à étudier, cette liste contient : la mémorisation, la curiosité, la vitesse, la motivation et la prudence. Dans un premier lieu nous cherchons d'étudier la mémorisation de l'élève. Pour cela nous suggérons de modéliser la trace du joueur en utilisant un réseau de pétri, et faire une comparaison avec la trace d'un expert (enseignant) pour tirer les caractéristiques, les préférences, la connaissance, plan et but du joueur.

A cet effet nous estimons que l'algorithme de colonies des abeilles peut facilite et régler le problème de guidage des apprenants dans leur apprentissage. Il permet la recommandation du chemin optimal de navigation, la prédiction de la performance des apprenants, et l'amélioration des algorithmes de classification. L'idée est de proposer des scénarios personnalisés basée sur l'algorithme des abeilles dans le but de préciser les ressources adéquates à chaque apprenant-joueur dans notre jeu adaptatif.

VI. Expériences

Pour les tests, une configuration expérimentale a été construite. Nous avons testé notre jeu GEO dans une école (CEM) pour les élèves de la 2ème année d'enseignement moyen et nous avons comparé le cours en utilisant le jeu sérieux GEO avec un cours traditionnel dans une salle de classe classique, le contenu du cours étant le même. Nous avons commencé par le premier cours (le globe terrestre) qu'il est présenté sous forme de narration, l'élève doit

répondre au quiz global qui est noté pour qu'à la fin il ait son bonus qui est « la découverte des planètes ». Nous avons suivi le joueur et sauvegardé sa trace en remarquant : Si le joueur répond avec une mauvaise réponse plusieurs fois => l'élève ne mémorise pas. Donc nous suggérons que le jeu devrait ralentir la narration et avancer lentement, en ajoutant d'autres propositions (utilisation un feed back formative, refaire le cours…)

VII. Résultats et discussion

On a commencé note expérience dans une salle de classe pour une durée de **45** minutes : « cours traditionnel »pour les élèves de la **2**^{eme} année moyen, le nombre d'élève : **34** élèves, le cours **" المعالم الجغرافية"**qui contient **15** questions, les résultats des élèves dans un cours traditionnel sont comme suit : Les bonnes/mauvaises réponses :

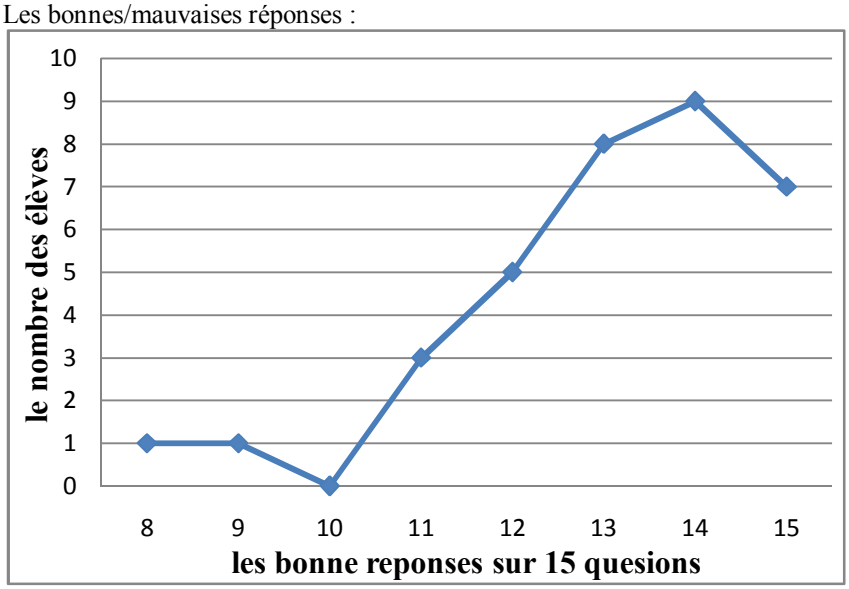


Fig.3 Les bonnes/mauvaises réponses par élèves dans un cours traditionnel

Par la suite on a testé notre jeu « GEO » pour une durée de **15** minutes pour les même élèves de la **2**^{eme} année moyen, le nombre d'élève : **23** élèves, le cours **" المعالم الجغرافية"**qui contient **15** questions, La préférence des élèves a été le cours avec **vidéo et narration,** les résultats des élèves dans le jeu GEO sont comme suit : les bonnes/mauvaises réponses :

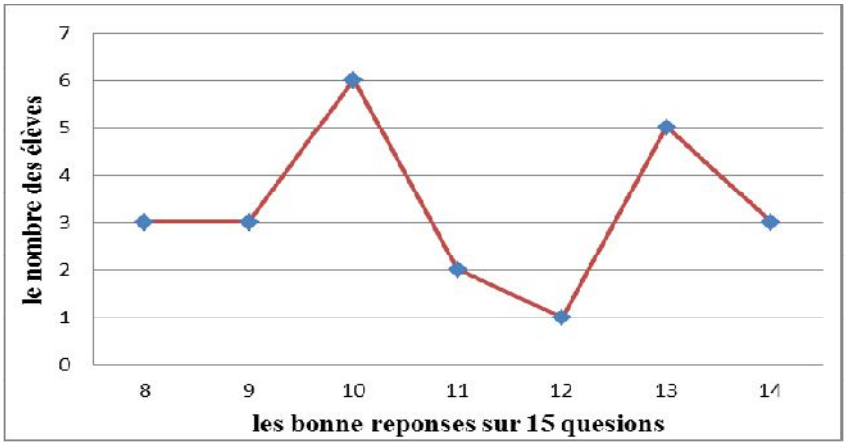


Fig.4 Les bonnes/mauvaises réponses par élèves dans le jeu sérieux GEO

Pendant le jeu, les paramètres du jeu ont été adaptés en fonction du niveau des élèves (débutants), Sur la figure 3, on montre les performances de chaque élèves individuel dans un cours traditionnel (salle de classe). On peut observer que les élèves montrent des réponses distinctes pour une durée de **45** minutes alors que sur la figure 4 on montre que les performances de chaque joueur dans le jeu (GEO). On peut dire que les résultats sont presque identiques par rapport au cours traditionnel mais avec une durée de **15** minutes seulement. Nous ne constatons aucune amélioration en termes de qualité d'apprentissage néanmoins la durée du cours a été divisée par 3.

Conclusion

Guider l'apprenant au cours de sa formation en lui proposant le parcours adapté à son profil est l'objectif principal de cette recherche. L'enjeu est de mobiliser le potentiel important des systèmes d'enseignements adaptatifs basé sur les profils des apprenants afin que les processus d'apprentissage et l'accès à la connaissance soient plus efficaces, plus solides, plus adaptés aux rapides évolutions auxquelles les personnes et les organisations doivent faire face.

References

1. Krassmann, AlianeLoureiro, Leo NatanPaschoal, AndressaFalcade, etRoseclea Duarte Medina. 2015.« Evaluation of Game-Based Learning Approaches through Digital Serious Games in Computer Science Higher Education: A Systematic Mapping In , 43‑51. IEEE.https://doi.org/10.1109/SBGames.2015.18.

2. Moh. Aries Syufagi, MauridhiHery P. and MochamadHariadi. 2011. «Modeling Serious Games based on Cognitive Skill Classification using Learning Vector Quantization with Petri Net »2011 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2011): Jakarta, Indonesia, 17 - 18 December 2011

3. Moh. Aries Syufagi, MauridhiHery P. and MochamadHariadi. 2011. «Modeling Serious Games based on Cognitive Skill Classification using Learning Vector Quantization with Petri Net »2011 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2011): Jakarta, Indonesia, 17 - 18 December 2011

4.Sabourin, Jennifer, Bradford Mott, et James C. Lester. 2011. « Modeling learner affect with theoretically grounded dynamic Bayesian networks ». In *International Conference on Affective Computing and Intelligent Interaction*, 286–295.Springer.

5. Andrade, Kleber de O., Guilherme Fernandes, Glauco A.P. Caurin, Adriano A.G. Siqueira, Roseli A.F. Romero, et Rogerio de L. Pereira. 2014. « Dynamic Player Modelling in Serious Games Applied to Rehabilitation Robotics ». In , 211‑16. IEEE. https://doi.org/10.1109/SBR.LARS.Robocontrol.2014.41.

6. Sebaha, Karim, etAarjMahmoodHussaan. 2014. « Architecture et modèles génériques pour la génération adaptative des scénarios de jeux sérieux. Application: Jeu d'évaluation et de rééducation cognitives ». *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation* 21 (1): 615–648.

7.Psaltis, Athanasios, KyriakiKaza, KiriakosStefanidis, Spyridon Thermos, Konstantinos C. Apostolakis, KosmasDimitropoulos, et Petros Daras. 2016. « Multimodal affective state recognition in serious games applications ». In *Imaging Systems and Techniques (IST), 2016 IEEE International Conference on*, 435–439. IEEE.

8.Grawemeyer, Beate, ManolisMavrikis, Wayne Holmes, Sergio Gutiérrez-Santos, Michael Wiedmann, etNikolRummel. 2017. « Affective Learning: Improving Engagement and Enhancing Learning with Affect-Aware Feedback ». *User Modeling and User-Adapted Interaction* 27 (1): 119‑58. https://doi.org/10.1007/s11257-017-9188-z.

Trees with unique minimum glolal offensive alliance sets

Mohamed Bouzefrane
Faculty of Technology, University of Médéa, Algeria
email: mohamedbouzefrane@gmail.com

Isma Bouchemakh
Faculty of Mathematics, Laboratory L'IFORCE,
University of Sciences and Technology Houari Boumediene (USTHB),
B.P. 32 El-Alia, Bab-Ezzouar, 16111 Algiers, Algeria
emails: isma_bouchemakh2001@yahoo.fr, ibouchemakh@usthb.dz

Mohamed Zamime
Faculty of Technology, University of Médéa, Algeria
email: zamimemohamed@yahoo.com

Noureddine Ikhlef-Eschouf
Faculty of Sciences, Department of Mathematics and Computer Science,
University of Médéa, Algeria
email: nour_echouf@yahoo.fr

Abstract

Let $G = (V, E)$ be a simple graph. A non-empty set $S \subseteq V$ is called a global offensive alliance if S is a dominating set and for every vertex v in $V - S$, at least half of the vertices from the closed neighborhood of v are in S. The global offensive alliance number is the minimum cardinality of a global offensive alliance in G. In this paper, we give a constructive characterization of trees having a unique minimum global offensive alliance.

Keywords: Domination, global offensive alliance.

1 Introduction

Throughout this paper, $G = (V, E)$ denotes a simple graph with vertex-set $V = V(G)$ and edge-set $E = E(G)$. Let G and H be two graphs with

two disjoint vertex sets. Their *disjoint union* is denoted by $G \cup H$, the disjoint union of k copies of G is denoted by kG and the disjoint union of a family of graphs $G_1, G_2, \ldots, G_k$ is denoted by $\cup_{i=1}^{k} G_i$. For every vertex $v \in V(G)$, the *open neighborhood* $N_G(v)$ is the set $\{u \in V(G) \mid uv \in E(G)\}$ and the *closed neighborhood* of v is the set $N_G[v] = N(v) \cup \{v\}$. The *degree* of a vertex $v \in V(G)$, denoted $d_G(v)$, is the size of its open neighborhood. A vertex of degree one is called a *leaf* and its neighbor is called a *support vertex*. If v is a support vertex of a tree T, then $L_T(v)$ will denote the set of the leaves attached at v. Let $L(T)$ and $S(T)$ denote the set of leaves and support vertices, respectively, in T, and let $|L(T)| = l(T)$. As usual, the *path* of order n is denoted by P_n, and the *star* of order n by $K_{1,n-1}$. A *double star* $S_{p,q}$ is obtained by attaching p leaves at an endvertex of a path P_2 and q leaves at the second one. A *subdivision* of an edge uv is obtained by introducing a new vertex w and replacing the edge uv with the edges uw and wv. A subdivided star denoted by SS_k is a star $K_{1,k}$ where each edge is subdivided exactly once. A *wounded spider* is a tree obtained from $K_{1,r}$, where $r \geq 1$, by subdividing at most $r-1$ of its edges. For a vertex v, let $C(v)$ and $D(v)$ denote the set of *children* and *descendants*, respectively, of v in a rooted tree T, and let $D[v] = D(v) \cup \{v\}$. The *maximal subtree* at v is the subtree of T induced by $D[v]$, and is denoted by T_v.

A *dominating set* of a graph G is a set D of vertices such that every vertex in $V - D$ is adjacent to some vertex in D. *The domination number* of G, denoted by $\gamma(G)$, is the minimum cardinality of a dominating set of G. The concept of domination in graphs, with its many variations, is now well studied in graph theory. For more details, see the books of Haynes, Hedetniemi, and Slater [19, 20].

Among the many variations of domination, we mention the concept of alliances in graphs that has been studied in recent years. Several types of alliances in graphs are introduced in [18], including the offensive alliance that we study here. A dominating set D with the property that for every vertex v not in D,

$$|N_G[v] \cap D| \geq |N_G[v] - D| \tag{1}$$

is called *global offensive alliance set of* G and abbreviated *GOA-set* of G. The *global offensive alliance number* $\gamma_o(G)$ is the minimum cardinality among all GOA-sets of G. A GOA-set of G of cardinality $\gamma_o(G)$ is called γ_o-set of G, or $\gamma_o(G)$-set. Several works have been carried out on global offensive alliances in graphs (see, for example, [2, 6], and elsewhere).

Graphs with unique minimum μ-set, where μ is a some graph parameter, is another concept to which much attention was given during the last

two decades. For example, graphs with unique minimum γ-set were first studied by Gunther et al. in [13]. Later this problem was studied for various classes of graphs including block graphs [7], cactus graphs [9], some cartesian product graphs [14] and some repeated cartesian products [15]. Several works on uniqueness related to other graph parameters have been widely studied, such as locating-domination number [1], paired-domination number [3], double domination number [4], roman domination number [5] and total domination number [17]. Further work on this topic can be found in [8, 10, 11, 12, 16, 21, 22, 23]

The aim of this paper is to characterize all trees having unique minimum global offensive alliance set. We denote such trees as *UGOA-trees*.

2 Preliminaries results

We give in this section the following observations. Some results are straightforward and so their proofs are omitted.

Observation 1 *Let T be a tree of order at least three and $u \in S(T)$. Then,*

- (i) *there is a $\gamma_o(T)$-set that contains all support vertices of T,*
- (ii) *if D is a unique $\gamma_o(T)$-set, then D contains all support vertices but no leaf,*
- (iii) *if $l_T(u) \geq 2$, then u belongs to any γ_o-set(T).*

Proof. (i) and (ii) are obvious. If (iii) is not satisfied, then all leaves attached at u would be contained in D, which is a contradiction with the minimality of D. ■

Observation 2 *Let T be a tree obtained from a nontrivial tree T' by joining a new vertex v at a support vertex u of T'. Let D and D' be $\gamma_o(T)$-sets of T and T', respectively. Then,*

- (i) $|D'| = |D|$,
- $ii)$ *$D \cap V(T')$ is a $\gamma_o(T')$-set,*
- (iii) *if T is a UGOA-tree such that u is in any $\gamma_o(T')$-set, then T' is a UGOA-tree.*

Proof. According to Observation 1 (*iii*), u must be in D since $l_T(u) \geq 2$.

i) D is clearly a GOA-set of T', and then $|D'| \leq |D|$. By Observation 1 (*i*), we can assume that $u \in D'$. Hence, D' can be extended to a GOA-set of T, which leads to $|D| \leq |D'|$. Thus equality holds.

ii) Since $D \cap V(T') = D$ is a GOA-set of T' with cardinality $|D| = |D'|$, we deduce that $D \cap V(T')$ is a $\gamma_o(T')$-set.

iii) Item (*i*) together with the fact that u belongs to any $\gamma_o(T')$ imply that D' can be extended to a $\gamma_o(T)$-set. Therefore, the uniqueness of D as a $\gamma_o(T)$-set leads to $D' = D$, which means that D' is the unique $\gamma_o(T')$. ■

Observation 3 *Let T be a tree obtained from a nontrivial tree T' different from P_2 by joining the center vertex y of the path $P_3 = x$-y-z at a support vertex v of T'. Let D and D' be $\gamma_o(T)$-sets of T and T', respectively such that each of them contains all support vertices. Then,*

(*i*) $|D'| = |D| - 1$,

(*ii*) $D \cap V(T')$ *is a* $\gamma_o(T')$*-set,*

(*iii*) *if T is a UGOA-tree, then T' is a UGOA-tree.*

Proof. *i*) Since $y \in D$ and $v \in D \cap D'$, it follows that $D - \{y\}$ is a GOA-set of T' and so $|D'| \leq |D| - 1$. Moreover, since $v \in D'$, D' can be extended to a GOA-set of T by adding y. Then $|D| \leq |D' \cup \{y\}| = |D'| + 1$ and equality holds.

ii) Since $D \cap V(T') = D - \{y\}$ is a GOA-set of T' with cardinality $|D| - 1 = |D'|$, $D \cap V(T')$ is a $\gamma_o(T')$-set.

iii) Let $B = \{y\}$. In view of item (*i*), D' can be extended to a $\gamma_o(T)$-set by adding the unique vertex of B. This and item (*ii*) together with the uniqueness of D imply that $D' = D \cap V(T')$ is the unique γ_o-set of T'. ■

Observation 4 *Let k be a positive integer and let T be a tree obtained from a nontrivial tree T' by adding kP_2 joining k pairwise non-adjacent vertices of kP_2 to the same leaf v of T'. Let w be the support vertex adjacent to v, and let D and D' be $\gamma_o(T)$-sets of T and T', respectively. If $w \in D \cap D'$, then the following three properties are satisfies.*

(*i*) $|D'| = |D| - k$,

(*ii*) $D \cap V(T')$ *is a* $\gamma_o(T')$*-set,*

(iii) if T is a UGOA-tree, then T' is a UGOA-tree.

Proof. Let $V(kP_2) = \{x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_k\}$ and $E(kP_2) = \{x_iy_i : i = 1, 2, \ldots, k\}$. Let v be a leaf of T' and w be the support vertex adjacent to v. We assume that for each $i \in \{1, \ldots, k\}$, y_i is adjacent to v in T.

i) Obviously, all vertices of $\cup_{j=1}^{k} \{y_j\}$ are support vertices in T. Hence, in view of Observation 1 (i), we can assume that D contains all vertices of $\cup_{j=1}^{k} \{y_j\}$. Therefore, since $w \in D$, $D - (\cup_{j=1}^{k} \{y_j\})$ is a GOA-set of T', which means that $|D'| \leq \left|D - (\cup_{j=1}^{k} \{y_j\})\right| = |D| - k$. Observe that since $w \in D'$, D' can be extended to a GOA-set of T by adding all vertices of $\cup_{j=1}^{k} \{y_j\}$. Hence $|D| \leq \left|D' \cup (\cup_{j=1}^{k} \{y_j\})\right| = |D'| + k$ and so equality holds.

ii) The proof is similar to that of Observation 3(ii), by taking $D \cap V(T') = D - (\cup_{j=1}^{k} \{y_j\})$.

iii) The proof is similar to that of (iii) of Observation 3(iii), by taking $B = \cup_{j=1}^{p} \{y_j\}$. ■

Observation 5 *Let $V(T')$ be the vertex-set of a nontrivial tree T', and let D' be a $\gamma_o(T')$-set such $V(T') - D'$ has a vertex w with degree $q \geq 2$ and $|N_{T'}(w) \cap (V(T') - D')| \leq 1$. Let p be a positive integer such that*

$$\begin{cases} p \leq q - 1 \text{ if } |N_{T'}(w) \cap (V(T') - D')| = 0, \\ \text{or} \\ p \leq q - 3 \text{ if } |N_{T'}(w) \cap (V(T') - D')| = 1. \end{cases} \tag{2}$$

Let T be a tree obtained from T' by adding p subdivided stars $SS_{k_1}, \ldots, SS_{k_p}$ ($k_i \geq 2$ for all i) with centers x_1, $x_2, \ldots, x_p$, respectively, and joining each x_i $(1 \leq i \leq p)$ at w. Let D be a γ_o-set of T. If w and $x_1, x_2, \ldots, x_p$ are not in D, then the following three properties are satisfied.

(i) $|D'| = |D| - \sum_{i=1}^{p} k_i$,

(ii) $D \cap V(T')$ *is a* $\gamma_o(T')$*-set,*

(iii) if T is a UGOA-tree, then T' is also a UGOA-tree.

Proof. For $i \in \{1, \ldots, p\}$, let $S(SS_{k_i})$ be a support vertex-set of SS_{k_i}.

i) Since w together with $x_1, x_2, \ldots, x_p$ are not in D, all vertices of $\cup_{i=1}^{p} S(SS_{k_i})$ must be in D. Therefore, $D \setminus \bigcup_{i=1}^{p} S(SS_{k_i})$ is a GOA-set of T', giving that

$|D'| \leq |D| - \sum_{i=1}^{p} k_i$.

On the other hand, let $A = \cup_{i=1}^{p} S\left(SS_{k_i}\right) \cup D'$. We have to show that A is a GOA-set of T. For this, it suffices to show that $|N_T[z] \cap A| \geq |N_T[z] - A|$ for each $z \in \{w, x_1, x_2, \ldots, x_p\}$. Indeed, we have to distinguish between two cases.

Case 1. $z = x_i$, for some $i \in \{1, \ldots, p\}$.
We have then

$$|N_T[z] \cap A| = |N_T[z] \cap \cup_{i=1}^{p} S\left(SS_{k_i}\right)| = k_i \geq 2,$$

and

$$|N_T[z] - A| = |\{z, w\}| = 2.$$

Case 2. $z = w$.
We have then

$$|N_T[z] \cap A| = \begin{cases} q & \text{if } |N_{T'}(w) \cap (V(T') - D')| = 0, \\ q-1 & \text{if } |N_{T'}(w) \cap (V(T') - D')| = 1. \end{cases}$$

and

$$|N_T[z] - A| = \begin{cases} p+1 & \text{if } |N_{T'}(w) \cap (V(T') - D')| = 0, \\ p+2 & \text{if } |N_{T'}(w) \cap (V(T') - D')| = 1. \end{cases}$$

According to (2), we have in each case $|N_T[z] \cap A| \geq |N_T[z] - A|$ for each $z \in \{w, x_1, x_2, \ldots, x_p\}$. Therefore A is a GOA-set of T, giving that $|D| \leq |A| = |D'| + \sum_{i=1}^{p} k_i$. Hence the equality holds.

ii) Using the fact that $D \cap V(T') = D \setminus \cup_{i=1}^{p} S\left(SS_{k_i}\right)$, this property follows in a similar manner as the proof of Observation 3(*ii*).

(*iii*) This property follows in a similar manner as the proof of Observation 3(*iii*), by taking $B = \cup_{i=1}^{p} S\left(SS_{k_i}\right)$. ■

3 The main result

In order to characterize the trees with unique minimum global offensive alliance, we define a family $\mathcal{F}$ of all trees T that can be obtained from a sequence $T_1, T_2, \ldots, T_r$ $(r \geq 1)$ of trees, where T_1 is the path P_3 centered at a vertex y, $T = T_r$, and if $r \geq 2$, T_{i+1} is obtained recursively fom T_i by one of the following operations. Let $A(T_1) = \{y\}$.

- Operation $\mathcal{O}_1$: Attach a vertex by joining it to any support vertex of T_i. Let $A(T_{i+1}) = A(T_i)$.

- Operation $\mathcal{O}_2$: Attach a path $P_3 = u$-v-w by joining v to any support vertex of T_i. Let $A(T_{i+1}) = A(T_i) \cup \{v\}$.

- Operation $\mathcal{O}_3$: Let w be a support vertex of T_i that satisfies one of the following two conditions.

  1. $l_{T_i}(w) \geq 3$,
  2. $|N_{T_i}[w] \cap A(T_i)| < |N_{T_i}(w) \cap (V(T_i) - A(T_i)|$ or
     - either $l_{T_i}(w) = 2$ and $N_{T_i}(w) - A(T_i)$ has a vertex w_t such that $|N_{T_i}(w_t) \cap A(T_i)| \leq |N_{T_i}[w_t] \cap (V(T_i) - A(T_i)| + 1$,
     - or $l_{T_i}(w) = 1$ and $N_{T_i}(w) - A(T_i)$ has two vertices w_p, w_q so that for $l = p, q$, $|N_{T_i}(w_l) \cap A(T_i)| \leq |N_{T_i}[w_l] \cap (V(T_i) - A(T_i)| +$ 1.

Let kP_2 be the disjoint union of $k \geq 1$ copies of P_2, and let B be a set of k pairwise non-adjacent vertices of kP_2. Add kP_2 and attach all vertices of B to a same leaf in T_i that is adjacent to w. Let $A(T_{i+1}) = A(T_i) \cup B$.

- Operation $\mathcal{O}_4$: Let $w \in V(T_i) - A(T_i)$ be a vertex of degree $q \geq 2$ in T_i such that $|N_{T_i}(w) \cap (V(T_i) - A(T_i))| \leq 1$. Attach $p \geq 1$ subdivided stars SS_{k_i} ($k_i \geq 2$ for $1 \leq i \leq p$) with support vertex-set $S(SS_{k_i})$ and of center x_i by joining x_i to w for all i such that

$$p \leq \begin{cases} q-1 & \text{if } |N_{T_i}(w) \cap (V(T_i) - A(T_i))| = 0, \\ q-3 & \text{if } |N_{T_i}(w) \cap (V(T_i) - A(T_i))| = 1. \end{cases}$$

Let $A(T_{i+1}) = A(T_i) \cup (\cup_{i=1}^{p} S(SS_{k_i}))$.

Before stating our main result, we need the following lemma.

Lemma 6 *If $T \in \mathcal{F}$, then $A(T)$ is the unique $\gamma_o(T)$-set.*

Proof. Let $T \in \mathcal{F}$. We proceed by induction on the number of operations, say r, required to construct T. The property is true if T is a path P_3 centered at y since $A(T) = \{y\}$ is the unique $\gamma_o(T)$-set. This establishes the base case.
Assume that for any tree $T' \in \mathcal{F}$ that can be constructed with $r-1$ operations, $A(T')$ is the unique $\gamma_o(T')$-set. Let $T = T_r$ with $r \geq 2$ and $T' = T_{r-1}$.

We distinguish between four cases.

Case 1. T is obtained from T' by using Operation $\mathcal{O}_1$.
Assume that T is obtained from T' by attaching an extra vertex at a support vertex u of T'. In view of Observation 1 (ii), $u \in A(T')$. Hence $A(T')$ can be extended to a GOA-set of T. By Observation 2 (i), $\gamma_o(T) = \gamma_o(T')$, implying that $A(T')$ is a $\gamma_o(T)$-set. Applying the inductive hypothesis to T', $A(T')$ is the unique $\gamma_o(T')$-set. It follows that $A(T) = A(T')$ is the unique $\gamma_o(T)$-set.

Case 2. T is obtained from T' by using Operation $\mathcal{O}_2$.
$A(T') \cup \{v\}$ is a GOA-set of T. By Observation 3 (i), $\gamma_o(T) = \gamma_o(T') + 1$, meaning that $A(T') \cup \{v\}$ is a $\gamma_o(T)$-set. The inductive hypothesis sets that $A(T')$ is the unique $\gamma_o(T')$-set. Thus $A(T) = A(T') \cup \{v\}$ is the unique $\gamma_o(T)$-set.

Case 3. T is obtained from T' by using Operation $\mathcal{O}_3$.
$A(T') \cup B$ is a GOA-set of T. Observation 4 (i) sets that $\gamma_o(T) = \gamma_o(T') + k$, which means that $A(T') \cup B$ is a $\gamma_o(T)$-set. By the inductive hypothesis, $A(T')$ is the unique $\gamma_o(T')$-set. Thus $A(T) = A(T') \cup B$ is the unique $\gamma_o(T)$-set.

Case 4. T is obtained from T' by using Operation $\mathcal{O}_4$.
$A(T') \cup (\cup_{i=1}^{p} S(SS_{k_i}))$ is a GOA-set of T. According to Observation 5 (i), we have $\gamma_o(T) = \gamma_o(T') + \sum_{i=1}^{p} k_i$, whence, $A(T') \cup (\cup_{i=1}^{p} S(SS_{k_i}))$ is a $\gamma_o(T)$-set. By the inductive hypothesis, $A(T')$ is the unique $\gamma_o(T')$-set. It follows that $A(T) = A(T') \cup (\cup_{i=1}^{p} S(SS_{k_i}))$ is the unique $\gamma_o(T)$-set. ■

Remark that in each case, $A(T_{i+1})$ is obtained from $A(T_i)$ by adding all support vertices in $T_{i+1} \backslash T_i$. Hence the following corollary is immediate.

Corollary 7 *Let $T \in \mathcal{F}$ and $S(T)$ be a set of support vertices in T. Then $\gamma_o(T) \geqslant |S(T)|$.*

Now we are ready to prove our main result.

Theorem 8 *A tree T is a UGOA-tree if and only if $T = K_1$ or $T \in \mathcal{F}$.*

Proof. It is obvious that $T = K_1$ is a UGOA-tree. Also, Lemma 6 states that any member of $\mathcal{F}$ is a UGOA-tree. Now, we prove the converse by induction on the number n of vertices of T. The converse holds trivially for $n = 1$ and 3 but not for $n = 2$ since P_2 is not a UGOA-tree. When $n = 4$, T is either a $K_{1,3}$ or a P_4. Clearly P_4 is not a UGOA-tree, whilst

$K_{1,3}$ is a UGOA-tree that can be obtained from a P_3 using operation $\mathcal{O}_1$, and so $K_{1,3} \in \mathcal{F}$. If $n = 5$, then T is either a double star $S_{1,2}$ which is not a UGOA-tree, or it is a $K_{1,4}$ or P_5 that are UGOA-tree since $K_{1,4}$ can be obtained from $K_{1,3}$ by using operation $\mathcal{O}_1$, and P_5 can be obtained from a P_3 by using operation $\mathcal{O}_3$. Therefore $K_{1,4}$ and P_5 are in $\mathcal{F}$. This establishes the base case.

Now, let $n \geq 6$ and assume that any tree T' of order $3 \leq n' < n$ with the unique $\gamma_o(T')$-set is in $\mathcal{F}$. Let T be a tree of order n with the unique $\gamma_o(T)$-set D and let $s \in S(T)$. By Observation 1 (ii), $s \in D$. If $l_T(s) \geq 3$, then let T' be the tree obtained from T by removing a leaf adjacent to s and let D' be a $\gamma_o(T')$-set. Then, clearly $n' = |V(T')| = n - 1 \geq 5$, and $l_{T'}(s) \geq 2$, so $s \in D'$ by Observation 1 (iii). According to Observation 2 (ii), T' is UGOA-tree. Applying the inductive hypothesis to T', we get $T' \in \mathcal{F}$. Thus T is obtained from T' by operation $\mathcal{O}_1$, implying that $T \in \mathcal{F}$. Assume now that

$$\text{for each } x \in S(T),\ l_T(x) \leq 2. \tag{3}$$

Root T at a vertex r of maximum eccentricity. Let u be a support vertex of maximum distance from r and let u' be a leaf adjacent to u. Let v and w be the parents of u and v, respectively, in the rooted tree. We consider two cases.

Case 1. $v \in D$.
If $l_T(u) = 1$, then $D \cup \{u'\} - \{u\}$ is a $\gamma_o(T)$-set, contradicting the uniqueness of D as a $\gamma_o(T)$-set. Hence by (3), $l_T(u) = 2$. We claim that $v \in S(T)$. Suppose not. Then either $w \in D$ and so $D - \{v\}$ is a GOA-set of T with cardinality less than $|D|$, contradicting the minimality of D, or $w \notin D$ and so $D - \{v\} \cup \{w\}$ is a $\gamma_o(T)$-set, contradicting the uniqueness of D as a $\gamma_o(T)$-set. This completes the proof of the claim. Let $T' = T - T_u$ and D' be a γ_o-set of T'. By Observation $1(i)$, we can assume that D' contains all support vertices in T'. Since $|V(T_u)| = 3$, it follows that $n' = |V(T')| = n - 3 \geq 3$ and so $T' \neq P_2$. By Observation $3(iii)$, T' is a UGOA-tree. Applying our inductive hypothesis, we get $T' \in \mathcal{F}$. Thus, T can be obtained from T' by operation $\mathcal{O}_2$ and so $T \in \mathcal{F}$.

Case 2. $v \notin D$.
According to Observation $1(ii)$, $v \notin S(T)$ and so $l_T(v) = 0$. Let $k = |N_T(v) - \{w\}|$. We have then $d_T(v) = k + 1$ and since $u \in N_T(v) - \{w\}$, we clearly deduce $k \geq 1$. For $i \in \{1, \ldots, k\}$, let $u_i \in N_T(v) - \{w\}$ such that $u_1 = u$. The choice of v sets that

$$u_i \in S(T),\ l_T(u_i) \geq 1 \text{ and so } u_i \in D \text{ for all } i. \tag{4}$$

Hence by (3), we have $1 \leq l_T(u_i) \leq 2$ for all i. Assume first that $l_T(u_j) = 2$ for some j in $\{1, \ldots, k\}$. Without loss of generality, let $j = 1$. Then u has a further neighbor $u'' \neq u'$ in T. Let $T' = T - \{u''\}$ and D' be any γ_o-set of T'. Clearly u' is the unique leaf of u in T'. We claim that $u \in D'$. Suppose not. Then u' and v must be in D' and therefore $D'' = (D' \backslash \{u'\}) \cup \{u\}$ is a further $\gamma_o(T)$-set other than D (since v belongs to D'' and not to D), a contradiction. This completes the proof of the claim. We have $n' = n-1 \geq 5$. By Observation 2(iii), T' is a UGOA-tree. Applying our inductive hypothesis to T', we get $T' \in \mathcal{F}$. Hence T is obtained from T' by operation $\mathcal{O}_1$, implying that $T \in \mathcal{F}$. Assume now that

$$l_T(u_i) = 1 \text{ and hence } d_T(u_i) = 2 \text{ for all } i. \tag{5}$$

For all $i \in \{1, \ldots, k\}$, let u'_i be the unique leaf adjacent to u_i (with $u'_1 = u'$). We distinguish between two subcases, depending on whether w belongs to D or not.

Case 2.1. $w \in D$.
In view of (5), $T_v - \{v\} = kP_2$ with $V(kP_2) = \{u_1, u_2, \ldots, u_k, u'_1, u'_2, \ldots, u'_k\}$ and $E(kP_2) = \{u_i u'_i : i = 1, 2, \ldots, k\}$. Let $T' = T - (T_v - \{v\})$. Clearly $v \in L(T')$ and $w \in S(T')$. If $n' = |V(T')| = 2$, then T is a wounded spider with exactly one non-subdivided edge and in this case, it is not difficult to see that such a graph is not a UGOA-tree. Hence assume that $n' \geq 3$. We claim the following:
If $l_T(w) \in \{0, 1\}$, then one of the two conditions holds:

C_1 : $|N_T[w] \cap D| \leq |N_T(w) \cap (V(T) - D|$.

C_2 : (i) either $l_T(w) = 1$ and $N_T(w) - D$ has a vertex w_t such that

$$|N_T(w_t) \cap D| \leq |N_T[w_t] \cap (V(T) - D)| + 1$$

(ii) or, $l_T(w) = 0$ and $N_T(w) - D$ has two vertices w_p, w_q such that for $l \in \{p, q\}$,

$$|N_T(w_l) \cap D| \leq |N_T[w_l] \cap (V(T) - D)| + 1.$$

Indeed, suppose that C_1 and C_2 are not satisfied. Assume first that $l_T(w) = 1$, so $L_T(w)$ has exactly one vertex, say w'. In this case $D - \{w\} \cup \{w'\}$ is a $\gamma_o(T)$-set different from D, a contradiction. Now, assume that $l_T(w) = 0$. Since C_2 is not fulfilled, item (ii) of C_2 is satisfied for at most one vertex

in $N_T(w) - D$, say w''. Then $D - \{w\} \cup \{w''\}$ is a $\gamma_o(T)$-set different from D, a contradiction. If no vertex in $N_T(w) - D$ for which item (ii) of C_2 is satisfied, then $D - \{w\} \cup \{v\}$ is a $\gamma_o(T)$-set different from D, which leads to a contradiction again. This complete the proof of the claim.
Observe that when $l_{T'}(w) \in \{1, 2\}$, the previous claim remain true by replacing D by D' and T by T'. Thus, according to Observation 4 (iii), T' is a UGOA-tree. By induction on T', we get $T' \in \mathcal{F}$. Since T is obtained from T' by using operation $\mathcal{O}_3$, we directly obtain $T \in \mathcal{F}$.

Case 2.2. $w \notin D$.
By Observation 1(ii), $w \notin S(T)$ and so $l_T(w) = 0$. Since v and w are in $V(T) - D$, v must have at least two neighbors in D. Hence $d_T(v) = k+1 \geq 3$. Let t be the parent of w, and let X, Y and Z be the following sets

$$Y = C(w) \cap S(T),\ X = C(w) - Y \text{ and } Z = D(w) \cap (S(T) - Y).$$

Observe that $v \in X$, $u \in Z$, $N_T(w) = \{t\} \cup X \cup Y$ and every vertex in Z plays the same role as u. Therefore by (4), we have $Z \subset D$ since $Z \subset S(T)$, and by (5), every vertex in Z has exactly two neighbors such that one of them is a leaf and the other one is in X. Furthermore, as $v \in X$, $u_i \in Z$ for all $i \in \{1, \ldots, k\}$, so $|Z| \geq k \geq 2$. Notice also that $|X| \geq 1$ since $v \in X$. Likewise $|Y| \geq 1$ since D is a $\gamma_o(T)$-set. It is clear that $Y \subseteq S(T)$ and thus $Y \subseteq D$ by Observation 1(ii). Setting

$$X = \{x_1, x_2, \ldots x_p\} (p \geq 1) \text{ with } x_1 = v \text{ and } |Y| = q - 1\ (q \geq 2).$$

Since every vertex in X plays the same role as v, $x_i \in V(T) - D$ for all $i \in \{1, \ldots, p\}$. Setting

$$p_i = |N_T(x_i) - \{w\}| \text{ for } i = 1, \ldots, p.$$

Then $p_1 = k$. Since for all $i \in \{1, \ldots, p\}$, x_i and w are in $V(T) - D$, x_i must have at least two neighbors in Z. Hence $d_T(x_i) = p_i + 1 \geq 3$. This means that for all $i \in \{1, \ldots, p\}$, $V(T_{x_i})$ induces a subdivided star SS_{p_i} of order $p_i + 1$ centered at x_i. Since $w \in V(T) - D$, inequality (1) is valid by replacing v with w. This gives

$$p \leq q - 1 \text{ if } t \in D, \text{ or } p \leq q - 3 \text{ otherwise.} \tag{6}$$

Let $T' = T - (\cup_{i=1}^{p} T_{x_i})$ and D' be a $\gamma_o(T')$-set. Observe that T' contains at least one P_3 as an induced subgraph, which means that $n' = |V(T')| \geq 3$.

For all $i \in \{1, \ldots, p\}$, let $S(SS_{p_i})$ be the support vertex-set of SS_{p_i}. Clearly $\cup_{i=1}^{p} S(SS_{p_i}) = Z$ and $N_{T'}(w) = Y \cup \{t\}$, so

$$d_{T'}(w) = q \geq 2.$$

According to Observation 1 (i), we can assume that $Y \subset D'$ since $Y \subset S(T')$. Then t is the only neighbor of w in T' that may not be in D', that is

$$\left|N_{T'}(w) \cap (V(T') - D')\right| \leq 1.$$

If $t \in D'$, then the minimality of D' sets that $w \in V(T') - D'$, because otherwise, we replace w by t in D'.
By Observation 5 (ii) and (iii), we have $D' = D \cap V(T')$. Hence $t \in D$ if and only if $t \in D'$. Notice that if $t \in D'$, then $N_{T'}(w) \cap (V(T') - D')$ is an empty-set, otherwise, t would be the unique vertex of $N_{T'}(w) \cap (V(T') - D')$. Thus (6) can be rewritten as follows.

$$\text{If } \left|N_{T'}(w) \cap (V(T') - D')\right| = 0, \text{ then } p \leq q - 1,$$

and

$$\text{if } \left|N_{T'}(w) \cap (V(T') - D')\right| = 1, \text{ then } p \leq q - 3.$$

Again Observation 5(iii) sets that T' is a UGOA-tree. Applying the inductive hypothesis to T', we deduce $T' \in \mathcal{F}$. Now since T can be obtained from T' by operation $\mathcal{O}_4$, and finally $T \in \mathcal{F}$. This completes the proof of Theorem 8. ■

4 Open Problems

The previous results motivate the following problems.

1- Characterize other UGOA-graphs.

2- Characterize trees with unique minimum defensive alliance sets (UGDA).

References

[1] M. Blidia, M. Chellali, R. Lounes and F. Maffray, Characterizations of trees with unique minimum locating-dominating sets, J. Combin. Math. Combin. Comput.76 (2011) 2011, 225-232.

[2] M. Bouzefrane, M. Chellali, On the global offensive alliance number of a tree, Opuscula Math. 29 (2009), 223-228.

[3] M. Chellali and T.W Haynes, Trees with unique minimum paired domination sets. Ars Comb. 73 (2004) 3-12.

[4] M. Chellali and T.W Haynes, A characterization of trees with unique minimum double domination sets, Util. Math., 83 (2010) 233-242.

[5] M. Chellali and N.J. Rad, Trees with unique Roman dominating functions of minimum weight, Discrete Math. Algorithm. Appl. 06, 1450038 (2014).

[6] M. Chellali, L. Volkmann. Independence and global offensive alliance in graphs, Australas. J. Combin., 47 (2010) 125-131.

[7] M. Fischermann, Block graphs with unique minimum dominating sets, Discrete Math. 240 (1-3) (2001), 247-251.

[8] M. Fischermann, D. Rautenbach and L. Volkmann, Maximum graphs with a unique minimum dominating set, Discrete Math. 260 (1-3) (2003), 197-203.

[9] M. Fischermann and L. Volkmann, Cactus graphs with unique minimum dominating sets, Util. Math. 63 (2003), 229-38.

[10] M. Fischermann and L. Volkmann, Unique independence, upper domination and upper irredundance in graphs, J. Combin. Math. Combin. Comput. 47 (2003), 237-249.

[11] M. Fischermann, L. Volkmann and I. Zverovich, Unique irredundance, domination, and independent domination in graphs, Discrete Math. 305 (1-3) (2005), 190-200.

[12] M. Fraboni and N. Shank, Maximum graphs with unique minimum dominating set of size two, Australas. J. Combin. 46 (2010), 91-99.

[13] G. Gunther, B. Hartnell, L. Markus and D. Rall, Graphs with unique minimum dominating sets, in: Proc. 25th S.E. Int. Conf. Combin., Graph Theory, and Computing, Congr. Numer. 101 (1994), 55-63.

[14] J. Hedetniemi, On unique minimum dominating sets in some cartesian product graphs, Discuss. Math. Graph Theory 34 (4) (2015), 615-628.

[15] J. Hedetniemi, On unique minimum dominating sets in some repeated cartesian products, Australas. J. Combin. 62 (2015), 91-99.

[16] J. Hedetniemi, On unique realizations of domination chain parameters, J. Combin. Math. Combin. Comput. 101 (2017), 193-211.

[17] T. W. Haynes and M. A. Henning, Trees with unique minimum total dominating sets. Discuss. Math. Graph Theory 22 (2002) 233-246.

[18] S.M. Hedetniemi, S. T. Hedetniemi, and P. Kristiansen, Alliance in graphs. J. Comb. Math. Combin. Comput. 48 (2004) 157-177.

[19] Haynes T W, Hedetniemi S T & Slater P J, 1998, Fundamentals of Domination in graphs, Marcel Dekker, New York.

[20] Haynes T W, Hedetniemi S T & Slater P J, (1998) (Eds.), Domination in graphs: Advanced Topics, Marcel Dekker, New York, 1998.

[21] G. Hopkins and W. Staton, Graphs with unique maximum independent sets, Discrete Math. 57 (1985) 245-251.

[22] W. Siemes, J. Topp and L. Volkmann, On unique independent sets in graphs, Discrete Math. 131 (1-3) (1994), 279-285.

[23] J. Topp, Graphs with unique minimum edge dominating sets and graphs with unique maximum independent sets of vertices, Discrete Math. 121 (1-3) (1993), 199-210.

A Logic-Based Approach to Reconstruct Web Services Protocols

Ali Khebizi[1] and Hassina Seridi[2]

[1] Department of Computer Science, 8 May 1945 University, P.O. Box 401, 24000 Guelma, -Algeria-, LabSTIC Laboratory, 8 May 1945 University, P.O. Box 401, 24000 Guelma, Algeria- khebizi.ali@univ-guelma.dz, ali.khebizi@gmail.com
[2] LabGed, Badji Mokhtar University, Annaba -Algeria- seridi@labged.net

Abstract. We investigate the problem of web services protocols reconstruction from a control flow perspective, *i.e., how to extract protocols specifications from execution traces ?*
We propose a comprehensive logic-based approach to reconstruct business protocol models of deployed web services. Recorded execution traces are transcribed to a facts base and a set of structure inference patterns is translated to corresponding production rules. The conceived knowledge base is explored by a reasoning engine to infer the various elements describing the finite state machine representing the target service protocol and thus, capturing the behavior contained in the execution log files.

Keywords: Web service, Service protocol, Protocol reconstruction, Execution traces, Inference patterns, Logic programming.

1 Introduction

Over the last few years web services are becoming the dominant technology for integrating distributed and heterogeneous information systems.

In the web service ecosystem, two elements are fundamental for providing a high interactivity level between service providers and service requesters. The first element is the service interface described via the WSDL standard which expresses the service localization and the allowed operations with their signature. The second one is the service protocol (*business protocol*) which reflects the provider's business process logic. A Service protocol is an abstract tool which describes the service external behavior by expressing constraints governing the operations invocation such as order and temporal constraints [4, 16].

Web services research literature has largely highlighted the usefulness of specifying service protocols and various models are proposed for their representation [4, 5]. Generally, service protocols are used to check the compatibility of customer's protocols with published ones and to verify their conformance with standardized processes. Also, they are inevitable during services composition. Thus, they constitute a cornerstone for the web services technology during the entire life-cycle: *i.e., design, enactment, management and analysis.*

Despite the importance of service protocols, various reasons can lead to their unavailability, such as: rapid service deployment, migrating legacy systems, automatic generation of the *WSDL file*... Furthermore, it frequently happens to have an obsolete service protocol version that does not reflect an evolved business logic. In such contexts, the service protocol must be reconstructed from executions logs, enhanced and advertised in the web services registries.

The problem of reconstructing business protocols basing on execution traces is well known and it raises several difficulties of different natures. Most of the related issues have been extensively investigated in the research literature and several works have addressed different facets of the problem [1, 2, 8, 9, 12, 14, 15].

In this paper, a fundamental paradigm shift is suggested to reconstruct service protocols from execution traces. The salient feature of the proposed work lies in a framework based on a declarative approach to support service providers in defining fine-grained protocol reconstruction by customizing a set of high-level abstractions. In the proposed approach, execution traces are transcribed to a facts base formalized in the first-order logic predicates and a set of **structures inference patterns** is used as a rules base for inferring various elements of the target protocol to be discovered.

Paper organization: We start by discussing related works in section 2. Section 3 presents the different facets of the proposed logic-based approach. In section 4, the system architecture is illustrated and the implementation and experimentation of a prototype based on the use of Prolog is exposed. Finally, conclusion and potential directions for future works are drawn at section 5.

2 Related Works

Recent literature is very rich in approaches that have addressed the various challenges related to the issue of business protocol discovery. Hereafter, we discuss different works that have coped with the problem from different perspectives.

- In [1–3], an approach based on establishing a causality relation in the set of execution traces is suggested to discover the corresponding process or workflow models. The proposed α-algorithm [1, 2] uses a rules set to specify relations between activities and a Petri net with special properties (*workflow nets*) is extracted from event logs. As the α-algorithm can't deal with short loops of length one and two, it was enhanced to α^{+}, α^{++} [13]. But, these last ones have residual problems with complex control-flow.

- In [14, 15], the authors propose efficient algorithms to deal with the problem of event correlation in service-based processes and characterize the set of events in service logs belonging to the same instance of a process. Such an approach is complementary to our work and it could be deployed during preliminary steps when the execution traces are not characterized by an unique identifier.

- The work presented in [8] proposes a formal framework to discover implicit timed transitions of conversation protocols. The concept of *implicit transitions* was formalized to express activities that can be triggered under time constraints.

In this approach, the working hypotheses are very restrictive and many conditions are imposed to model the associated constraints (*complete traces, no timed transitions leading to final states, no noise in logs, ...*).

- In [19], the authors introduce an heuristics driven process mining algorithm to discover the main behavior registered in an event log. The frequency of activities is used as a base to construct a dependency graph basing on a metric which indicates the degree of dependence between two events. This approach is applicable only to specific data having no too many different events.

- Other significant algorithms for business process discovery have been proposed recently (fuzzy [9], genetic [12]). In [10], a technique is suggested to evaluate these algorithms efficiently, and thus, to allow business managers selecting the appropriate algorithm that is most suitable for a given data-set.

3 Approach for Extracting Service Protocols from Traces

We present below the different facets of our logic-based approach and we expose the underlying models. *(i)* First, we introduce the explicit choices on formal models used for representing service protocols and execution traces. *(ii)* The components of the used knowledge base are addressed at a conceptual level by formalizing execution traces and inference patterns. *(iii)* A reasoning engine is deployed to infer the target service protocol from the conceived knowledge base.

In what follows, these steps are deeply discussed and illustrated.

3.1 Modeling protocols and traces

We introduce below the formal protocol model manipulated throughout the approach and we describe execution traces specification.

3.1.1 The service protocol model A service business protocol, (shortly *the service protocol*) describes the external visible behaviors of a given web service by specifying the constraints (*e.g., ordering of messages, ...*) that customers must comply with in order to correctly interact with the service [4, 5].

While various models of different levels of expressiveness have been proposed in the literature to capture different kinds of abstractions of service protocols, in our work a basic version of the model basing on automaton is used. It describes the ordering constraints that govern the activities' execution [4, 6]. The choice of finite state machines is motivated by the important role of this formalism to represent the behavior of dynamic systems and to support formal analysis of business processes [11]. In the other hand, other existing models such UML diagrams, PetriNets and BPMN diagrams can be translated to such models by using adequate transformation techniques [7, 18].

Definition 1 *A (web service) business protocol is a tuple* $\mathcal{P} = (S, s_0, \mathcal{F}, \mathcal{M}, \mathcal{R})$*; where:* S *is a finite set of states;* $s_0 \in S$ *is the initial state of the protocol;* $\mathcal{F} \subseteq S$ *is the set of final states;* $\mathcal{M}$ *is a finite set of abstract activities;* $\mathcal{R} \subseteq S \times \mathcal{M} \times S$

is a transition relation. Each element $(s, m, s') \in \mathcal{R}$ *represents a transition from a source state* s *to a target one* s' *upon the execution of the abstract activity* m*.*

According to this definition, states represent the different phases that a service may go through while transitions represent activities that a service can perform to move from one step to another [5].

Example 1. A real-world example of a retirement service protocol is depicted in **Fig. 1**. The protocol could be deployed as a web service by a nationwide network of social security centers for managing citizen's applications for pension benefits.

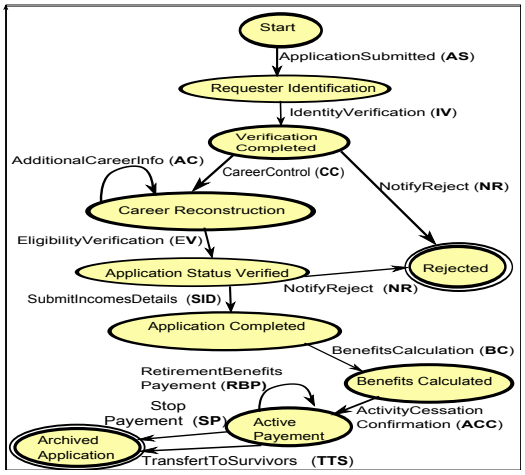


Fig. 1. The retirement service protocol

In such a protocol, state names (e.g. Rejected) are meaningless symbols that do not affect the operational usage of the service. The transitions' labels are meaningful and correspond to the executed activities. In the sequel, the names of the activities are abbreviated as indicated in the figure (*e.g., the activity ApplicationSubmitted is abbreviated as AS*).

It should be noted that the presented service protocol is in fact an over simplified version of a real life retirement business processes but which is however, sufficient to illustrate our approach.

3.1.2 Formalizing execution traces Each invocation of the web service by a particular customer corresponds to an execution of a separated instance of the service that generates an execution trace. We introduce bellow the concept of *execution path* which is necessary for representing execution traces of service instances.

Definition 2 *An **execution path** of a protocol $\mathcal{P} = (S, s_0, \mathcal{F}, \mathcal{M}, \mathcal{R})$ is an alternating sequence $c = s_k.m_k.s_{k+1}.m_{k+1} \ldots s_n.m_n.s_{n+1}$ of states and activities of $\mathcal{P}$, that (i) starts at a state s_k of $\mathcal{P}$, (ii) ends at a state s_{n+1} of $\mathcal{P}$, and (iii) is consistent with the transition relationship of $\mathcal{P}$, i.e., $(s_i, s_{i+1}, m_i) \in \mathcal{R}, \forall i \in [0, n]$.*

An execution path c is called **complete** if it starts from the initial state of P (*i.e*, $s_k = s_0$) and ends at one of its possible final states (*if* $s_{n+1} \in \mathcal{F}$).

According to the previous definition, an execution trace $\mathcal{T}_i$ of an instance I represents the sequence of historical activities performed by I, from the beginning of the service invocation to its current state. More formally;

Definition 3 *An **execution trace** $\mathcal{T}_i$ of an instance I is a finite sequence of activities $m_0.m_1 \ldots m_n$ obtained by removing the state names from the associated execution path $c = s_0.m_0.s_1 \ldots s_n.m_n.s_{n+1} \in \mathcal{P}$ followed by the instance I. It is called **complete** if it is originated from a complete execution path.*

We denote by $\mathcal{T}(\mathcal{P})$, the set of all execution traces $\mathcal{T}$ of a service protocol $\mathcal{P}$ and $|\mathcal{T}(\mathcal{P})|$ designates the cardinality of this set. Each element in $\mathcal{T}(\mathcal{P})$ is expressed with $\mathcal{T}_i(\mathcal{P})$, for $i = 1 \ldots |\mathcal{T}(\mathcal{P})|$ and the length of $\mathcal{T}_i(\mathcal{P})$ is noted $|\mathcal{T}_i(\mathcal{P})|$, *i.e., the number of activities contained in the trace $\mathcal{T}_i(\mathcal{P})$.*

Example 2. Hereafter, we show four execution traces of different instances belonging to the retirement protocol of **Fig. 1**.
(1) $\mathcal{T}_{22}(retirement)$=AS. IV. NR, (2) $\mathcal{T}_{23}(retirement)$= AS. IV. CC.EV,
(3) $\mathcal{T}_{24}(retirement)$=AS. IV. CC. AC. EV. SID, and
(4) $\mathcal{T}_{25}(retirement)$=AS. IV. CC. AC. AC. EV. NR.

Among the previous traces, $\mathcal{T}_{22}$ and $\mathcal{T}_{25}$ are complete, while $\mathcal{T}_{23}$ and $\mathcal{T}_{24}$ are incomplete ones. The length $|\mathcal{T}_{25}(retirement)|$ of the trace $\mathcal{T}_{25}$ is 7.

Before ending this section, we assume that traces are classified in a manner that it's possible to dissociate them from one service to another by using a service ID or a service name. From another point of view, traces may be characterized by other attributes, such as time-stamps and activities cost. However, in the service reconstruction context we focus only on the activities names, which are sufficient to illustrate our approach.

3.2 Knowledge base specification

The knowledge base underlying the proposed approach is articulated around, (i) a facts base containing execution traces, and (ii) a set of production rules expressing inference patterns. These two elements are addressed below.

3.2.1 Translating traces to a facts base We formalize each activity of a given execution trace as a first-order logic predicate, expressed as follows.

$$\mathcal{A}\ (Type, Order, Id, Sstate, Aname, Tstate)\ . \tag{1}$$

Where $\mathcal{A}$ is a first-order predicate symbol of valence 6 and the semantics of the associated attributes is as follows.

- Type: this attribute characterizes the activity's type. The used values are 1: if the activity is the first one in the trace, 2: for a last activity and the value 0 is used to express intermediate activities.
- Order: specifies the order of the activity; *i.e., its rank in the trace.*
- Id: After splitting a trace to a set of separate facts, the trace's Id is integrated as an attribute for all the facts belonging to the same trace.
- Aname designates the abstract name of the activity in the trace;
- Sstate is the source state $s \in \mathcal{P}$ from which the activity Aname starts;
- Tstate is the target state $s' \in \mathcal{P}$ to which the activity *Aname* ends.

It's forth noting that the predicate $\mathcal{A}(Attribute)$ applied to a term of the predicate $\mathcal{A}$ returns the value taken by the input parameter Attribute.

According to this specification, an execution trace $\mathcal{T}_i(\mathcal{P})$ of length $l = |\mathcal{T}_i(\mathcal{P})|$ generates a set of l separate facts. The total facts number obtained after transforming all the existing execution traces $\mathcal{T}(\mathcal{P})$ of the protocol $\mathcal{P}$ leads to a facts base, noted $\mathcal{FB}(\mathcal{P})$ having a size $|\mathcal{FB}(\mathcal{P})|$.

Example 3. According to equation (1), the traces $\mathcal{T}_{22}(retirement)$, $\mathcal{T}_{23}(retirement)$ and $\mathcal{T}_{25}(retirement)$ of example (2) are converted to the following facts base. For each fact $A_{i,j}$, i is the trace identifier and j corresponds to the attribute Order, while $S_{i,j}$ and $T_{i,j}$ are, respectively, source and target states of the activity.

- $A_{22,1}$(1,1,22,$S_{22,1}$, AS, $T_{22,1}$)
- $A_{22,2}$(0,2,22,$S_{22,2}$, IV, $T_{22,2}$)
- $A_{22,3}$(2,3,22,$S_{22,3}$, NR, $T_{22,3}$)
- $A_{23,1}$(1,1,23,$S_{23,1}$, AS, $T_{23,1}$)
- $A_{23,2}$(0,2,23,$S_{23,2}$, IV, $T_{23,2}$)
- $A_{23,3}$(0,3,23,$S_{23,3}$, CC, $T_{23,3}$)
- $A_{23,4}$(2,4,23,$S_{23,4}$, EV, $T_{23,4}$)
- $A_{25,1}$(1,1,25,$S_{25,1}$, AS, $T_{25,1}$)
- $A_{25,2}$(0,2,25,$S_{25,2}$, IV, $T_{25,2}$)
- $A_{25,3}$(0,3,25,$S_{25,3}$, CC, $T_{25,3}$)
- $A_{25,4}$(0,4,25,$S_{25,4}$, AC, $T_{25,4}$)
- $A_{25,5}$(0,5,25,$S_{25,5}$, AC, $T_{25,5}$)
- $A_{25,6}$(0,6,25,$S_{25,6}$,EV, $T_{25,6}$)
- $A_{25,7}$(2,7,25,$S_{25,7}$,NR, $T_{25,7}$)

In the previous facts base, the predicate $A_{23,3}(Aname)$ applied to the parameter *Aname* returns the value CC.

3.2.2 Identifying and formalizing inference patterns We propose a set of generic **Inference Patterns (IP)** to capture descriptive elements of the target protocol by exploring execution traces. The proposed patterns define recurrent situations occurring during protocol invocation and they are intended to combine in a single specification, both the syntactic elements associated to execution traces (*activities*) with the structural concepts manipulated at a high abstract level and expressing protocol schema *(transitions, nodes, split, join . . .)*.

Patterns specification and semantics Inference patterns are described with a pattern name, a formal specification and a textual description.

Pattern specification. Let $\mathcal{T}$ be a collection of traces represented as a set of facts and let $\mathcal{P}$ be the target service protocol to be reconstructed. A general specification of an inference pattern is as follows.

$$TargetElement \models IP(Scope, Param), \quad where: \tag{2}$$

- $TargetElement$ constitutes the logic conclusion inferred from premises expressed through the predicate IP. This goal specifies a particular element of the protocol (*states, sequences, loops,...*) that to be discovered.
- IP is a predicate in the first-order logic which captures the semantics of the pattern type.
- *Scope* is a constraint over the execution traces set $\mathcal{T}$, i.e., execution traces explored using this pattern.
- $Param$ is a set of optional parameters expressing values related to the protocol identifier, the activities' name or other traces attributes.

Pattern semantics. The inference pattern specification stipulates that a subset of execution traces satisfying the pattern's *Scope* is used in input to evaluate the predicate $IP(Scope, Param)$, while considering the set of parameters *Param*. If this predicate is evaluated to True then the goal is satisfied and a corresponding *TargetElement* is identified as a structure of the target protocol.

In the following eight inference patterns are identified and formalized. The first three ones are intended to extract static components (*states*) while the five last ones concern dynamic structures (*transitions*). For each identified pattern, we give its formal specification, we explain its semantics and we illustrate its usage through an example.

Initial state pattern (IP1) In order to identify the initial state of a protocol $\mathcal{P}$, the set of first activities contained in the associated facts base $\mathcal{T}(\mathcal{P})$ is filtered out. The constraints governing the initial state specification are expressed by the following inference pattern.

$$s_0 \models InitialState(\mathcal{T}, \mathcal{P}). \tag{3}$$

The semantics of the pattern stipulates that the scope of the pattern is given by the subset of traces $\mathcal{T}$ of $\mathcal{P}$; i.e., $\mathcal{T}(P)$. After transforming the filtered traces to a facts base $\mathcal{FB}(\mathcal{P})$, the initial state s_0 of the protocol is discovered if the predicate $InitialState(\mathcal{T}, \mathcal{P})$ is evaluated to True. This predicate is evaluated to True if the following condition holds.

$$\exists\ i, j, \text{ such that: } A_{i,j} \in \mathcal{FB}(\mathcal{P}), \text{ and } A_{i,j}(Type) = 1.$$

As an illustration, the previous pattern is satisfied for the following three facts of the facts base of example (3).
•$A_{22,1}(1, 1, 22, S_{22,1}, \text{AS}, T_{22,1})$; •$A_{23,1}(1, 1, 23, S_{23,1}, \text{AS}, T_{23,1})$;
•$A_{25,1}(1,1,25,S_{25,1}, \text{AS}, T_{25,1})$.
Furthermore, as the three activities' names are identical; (i.e, AS), an unique execution path starting from the state s_0 is constructed in the target automaton.

Final states pattern (IP2) The inference pattern specifying the final states $F_l \subseteq S$ of a target protocol is formalized as follows.

$$F_l \models FinalStates(\mathcal{T}, \mathcal{P}),\ with\ l \geq 1\ . \tag{4}$$

The final states of $\mathcal{P}$ are recognized by focusing on target states of last activities. Thus, a fact $A_{i,j}$ expressing a transition $(s, m, s') \in \mathcal{R}$ is ending at a final state, only if the predicate $FinalStates(\mathcal{T}, \mathcal{P})$ is evaluated to True. This predicate is satisfied if the following condition holds.

$$\exists\ i, j, \text{ such that: } A_{i,j} \in \mathcal{FB}(\mathcal{P}), \text{ and } A_{i,j}(Type) = 2.$$

Whenever a final state F_l is discovered, it is added to the already constructed final states set; *i.e.,* $\mathcal{F} = \mathcal{F} \cup\ F_l$.

As an illustration, the deployment of the pattern **IP2** of equation (4) to the facts base of example (3) produces the following subset of facts.
$\bullet A_{22,3}(2, 3, 22, S_{22,3}, \text{NR}, T_{22,3})$, $\bullet A_{23,4}(2, 4, 23, S_{23,4}, \text{EV}, T_{23,4})$,
$\bullet A_{25,7}(2, 7, 25, S_{25,7}, \text{NR}, T_{25,7})$
Thus, three final states F_1, F_2 and F_3 are discovered and added to the final states set $\mathcal{F}$ of the protocol $\mathcal{P}$; i.e;, $\mathcal{F} = \mathcal{F} \cup \{\ F_1, F_2, F_3\}$.

Intermediate states pattern (IP3) From a structural point of view, an intermediate state is a target state for an ingoing activity and a source state for an outgoing one. The corresponding inference pattern is formalized as follows.

$$S_k \models IntermdediateState(\mathcal{T}, \mathcal{P}),\ with\ k \geq 1. \tag{5}$$

The Pattern semantic indicates that a new intermediate state S_k is discovered in the protocol $\mathcal{P}$, if the predicate $IntermediateState(\mathcal{T}, \mathcal{P})$ applied to the facts base $\mathcal{FB}(\mathcal{P})$ is evaluated to True. This predicate is True for two facts $A_{i,j}$ and $A_{i,j'} \in \mathcal{FB}(\mathcal{P})$ if their related activities are consecutive ones. More formally.

$$A_{i,j'}(Order) = A_{i,j}(Order) + 1;\ for\ i \in [1, |\mathcal{T}(\mathcal{P})|] \text{ and } \ j, j' \in [1, |\mathcal{T}_i(\mathcal{P})|].$$

During the intermediate states reconstruction process, the target state $T_{i,j}$ of the first activity and the source state $S_{i,j+1}$ of the consecutive one are renamed with the same state name $\mathcal{S}_k$.

As an example, when applying the previous pattern to the facts base of example (3) with the particular value of $i = 22$, two intermediate states are discovered. After the states extraction, the variables $T_{22,1}$ and $S_{22,2}$ in example (3) are renamed to $\mathcal{S}_1$. In a similar fashion the states $T_{22,2}$ and $S_{22,3}$ are renamed to $\mathcal{S}_2$. The discovered two states $\mathcal{S}_1$ and $\mathcal{S}_2$ are added to the set of states: $\mathcal{S} = \mathcal{S} \cup \{\mathcal{S}_1, \mathcal{S}_2\}$.
By taking into account the already discovered initial and final states, the improved facts associated to the trace $\mathcal{T}_{22}(retirement)$ become as follows.
$\bullet A_{22,1}(1, 1, 22, S_0, \text{AS}, S_1)$ $\bullet A_{22,2}(0, 2, 22, S_1, \text{IV}, S_2)$ $\bullet A_{22,3}(2, 3, 22, S_2, \text{NR}, F_1)$.

Self-loop pattern(IP4) The existence of self-loop structures is manifested by execution traces that exhibit activities sequences of the form $m.m.m\ldots$. The following inference pattern allows detecting states S_j that exhibit such a behavior.

$$S_j \models Loops(\mathcal{T}, (\mathcal{P}, A)),\ with\ j \geq 0. \tag{6}$$

The self-loop inference pattern is customized with two parameters; the protocol $\mathcal{P}$ and the activity A concerned by the loop test. A state S_j holds a loop structure upon the execution of an activity A, if the predicate $Loops(\mathcal{T}, (\mathcal{P}, A))$ is evaluated to True. This predicate is True if:

$$\exists\ i, j\ (j \in [2, |\mathcal{T}_i(\mathcal{P})| - 1]), \text{ such that: } A_{i,j}, A_{i,j+1} \in \mathcal{FB}(\mathcal{P}) \text{ and } A_{i,j}(Aname) = A_{i,j+1}(Aname) = A$$

In the previous condition, the constraint $j \in [2 \ldots |\mathcal{T}_i(\mathcal{P})| - 1]$ ensures that loops can't be built on the initial state ($j \neq 1$) and final ones ($j \in |\mathcal{T}_i(\mathcal{P})| - 1$).

The exploration of the facts base of example (3) basing on the self-loop pattern of equation (6) leads to the two following facts that satisfy the predicate $Loops(\mathcal{T}, (retirement, AC))$ upon the execution of the activity AC.
$\bullet A_{25,4}(0, 4, 25, S_{25,4}, \text{AC}, T_{25,4})$ $\bullet A_{25,5}(0, 5, 25, S_{25,5}, \text{AC}, T_{25,5})$
After renaming states as follows: $S_4 = S_{25,4} = T_{25,4} = S_{25,5} = T_{25,5}$, the two previous facts become. $\bullet A_{25,4}(0, 4, 25, S_4, \text{AC}, S_4)$ $\bullet A_{25,5}(0, 5, 25, S_4, \text{AC}, S_4)$.

Activities sequence pattern (IP5) A sequence structure links chronologically, at least, two activities $\mathcal{A}_{i,j}$ and $\mathcal{A}_{i,j'}$ by a common state $\mathcal{S}_k$. Such structures are expressed by two transitions $(s, m, \mathcal{S}_k)$ and $(\mathcal{S}_k, m', s') \in \mathcal{R}$. The following inference pattern allows extracting sequences from the activities facts base.

$$(A_{i,j}, A_{i,j'}) \models Sequence(\mathcal{T}, (\mathcal{P}, S_k)). \tag{7}$$

The semantics of the pattern is interpreted as follows. First, only traces of the input protocol $\mathcal{P}$ are handled. When exploring the corresponding facts base $\mathcal{FB}(\mathcal{P})$, for each state S_k introduced as an input parameter of the pattern, the pairs of consecutive activities $(A_{i,j}, A_{i,j'})$ are searched by locating incoming activities $A_{i,j}$ and the outgoing ones $A_{i,j'}$, while evaluating the predicate $Sequence(\mathcal{T}, (\mathcal{P}, S_k))$. For a given state S_k this predicate is evaluated to True, for all the facts $A_{i,j}$ and $A_{i,j'} \in \mathcal{FB}(\mathcal{P})$ satisfying the condition.

$$\forall\ i, j, j', \text{ such that: } A_{i,j}, A_{i,j'} \in \mathcal{FB}(\mathcal{P}); \quad A_{i,j}(Tstate) = A_{i,j'}(Sstate) = S_k\ (S_k \notin \mathcal{F})$$

By applying the pattern **IP5** to the facts sub-base presented at the end of **IP3** section with S_2 as a parameter, the predicate $Sequence(\mathcal{T}, (retirement, S_2))$ is evaluated to True for the two activities IV and NR *(j=2 and j'=3)*, because $A_{22,2}(Tstate) = A_{22,3}(Sstate) = S_2$. Thus, it's inferred that the state S_2 links the two activities IV and NR. Consequently, a sequence structure is deduced.

Join structure pattern (IP6) Join structures express the convergence of several activities to an unique state of the automaton. The following pattern is used to detect activities involving a join structure over a state S_k of $\mathcal{P}$.

$$A_{i_1,j_1}, A_{i_2,j_2}, \ldots, A_{i_n,j_m} \models Join(\mathcal{T}, (\mathcal{P}, S_k)). \tag{8}$$

The pattern semantics indicates that a state S_k introduced as an input parameter of the pattern represents a join node for a sub-set of activities performed in execution traces $\mathcal{T}$ of a protocol $\mathcal{P}$, if the predicate $Join(\mathcal{T}, (\mathcal{P}, S_k))$ is evaluated to True. This predicate is True if the following condition holds.

$$\exists\ i_1, i_2, \ldots, i_n, j_1, j_2, \ldots, j_m\ \text{, such that: } A_{i_1,j_1}, A_{i_2,j_2}, \ldots, A_{i_n,j_m} \in \mathcal{FB}(\mathcal{P}) \text{ and } A_{i_1,j_1}(Tstate) = A_{i_2,j_2}(Tstate), \ldots, = A_{i_n,j_m}(Tstate) = S_k$$

For instance, assume that four facts $A_{20,2}$, $A_{30,3}$, $A_{40,5}$ and $A_{50,8}$ of a purchase protocol are described according to the facts predicate of equation (1).
$\bullet A_{20,2}(0, 2, 20, S_2,$ Select, $T_2)$; $\bullet A_{30,3}(0, 3, 30, S_3,$ Validate, $T_3)$;
$\bullet A_{40,5}(0, 5, 40, S_5,$ Purchase, $T_4)$; $\bullet A_{50,8}(0, 8, 50, S_8,$ Payment, $T_4)$.
According to equation (8), the predicate $Join(\mathcal{T}, (retirement, T_4))$ is evaluated to True for the two last fact $\{A_{40,5}, A_{50,8}\}$, $(i_1 = 40, i_2 = 50$ and $j_1 = 5, j_2 = 8)$ and the state T_4 constitutes a join node for the activities Purchase and Payment.

Split structure pattern (IP7) A split structure represents a subset of distinguished transitions which are outgoing from a same source state S_k. The specification of the inference pattern for discovering split nodes in the facts base $\mathcal{T}(\mathcal{P})$ is formalized as follows.

$$A_{i_1,j_1}, A_{i_2,j_2}, \ldots, A_{i_n,j_m} \models Split(\mathcal{T}, (\mathcal{P}, S_k)). \tag{9}$$

The semantic carried by this pattern is interpreted as follows. A state S_k introduced as a parameter of the pattern constitutes a split node for the outgoing activities $A_{i_1,j_1}, A_{i_2,j_2}, \ldots, A_{i_n,j_m}$, if the predicate $Split(\mathcal{T}, (\mathcal{P}, S_k))$ is evaluated to True. This predicate is satisfied if the following constraint holds.

$$\exists\ i_1, i_2, \ldots, i_n, j_1, j_2, \ldots, j_m\ \text{, such that: } A_{i_1,j_1}, A_{i_2,j_2}, \ldots, A_{i_n,j_m} \in \mathcal{FB}(\mathcal{P}) \text{ and } A_{i_1,j_1}(Sstate) = A_{i_2,j_2}(Sstate), \ldots, = A_{i_n,j_m}(Sstate) = S_k$$

As an illustration, let us consider the two following facts of example (*3*).
$\bullet A_{22,3}(2, 3, 22, S_{22,3},$ NR, $T_{22,3})$ $\bullet A_{23,3}(0, 3, 23, S_{23,3},$ CC, $T_{23,3})$.
According to the intermediate state pattern **IP3**, the states $S_{22,3}$ and $S_{23,3}$ are renamed with the state S_3 and the resulting facts are:
$\bullet A_{22,3}(2, 3, 22, S_3,$ NR, $T_{22,3})$ $\bullet A_{23,3}(0, 3, 23, S_3,$ CC, $T_{23,3})$.
The predicate $Split(\mathcal{T}, (\mathcal{P}, S_k))$ is evaluated to True for the state S_3 which constitutes a split node for the two activities NR and CC.

Cycles on sub-structures (IP8) Instead of looping on simple states, service instances may iterate a whole activities sequence during their executions. More formally, consider that I is an instance of a web service, $\mathcal{T}_i$ its trace and let $[A_{i,1}.A_{i,2}.\ldots.A_{i,l}]$ be an activities sequence of length l (with $l > 1$) which is performed by I during its interaction with the service. If the previous activities sequence occurs more than once in the same trace, it is deduced that a cycle on a sub-structure exists in the protocol specification. The following pattern is deployed to extract potential existing cycles on sub-structures of the protocol.

$$[A_{i,j}.A_{i,j+1}.\ldots.A_{i,j+l}] \models Cycle(\mathcal{T}, \mathcal{P}). \tag{10}$$

The semantics of the pattern is interpreted as follows. A cycle is detected for an activities sequence $[A_{i,j}.A_{i,j+1}.\ldots.A_{i,j+l}]$, if the predicate $Cycle(\mathcal{T},\mathcal{P})$ is evaluated to True. This predicate is satisfied for the previous sequence if:

$$\exists\ i,j\ \text{, such that: } A_{i,j}, A_{i,j+1}, \ldots, A_{i,j+l} \in \mathcal{FB}(\mathcal{P}) \text{ and } Occurrence([A_{i,j}, A_{i,j+1}, \ldots, A_{i,j+l}]) > 1.$$

For identifying potential cycles in the target protocol, possible duplicated occurrences of the activities sequence in a trace $\mathcal{T}_i$ are examined depending on different values of the length $1 < l \leq |\mathcal{T}_i|$.

As an illustration, consider an instance $I = 100$ with its execution trace $\mathcal{T}_{100}(\mathcal{P}) = a.b.c.d.b.c.e$, having the two redundant activities b and c. According to the facts representation, (*see subsection (3.2.1)*), the trace $\mathcal{T}_{100}(\mathcal{P})$ is transformed to the following subset of facts.

- $A_{100,1}(0, 1, 100, S_{100,1}, \text{a}, T_{100,1})$;
- $A_{100,2}(0, 2, 100, S_{100,2}, \text{b}, T_{100,2})$;
- $A_{100,3}(0, 3, 100, S_{100,3}, \text{c}, T_{100,3})$;
- $A_{100,4}(0, 4, 100, S_{100,4}, \text{d}, T_{100,4})$;
- $A_{100,5}(0, 5, 100, S_{100,5}, \text{b}, T_{100,5})$;
- $A_{100,6}(0, 6, 100, S_{100,6}, \text{c}, T_{100,6})$.
- $A_{100,7}(0, 7, 100, S_{100,7}, \text{e}, T_{100,7})$.

For the same trace $\mathcal{T}_{100}$, the activities b and c appear more then once in the facts base. Further, for the value $l = 2$, the sequence b.c is redundant in the facts base and thus, it is deduced that a sub-structure performing a cycle on the activities sequence b.c exists in the service protocol.

3.3 Exploring the knowledge base

A multi-stages process is conducted to explore the conceived knowledge base in order to extract the structural elements of the target protocol. These steps are:

1. Once the traces database is imported from providers' information systems, a pre-processing phase is initiated. It consists in removing noises and unnecessary attributes, such as information about service quality (QoS) and time-stamp. Further, in order to optimize the overall efficiency of the proposed approach redundant and included traces are removed. Thus, if the traces $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n$ are contained in a trace $\mathcal{T}_m$, then only $\mathcal{T}_m$ is exported to the cleaned traces database.
2. Cleaned execution traces are exported to an adequate database that is ready to be exploited as input of the knowledge base reasoning system.
3. Each execution trace $\mathcal{T}_i$ of length l is transcribed to l separate facts which are stored in the facts base and the formalized inference patterns are implemented as production rules. Facts and rules are expressed in a low level language such as *Prolog*.
4. An inference engine using backward chaining (*Prolog*) is deployed to produce the descriptive elements of the target protocol by assessing the production rules. The attributes of inferred structures are stored as XML elements expressing *state names and types, transition names and attributes,...*. Additionally, conformance checking actions are operated to avoid errors and inconsistency that can occur in the protocol, such as *unreachable states, absence of final states.*

5. For ergonomic viewing reasons, a graphical representation of the reconstructed protocol is provided to end users to allow them refining and enhancing the produced protocol in a visual manner.

4 Implementation and Experiments

This section briefly describes a prototype, named **L**ogical **B**usiness **P**rotocol **R**econstructor (**LBPR**) which implements the proposed approach. First, the system architecture is presented then experimental results are discussed.

4.1 System Architecture and functionalities

Java environment integrating useful Eclipse plug-ins, such as MySQL-connector and **JPL** (Java Programming Logic) [17] have been used to implement the prototype **LBPR**. The **JPL** library allows incorporating Prolog in Java and it's used by Java for interacting with SWI-Prolog [20]. As illustrated in **Fig. 2**, the prototype **LBPR** is organized around three main components interacting with the conceived knowledge base.

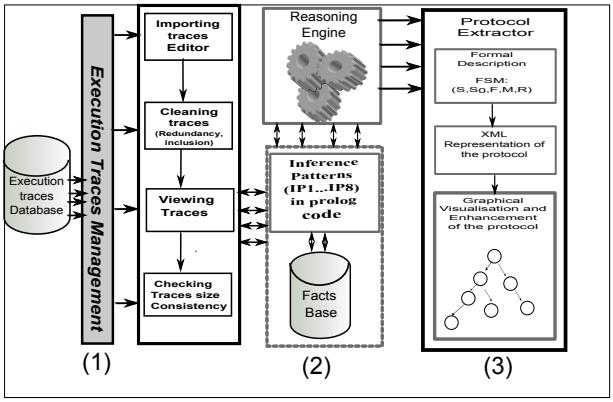


Fig. 2. System Arhitecture and functionalities

The first component of **LBPR** is the Execution Traces Management module which allows users to connect to the log files database in order to select and to import execution traces of the concerned web service. After that, a pre-processing step is initiated. It consists in removing redundant and included traces (module: cleaning traces). At this stage the relevance and the representativeness of imported traces are evaluated by calculating the rate ρ of the complete traces with regard to the total number of the imported traces: $\rho = C/P$; *with* $C =$ *Complete traces and* $P =$ *Imported traces*. This verification is ensured by the module *checking traces size and consistency*.

The second component of **LBPR** enables constructing the facts base in accordance to the facts predicate model of equation (**1**). Once the facts base is generated, the system performs automatically the reconstruction of the protocol as a sequence of chronological actions. In each step, a particular structure inference pattern is activated. The protocol reconstruction process is controlled through parameters settings which determine the inference pattern that should be performed, some adjustable thresholds and the levels of details for writing the reasoning engine logs. The user gets a detailed log of all mining steps expressed by native prolog traces.

The last component of **LBPR** allows different representation of the produced protocol. In fact, upon patterns execution the induced goals constitute the elements describing the target protocol. The discovered structures are managed in XML-format and are stored in adequate files describing the target automaton. Each XML element in the output file represents either a protocol state or a transition between states. The attributes associated to XML elements express values of properties characterizing states names, transitions names, states types and positions manipulated during graphical visualization of the target protocol.

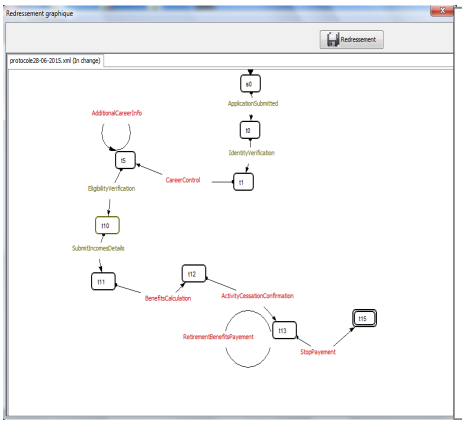


Fig. 3. Target protocol graphical browsing

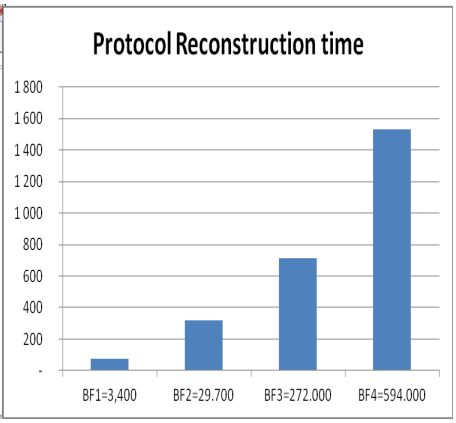


Fig. 4. Variation of reconstruction time

For ergonomic reasons, the prototype **LBPR** was consolidated by a graphical editor that allows viewing and browsing the extracted protocols. As depicted in **Fig. 3**, a partial reconstructed protocol retirement is produced by **LBPR** from execution traces data-set. Furthermore, in order to enhance the protocol specification and visualization, the graphical tool box provides useful functions, like moving states/activities from one position to another, adding or renaming states/activities and finally removing elements. Another useful feature of the system is the conformance-checking tool which enables ensuring the verification of a set of correctness criteria (*initial state, final and unreachable states ...*).

4.2 Experiments

To evaluate the scalability and the performance of the proposed approach, we conducted experiments with the prototype **LBPR** over synthetic data-sets originating from various fields (*pension, e-commerce, booking flights*) and containing respectively: DS1=1.000, DS2=10.000, DS3=100.000 and DS4=200.000.

In a first experiment, we focused on the pre-processing step by evaluating the rates of redundant and included traces, while recording the elapsed time during data cleaning. To this end, the data-set DS3 containing 100.000 traces was chosen and treated. It was observed that the rate of redundant execution traces represents 46 % of the total size of the introduced data-set. Once duplicated traces were removed, only filtered ones are submitted to an inclusion checker module which tests traces inclusion and discards the included traces for the next experimental steps. Empirical observations show that average 38 % of the total traces are included in other ones, because we are dealing with real business processes characterized by a bounded number of execution paths. After the pre-processing step, execution traces are reduced approximately to 34 % of their original size. These proportions are largely suitable for the remainder of the experiment process. In the other hand, the observation of the manipulated data shows that the average of activities is generally around 10 activities per trace. Consequently, after the pre-processing step, the size of corresponding facts in each data-set approximates, respectively, $\mathcal{BF}1$=3.400, $\mathcal{BF}2$=29.700, $\mathcal{BF}3$=272.000 and $\mathcal{BF}4$=594.000.

To examine the efficiency and the relevance of the proposed patterns, in a second experiment the time spent during the reasoning step is evaluated. Hence, when chronologically activating the various inference patterns, the time consumed by the reasoning engine to resolve Prolog clauses was calculated. As shown in **Fig. 4**, this time grows linearly according to the size of the facts base introduced in input. As the protocol reconstruction process operates on cleaned data, the obtained results are very satisfactory and turn around 25 minutes (1532 seconds) for the largest data-set DS4 ($BF \simeq$ *594.000 facts*). These results encourage the system deployment for handling large-public applications. Nevertheless, when viewing the resulting protocol some display problems arise due the aleatory attributes associated to the automaton states in the XML files. In fact, we associate to each discovered state two random coordinates (x,y) to display it on the screen. To overcome this limitation, while enhancing protocol's viewing, user's manual intervention is required.

5 Conclusion and Future Work

In this work an approach based on first-order logic predicates is proposed to reconstruct web services business protocols. Activities expressing log events are transcribed to first-order predicates and implemented as Prolog clauses and a set of inference patterns has been identified and formalized as a key concept for business protocol reconstruction. The discovered protocols are complete; i.e., all recorded traces are covered by the extracted protocol, and minimal; i.e.,

future interactions can't be predicted and handled by the system. Beyond the benefits from logic paradigm (*extensibility, traceability, easy implementation and best performances...*), the main contributions of this work are: (i) a logic formal framework for extracting web services protocols is suggested. (ii) a technique based on redundant and included traces is adopted to reduce the search space. (iii) the proposed approach is declarative, configurable and customizable. (iv) the proposed approach is implemented and experimented using real-life data.

In future works, we plan to improve the output protocol viewing. As a potential direction, we project to deploy and experiment the approach on big data originating from social networks.

References

1. Aalst, W.V.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. 1st edn. (2011)
2. Aalst, W.V., van Dongen, B.F., Herbst, J., Maruster, L., Schimm, G., Weijters, A.: Workflow mining: A survey. D.K.E. **47**(2) (2003)
3. Aalst, W.V., Weijter, A., Maruster, L.: Workflow mining: Discovering process models from event logs. IEEE Transactions on KDE. (2003)
4. Benatallah, B., Casati, F., Toumani, F.: Web service conversation modeling: A cornerstone for e-business automation. IEEE Internet Computing **8**(1) (2004)
5. Benatallah, B., Casati, F., Toumani, F.: Representing, analysing and managing web service protocols. Data Knowl. Eng. **58**(3), 327–357 (2006)
6. Berardi, D., Cheikh, F., Giacomo, G.D., Patrizi, F.: Automatic service composition via simulation. IJFCS **19**(2), 429–451 (2008)
7. Cassez, F., Roux, O.H.: Structural translation from time petri nets to timed automata. E.N.T.C.S. **128**(6), 145 – 160 (2005), proceedings AVoCS 2004
8. Devaurs, D., Marchi, F.D., Hacid, M.S.: Caractérisation des transitions temporisées dans les logs de conversation de services web. In: EGC (2007)
9. Günther, C.W., Aalst, W.V.: Fuzzy mining: Adaptive process simplification based on multi-perspective metrics. In: BPM'07. Berlin (2007)
10. Gupta, E.P.: Process mining a comparative study. IJARCCE **3**(11) (2014)
11. Klai, K., Tata, S., Desel, J.: Symbolic abstraction and deadlock-freeness verification of inter-enterprise processes. In: BPM. pp. 294–309 (2009)
12. Medeiros, A.K., Weijters, A.J., Aalst, W.V.: Genetic process mining: An experimental evaluation. D. M. K. D. **14**(2), 245–304 (Apr 2007)
13. Medeiros, A.K.A.D., van Dongen, B.F., Aalst, W.V., Weijters, A.J.M.M.: Process mining: Extending the alpha-algorithm to mine short loops. In: Eindhoven University of Technology, Eindhoven (2004)
14. Motahari-Nezhad, H.R., S-Paul, R., Casati, F., Benatallah, B.: Event correlation for process discovery from web service interaction logs. VLDB **20**(3) (2011)
15. Nezhad, H.R.M., Saint-Paul, R., Benatallah, B., Casati, F.: Protocol discovery from imperfect service interaction logs. In: ICDE. pp. 1405–1409 (2007)
16. Ponge, J., Benatallah, B., Casati, F., Toumani, F.: Fine-grained compatibility and replaceability analysis of timed web service protocols. In: ER (2007)
17. Singleton, P.: Jpl: A java interface to prolog (September 2012)
18. Wang, X., Miao, H., Guo, L.: Towards automatic transformation from uml model to fsm model for web applications. JSEA **1**(1), 68–75 (2008)
19. Weijters, A., Medeiros, A.A.D.: Process mining with the heuristics miner algorithm
20. Wielemaker, J.: Swi-prolog home page. http://www.swi-prolog.org/

Clustering methods evaluation by a new test case generator for bivariate correlated data*

Radhwane Gherbaoui[1], Mohammed Ouali[2], and Nacéra Benamrane[1]

Lab. SIMPA, Département d'Informatique, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf USTO-MB, BP 1505 Elmenaouer, Oran, 31000, Algrie. radhwan.gherbaoui@univ-usto.dz

Abstract. In this paper, we will present a new algorithm for generating bivariate Gaussian mixture data. This algorithm allows to generate mixture with any number of components and the more important is that is able to control the degrees of overlap between the generated clusters. The artificial data will be used to evaluate clustering methods and the validity indices by comparing the resulted clustering data and the ground-truth model resulted from the generator.

Keywords: Artificial Data · Clustering Methods · Test Case Generator.

1 Introduction

Clustering methods give different outcome for the same set of observations to extend that the same method gives various structure according to initial parameters. It is difficult to analytically measure the propriety of these methods. Many authors use constructed data with knowing structures for evaluating clustering methods [11], [3], [6]. In other side, mixture model is strongly used as a generic model for artificial data, especially Gaussian models. The constructed data must satisfy some intuitive conditions mainly resumed in two criteria: the internal cohesion and the external isolation. *Internal cohesion*, as defined by Milligan, means that entities belonging to the same clusters have a big resemblance. *External isolation* assures a high dissimilarity between observations belonging to different clusters. In order to ensure that constructed data have structures of clusters, three categories of techniques are used, the first is based on employing distance to measure similarity-dissimilarity between clusters, others measure the probability between clusters and the third one is by adjusting geometrical proprieties of the data structures. In [8], Dasgupta propose a test case generator based on a statistical model but he did not involve al the mixture parameter. Underhill and Henson [8] proposed a methodology for constructing qualitative clusters, but it cannot be extended to high dimensions without losing substantial information. Other in [16] propose method for generating data in one dimension and extend his method to high dimension by projection on the most *presentative* axis but even in this method, the projection conducts to lose substantial

*

information. Everitt [9] argue that a large separation between components of the mixture conduct to build clusters by nature and any method allows to easily find the true structure, therefore, it not useful to evaluate clustering methods with such structures.
In these lasts years, OCLUS is suggested. For each pair of components and for each dimension, the overlap first is modeled to be the cumulative probability from the intersection point between pair of components. After the degree of overlap is presented to be the product of overlap of each dimension. This generator does not take on consideration the correlation between components. Indeed, different structure can be built with the same overlap degree. To get rid of this problem, MixSim is proposed since the degree of overlap is formulated by employing cumulative probability for all dimensions at the same time for two components of mixtures. Indeed, both methods cannot be employed to assess the clustering process in dealing with the problem of cluster sensitivity. In this papers, we introduce Gaussian mixtures and the notion of total overlap. After we suggest a formal quantification of the overlap based on visual inspection. We inherited from the formal quantification a methodology for generating artificial data. The new generator will be advocated to assess the performance of partition method and the most validity indices used in the literature.

2 Bivariate correlated Gaussian Mixture model

Mixtures models are emmployed in many field because many clustering algorithms are based on such distributions [9, 2, 3, 14, 17]. Gaussian mixture is more approved to asses clustering methods. A mixture of M Gaussian 2D components is given by

$$P(x,y) = \sum_{j=1}^{M} \kappa_j \Gamma_j(x,y,\theta_j), \tag{1}$$

where $\sum_{j=1}^{M} \kappa_j = 1$ and $\theta_i = (\mu_{xj}, \mu_{yj}, \sigma_{xj}, \sigma_{yj}\rho_j)$ denotes the parameters of the j^{th} distribution Γ_j where (μ_{xj}, μ_{yj}) expose the component center coordinates for the first and the second dimension respectively. σ_{xj} and σ_{yj} are the standard deviations of the first and second dimension respectively where ρ_j is the coefficient of correlation between the two dimensions; X and Y. Γ_j is the Γ^{th} distribution given by

$$\Gamma_j(x,y) = Ae^{\left(-\frac{1}{2(1-\rho_j^2)}\left[\left(\frac{(x-\mu_{xj})^2}{\sigma_{xj}^2}+\frac{(y-\mu_{yj})^2}{\sigma_{yj}^2}\right) - \frac{2\rho_j(x-\mu_{xj})(y-\mu_{yj})}{\sigma_{xj}\sigma_{yj}}\right]\right)} \tag{2}$$

where $A = \frac{1}{2\pi\sigma_{xj}\sigma_{yj}\sqrt{1-\rho_j^2}}$ is a real strictly positive.

2.1 Why controlling the overlap ?

The problem is when we want to generate a large set of the data. We must be sure that the generated mixture components are not in a case of total overlap.

A components in a case of total overlap means that the two criteria of internal cohesion and external isolation are not respected. To know what we mean by total overlap, we examine the example of Fig.1. In Fig.1 (a), mixture is constituted of three components but only two are visible. It represents the case of total overlap. A case of maximal overlap is shown in Fig.1 (b). We can approximatively distinguish that there are three components. In Fig.1 (c), there are a partial overlap between the three components of the mixture. It is clear that the mixture is composed of three components.

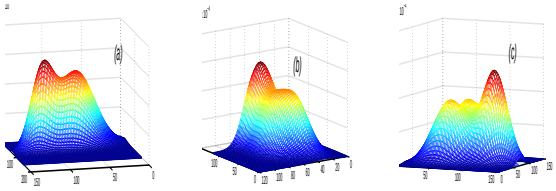


Fig. 1. Overlap between three components of the mixture in the three cases. (a) total overlap; (b) Maximal overlap; (c) partial overlap.

In the rest of this paperL, we will use the notation $\Gamma_i(\kappa_i, \mu_{xi}, \mu_{yi}, \sigma_{xi}, \sigma_{yi}, \rho_i)$ to describe the parameters of the i^{th} component Γ_i where κ_i represents the mixture coefficient, (μ_{xi}, μ_{yi}) is the coordinate of the center of the components. σ_{xi} and σ_{yi} are the standard deviations of the first and the second dimension respectively. ρ_i is the coefficients of correlation between the two component dimensions.

2.2 Overlap principe

In order to control overlap, a formal quantification is strangely recommended. In [20], the presence of false edge between the steps of an staircase edges is treated as a problem of univariate gaussian mixture. They assume that there are a false edge between two equivalent components (maximal overlap) if the difference of the center's component is higher than the sum of the standard deviations $|\mu_1 - \mu_2| > 2\sigma$ (see Fig.2 (a)). In the case where the difference is equal to the sum of the deviations, there are a straight line between the center of the two components (see Fig.2 (b)). It considered as the case of maximal overlap. Figure 2.(c) represent the case where $|\mu_1 - \mu_2| < 2\sigma$.

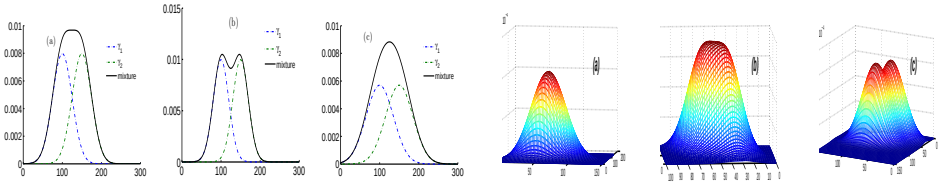
This problem is treated as a problem of mixture of two components which have the same standard deviation σ. In [1][14], this condition is exploited to propose three definitions to characterize the tree cases of overlap.

It is difficult to use the same formulas, to quantify the overlap between two multivariate components. We see from those studies, that the value intersection

point is a good parameter for quantifying overlap. Even the intersection point has a small value the component are more separated. There are generally two intersection points where $\Gamma_1(\mu_1) > \Gamma_2(\mu_1)$ and $\Gamma_1(\mu_2) < \Gamma_2(\mu_2)$, one of them is located between the two component centers (the higher). The concerned point that quantify the overlap is the higher. We use the same concept to characterize overlap between two Gaussian components for a set of observations of two patterns. Intersection of two Gaussians in those case results an infinity of intersection points whereas there is one highest point. From the condition of maximal overlap in 1D (in a space of one dimension), we conclude that the height of the interaction point equal 60% of the components high at the center. We save the same concept for the bivariate component such that

$$\Gamma_1(x_{int}, y_{int}) = 0.6\Gamma_1(\mu_1) \tag{3}$$

where x_{int}, y_{int} are the coordinates of the higher intersection point.



(a) (b) (c) (d) (e) (f)

Fig. 2. Illustration of Tabbone's condition for 1D Gaussian in (a),(b),(c); and generalization for 2D mixtures in (e),(f),(g).

Figure 2 (e),(f),(g) illustrates the three situations related to our condition. In Fig.2 (e), the value of the intersection point is higher then the value given in 3. Hear, we are in the case of total overlap, we cannot visually distinguish the mixture components. In Fig.2 (f), the intersection point respects the condition 3. As in Fig.2 (a) in the univariate space, we observe that there are approximatively straight line between the centers of the two components. We consider this situation as the case of maximal overlap or a limited case between the total(forbidden) and the partial(authorized) overlap.

3 Formal quantification of overlap

To evaluate clustering method, a formal quantification of the overlap is strongly recommended. We define *the maximal overlap between two adjacent components such that the highest intersection point satisfy* $\Gamma_1(x_{int}, y_{int}) = 0.6 * \min(\Gamma_1(\mu_1), \Gamma_2(\mu_2))$. where x_{int} and y_{int} are the coordinate of the highest intersection point.

Maximal overlap is considered as the highest high at which the high of the intersection point is limited. Less than the height of the intersection point the

component are partially overlapped. to quantify formally the overlap degree between cluster component, *we propose the definition of well separated clusters noted by λ to be the ratio of the higher intersection point to the value of the higher intersection point in the case of maximal overlap.* Formally, the rate of overlap can be expressed as

$$\lambda = \frac{0.6 \times \Gamma_1(x_{int}, y_{int})}{min(\Gamma_1(\mu_1), \Gamma_2(\mu_2))}. \tag{4}$$

A Structure of well separated clusters is also needed to evaluate clustering methods. Well separated clusters are a structure in which an empty space appears between clusters where, practically, there is no overlap between components. Noted in [14] by minimal overlap for 1D data, it is defined such that the intersection point is far from each center by 4 standard deviations, we can easily conclude that the hight of the intersection point can be used as a metric for measuring the degree of overlap. From the characteristics of Gaussian distribution, we find that is sufficient to have a structure of well separated clusters by a low value λ. In our experiments, we choose λ to take value 0.001 to have well separated cluster. Analytically, to have empties spaces between clusters, we must take into consideration the number of the observation, mixture coefficients and the chosen overlap degree.

4 Controlling mixture component overlap

By using definitions characterizing overlap, we propose an algorithm for generating artificial data. So for two adjacent components Γ_1 and Γ_2, we generate randomly the parameters of the first component, the mixture coefficients, the standard deviations and the coefficients of correlation of the others components. We also randomly introduce angles of intersection between components. After, we fixed the centers of the components one by one according to the rate of overlap.We apply the definition of rate of overlap on the two components, we arrive at two cases; case 1: $\Gamma_1(\mu_1) > \Gamma_2(\mu_2)$ and case 2: $\Gamma_1(\mu_1) < \Gamma_2(\mu_2)$.

Case 1: For Γ_1, after applying the definition of rate of overlap, we find

$$\Gamma_1(x_{int}, y_{int}) = 0.6 * \Gamma_2(\mu_2).$$

After some transformations, we arrive

$$\begin{aligned} &a_1(x_{int} - \mu_{x1})^2 + b_1(y_{int} - \mu_{y1})^2 \\ &+c_1(x_{int} - \mu_{x1})(y_{int} - \mu_{y1}) - 1 = 0, \end{aligned} \tag{5}$$

where

$$\begin{cases} e_1 = 1 - 2\ln\left(\frac{\lambda\kappa_2\sigma_{x1}\sigma_{y1}\sqrt{1-\rho_1^2}}{\kappa_1\sigma_{x2}\sigma_{y2}\sqrt{1-\rho_2^2}}\right) \\ a_1 = \frac{1}{(1-\rho_1^2)\sigma_{x1}^2 e_1} \\ b_1 = \frac{1}{(1-\rho_1^2)\sigma_{y1}^2 e_1} \\ c_1 = -\frac{2\rho_1}{\sigma_{x1}\sigma_{y1}(1-\rho_1^2)e_1}. \end{cases} \tag{6}$$

For the second component and by suiting the same reasoning, we find that

$$\begin{array}{l} a_2(x_{int}-\mu_{x2})^2+b_2(y_{int}-\mu_{y2})^2+ \\ c_2(x_{int}-\mu_{x2})(y_{int}-\mu_{y2})-1=0, \end{array} \tag{7}$$

where

$$\begin{cases} e_2 = 1-2\ln(\lambda) \\ a_2 = \frac{1}{(1-\rho_2^2)\sigma_{x2}^2 e_2} \\ b_2 = \frac{1}{(1-\rho_2^2)\sigma_{y2}^2 e_2} \\ c_2 = -\frac{2\rho_2}{\sigma_{x2}\sigma_{y2}(1-\rho_2^2)e_2}. \end{cases} \tag{8}$$

Case 2: In this case, we find the same equation 5 and 7 with the parameters

$$\begin{cases} e_1 = 1-2\ln(\lambda) \\ a_1 = \frac{1}{(1-\rho_1^2)\sigma_{x1}^2 e_1} \\ b_1 = \frac{1}{(1-\rho_1^2)\sigma_{y1}^2 e_1} \\ c_1 = -\frac{2\rho_1}{\sigma_{x1}\sigma_{y1}(1-\rho_1^2)e_1} \end{cases} \tag{9}$$

for the first component and for the second component, we find

$$\begin{cases} e_2 = 1-2\ln\left(\frac{\lambda\kappa_1\sigma_{x2}\sigma_{y2}\sqrt{1-\rho_2^2}}{\kappa_2\sigma_{x1}\sigma_{y1}\sqrt{1-\rho_1^2}}\right) \\ a_2 = \frac{1}{(1-\rho_2^2)\sigma_{x2}^2 e_2} \\ b_2 = \frac{1}{(1-\rho_2^2)\sigma_{y2}^2 e_2} \\ c_2 = -\frac{2\rho_2}{\sigma_{x2}\sigma_{y2}(1-\rho_2^2)e_2} \end{cases} \tag{10}$$

In the plan defined by equation $(T) : z = \min(\Gamma_1(\mu_1), \Gamma_2(\mu_2))$, equations 5 and 7 are characteristic equations of two ellipses. This means that fixing the center of the second component return to fix the center of the second ellipse. First, we compute the value of the intersection point after, we compute value of the second component center. We proceed to some transformations of the plan marker, we will translate the marker after we will rotate it such that the big axis of the ellipse will be parallel to the X axis of the new marker. Let R the marker in the plan (T) (see Fig.3).

We translate the marker by the vector $\boldsymbol{m}(\mu_{x1}, \mu_{y1})$. Equation 5 will

$$a_2 x_{int}^2 + b_2 y_{int}^2 + c_2 x_{int} y_{int} - 1 = 0. \tag{11}$$

Now, we obtain an ellipse where it center is the center of the marker. After translating the marker, we proceed to the rotation in which the big axis of the ellipse will be parallel to the X axis, let call this new marker by R_1. Figure 3 illustrates the markers and the angles used for the rotation. We consider the rotation angle ϕ_1. The application of a rotation in a bivariate space is given by

$$\begin{cases} \dot{x} = x\cos(\phi_1) - y\sin(\phi_1) \\ \dot{y} = x\sin(\phi_1) + y\cos(\phi_1) \end{cases} \tag{12}$$

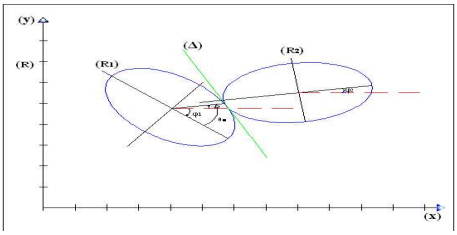


Fig. 3. Illustration of ellipses intersection, the different markers and angles used for rotation and the translation of the deferent markers.

where $\grave{x}, \grave{y}$ are the new coordinates in the new marker. The first ellipse equation after the rotation is given by

$$a_1\grave{x} + b_1\grave{y} - 1 = 0. \tag{13}$$

where a_1 and b_1 are reals strictly positives. After some transformation, we arrive

$$\begin{cases} \phi_1 = 0.5 \arctan(\frac{c_1}{b_1 - c_1}) \\ \grave{b_1} = \frac{b_1 \cos^2(\phi_1) - a_1 \sin^2(\phi_1)}{\cos(2\phi_1)} \\ \grave{a_1} = \frac{a_1 \cos^2(\phi_1) - b_1 \sin^2(\phi_1)}{\cos(2\phi_1)} \end{cases} \tag{14}$$

where angle θ is given by the users represents the deviation of the intersection point from the X axis in the marker R. We conclude that the intersection angle in the marker R_1 is

$$\theta_{int} = \phi_1 + \theta. \tag{15}$$

Using the parametrical equation of the ellipse, we conclude the value of the intersection point in the marker R_1.

$$\begin{cases} t = \arctan(\sqrt{\frac{\grave{b_1}}{\grave{a_1}}} \tan(\theta_{int})) \\ x_1 = \frac{\cos(t)}{\sqrt{\grave{a_1}}} \\ y_1 = \frac{\sin(t)}{\sqrt{\grave{b_1}}}. \end{cases} \tag{16}$$

In order to compute the value of the center of the second component, we need the value of the obliquity in the intersection point, hear we have three cases, case 1; $\theta_{int} \in]0, \pi/2[$, case 2: $\theta_{int} \in]-\pi/2, 0[$ and case 3: $\theta_{int} = 0$.
For the first case, the value of the obliquity of the tangent in the intersection point is given by

$$\delta_1 = -\frac{a_1\grave{x_1}}{b_1\sqrt{1 - a_1\grave{x_1}^2}} \tag{17}$$

For the second case, we find that

$$\delta_1 = \frac{a_1 \grave{x_1}}{b_1 \sqrt{1 - a_1 \grave{x_1}^2}} \tag{18}$$

Now, we compute the intersection point value (x_0, y_0) and the obliquity of the tangent δ_0 in the marker (R). First, We proceed to the rotation, after, we translate the resulted marker such that

$$\begin{cases} x_0 = x_1 \cos(-\phi_1) - y_1 \sin(-\phi_1) + \mu_{x1} \\ y_0 = x_1 \sin(-\phi_1) + y_1 \cos(-\phi_1) + \mu_{y1} \\ \delta_0 = \frac{\sin(-\phi_1) + \delta_1 \cos(-\phi_1)}{\cos(-\phi_1) - \delta_1 \sin(-\phi_1)} \end{cases} \tag{19}$$

For the second ellipse, we draw the plan (R) by the vector $\boldsymbol{v}(\mu_{x2}, \mu_{y2})$ after, as we do with the first ellipse, we compute the angle θ_2 such that the resulted marker has an axis X parallel to the big axis of the second ellipse. So we arrive

$$\begin{cases} \phi_2 = 0.5 \arctan(\frac{c_2}{b_2 - c_2}) \\ \grave{b_2} = \frac{b_2 \cos^2(\phi_2) - a_2 \sin^2(\phi_2)}{\cos(2\phi_2)} \\ \grave{a_2} = \frac{a_2 \cos^2(\phi_2) - b_2 \sin^2(\phi_2)}{\cos(2\phi_2)} \end{cases} \tag{20}$$

The value of the obliquity of the tangent in the new marker (R_2) is given by

$$\begin{cases} \delta_2 = \frac{\sin(\phi_2) + \delta_0 \cos(\phi_2)}{\cos(\phi_2) - \delta_0 \sin(\phi_2)} \\ \delta_2 = -\frac{\cos(\phi_2)}{\sin(\phi_2)}, \text{ if } (\theta_{int} = 0). \end{cases} \tag{21}$$

We conclude the value of the intersection point coordinates (x_2, y_2) in the marker R_2.

$$\begin{cases} x_2 = -\sqrt{\frac{\delta_2^2 \grave{b}}{\delta_2^2 \grave{b_2} \grave{a_2} + \grave{a_2^2}}} \\ y_2 = \sqrt{\frac{1 - \grave{a_2} x_2^2}{\grave{b_2}}} \text{ if } \delta_2 < 0 \\ y_2 = -\sqrt{\frac{1 - \grave{a_2} x_2^2}{\grave{b_2}}} \text{ if } \delta_2 > 0. \end{cases} \tag{22}$$

x_2 can take a negative value but in order to assure that there are no total overlap between the no adjacent components, we chose the negative values.

We compute the coordinates x_2 and y_2 in R on function of μ_{x2} and μ_{y2} after we conclude the values μ_{x2} and μ_{y2} where

$$\begin{cases} \mu_{x2} = x_2 \cos(-\phi_2) - y_2 \sin(-\phi_2) + x_0 \\ \mu_{y2} = x_2 \sin(-\phi_2) + y_2 \cos(-\phi_2) + y_0 \end{cases} \tag{23}$$

For the maximal overlap, we can substitute $\lambda = 1$ and we said that the maximal overlap is a special case of the partial overlap. the complexity of the algorithm is quadratic. According to Milligan [13], controlling the overlap is sufficient for one dimension to have naturel cluster. For our methods, we control the overlap in 2D space which means that our method gives more accurate structures of clusters.

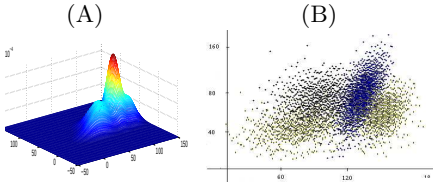
5 Experimental results

For each value of the overlap rate $\lambda \in \{0, 0.5, 0.75, 1\}$, we employed an artificial mixture having a number of component between 2 and 7. For clustering methods, we choose as minimal number of clusters 2 and as maximal 10.Figures of Table1 (A) and (C) represent the distributional and observations of mixtures samples of the served constructed data respectively. As in the example of 4 components, for each components number, we keep the same mixtures parameters and we conclude the value of centers by using the generation method. As it is seen in table 1(C), we evaluate fussy c-means(FCM), FCM based splitting algorithm (FBSA) [12, 18]. FBSA is proposed to provide an automatic routine for the models initialization. For the first time, the centers are randomly initialized. After, the cluster which has the most dispersed observations to its number is split to two such that the new centers are far the maximum each to other and to the others components' centers. The new centers must verify also that at least 10% of the cluster elements devised are nearest to each new centers. The determination of the optimal number of clusters can be processed as a problem of model selection by using a cross validation technique. The main validating algorithm is deduced by using a clustering routines for various numbers of clusters which allows us to have many models representing the data structure. For each model, we apply a validity index. According to the values of the index, the optimal number of cluster is chosen. The indices of validity belong to one of the three classes. Some indices exploit the obtained structure by the clustering methods like Davies-Bouldin index (DB) [4] and R-squared (RS) [7], others use matrix of computed memberships by the fuzzy methods as Partition coefficient (PC) and Classification Entropy (CE) [7, 5]. The thirds classes are hybrid and they employ the structure of the data and the membership's matrix like Wang-Sun-Jian (WSJ)[18] and Xie-Benie indices(XB)[21]. K-Means can only use DB and RS because the others indices require the memberships computed by the fuzzy methods. For each components number, we illustrate the result in group according to the rate of overlap λ.

From the results, we can easily observe that the indices RS, PC and WSJ often give the same results which means that these indices suffer from the problem of monotony tendency. For a number of observations sufficiently and relatively high to components number, these indices give no satisfied results. In [15, 10], we arrive at the same results for the uni-variant and the no-correlated data. In [19], a study concerning the WSJ is shown. It based on the Bersdak suggestion, where the observations number $N = \sqrt{C_{max}}$. The proportion N/C_{max}, in that study, assure a good results but in our case where the number of observations $N = 3000$, it is clear that WSJ is no monotone. For the same reason, PC and RS have a monotony tendency.

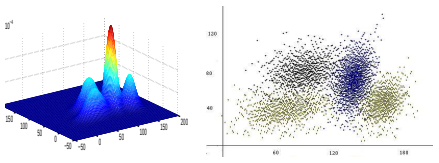
We also confirm that with increasing the components number or the overlap rate, the results became less exact. From the results, validity indices DB, XB and PC determine the components number but, with the same rate in mixture with large components number, the above validity indices has not the ability to identify the true component numbers. A large components number means that

(A) (B) (C)



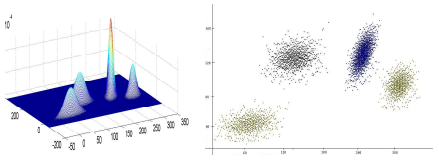
	RS	PC	DB	XB	WSJ	CE
FBSA	2	10	2	2	8	2
FCM	2	10	2	2	10	2
			DB		RS	
K-means			2		2	

a) $\lambda = 1$



	RS	PC	DB	XB	WSJ	CE
FBSA	2	10	4	10	10	2
FCM	2	10	4	9	10	2
			DB		RS	
K-means			2		2	

b) $\lambda = 0.5$



	RS	PC	DB	XB	WSJ	CE
FBSA	2	10	4	4	10	4
FCM	2	10	4	4	10	4
			DB		RS	
K-means			3		2	

c) $\lambda = 0.001$

Table 1. Variation of the rate of overlap between 4 components of the mixture and the result of the process of clustering.

there are relatively a large global minimum to locate, such that, between these global minimum a large set of local minimum where clustering methods can be converged. Form Table 1(C), we can easily detect that the determination of the component number became less exact by increasing the overlap rate. From the results, we locate that all the no monotonous validity indices resolve the components number with a rate $\lambda = 0.001$ but none of them can conclude the exact components number with $\lambda = 1$. These results are confirmed by examining the other table. We can conclude the influence of the overlap in the set of results.

Comparatively to the experiences concerning the 1D data cited in [15], the process of clustering, at this time, cannot converge frequently to the true models as in the case of 1D. Here, the influence of the curse of dimensionality clearly appears for two big reasons. The first one concerns the frequencies dispersions of the data. For the same observations number, in 1D space, the data is dispersed only on one dimension which conduct the data to be more compact and the *pdf* appear like it is a continuous function with less local minimum. Contrary for the case of 2D data, the data is distributed on two axes such that a small empty space more larger then in 1D. Analytically, these spaces are viewed as locals

minimum. In the *pdf* representation, it appear as noises. The second reason concerns overlap between more then two adjacent components. In [15] for the 1D data, we confirmed that the worse situations in which clustering methods find a big difficulties to determine the exact components number is where there is a component with a small deviation between two components covering large standard deviations. In these circumstances, the first components overlap beyond the second adjacent component and reached the third component. In 2D, the ternary overlap is more occurred. Suppose for example that we have three 2D components such that the intersection angle $\theta_{int} = 0.45\pi$, between the second and the third component $\theta_{int} = -0.45\pi$. The first component in this situation is near strongly to the third components such that we can find that the first and the third components are in case of total overlap. For this reason, we limited the intersections angles to be included in $]-\pi/3, \pi/3[$, despite, the situation stay more complicate than in the case of 1D data. Another reason, which make the two first more intricate, depend to the initialization, the initial centers which we suggest for the 1D give more chance to the process of clustering to regain the true centers of the data. In the case of 2D, the large space constituting by the two axes make difficult to easily reach the true model centers. In [10], we practically arrived to the same results with no correlated data. We conclude that the correlation between the data has not attraction to results and the major factors that affect clustering process stay: the overlap degree, the components number and the dimensionality of the data.

6 Conclusion

In this paper, we proposed a new algorithm for generating artificial correlated bivariate gaussian data. Our generator is based on the set of interesting definitions. This definition respects the two criteria of internal cohesion and external isolation. this two criteria can be verified by using no automatical method as visual inspection. So our generator present a formal methodologies for controlling the overlap between the component of the mixture. The figures presented in this papers show the effectiveness of this method to control the overlap. It can be used in many problem related to the bivariate gaussian simulated data.

References

1. E.M. Aitnouri, F. Dubeau, S. Wang, and D. Ziou. Controlling mixture component overlap for clustering algorithms evaluation. *Pattern Recognition and Image Analysis*, 12(4):331–346, 2002.
2. E.M. Aitnouri, S. Wang, and D. Ziou. On comparison of clustering techniques for histogram pdf estimation. *Pattern Recognition Image Analysis*, 10(2):206–217, 2000.
3. M.R Anderberg. *Cluster Analysis for Applications.* New York: Academic Press, 1973.

4. O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–246, 2013.
5. J.C. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3):58–72, 1974.
6. R.M. Cormack. A review of classification. *Journal of the Royal Statistical Society*, 134(3):321–367, 1971.
7. A.K. Das and J. Sil. Cluter validation methods for stable cluster formation. *Canadian Journal of Artificial Intelligence, Machine Learning and pattern recognition*, 1(3):26–41, july 2010.
8. S. Dasgupta. Learning mixtures of gaussians. *IN Proceeding of the IEEE Symposium on Fondations of Computer Science, New York*, pages pp. 633–644, 1999.
9. B.S. Everitt. *Cluster Analysis*. Heinemann Educational [for] the Social Science Research Council, 1974.
10. R. Gharbaoui, M. Ouali, and E. Aitnouri. A mixture model-based 2d data generator for performance with controlled overlap for performance evaluation. *Engineering and Technology*, 78:73–80, 2011.
11. S. Jahirabadkar and P. Kulkarni. Algorithm to determine ε-distance parameter in density based clustering. *Expert Systems with Applications*, 41:2939–2946, 2014.
12. L.Szilagyi, S.M.Szilgyi, and C.Enachescu. A study on cluster size sensitivity of fuzzy c-means algorithm variants. *Neural Information Processing*, 99(48):470–478, 2016.
13. G.W. Milligan. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3):325–342, 1980.
14. M. Ouali and E.M. Aitnouri. Performance evaluation of clustering technique for image segmentation. *Computer Science Journal of Maldova*, 18(03):271–302, 2011.
15. M. Ouali, R. Gharbaoui, and E. Aitnouri. Benchmarking taxonomy for 1d clustering algorithms. In *System, Signal processing and thier application(WOSSPA 2011)*, pages 151–154, May 2011.
16. W. Qui and H. Joe. Separation index and partial membership for clustering. *Computational Statistics and Data Analysis*, 50(3):585–603, 603 2006.
17. A. S. Salem and K. A. Nandy. Developpement of assessment criteria for clustering algorithms. *Pattern Analysis and Application*, 12:79–98, 2009.
18. H. Sun, S. Wang, and Q. Jiang. Fcm-based model selection algorithm for dermining the number of clusters. *Pattern Recognition*, 37(10):2027–2037, 2004.
19. H. Sun, S. Wang, and Q. Jiang. Fcm-based model selection algorithm for determinig the number of cluster. *Pattern Recognition*, 37(10):2027–2037, 2004.
20. S. Tabbone. *Détection multi-échelle de contours sous-pixel et de jonctions*. PhD thesis, Institut National Polytechnique de Lorraine, France, 1994.
21. Y. Tang and F. Sun. Improved validy index for fuzzy clustering. *American Control Conference*, 2:1120–1125, 2005.

Le couplage généralisé dans un graphe biparti*

Dj. TALEM[1] and B. SADI[2]

[1] Laboratoire de Recherche Opérationnelle et Mathématiques de Décision, Université Tizi Ouzou, Algerie
vouleze@yahoo.fr

[2] Laboratoire de Recherche Opérationnelle et Mathématiques de Décision, Université Tizi Ouzou, Algerie
sadibach@yahoo.fr

Résumé Soient $P = (E, \leq_P)$ un ensemble partiellement ordonné et $G = (X, Y, E)$ un graphe biparti dont les arêtes sont isomorphes aux éléments de P. Un sous ensemble d'arêtes M dans G est appelé P-couplage ou couplage généralisé de G si pour toute antichaîne A de P, M/A est un couplage dans G ; autrement dit, pour toute antichaîne A de P,le sous ensemble de M constitué uniquement par les arêtes qui sont isomorphes aux éléments de A est un couplage. Il s'agit de trouver un P-couplage avec un maximum possible d'éléments. Il est facile de voire que si P est une antichaîne, alors M est un couplage, et donc le problème est facile. Dans ce papier, nous présentons quelques classes d'ordres et aussi quelques classes de graphes biparti pour lesquelles un P-couplage maximum peut être calculé en un temps polynomiale.

Keywords: Couplage · Couplage généralisé · Ensemble partiellement ordonné Ordre d'intervalle.

1 Préliminaire

Nous supposons que les ordres considérés dans cet article sont finis. Un ensemble ordonné ou un ordre (poset) est un couple $P = (E, \leq_p)$ où E est un ensemble non vide appelé la base de P et " $\leq_p$ " est une relation binaire sur E réflexive, antisymétrique et transitive. L'ordre P est fini lorsque la base E est finie. Deux éléments $x, y \in E$ sont dits comparables si, $x \leq_P y$ ou $y \leq_P x$; on écrit $x \sim_P y$. Sinon, ils sont dits incomparables et on écrit $x \| _P y$. Un couple $(x, y) \in E \times E$ est une couverture et on écrit $x \prec_P y$ si $x <_P y$ et il n'existe aucun élément $z \in E$ vérifiant $x <_P z <_P y$. Une chaîne de P est un sous-ensemble d'éléments de E deux à deux comparables ; sa longueur est le nombre de ses éléments moins un. Une antichaîne est un sous-ensemble d'éléments de E deux à deux incomparables.
Pour x, $y \in E$, la relation $x <_P y$ signifie aussi que y est un successeur de x, donc x est un prédécesseur de y ; $Succ(x) = \{y \in E, x <_P y\}$ désigne l'ensemble de tous les successeurs de x ; $Pred(x) = \{y \in E, y <_P x\}$ désigne l'ensemble de tous les prédécesseurs de x.

⋆. Supported by organization x.

Un élément x est maximal dans P si $Succ(x) = \emptyset$; un élément x est minimal dans P si $Pred(x) = \emptyset$. On note $Max(P)$ (resp. $Min(P)$) l'ensemble des maximaux (resp. des minimaux) de P.
Pour tout $F \subseteq E$, le couple $(F, \leq_F)$ où " $\leq_F$ " est la restriction de la relation " $\leq_P$ " sur F est un sous-ordre induit par F. Pour tout ordre fini $P = (E, \leq_P)$, **rang** est l'application de E dans l'ensemble des entiers naturels $\mathbf{N}$ définie par : $\forall x \in E$,

$$\text{rang(x)}=\begin{cases} 0 & \text{si } x \in Min(P), \\ 1 + max_{\{y \in Pred(x)\}} rang(y) & \text{sinon} \end{cases} \quad (1)$$

Le i-ième niveau de P est le sous-ensemble $N_{i-1} = \{x \in E / rang(x) = i-1\}$. Les éléments d'un même niveau sont deux à deux incomparables. De plus, $\forall x \in N_i$, $\exists x_0, x_1 \ldots, x_i = x$ tels que $x_0 \prec_p x_1 \prec_p \ldots \prec_p x_i = x$.

Un graphe G est un couple $G = (V(G), E(G))$, où $V(G)$ est l'ensemble des sommets de G et $E(G) = \big\{\{u, v\} : u, v \in V(G)\big\}$ est l'ensemble des arêtes de G. S'il n y a aucune confusion, on écrit $G = (V, E)$. Les sommets $u, v \in V$ sont adjacents si $\{u, v\} \in E$; dans ce cas on dit que l'arête $\{u, v\}$ est incidente aux sommets u et v. Pour tout sommet $v \in V$, $N(v) = \big\{u, \{u, v\} \in E\big\}$ dénote l'ensemble des sommets de G qui sont adjacents à v. Pour tout ensemble $S \subset V$, $G[S]$ dénote le sous graphe de G induit par S. Un graphe G est complet si ses éléments sont deux à deux adjacents. Un graphe $G = (V, E)$ est dit biparti si l'ensemble de ses sommets peut être partitionné en deux sous ensembles $V = X \cup Y$ tel que les éléments de X (resp. Y) sont deux à deux non adjacents ; dans ce cas, on écrit $G = (X, Y, E)$. Un graphe biparti G est dit complet si tout élément de X est adjacent à tous les éléments de Y ; la notation $K_{n,m}$ désigne un graphe biparti complet, avec $|X| = n, |Y| = m$. Un graphe biparti complet $K_{1,m}$ est appelé étoile. Une sequence de sommets $(v_1, v_2, \ldots, v_k)$ forment un chemin dans G si, pour $i = 2 \ldots k$, $\{v_{i-1}, v_i\}$ est une arête dans G. La longueur de ce chemin est $k-1$. Un graphe $G = (V, E)$ est connexe si pour toute paire de deux sommets u et v il existe un chemin dans G allant de u à v. Le chemin $(v_1, v_2 \ldots v_k)$ forme un cycle dans G si $\{v_1, v_k\}$ est une arête dans G. La longueur de ce cycle est k. Une arête $\{v_i, v_j\}$ est une corde dans le chemin $(v_1, \ldots, v_k)$ si $|i - j| > 1$. $\{v_1, v_k\}$ est une corde dans le chemin $(v_1, v_2 \ldots, v_k)$, mais ce n'est pas une corde dans le cycle $(v_1, v_2 \ldots, v_k)$. Deux graphes G et H sont isomorphes s'il existe une bijection f entre $V(G)$ et $V(H)$ telle que, $\forall u, v \in V(G)$, $\{u, v\} \in E(G)$ si et seulement si $\{f(u), f(v)\} \in E(H)$. Pour une famille Γ de graphes, un graphe G est Γ-libre si G n'a aucun sous graphe isomorphe à un graphe de Γ. C_k désigne un cycle de k sommets sans corde ou un trou de longueur k.

2 Définition et propriétés d'un P-couplage

Soient $P = (E, \leq_P)$ un ensemble partiellement ordonné et $G = (X, Y, E)$ un graphe biparti associé à P dont les arêtes sont isomorphes aux éléments de P.

Dans toute la suite, pour une antichaîne A de P et un sous ensemble d'arêtes M de G, la notation M/A qui est la même que $M \cap A$ désigne l'ensemble des arêtes qui sont isomorphes aux éléments de A.

Définition 1. *On appelle P-couplage ou couplage généralisé tout ensemble d'arêtes M dans G pour lequel M/A est un couplage dans G pour toute antichaîne A de P.*

Le problème de P-couplage maximum est défini comme suit :
(PROBLEME DU P-COUPLAGE MAXIMUM)
Instance : Un poset P et un graphe biparti $G = (X, Y, E)$
Objectif : Déterminer un P-couplage maximum de G.
La proposition ci-dessous donne une caractérisation des éléments d'un P couplage.

Proposition 1. *Soient $P = (E, \leq_P)$ un ensemble partiellement ordonné et $G = (X, Y, E)$ son graphe biparti associé. Une condition nécessaire et suffisante pour que deux éléments u et v soient dans un même P-couplage M est que u et v soient disjointes dans G ou comparables dans P.*

Démonstration. Soient u et v deux arêtes de G. Si u et v sont disjointes, il est évident qu'elles peuvent appartenir à un même P-couplage ; si elles sont comparables dans P, alors elles ne peuvent pas appartenir à une même antichaîne A, et donc M/A contient au plus l'une des deux arêtes, ce qui veut dire qu'elles peuvent appartenir toutes les deux à M. D'où, la condition est suffisante. Réciproquement, si u et v sont incomparables dans P et non disjointes dans G, alors il existe une antichaîne A telles que $u, v \in A$, et donc $u, v \in A \cap M$. Mais ceci contredit la définition d'un P-couplage. Par conséquent, la condition est nécessaire. □

Nous allons voir maintenant que le calcul d'un P-couplage maximum est équivalent au calcul d'une clique maximum dans un graphe associé à G. En effet, soit $H = (E, E(H))$ le graphe dont les sommets sont les arêtes de G et deux arêtes u et v de G sont adjacentes dans H si et seulement si elles peuvent appartenir à un même P-couplage M. Il en résulte que tout P-couplage de G est une clique dans H, et vice versa.
Le long de ce papier, P est un ordre, G est un graphe associe à P et H est un graphe associe à G comme indiqué précédement.

Remarque 1. *1. Notons que si u et v sont deux arêtes disjointes dans G, alors le nombre maximum d'arêtes qui puissent intersecter simultanément u et v est deux. Dans ce cas les 4 arêtes définissent un C_4 dans G.*
2. Par définition, deux éléments de P qui ne sont pas adjacents dans H sont non disjoints dans G. la réciproque est fausse : les éléments comparables dans P et non disjoints dans G sont adjacents dans H. Ainsi, il est facile de voir qu'un ensemble d'arêtes dans G non connexe avec au moins 5 arêtes ne peut pas induire un trou dans H.

Proposition 2. *Soient a et b deux arêtes disjointes dans G. Alors le plus grand trou qui passe par les sommets a et b dans H est de longueur $n \leq 6$.*

Démonstration. Soient a et b deux arêtes disjointes dans G. Par définition d'un P-couplage, a et b sont adjacentes dans H. Notons que les arêtes a et b peuvent appartenir à un C_3, il suffit de considérer une troisième arête dans G qui soit disjointe avec a et b ; la configuration de la Figure 1(a), montre que les arêtes a et b peuvent appartenir à un C_4, il suffit de considérer le sous ordre defini par les deux couvertures $a \prec b$, $d \prec c$; la configuration de la Figure 1(b), montre que les arêtes a et b peuvent appartenir à un C_5, il suffit de considérer le sous ordre défini par les couvertures $a \prec e, b \prec c, d \prec c, d \prec e$; en fin, La configuration de la Figure 1(c), montre que les arêtes a et b peuvent appartenir à un C_6, il suffit que les arêtes a, b, c, d, e, f induisent une antichaîne dans P. Supposons maintenant qu'il existe un trou $\mu = (a, b, c, d, e, f, g)$ de longueur 7. Dans ce cas, les arêtes d, e et f devraient intersecter simultanément les arêtes a et b. Or d'après la Remarque 1, ceci est impossible. □

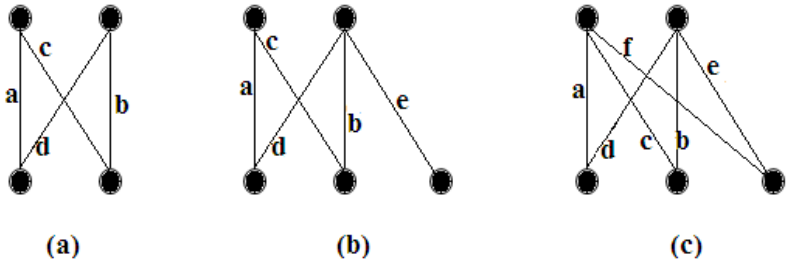


Figure 1.

Proposition 3. *Si P est une antichaîne alors le plus long trou dans H est C_4.*

Démonstration. Puisque P est une antichaîne alors, d'après la Proposition 1, deux éléments de P sont adjacents dans H (i.e appartiennent à M) si et seulement si ils sont disjoints dans G. D'après la Proposition 2, la longueur d'un plus grand trou ne dépasse pas 6. Supposons maintenant qu'il existe un trou $\mu = (a, b, c, d, e)$ (voir la Figure 2(a)). Il en résulte que dans G : $a \cap b = \emptyset$, d intersecte simultanément a, b. L'arête c intersecte a et n'intersecte aucune des arêtes b et d (voir la Figure 2(b) qui en donne une configuration à isomorphisme près). Pour compléter, l'arête e doit intersecter simultanément b et c, la Figure 2(c) donne les deux configurations possibles de μ dans G. Or, l'arête e doit être disjointe avec les arêtes a et d. Absurde.□

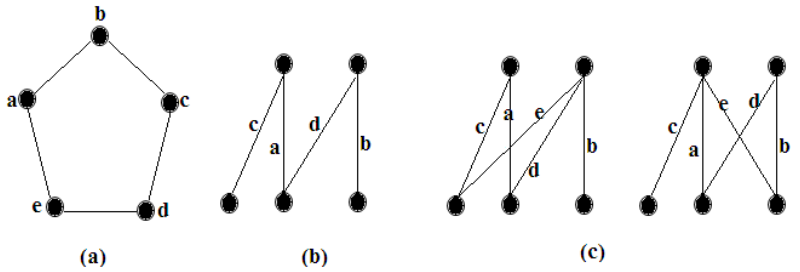


Figure 2.

Dans le graphe complémentaire de H, noté H^c, deux éléments u et v sont adjacents si et seulement si u et v sont incomparables dans P et non disjoints dans G. Il en résulte que les arêtes disjointes dans G ne sont pas adjacentes dans H^c. Cependant, la réciproque est fausse, car deux sommets non adjacents dans H^c peuvent correspondre aux arêtes non disjointes dans G et qui sont comparables dans P.

Proposition 4. *Si P est une antichaîne alors H^c est un graphe sans trou impair.*

Démonstration. Soit $\mu = (a_1, a_2, \ldots, a_k), k \geq 5$ un cycle de longueur k. Supposons que k est impair et μ est un trou.
Puisque P est une antichaîne, alors deux éléments u et v sont adjacents dans H^c si et seulement si ils ne sont pas disjoints dans G. Ainsi, les deux configuration possibles de μ dans G sont données par la figure 3.

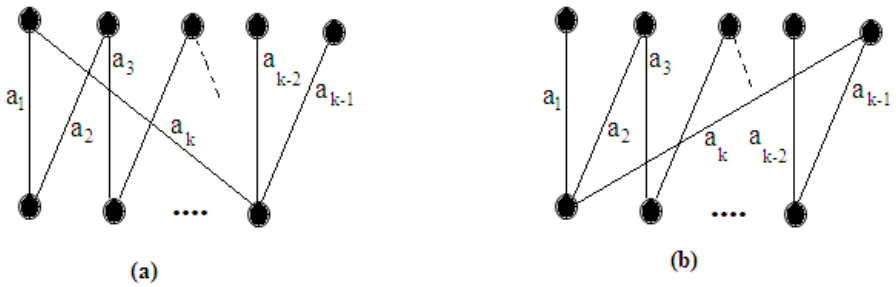


Figure 3.

Mais d'après cette figure, on voit bien que $d\mu(a_k) = 3$ (le degré de a_k dans le sous graphe induit par les éléments de μ), et ceci implique que μ ne peut pas être un trou. □

3 Classes polynomiales d'un P-couplage

Dans la section précédente, on a vu que le problème du P-couplage maximum dans le graphe G est équivalent au problème de la clique maximum dans le graphe H. Les graphes parfaits sont les graphes ne contenant ni trou ni anti-trou (complémentaire d'un trou) impair comme sous graphe induit, ils ont été définis par Claude Berge au debut des années 60 [9]. Un graphe faiblement triangulé est un graphe ne contenant ni trou ni anti-trou de longueur supérieure ou égale à 5. Il est claire que les graphes faiblement triangulés, qui ont été introduits par Hayward en 1985 [7], sont parfaits. Il est connu que le problème de la clique maximum, qui est NP-complet en générale [8], peut être résolu en un temps polynomiale pour la classe de graphes parfaits et donc pour la classe de graphes faiblement triangulés [6]. Nous allons voir maintenant que le graphe H associé à G est un graphe parfait si P est un ordre faible, il est un graphe faiblement triangulé si G est sans cycle.
Comme cas particuliers, nous avons deux cas :
1. Si P est une antichaîne, alors les éléments d'un P-couplage M doivent être deux à deux disjoints. Dans ce cas, un P-couplage maximum est un couplage maximum, dans le graphe biparti G, qui peut être calculé en un temps polynomiale grâce l'algorithme d'Edmonds [5].
2. Si G est une étoile, alors les éléments d'un P-couplage M doivent êtres deux à deux comparables dans P. Ici, H est exactement le graphe de comparabilité de P (un graphe dont les sommets sont isomorphes aux éléments de P et deux sommets sont adjacents dans H si et seulement si ils sont comparables dans P). Là aussi, un P-couplage maximum peut être calculé en un temps polynomiale [1].

Définition 2. *Un ordre faible (weak order) est un ordre obtenu par composition séries d'antichaînes. Ainsi, les seules antichaînes d'un ordre faible sont ses niveaux.*

Théorèm 1. *Si P est un ordre faible, alors H est un graphe parfait.*

Démonstration. Soit $A = \{a_1, a_2, \ldots, a_k\}$ une antichaîne maximale dans P. Alors, d'après les Propositions 3 et 4, ni $H[A]$ ni $H^c[A]$ ne peut être un trou impair. Soit maintenant $x \in P \setminus A$. Puisque P est faible, alors x est comparable à tous les éléments de A. Ceci implique que le sommet x est adjacent à tous les éléments de A dans H et x n'est adjacent à aucun élément de A dans H^c. et donc ni $H[A \cup \{x\}]$ ni $H^c[A \cup \{x\}]$ est un trou. Par conséquent ni H ni H^c ne peut contenir de trous impairs, c'est-à-dire H est parfait. □

Lemme 1. *Soient a et b deux éléments de P qui sont disjoints dans G. Si le graphe G est sans cycle alors*
1. Aucun cycle ne passe par a et b dans H^c.
2. Aucun trou de longueur supérieur à 4 ne passe par a et b dans H.

Démonstration. 1. Remarquons que si G est un graphe sans cycle, alors deux arêtes quelconque sont connectées par au plus un chemin. Supposons que dans H^c il existe un cycle passant par les sommets a et b. Donc a et b sont connectés dans H^c par deux chemins $\mu_1 = (a, x_1, x_2, \ldots, x_k, b)$ et $\mu_2 = (b, y_1, \ldots, y_l, a)$. Or, par définition de H^c, deux éléments consécutifs dans μ_1 (resp. μ_2) ne sont pas disjoints dans G, et donc μ_1 et μ_2 sont deux chemins différents qui passent par a et b dans G. Impossible, car G est sans cycle.
2. D'après la Proposition 2, il suffit de démontrer que H est un graphe C_5-libre. Supposons le contraire et soit $\mu = (a, b, c, d, e)$ un trou dans H de longueur 5 comme dans la Figure 4(a).

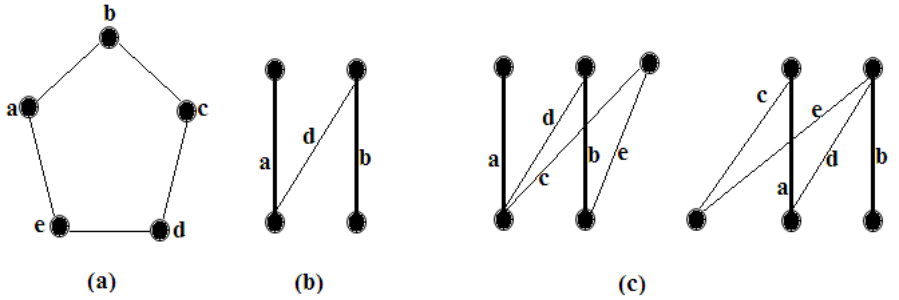


Figure 4.

Dans ce cas, par définition d'un P-couplage, l'arête d intersecte simultanément les arêtes a et b dans G (voir la Figure 4(b)). Par définition d'un P-couplage, nous avons aussi dans G, $a \cap c \neq \emptyset$, $b \cap e \neq \emptyset$, $e \cap c \neq \emptyset$. Il en résulte que les seules configurations possibles pour μ dans G sont données par la Figure 4(c). Mais, ceci contredit le fait que G est un graphe biparti sans cycle. □

Théorèm 2. *Si le graphe G est sans cycle, alors H est un graphe faiblement triangulé.*

Démonstration. D'après la Proposition 2, il suffit de montrer que H et H^c sont C_5-libres.
Soit $\mu = \{a, b, c, d, e\} \subset E$. D'après les Propositions 3 et 4, si les éléments de μ sont deux à deux incomparables dans P, alors μ ne peut pas être un trou de longueur 5 ni dans H ni dans H^c. Si le sous graphe défini par μ dans G est une étoile, alors le sous graphe $H[\mu]$ (resp. $H^c[\mu]$) induit par μ dans H (resp. H^c) est un graphe de comparabilité (resp. incomparrabilté) et donc ne peut pas

être un trou (les graphes de comparabilités et d'incomparabilités sont parfaits [1]). Enfin, d'après le Lemme 1, si μ contient au moins deux éléments qui sont disjoints dans G, là aussi, il ne peut pas induire un trou ni dans H ni dans H^c. Par conséquent, H est un graphe faiblement triangulé. □

Références

1. M.C. Golumbic Algorithmic Graph Theory and Perfect Graphs, Elsevier, Second Edition (2004).
2. N. Caspard, B. Leclerc, B. Monjardet, Finite Ordered Sets : Concepts, Results and Uses. Springer, Berlin Heidelberg New York (2007).
3. D. Jungnickel, Graphs, Networks and Algorithms, Springer, Berlin Heidelberg New York Dordrecht London (2013).
4. A. Brandstädt, V.B. Le., and J. Spinrad. Graph classes : a survey. SIAM, Philadelphia, (1999).
5. J. Edmonds : Paths, trees, and flowers. Canadian Journal of Mathematics, 17(3) : 449-467, (1965).
6. A. Pecher, A.K. Wagler, Clique and chromatic number of circular-perfect graphs, Proceedings of ISCO 2010 - International Symposium on Combinatorial Optimization, Elec. Notes in Discrete Math 36 199-206 (2010)
7. R.B. Hayward, Weakly triangulated graphs, J. Combin. Theory B, 39, 200-208 (1985).
8. R.Karp, Reducibility among combinatorial problems, complexity of computer computation 85-103 plenum pres (1972).
9. C. Berge and J. L. Ramrez Alfonsn. Origins and Genesis. In Ramrez Alfonsn and Reed, pages 1-12 (1994).

b-coloration des arêtes de certains graphes

Amel BENDALI-BRAHAM[1], Noureddine IKHLEF ESCHOUF[2], and Mostafa BLIDIA[3]

[1] Laboratoire de Mécaniques, Physiques et Modélisation Mathématique,
Département de Mathématiques et Informatique,
Faculté des Sciences, Université de Médéa, Algérie.
bendali-braham@hotmail.fr
[2] Département de Mathématiques et Informatique,
Faculté des Sciences, Université de Médéa, Algérie.
nour_echouf@yahoo.fr
[3] Laboratoire LAMDA-RO,
Département de Mathématiques,
Faculté des Sciences, Université de Blida 1, B.P. 270, Blida, Algérie.
m_blidia@yahoo.fr

Résumé Une *b*-coloration des arêtes d'un graphe G est une coloration propre des arêtes de G de telle sorte que chaque classe de couleur possède au moins une arête incidente à au moins une arête dans chacune des autres classes de couleur. Une telle arête est dite *b*-arête. L'indice *b*-chromatique, noté b'(G), est le nombre maximum de couleurs pour lequel G admet une *b*-coloration des arêtes avec b'(G) couleurs. Dans ce travail, nous donnons des valeurs exactes et des bornes de l'indice *b*-chromatique de certains graphes, comme la chaîne, le cycle, la roue, la double étoile, le graphe milieu et le graphe total d'une chaîne et d'un cycle.

Keywords: *b*-coloration des arêtes, indice *b*-chromatique.

1 Introduction

Soit G = (V, E) un graphe simple et non orienté. Le *voisinage ouvert d'un sommet* u de V est $N(u) = \{v \in V : uv \in E\}$, dans ce cas u est de *degré* $d(u) = |N(u)|$. Le *voisinage ouvert d'une arête* $e = uv$ de E, noté $N(e)$, est l'ensemble des arêtes différentes de e ayant u ou v comme extrémité. Le degré de l'arête e est $d(e) = |N(e)|$.

Une *coloration propre des sommets* de G consiste à affecter une couleur à chaque sommet de V de telle manière que deux sommets adjacents ne reçoivent pas la même couleur. Une classe de couleur est un sous-ensemble stable de sommets de V qui sont de même couleur. Le *nombre chromatique* de G, noté χ(G), est le nombre minimum de classes de couleur qui partitionnent V.

Une *b-coloration des sommets* de G est une coloration propre des sommets de G telle que toute classe de couleur possède au moins un sommet qui a un voisin dans toutes les autres classes de couleurs. Un tel sommet est dit sommet *b-dominant* de couleur (ou en bref *b-sommet*). Le nombre *b-chromatique* est le nombre maximum de classes de couleur dans une *b*-coloration. Ce concept a été introduit par R.W. Irving and D.F. Manlove [1,2]. Ils ont ont prouvé que le problème de décision lié

au nombre b-chromatique est NP-complet en général, même si on se restreint aux graphes bipartis [3], tandis qu'il est polynomial pour les arbres [1,2]. En présence de ces résultats de NP-complétude, beaucoup de travaux ont été menés pour définir des bornes pour le nombre b-chromatique d'un graphe arbitraire ou bien pour déterminer sa valeur exacte pour certaines classes de graphes (voir [4]). Irving et Manlove ont défini le m-degré d'un graphe G, noté m(G) comme le plus grand entier k, tel que G possède au moins k sommets de degré au moins $k-1$, et ils ont prouvé que $b(\mathrm{G}) \leq m(\mathrm{G})$ [1,2].

Une *coloration propre des arêtes* de G consiste à attribuer à chaque arête de G une couleur de sorte que deux arêtes de E ayant une extrémité commune ne reçoivent pas la même couleur. Un ensemble d'arêtes qui ne sont pas adjacentes s'appelle un *couplage*. Puisque chaque ensemble d'arêtes colorées avec la même couleur est un couplage, une coloration des arêtes d'un graphe est en effet une partition de E en couplages. *L'indice chromatique*, noté $\chi'(\mathrm{G})$, est le plus petit entier k pour lequel le graphe G admet une coloration des arêtes propre de G avec k couleurs. L'indice chromatique d'un graphe G est le nombre chromatique de son graphe adjoint, autrement dit si $\mathscr{L}(\mathrm{G})$ est le graphe adjoint de G, alors $\chi'(\mathrm{G}) = \chi(\mathscr{L}(\mathrm{G}))$. Vizing [13] a montré que $\Delta(\mathrm{G}) \leq \chi'(\mathrm{G}) \leq \Delta(\mathrm{G}) + 1$.

Une *b-coloration des arêtes* d'un graphe G est une coloration propre des arêtes de G de telle sorte que chaque classe de couleur possède au moins une arête adjacente à au moins une arête dans chacune des autres classes de couleur. Une telle arête est dite *b-arête*. L'*indice b-chromatique* de G, noté $b'(\mathrm{G})$, est le nombre maximum de couleurs pour lequel G admet une b-coloration d'arêtes avec $b'(\mathrm{G})$ couleurs.

M. Jakovac et I. Peterin sont les premiers qui ont étudié la b-coloration des arêtes d'un graphe [7], où ils ont remarqué que pour un graphe G admette une b-coloration des arêtes avec k couleurs, il doit avoir au moins k arêtes de degré au moins $k-1$, où ils ont défini le m'-degré de G, noté $m'(\mathrm{G})$, comme le plus grand entier m' pour lequel G possède au moins k arêtes de degré au moins $m'-1$. Une arête de degré au moins $m'(\mathrm{G})-1$ est dite *dense*. Et dans le même papier [7], les auteurs ont prouvé que tout graphe G satisfait l'inégalité suivante :

$$\Delta(\mathrm{G}) \leq \chi'(\mathrm{G}) \leq b'(\mathrm{G}) \leq m'(\mathrm{G}) \leq 2\Delta(\mathrm{G}) - 1. \tag{1}$$

De point de vue complexité Campos et al. [5] ont prouvé que le problème de décider si $m'(\mathrm{G}) = b'(\mathrm{G})$ est NP-complet. L'absence d'algorithme polynomial pour déterminer l'indice b-chromatique d'un graphe a incité les chercheurs à établir des bornes qui encadrent ce paramètre ou bien calculer sa valeur exacte pour certaines classes de graphes comme les chenilles, graphes d-réguliers le produit cartésien et le produit direct de certains graphes, etc... (voir [5,6,7,8,10]).

Dans ce papier, nous donnons la valeur exacte de l'indice b-chromatique de certains graphes, comme la chaîne, le cycle, la roue, la double étoile, le graphemilieu et le graphe total d'une chaîne et d'un cycle.

Notons que, dans la suite de ce papier, toutes les arêtes dessinées en lignes pointillées représentent les b-arêtes.

2 L'indice *b*-chromatique de certains graphes classiques

Dans cette section, nous déterminons la valeur exacte de l'indice b-chromatique dans le cas d'un cycle et d'une chaîne, d'une roue et d'une double étoile.

Théorème 21 *Soit* P_n *une chaîne à n sommets. Alors,*

$$b'(P_n) = \begin{cases} 1 \text{ si } n = 2 ; \\ 2 \text{ si } n \in \{3,4,5\} ; \\ 3 \text{ si } n \geq 6. \end{cases}$$

Démonstration. Il est facile de vérifier que le graphe adjoint d'une chaîne d'ordre n est une chaîne d'ordre $(n-1)$. Ainsi, vu que $b'(P_n) = b(\mathscr{L}(P_n)) = b(P_{n-1})$, il s'ensuit alors que $b'(P_2) = b(P_1) = 1$, $b'(P_n) = b(P_{n-1}) = 2$, pour $n \in \{3,4,5\}$ et $b'(P_n) = b(P_n) = b(P_{n-1}) = 3$ pour $n \geq 6$ [9].

Théorème 22 *Soit* C_n *un cycle d'ordre n. Alors,*

$$b'(C_n) = \begin{cases} 2 \text{ si } n = 4 ; \\ 3 \text{ si } n = 3 \text{ ou } n \geq 5. \end{cases}$$

Démonstration. Puisque le graphe adjoint de C_n est un cycle à n sommets, alors $b'(C_n) = b(C_n)$. De ce fait $b'(C_n) = 2$ si $n = 4$ et $b'(C_n) = 3$ sinon, [9].

Maintenant, nous allons déterminer la valeur exacte de l'indice b-chromatique de la roue W_n ($W_n = C_n \vee K_1$).

Théorème 23 *Soit* $W_n = (V,E)$ *une roue d'ordre* $n+1 \geq 5$. *Alors,*

$$b'(W_n) = \begin{cases} 5 \text{ si } n = 4 ; \\ n \text{ si } n \geq 5. \end{cases}$$

Démonstration. Pour $n = 4$, il est facile de vérifier que $m'(W_n) = 5$. En utilisant (1), on constate que $b'(W_n) \leq 5$, et l'égalité est obtenue en exhibant une b-coloration d'arêtes de W_4 utilisant 5 couleurs (Voir Figure 1).

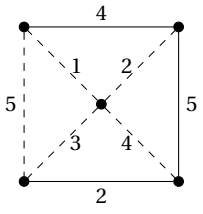


FIGURE 1. Une b-coloration des arêtes de W_4 avec 5 couleurs.

Pour $n \geq 4$, on a $m'(W_n) = \Delta(W_n) = n$; par conséquent (1) donne $b'(W_n) = n$.

Une *double étoile,* notée $S_{r,s}$ est le graphe obtenu en ajoutant une arête reliant deux centres de deux étoiles distinctes $K_{1,r}$ et $K_{1,s}$.

Théorème 24 *Soit* $S_{l,k}$ *une double étoile avec* $l \geq k \geq 1$. *Alors* $b'\left(S_{l,k}\right) = l+1$.

Démonstration. Le résultat est immédiat d'après (1) puisque $m'\left(S_{l,k}\right) = \Delta\left(S_{l,k}\right) = l+1$.

3 L'indice *b*-chromatique du graphe milieu et total d'un cycle et d'une chaîne

Dans cette section, nous donnons les valeurs exactes de l'indice b-chromatiques du graphe milieu et du graphe total d'une chaîne P_n, d'un cycle C_n.

Définition 31 *(Hamada et Yoshimura)[11] Soit* $G = (V,E)$ *un graphe simple. Le graphe milieu de* G, *noté* M(G) *est défini comme suit : L'ensemble des sommets de* M(G) *est* $V(G) \cup E(G)$, *et deux sommets* x, y *de* M(G) *sont adjacents si l'une des conditions suivantes est vérifiée :*

— $x; y \in E(G)$ *et* x *est adjacent à* y *dans* G.
— $x \in V(G)$, $y \in E(G)$ *et* x *est incident à* y *dans* G.

Définition 32 *(Behzad)[12] Le graphe total d'un graphe* $G = (V,E)$, *noté* T(G), *est défini comme suit. L'ensemble des sommets de* T(G) *est* $V(G) \cup E(G)$, *et deux sommets* x, y *de* T(G) *sont adjacents si l'une des conditions suivantes est vérifiée :*

— $x, y \in V(G)$ *et* x, y *sont adjacents dans* G.
— $x; y \in E(G)$ *et* x *est adjacent à* y *dans* G.
— $x \in V(G)$, $y \in E(G)$ *et* x *est incident à* y *dans* G.

Théorème 31 *Soit* P_n *une chaîne d'ordre* n. *Alors,*

$$i)\ b'(M(P_n)) = \begin{cases} n \text{ si } n \in \{2,3,4\}\ ; \\ 5 \text{ si } n \in \{5,6,7\}\ ; \\ 6 \text{ si } n \in \{8,9,10\}\ ; \\ 7 \text{ si } n \geq 11. \end{cases},\ ii)\ b'(T(P_n)) = \begin{cases} 3 \text{ si } n = 2\ ; \\ 4 \text{ si } n = 3\ ; \\ 5 \text{ si } n = 4\ ; \\ 6 \text{ si } n = 5\ ; \\ 6 \text{ ou } 7 \text{ si } n = 6\ ; \\ 7 \text{ si } n \geq 7. \end{cases}.$$

Démonstration. Posons $V(P_n) = \{v_1, ..., v_n\}$ et $E(P_n) = \{e_1, ..., e_{n-1}\}$ tel que $e_j = v_j v_{j+1}$ pour tout j, $1 \leq j \leq n-1$. Il est clair que $V(M(P_n)) = V(T(P_n)) = V(P_n) \cup E(P_n)$.

i) Pour $n = 2$, il est facile de voir que $M(P_2)$ est isomorphe à P_3, donc d'après le Théorème 21 $b'(M(P_2)) = 2$.

Pour $n \in \{3,4\}$. Il n'est pas difficile voir que $m'(M(P_n)) = n$ et que $\Delta(M(P_n)) = n$. Donc, d'après (1), il en résulte que $b'(M(P_n)) = n$ pour $n \in \{3,4\}$.

Pour $n \in \{5,6,7\}$. Dans un premier temps, montrons que $b'(M(P_n)) \geq 5$. Pour cela, voir la Figure 2. Il reste de prouver que $b'(M(P_n)) \leq 5$. Il est facile de voir que $m'(M(P_n)) = 5$, alors, d'après (1), $b'(M(P_n)) \leq 5$. D'où, $b'(M(P_n)) = 5$.

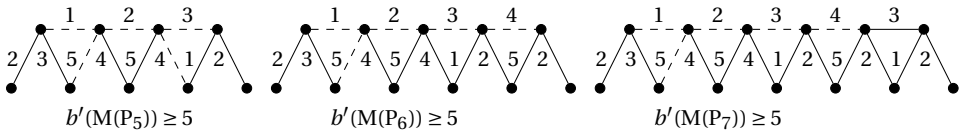


FIGURE 2. Une b-coloration des arêtes de $M(P_n)$ pour $n \in \{5,6,7\}$.

Pour $n \in \{8,9,10\}$. Afin de montrer que $b'(M(P_n)) \geq 6$, il suffit de colorer les arêtes de $M(P_n)$.Voir la Figure 3. Il reste de prouver que $b'(M(P_n)) \leq 6$. Il est facile de voir que $m'(M(P_n)) = 6$, alors, d'après (1), $b'(M(P_n)) \leq 6$. Par conséquent, $b'(M(P_n)) = 6$.

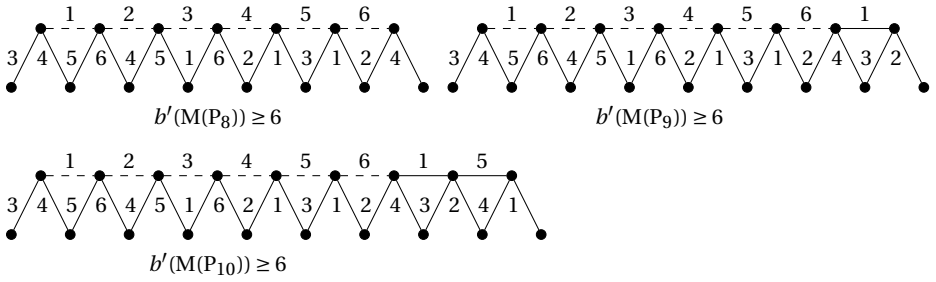


FIGURE 3. Une b-coloration des arêtes de $M(P_n)$ pour $n \in \{8,9,10\}$.

Pour $n \geq 11$. Il est facile de voir que $m'(M(P_{11})) = 7$, alors, d'après (1), $b'(M(P_n)) \leq 7$. D'autre part, la Figure 4 présente une b-coloration des arêtes de $M(P_{11})$ avec 7 couleurs. Par conséquent, $b'(M(P_{11})) = 7$. Soit $H = M(P_{11})$ un sous-graphe de $M(P_n)$ pour $n \geq 12$ et considérons une b-coloration c des arêtes de H avec 7 couleurs (coloration partielle de $M(P_n)$). Comme pour $n \geq 11$, le degré maximum des arêtes est $2\Delta(G) - 2 = 6$ et $M(P_n)$ admet une b-coloration des arêtes avec $2\Delta(M(P_n)) - 1$ couleurs, alors c peut s'étendre à $M(P_n)$. Donc, $b'(M(P_n)) = 7$.

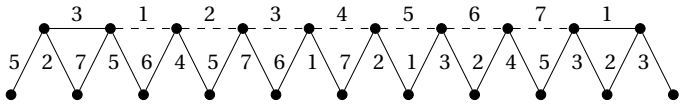


FIGURE 4. Une b-coloration des arêtes de $M(P_{11})$.

i i) Pour $n = 2$, $T(P_n)$ est isomorphe à C_3, et donc d'après le Théorème 22, $b'(T(P_2)) = 3$. Pour $n = 3$, on a $m'(T(P_3)) = \Delta(T(P_3)) = 4$, et d'après (1), on obtient $b'(T(P_3)) = 4$.

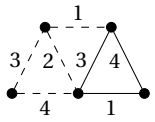


FIGURE 5. Une b-coloration des arêtes de $T(P_3)$.

Pour $n = 4$, montrons d'abord que $b'(T(P_n)) \geq 5$. Pour cela, il suffit de donner une b-coloration d'arêtes de $T(P_4)$ avec 5 couleurs (Voir Figure 6). Il reste de montrer que $b'(T(P_n)) \leq 5$. Supposons au contraire que $b'(T(P_4)) \geq 6$. Vu que $m'(T(P_4)) = 6$, alors (1) donne $b'(T(P_n)) = 6$. Comme $T(P_4)$ comporte 7 arêtes denses parmi $|E(T(P_n))| = 11$ arêtes, alors il existe au moins une classe de couleur qui possède exactement une seule arête, disons e. Ceci signifie que e est une b-arête de couleur non répétée. Donc, toutes les autres b-arêtes , de couleur différentes à celle de e, sont adjacentes à e. Montrons que $e \in \{e_2v_2, e_2v_3\}$. Supposons le contraire. Alors, soit $e = v_2v_3$ ou bien $e \in \{e_1v_2, e_3v_3, e_1e_2, e_2e_3\}$. Si $e = v_2v_3$ ou $\{e_1e_2, e_2e_3\}$ (resp., $e \in \{e_1v_2, e_2v_3\}$), alors il existe deux (resp., trois) arêtes denses non adjacentes à e. Comme $b'(T(P_4)) = 6$, alors au moins l'une de ces arêtes est une b-arête, ce qui contredit le fait que toutes les b-arêtes sont adjacentes à e. Par conséquent $e \in \{e_2v_2, e_2v_3\}$. Par symétrie et sans perte de généralité, supposons que $e = e_2v_2$ est une b-arête de couleur 1 et $c(e_1v_2) = 2$, $c(e_1e_2) = 3$, $c(e_2e_3) = 4$, $c(e_2v_3) = 5$ et $c(v_2v_3) = 6$. Vu que les arêtes denses adjacentes à e sont des b-arêtes, alors e_1v_2 et e_1e_2 sont des b-arêtes de couleurs 2 et 3 respectivement. Alors, les arêtes e_1v_1 et v_1v_2 peuvent prendre les couleurs 4 et 5 respectivement, mais dans ce cas e_1e_2 aura une couleur manquante dans son voisinage, contradiction. Par conséquent, $b'(T(P_n)) = 5$.

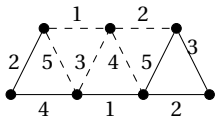


FIGURE 6. Une b-coloration des arêtes de $T(P_4)$.

Pour $n = 5$, la Figure 7 présente une b-coloration d'arêtes de $T(P_5)$ avec 6 couleurs. De ce fait $b'(T(P_n)) \geq 6$.

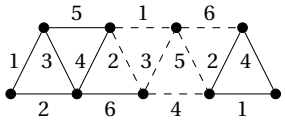


FIGURE 7. Une b-coloration des arêtes de $T(P_5)$ avec 6 couleurs.

Il reste de prouver que $b'(T(P_5)) \leq 6$. Il est facile de voir que $m'(T(P_5)) = 7$, car $T(P_5)$ a exactement 7 arêtes de degré 6, mais il n'a pas 8 arêtes de degré 7. Puisque $T(P_5)$ possède une b-coloration d'arêtes avec 6 couleurs, alors $6 \leq b'(T(P_5)) \leq 7$. Supposons que $b'(T(P_5)) = 7$. Alors, dans ce cas $T(P_5)$ possède exactement 7 b-arêtes. Ces arêtes (dessinées en gras dans la Figure 8) sont colorées différemment.

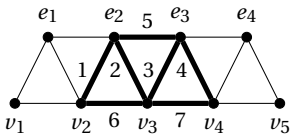


FIGURE 8. Une b-coloration des arêtes de $T(P_5)$.

L'arête v_3v_4 doit avoir dans son voisinage la couleur 5, mais dans ce cas l'arête e_3v_4 aura dans son voisinage deux arêtes colorées avec la même couleur 5, donc l'arête e_3v_4 ne peut pas être une b-arête, contradiction. D'où, $b'(T(P_5)) = 7$.

Pour $n = 6$, la Figure 9 présente une b-coloration d'arêtes de $T(P_6)$ avec 6 couleurs. De ce fait $b'(T(P_6)) \geq 6$.

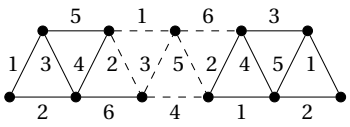


FIGURE 9. Une b-coloration des arêtes de $T(P_6)$ avec 6 couleurs.

Pour $n = 7$, on a $m'(T(P_7)) = 7$, et d'après (1), on constate que $b'(T(P_7)) \leq 7$. D'autre part, la Figure 10 présente une b-coloration des arêtes de $T(P_7)$ avec 7 couleurs. Par conséquent, $b'(T(P_7)) = 7$. Supposons maintenant que $n \geq 8$ et soit $H = T(P_7)$ un sous-graphe de $T(P_n)$. Considérons une b-coloration c des arêtes de H avec 7 couleurs (coloration partielle de $T(P_n)$). Comme le degré maximum des arêtes de $T(P_n)$ est $2\Delta(T(P_n)) - 2 = 6$, alors c peut s'étendre à $T(P_n)$ avec 7 couleurs. D'où $b'(T(P_n)) = 7$.

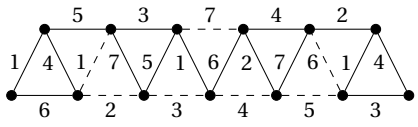


FIGURE 10. Une b-coloration des arêtes de $T(P_7)$ avec 7 couleurs.

Théorème 32 *Soit* C_n *un cycle d'ordre n. Alors,*

$$i)\ b'(M(C_n)) = \begin{cases} 5\ si\ n \in \{3,4,5\}\ ; \\ 6\ si\ n = 6; \\ 7\ si\ n \geq 7. \end{cases} , ii)\ b'(T(C_n)) = \begin{cases} 6\ si\ n = 3\ ; \\ 6\ ou\ 7\ si\ n = 4\ ; \\ 7\ si\ n \geq 5. \end{cases}.$$

Démonstration. Posons $V(C_n) = \{v_1, ..., v_n\}$ et $E(C_n) = \{e_1, ..., e_n\}$ tel que $e_j = v_j v_{j+1}$ pour tout $j, \leq j \leq n-1$. Il est clair que $V(M(C_n)) = V(T_n) = V(C_n) \cup E(C_n)$.

i) Soit c une b-coloration d'arêtes de $M(C_n)$ (resp. $T(C_n)$) avec k couleurs.

Pour $n \in \{3,4,5\}$. Pour montrer que $b'(M(C_n)) \geq 5$, il suffit de donner une b-coloration d'arêtes de $M(C_n)$ en utilisant 5 couleurs (Voir Figure 11). Il reste à montrer que $b'(M(C_n)) \leq 5$. En effet, vu que $m'(M(C_n)) = 5$, (1) donne $b'(M(C_n)) \leq 5$, et par conséquent $b'(M(C_n)) = 5$.

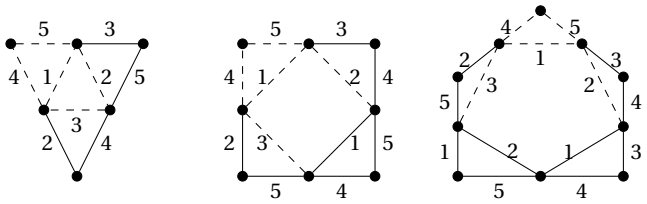


FIGURE 11. Une b-coloration des arêtes avec 5 couleurs de $M(C_n)$ pour $n \in \{3,4,5\}$.

Pour $n = 6$, en exhibant une b-coloration d'arêtes de $M(C_n)$ avec 6 couleurs (Voir Figure 12), on constate que $b'(M(C_n)) \geq 6$. D'autre part puisque $m'(M(C_n)) = 6$, on conclut que $b'(M(C_n)) = 6$.

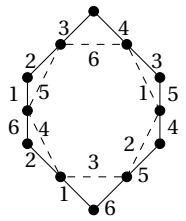


FIGURE 12. Une b-coloration des arêtes de $M(C_6)$.

Pour $n \geq 7$. Pour prouver que $b'(M(C_n)) \geq 7$, nous donnons une b-coloration des arêtes de $M(C_n)$ avec 7 couleurs. (Voir Figure 13 pour $n \in \{7,8\}$).

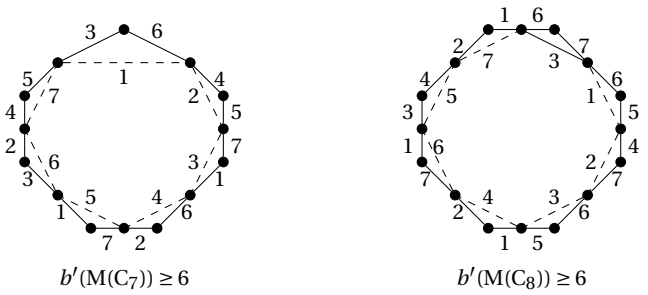


FIGURE 13. Une b-coloration des arêtes de $\mathrm{M}(\mathrm{C}_n)$ pour $n \in \{7,8\}$.

Pour $n \geq 9$, on commence d'abord par une coloration partielle des arêtes de $\mathrm{M}(\mathrm{C}_n)$ (pour $n = 9$) avec 7 couleurs ensuite on étend cette coloration à toutes les arêtes de $\mathrm{M}(\mathrm{C}_n)$; cette extension est possible puisque le degré maximum des arêtes de $\mathrm{M}(\mathrm{C}_n)$ est égal à 6. Soit $\mathrm{A} = \{e_i, v_i : 1 \leq i \leq 8 \text{ et } 1 \leq j \leq 9\}$. Posons $\mathrm{H} = \mathrm{G}[\mathrm{A}]$ un sous-graphe de $\mathrm{M}(\mathrm{C}_n)$ engendré par A. La coloration partielle est donnée comme suit. On affecte la couleur i aux arêtes $e_i e_{i+1}$ pour tout $i \in \{1,2,3,4,5,6,7\}$. Les autres arêtes sont colorées comme suit. $c(e_4 v_4) = c(e_6 v_6) = c(e_8 v_8) = 1$, $c(e_5 v_5) = c(e_7 v_8) = c(v_1 e_n) = 2$, $c(e_1 e_n) = c(e_6 v_7) = c(e_8 v_9) = 3$, $c(e_2 v_3) = c(e_7 v_7) = 4$, $c(e_1 v_2) = c(e_3 v_3) = c(e_8 e_9) = 5$, $c(e_2 v_2) = c(e_4 v_5) = 6$, $c(e_1 v_1) = c(e_3 v_4) = c(e_5 v_6) = 7$. Notons que les b-arêtes de c sont $\{e_i e_{i+1} : i \in \{1,2,3,4,5,6,7\}\}$. En conséquence, $b'(\mathrm{M}(\mathrm{C}_n)) \geq 7$. Vu que $m'(\mathrm{M}(\mathrm{C}_n)) = 7$, alors $b'(\mathrm{M}(\mathrm{C}_n)) = 7$.

ii) Pour $n \in \{3,4\}$, on a $b'(\mathrm{T}(\mathrm{C}_n)) \geq 6$ puisque la Figure 14 présente une b-coloration d'arêtes de $\mathrm{T}(\mathrm{C}_n)$ avec 6 couleurs.

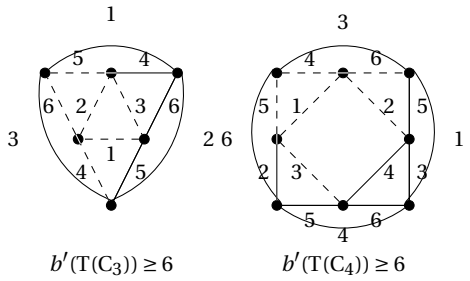


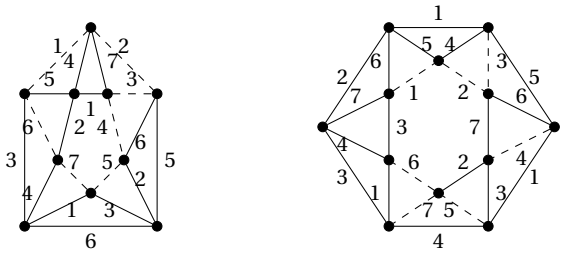
FIGURE 14. b-coloration d'arêtes avec 6 couleurs de $\mathrm{T}(\mathrm{C}_n)$ pour $n = 3,4$.

Montrons que pour $n = 3$, on a $b'(\mathrm{T}(\mathrm{C}_3)) \leq 6$. Supposons, au contraire, que pour $b'(\mathrm{T}(\mathrm{C}_3)) \geq 7$. Comme $m'(\mathrm{T}(\mathrm{C}_3)) = 7$, alors d'après (1), on a $b'(\mathrm{T}(\mathrm{C}_3)) = 7$. Puisque $\mathrm{T}(\mathrm{C}_3)$ possède 11 arêtes, alors il existe une classe de couleur qui contient exactement une seule arête, disons e. Clairement e est une b-arête de couleur non répétée. De ce

fait, toutes les autres b-arêtes, autre que e, sont adjacentes à e. Supposons d'abord que $e = e_1e_3$ et que $c(e) = 1$. Supposons sans perte de généralité que les autre b-arêtes e_1v_1, e_1e_2, e_1v_2, e_3v_1, e_2e_3 et e_3v_3 sont colorées 2, 3, 4, 5, 6 et 7 respectivement. Comme e_1v_1 est une b-arête de c, alors cette arête est besoin des couleurs 6 et 7 dans son voisinage.

Alors clairement $c(v_1v_2) = 7$ et $c(v_1v_3) = 6$ (sinon la coloration ne sera plus propre). Comme $T(P_3)$ est un graphe 6-arête régulier, alors toute couleur peut appraître une seule fois dans le voisinage de chaque b-arête. Cependant, l'arête e_3v_1 de couleur 5 a deux couleurs répétées dans son voisinage (couleurs 6 et 7), contradiction. Par conséquent, $b'(T(C_3)) = 6$.

Pour $n \geq 5$. Afin d'obtenir une b-coloration des arêtes de $T(C_n)$ avec 7 couleurs, il suffit de donner une b-coloration avec 7 couleurs. (Pour $n \in \{5,6\}$, voir Figure 15).



$b'(T(C_5)) \geq 7$ $\qquad$ $b'(T(C_6)) \geq 7$

FIGURE 15. Une b-coloration des arêtes de $T(P_5)$ et $T(P_6)$.

Pour $n \geq 7$, on a $V(M(C_n)) = V(T(C_n))$ et $T(C_n)$ peut être obtenu à partir de $M(C_n)$ en ajoutant les arêtes $\cup_{i=1}^{n-1}\{v_iv_{i+1}\}\cup\{v_1v_n\}$. Soit π une b-coloration d'arêtes de $M(C_n)$ avec 7 couleurs. Clairement π est une coloration partielle de $T(C_n)$. Puisque le degré maximum de $T(C_n)$ est égal à 6, alors on peut étendre π aux autres arêtes de $T(C_n)$. D'où $b'(T(C_n)) \geq 7$. Sachant que $m'(T(C_n)) = 7$, on conclut que $b'(T(C_n)) \leq 7$, et par conséquent, $b'(T(C_n)) = 7$.

Références

1. R.W. Irving, D.F. Manlove. The b-chromatic number of graphs. Discrete Appl. Math. 91, 127 – 141, 1999.
2. D.F. Manlove. Minimaximal and maximinimal optimization problems : a partial order-based approach. PhD thesis, technical report tr-1998 – 27 of the Computing Science Department of Glasgow University, 1998.
3. J. Kratochvíl, Z. Tuza, M. Voigt. On the b-chromatic number of graphs. Lecture Notes in Computer Science (Graph-Theoretic Concepts in Computer Science : 28th International Workshop, WG 2002) 2573, 310 – 320, 2002.
4. M. Jakovac, I. Peterin. The b-chromatic number and related topics-a survey. Preprint.

5. V. A. A. Campos, C. V. Lima, N. A. Martins, L. Sampaio, M. C. Santos, A. Silva, *b*-chromatic index of graphs. Discrete Mathematics. 338 (2015) 2072-2079.
6. V. Campos. A. Silva. Edge-*b*-Coloring Trees. Algorithmica Algorithmica, 80 (1), pp 104–115, 2018.
7. M. Jakovac, I. Peterin. The b-chromatic index of a graph. Bulletin of the Malaysian Math. Sc. Society 38 (4) (2014) 1375–1392.
8. A. Silva. Trees with small b-cromatic index. arXiv :1511.05847v1 [cs.DM] 18 Nov 2015.
9. M. Kouider, M. Mahéo. Some bounds for the b-chromatic number of a graph. Discrete Math. 256 : 267 - 277 ; 2002 :
10. I. Koch, I. Peterin. The b-chromatic index of direct product of graphs. Discrete Applied Mathematics 190–191 (2015) 109–117.
11. T. Hamada, I. Yoshimura : Traversability and connectivity of the middle graph of a graph. Discrete Math. 14 : 247 – 256 ; 1976.
12. M. Behzad : A criterion for the planarity of the total graph of a graph. Proc. Camb.Philos. Soc. 63 : 679 - 681 ; 1967.
13. V.G. Vizing. On an estimate of the chromatic class of a graph, Diskret. Anal.,pp. 25-30, 3 (1964).

The Use of Model Driven Architecture to Describe Bio-Inspired System: Case of Artificial Neural Network

S.E. Mili [1], D. Meslati [2] and V.Rodin [3]

[1] Ecole normale superieure Constantine, LISCO , Badji Mokhtar-Annaba, Algeria
[2] LISCO Laboratory, Badji Mokhtar-Annaba University, Algeria
[3]Department of Computer Science University of Western Brittany, France

Abstract. The engineering models and especially MDA (Model Driven Architecture) aims to provide a conceptual framework in which technological and methodological models are the axe of software engineering activities. The classic use of MDA is the description of the systems as templates and development as model transformations. The use of MDA to describe the operation itself of a system is not common or easy to perform. In this paper, we describe the behavior of ANN system using MDA in the purpose of make a conceptual framework who is well-suited for sophisticated biological models and well-founded analytical principles.

Keywords: Model Driven Architecture, Artificial Neural Network, Bio-Inspired System, Transformation Patterns, ATL Language.

1 Introduction

The model-driven engineering and specifically the MDA represents a new software engineering paradigm[1]. It is essentially based on the use of models to represent the different steps of software systems development process. OMG has defined a set of standards and tools to implement the MDA approach [2]. They offer resources and techniques of great wealth , providing a high level of coverage of modeling needs. They allow the creation, manipulation, management and processing models.
Artificial neural networks (ANN) are a strong current to model complex phenomena. ANN is a structure composed of entities that can perform calculations and interacting with each other . It can process problems of different nature than conventional tools are struggling to solve. Indeed, its operation is based on that of biological neural cells , and is therefore different methods of analytical calculation which is commonly used . It is very powerful in recognition problems classification, approximation or prediction [3].
In this paper we focus in two main computing domaines : Artificial neural networks (ANN) and software engineering. The new paradigm of software engineering have proved her efficiency in the industry of computing systems . paradigms like oriented aspect, oriented component programming and model driving architecture (MDA) makes developing and evolving software an easer task. The use of MDA is increasing

2

in academic world (research and university laboratory), and also in industry world. We aim in this paper to use principal's techniques of MDA in context of modeling and describing behavior of Artificial Neural Network (ANN); in the purpose to make a conceptual framework which is able to produce several models inspired by nature. A generic framework that is not restricted to a particular bio inspired system.
This paper is organized as follows. We start with some related work in section 2, Fundamentals of ANNs are given in Section 3. In Section 4, the definition of MDA is reviewed. In section 5 we make a proposition to use MDA to model and describe ANN system. A case study of proposition use is presented in Section 6. Conclusion and some discussion are given in Section 7

2 Related work

Nowadays they are panoply of bio-inspired systems to face the increasing complexity of solving problems; the authors of [10] give an idea of the development of these systems.
In literature, it seems to be common that naive approaches to extract metaphors from the natural biological system have been taken, this naivety often blocks understanding, development, and analysis of the computations but this has not always been the case[10]. However, more recently, work on bio inspired system has drifted away from the more biologically-appealing models and attention to biological detail, with a focus on more engineering-oriented approach. This has led to systems that are examples of the "reasoning by metaphor". These include simple models of immune network, clonal selection and negative selection [11][12][13][14][15], artificial neural networks [16], genetic algorithm [17][18], and other models like bee colony[19], bat algorithm[20], flower pollination algorithm[21]. We are pointing out that these may benefit from not only closer interaction with biologists, but also a more principled mechanism for the extraction, articulation, and application of the underlying computational metaphor. Indeed, this multiplicity of algorithms can lead to a multiple representation of the same concepts.
The work of [24][26] has attracted our attention, the authors try to describe a conceptual frameworks that can made bio inspired algorithm by probes biological phenomenon, the idea is good but the framework is too abstract to be exploit effectively (modelling part of the framework not well defined).The authors of [25] try to continue the principle of [24] [26] and tend to outline a generic framework that captures a collection of population-based algorithms, allowing common properties to be factored out, and properties previously thought particular to one class of algorithms to be applied uniformly across all the algorithms but the specification of the framework is basic and was reduced to a simple algorithm
The authors of [22] attempt to describe a framework for creating a model for each domain of use that captures biological concepts and transform them into UML diagrams, here in this work the main model is presented as a component of objects that communicate flow information, this is a very interesting work, but it is the reverse of

what we propose (based on biology to create computer systems instead of using computer systems to describe biological phenomena).
Another work that can be cited even if it is a trivial one is that of [23] which speaks of a methodology in software engineering based on bio-inspired systems. In this paper, the author give the principles pillars to go to systems inspired by biology and that integrates the power of software engineering. This work is related to our by integrating the power of the software engineering and the effectiveness of the POE model, but where we diverge it is in the concrete application of these paradigms.
In the rest of the paper, we propose that bio-inspired algorithms are best developed and analyzed in the context of a model driven engineering. This new patterns is well-suited for sophisticated biological models and well-founded analytical principles.

3 Artificial Neural Networks Fundamentals

An artificial neural network consists of simple processing units, the neurons, and directed, weighted connections between those neurons. Here, the strength of a connection (or the connecting weight) between two neurons i and j is referred to as wi,j. Data are transferred between neurons via connections with the connecting weight being either excitatory or inhibitory. Looking at a neuron j, we will usually find a lot of neurons with a connection to j, i.e. which transfer their output to j. For a neuron J the propagation function receives the outputs oi1 , . . . , oin of other neurons i1, i2, . . . , in (which are connected to j), and transforms them in consideration of the connecting weights w i,j into the network input netj that can be further processed by the activation function. Thus, the network input is the result of the propagation function [4].
Based on the model of nature every neuron is, to a certain extent, at all times active, excited or whatever you will call it. The reactions of the neurons to the input values depend on this activation state. The activation state indicates the extent of a neuron's activation and is often shortly referred to as activation.

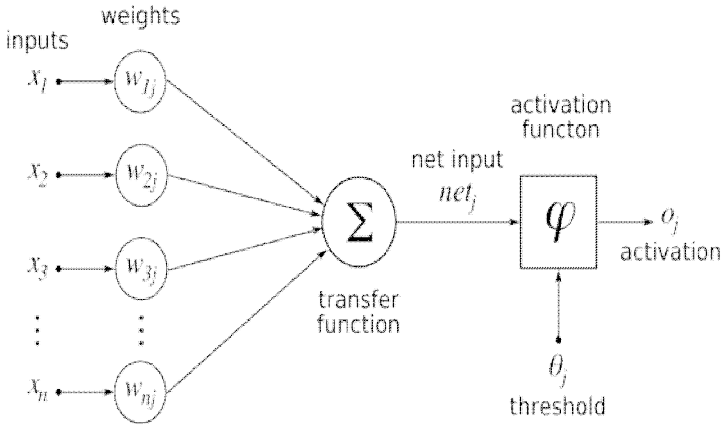


Fig.1. Formal Neural [4]

4

4 MDA and Architectural Units

One of the most important principles to cope with the complexity in software engineering is the separation of concerns principle. This principle states that a given problem involves different kinds of concerns, which should be identified and separated to cope with complexity, and to achieve the required engineering quality factors such as robustness, adaptability, maintainability, and reusability [5].
Key to MDA is the importance of models in the software development process. Within MDA the software development process is driven by the activity of modelling the business software system. The MDA development process does not look very different from a traditional lifecycle, containing the same phases (requirements, analysis, low level design, coding, testing, and deployment). One of the major differences to traditional development processes lies in the nature of the artefacts that are created during the development process. These artefacts are formal models, i.e. models that can be understood by computers and finally be transformed into a representation that lends itself to execution [6].
Architectural units blocs (AU) consists of a number n of input models and a transformation that produces the k output models. Transformations can have attributes and operators that are applied to produce the output models (see figure 2). Models as well as transformations can be of various types. The environment supplies diverse stimuli such as events that help in triggering or stopping the transformation (more details in [7]) this is due to the work they have done in terms of taxonomy [27].

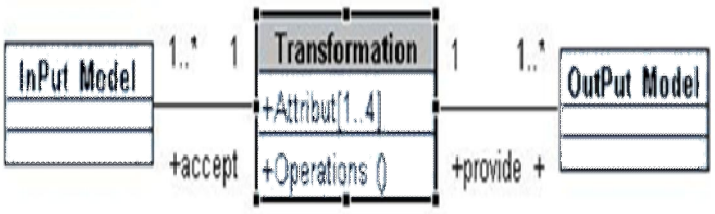


Fig.2. Independent metamodel of architectural unit [28]

5 Artificial Neural Network behavior using MDA

The bio-inspired systems can be characterized by several criteria. ANN are bio inspired systems can be characterized by epigenesis process[7]. Any system based on an epigenetic process is characterized by easily alterable structure and learning ability.
The Architectural Unit (AU) and environment in which the system operates are two main parties involved in the ANN systems. The AU of a ANN system defines the components of the system and the links between them. The environment generates stimuli that will be processed by the system (figure 3).

5

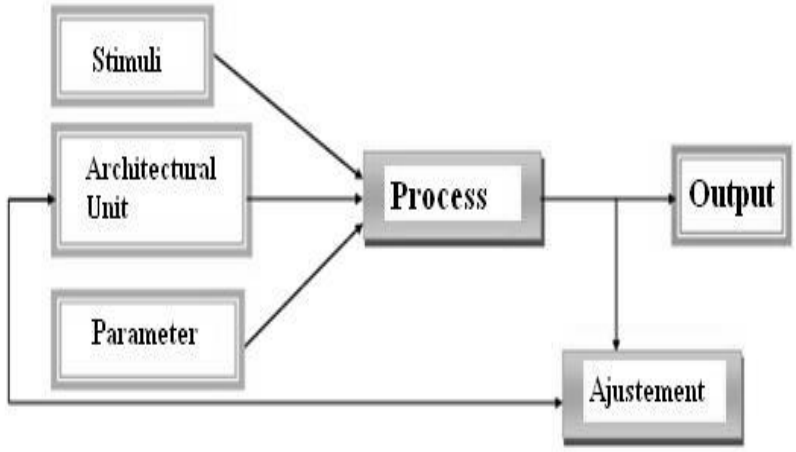


Fig.3. Main Units operating in Artifical Neural Network

The environment is represented by the unit of stimuli. It describes the set of data collected by the system, this set can be used to train, testing the network or to get output. The following expression define the stimuli:

$$S=\{I,O\} \quad (1)$$

Where I represent the input data set and O the output data set. The AU contains the topology of the system and the definition of its components . The topology is defined by three concepts: The layers , nodes (neurons) and connections (synapses) between nodes. The following expressions define the topology T:

$$T=\{C,L\} \quad (2)$$

$$C=\{c1, c2,c3,\ldots cn\} \quad (3)$$

$$Ci=\{nij\} \quad (4)$$

The system consists of a set of layer C and a set of links L. Each layer C is composed of a set of nij nodes . Each node is identified by layer i and its position j in this layer. The parameter unit contains a set of parameter used by the system (learning phase)

With the stimuli and parameter units, the ANN System defined by AU carry out and product an output result. The output result is used to adjust the system (learning phase), in case of errors. The adjustment affects weight connections or system architecture . We therefore distinguish two essential steps: a treatment step and a step of adjusting . These two steps succeed. In learning phase, there may be several iterations of this estate.

We call the artifact a data structure which is a component of the system. In ANN system two essential elements form a network : the artificial neurons (nodes) and links between them. Figure 3 represents a metamodel which enables the development of various types of artificial neural networks.

6

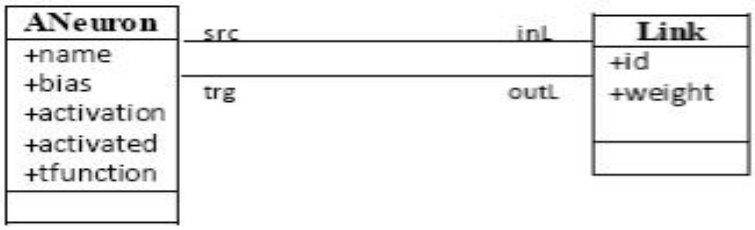


Fig.4. Metamodel of ANN

This metamodel is a kind of class diagram in which each fundamental concept is described with a class and each existing relationship between concepts using an association. It contains two classes, ANeuron (Artificial Neuron) and Link. The ANeuron class represents artificial neurons and has four attributes has two references. The Link class represents the links between artificial neurons. It has two attributes and tow references. figure 4.

The metamodel of Figure 4 is very abstract and does not explain the properties that characterize some ANN. We develop a metamodel type of neural network used to compute multilayer perceptron . This type of network is characterized by a concept that is not included in the previous metamodel. We talk about layer which is structured by nodes. The multilayer perceptron are composed of an input layer , an output layer and one or more hidden layers. The metamodel MMSystem (for System Meta Model) shown in Figure 5 corresponds to the systems based on a perceptron , whether monolayer or multilayer.

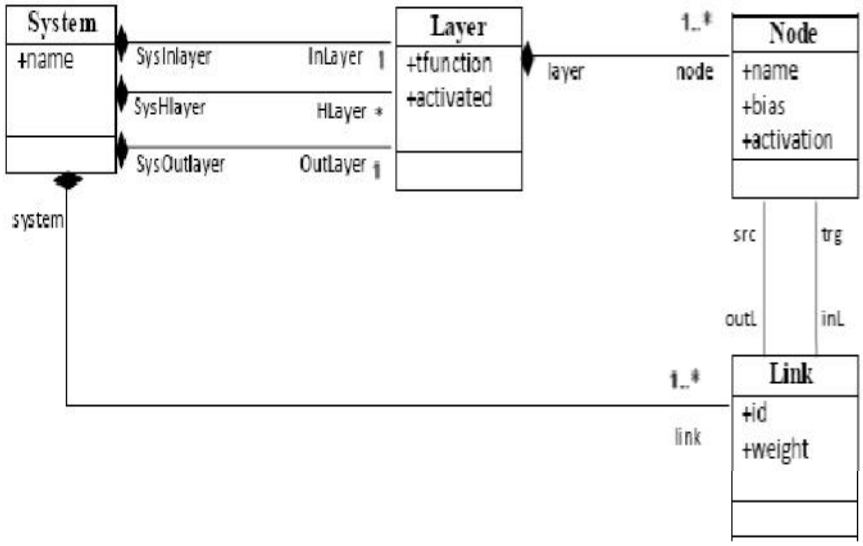


Fig.5. Metamodel of ANN System based on perceptron (MMSystem).

This metamodel consists of four classes. The System class represents the ANN system and contains an attribute named name. The Layer class represents the concept of layer

and contains an attribute named tfunction and an attribute named activated. The Node class is identical to ANeuron class in previous metamodel. The last class is the Link class. It is completely similar to the metamodel Figure 4. Each class in the metamodel of Figure 5 contains references .
Many differences can be noticed between the metamodel in Figure 4 and that of Figure 5. Besides the addition of the concept of the layer, some attributes are moved from the ANeuron class (node) to the Layer class. The replacement of the attribute tfunction (transfer function) and activated (the attribute that indicates whether the node is enabled or not) is greatly linked with the characteristic of perceptrons. We have said that the transfer function adopted by the artificial neurons of a layer is the same. In other words, the tfunction attribute of a layer contains the type of transfer function used by all nodes that comprise this layer. In addition, the parallelism is a fundamental characteristic in the perceptrons. It means that the nodes perform parallel processing, the propagation of data from one layer to another in the perceptron require the simultaneous activation of all nodes in the receiving layer.

6 Case study

To illustrate the operation of a system described above, we rely on the famous classification system that solves the XOR problem. In this context, the learning phase is not taken into account, and the system is considered valid. We focus on the system operation , i.e, the propagation of data from one layer to another.
After a learning and validation phase , the Figure 6 shows the network topology and the threshold of each node represented as an additional input

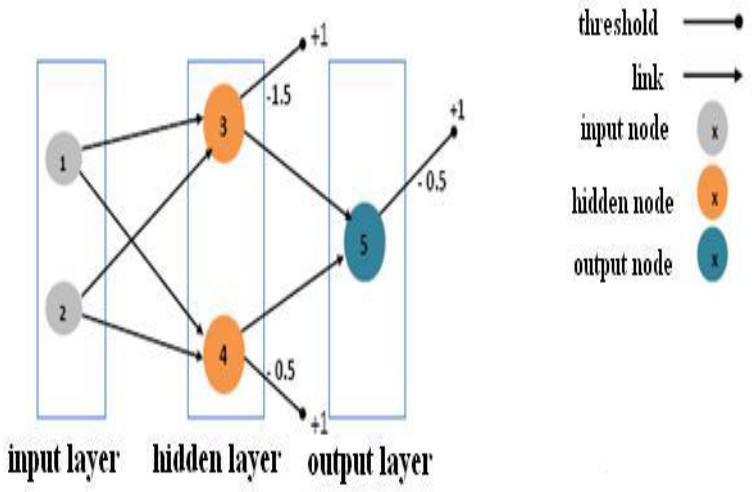


Fig.6. ANN System of XOR function

8

Figure 7 illustrates the relationship between the metamodel Figure 5 and model representing the connectionist system that allows the calculation of the XOR function. The dotted arrows represent the relationships between the elements of the metamodel and system elements. These relationships can be seen as instantiation relations.

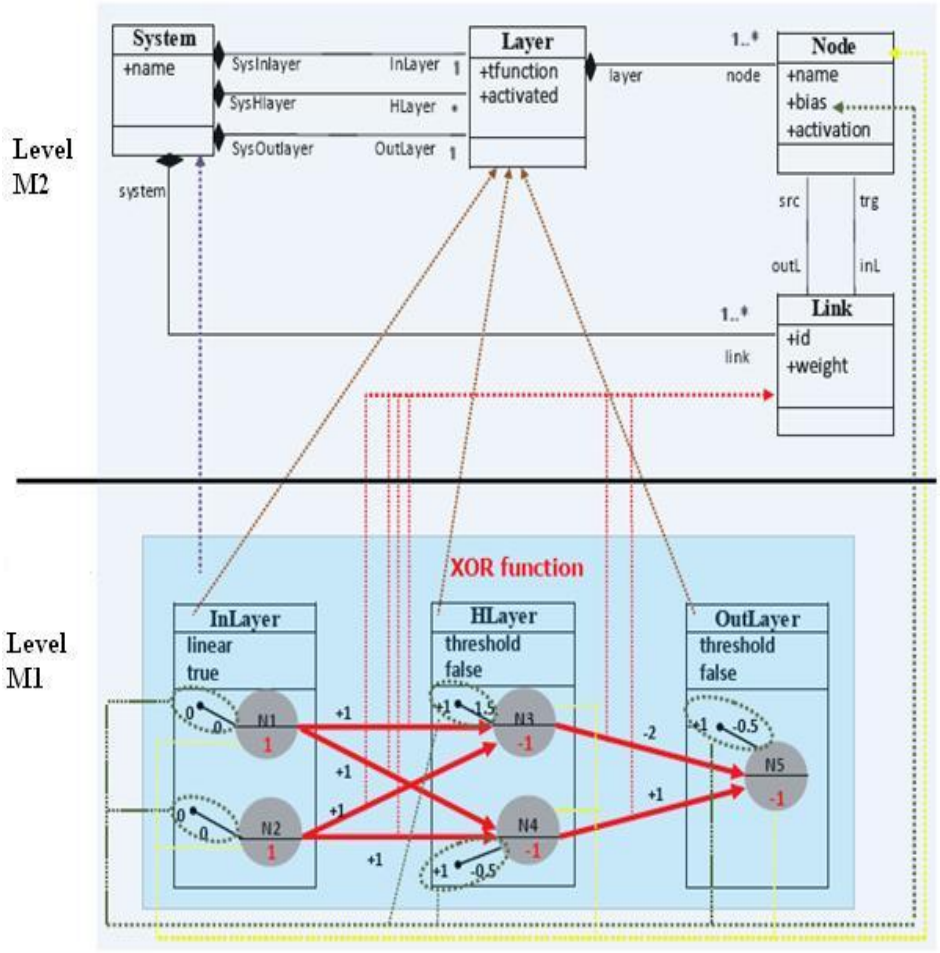


Fig.7. Relachanship between ANN XOR system in initial state and her metamodel

To describe the system "XOR function" in the form of models and model transformation, we must define the states in which it is located during operation and the kind of transformation that allows the passage from one state to another. First, it is important to understand how the system works before described. Any system based on a multilayer perceptron is to propagate data (stimuli) from one layer to another until they reach the output layer. Treatments will be performed on these data all along the propagation. These treatments allow to make changes to the data, which causes the change of the system state. We can deduce that the processing is in fact only a data propaga-

tion from layer to another. Figure 8 shows the application of two successive transformations on the system model "XOR function" which is in its initial state.

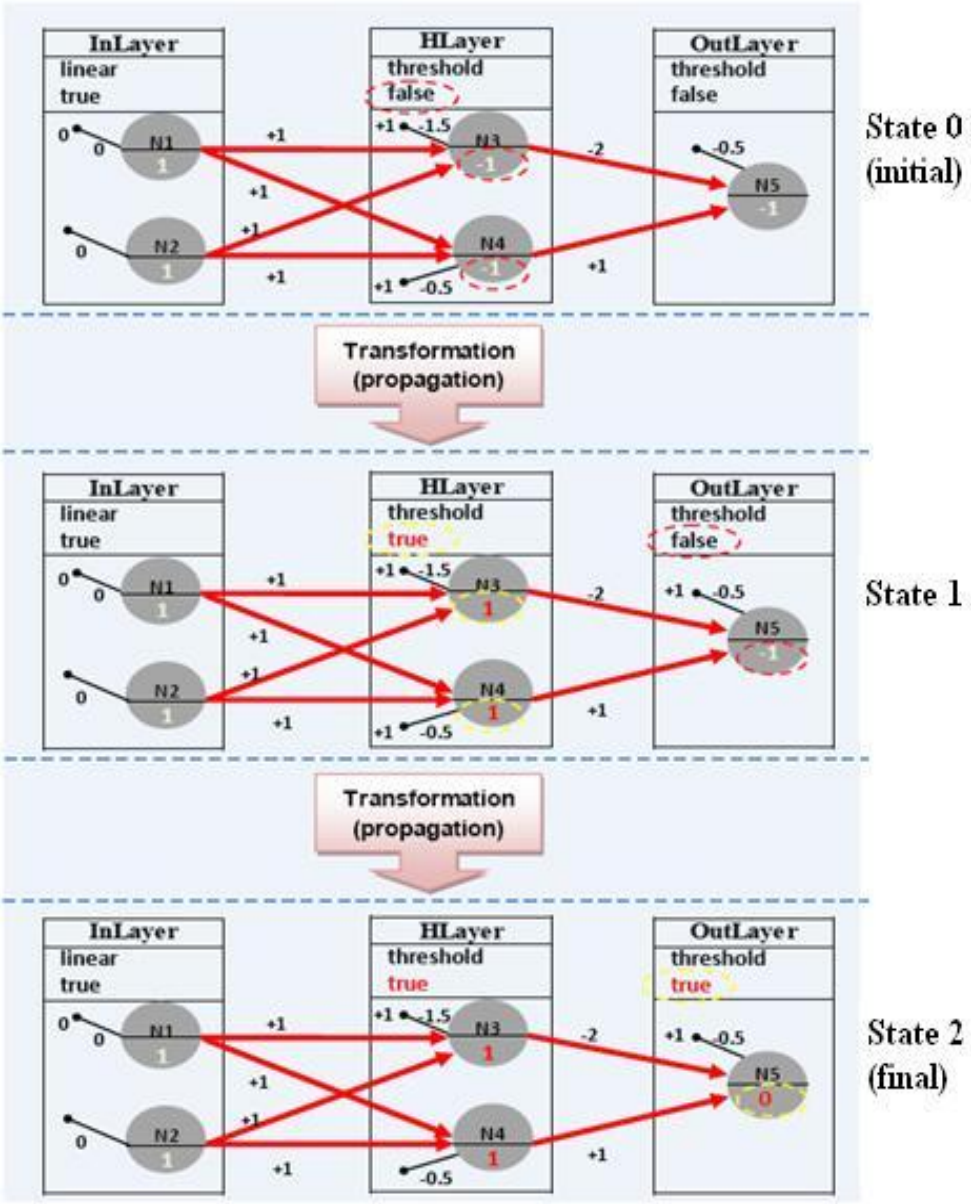


Fig.8. Performance of XOR ANN system with MDA

10

The transformation that we propose consists of a set of treatments to be applied to the source model of transformation. We can list the following treatments:

- Determination of system state
- Propagation data to the next layer if the system has not yet reached the final state , otherwise no changes will be made

To do this we use the ATL transformation language [8],[9]. we present below (figure 9) a part of ATL model transformation shown above.

```
module Propagation;
create OUTSystem : MMSystem refining INSystem : MMSyst
helper context MMSystem!Layer def:isTheInLayer() :Bool
if not self.SysInlayer-> oclIsUndefined()
then true else false endif;
helper context MMSystem!Node def:inSignal(): Real =
self.inL->iterate (l; sum : Real = 0 |
(l.weight*l.src.activation)+ sum)+self.bias;
rule Node2Node {
from
s : MMSystem!Node
to
t :MMSystem!Node (
      activation <-s.tFunction()
 )
 }
```

Fig.9. Listing part of ATL model transformation

We have chose Eclipse envirenment for implimenting the study case, for its flexibility and its ability to integrate plug-ins created specifically for the modeling domain. We use ADT, EMF and Ecore Tools plug ins for manipulation , serialization, graphic representation, and models transformations.

7 Conclusion and discussion

The dream of creating a machine with a form of intelligence is present for a long time in the human imagination. Scientists have conducted research by biological organisms inspiration whose objective is the creation of artificial systems with learning abilities and reasoning close to those of humans. This research led to the emergence of ANN approaches

In this paper, we focused on the MDA approach and the notions it brings to describe ANN approaches by model transformations . This approach has succeeded in modeling the systems development process by software models and model transformations . It relies on the use of models (CIMs , PIMs and PSMs) to represent the different

artifacts produced in the various phase of the process, and the use of model transformations to go from one step to another until the executable code generation.
Contrary to the conventional use of MDA in the context of our work, the description of the models do not regard the development process but the operation itself of a system based on ANN approach. This non- standard application of MDA in the field of modeling generates several problems related , first , to the complexity of operation of ANN systems , and , secondly , the need to adapt the use classic MDA aimed at achieving our goal. Thus, the choice of the modeling technique and processing the most appropriate models for describing the operation of an ANN system is very difficult. The same with regard to the tools used for modeling and implementing transformations models.

References

1. C.E Montenegro-Marin and all, Domain Specific Language for the generation of Learning Management Systems modules, Journal of Web Engineering Volume 11, Issue 1, 2012, Pages 23-50
2. F.Z. Belouadha, A model-driven approach for composing SAWSDL semantic Web Services, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 2, No 1, March 2010
3. A.B.Van Wyk and all Lambda-gamma learning with feedforward neural networks using particle swarm optimization, 2011 IEEE Symposium on Swarm Intelligence, SIS 2011, Paris, France, April 12-13, 2011
4. P. Škoda and all, FPGA kernels for classification rule induction, 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2016, Opatija, Croatia, May 30 - June 3, 2016.
5. B. Tekinerdogan, M. Aksit and F. Henninger. Impact of Evolution of Concerns in the Model-Driven Architecture Design Approach. 2sd International Workshop on Aspect-Based and Model-Based Separation of Concerns in Software Systems, Bilbao (Spain), 10 July 2006. Electronic Notes in Theoretical Computer Science, Elsevier, Vol. 163, No. 2, pages 45-64, April 2007
6. B. Bauer and J. Odell. UML 2.0 and agents: how to build agent-based systems with the new UML standard. Journal of Engineering Applications of Artificial Intelligence, Elsevier, Vol. 18, No. 2, Special issue on Agent-oriented Software Development, pages 141-157, March 2005.
7. S.E.Mili, D.Meslati, Multi Dimensional Taxonomy of Bio-inspired Systems Based on Model Driven Architecture, The International Arab Journal of Information Technology, Vol. 12, No. 3, May 2015.
8. M. Biehl, Literature Study on Model Transformations. Royal Institute of Technology Stockholm. Sweden, July2010.
9. ATLAS, Complex data management in distributed systems, http://www.sciences. univ-nantes. fr/lina/ATLAS/, 2006.
10. A. KumarKar, “Bio inspired computing – A review of algorithms and scope of applications”, Expert Systems with Applications An International Journal Elsevier.,59: 20–32, 2016.

12

11. L. N. de Castro, F. J. Von Zuben, "aiNet: An Artificial Immune Network for Data Analysis". In H. A. Abbass, R. A. Sarker, C. S. Newton (eds) Data Mining: A Heuristic Approach, Chapter XII. Idea Group Publishing, 2001.
12. J. Timmis. "Artificial Immune Systems: A Novel Data Analysis Technique Inspired by the Immune Network Theory", Ph.D. Dissertation, Department of Computer Science, University of Wales, September. 2000
13. M. Neal. "Meta-stable Memory in an Artificial Immune Network". In J. Timmis, P. Bentley, E. Hart (eds) ICARIS 2003, LNCS 2787, 168-180. Springer, 2003.
14. D. Taylor, D. Corne. "An Investigation of the Negative Selection Algorithm for Fault Detection in Refrigeration Systems". In J. Timmis, P. Bentley, E. Hart (eds) ICARIS 2003, LNCS 2787, 34-45. Springer, 2003.
15. D. W. Bradley, A. M. Tyrrell. "Immunotronics: Novel Finite State Machine Architectures with Built in Self Test using Self-Nonself Differentiation". IEEE Transactions on Evolutionary Computation, 6(3) 227–238, June 2002
16. A. K. Kar. "A hybrid group decision support system for supplier selection us- ing analytic hierarchy process, fuzzy set theory and neural network". Journal of Computational Science, 6 , 23–33, 2015.
17. A. Gogna, A. Tayal, "Metaheuristics: Review and application". Journal of Experimental & Theoretical Artificial Intelligence, 25 (4), 503–526, 2013.
18. D. E. Goldberg, "Genetic algorithms". India: Pearson Education, 2006.
19. D. Karaboga, B. Gorkemli, , C. Ozturk, N. Karaboga. "A comprehensive sur- vey: Artificial bee colony (ABC) algorithm and applications". Artificial Intelligence Review, 42 (1), 21–57, 2004.
20. X. B. Meng, and all, "A novel bat algorithm with habi- tat selection and Doppler effect in echoes for optimization". Expert Systems with Applications, 42 (17), 6350–6364, 2015.
21. X. S. Yang, and all. "Flower pollination algorithm: A novel approach for multi-objective optimization". Engineering Optimization, 46 (9), 1222–1237, 2014.
22. M.Read and all, "Modelling biological behaviours with the unified modelling language: an immunological case study and critique", Journal of the Royal Society Interface 11(9), 2014.
23. S.Ghoul, "A road map to bio-inspired software engineering". Res. J. Inform. Technol., 8: 75-81, 2016.
24. S. Stepny and all, "Conceptual Frameworks forArtificial Immune Systems", Int. Journ. of Unconventional Computing, Vol. 1, pp. 00–00, 2005.
25. J. Newborough and S. Stepney, "A generic framework for population-based algorithms, implemented on multiple FPGAs", 4th International Conference on Artificial Immune Systems, ICARIS 2005, Banff, Alberta, Canada, LNCS 3627 Springer, 2005.
26. S. Stepney and all, "Towards a Conceptual Framework for Artificial Immune Systems",3th International Conference on Artificial Immune Systems, ICARIS 2004 Catania, Sicily, Italy, LNCS 3239 Springer 2004.
27. M. Sipper, E. Sanche, D. Mange, M. Tomassini, A. Pérez-Uribe, and A. Stauffer, "A Phylogenetic, Ontogenetic and Epigenetic View of Bio-Inspired Hardware Systems," IEEE Transactions on Evolutionary Computation, vol.1, no. 1, pp. 83-97, 1997.
28. S.E.Mili,D.Meslati, "Modelisation of bio inspired systems using MDA", International Conference on Information Technology and e-Services , ICITeS.2012, Sousse, Tunisia.

Papiers posters

Determining a Global Optimum of a Nonconvex Function in R^n Box

Fadila Leslous[1] and Mohand Ouanes[2]

[1] **LAROMAD, Mouloud Mammeri University, Tizi-Ouzou, Algeria**
[2] **LAROMAD, Mouloud Mammeri University , Tizi-Ouzou, Algeria**
fadila.leslous@yahoo.fr
ouanes_mohand@yahoo.fr

Abstract. This paper presents a generalization of the proposed method [1] for nonconvex functions, using the difference of convex functions algorithm and the minimum of the average of approximations of the function from the vertices of the box. This strategy has the advantage of giving in general a minimum to be situated in the attraction zone of the global minimum searched. After applying the difference of convex functions algorithm from this minimum we arrive, certainly, at the global minimum searched.

Keywords: Optimization DC and DCA · Global optimization · Nonconvex optimization.

1 Introduction

In recent years, global optimization [1-4] has been the subject of several studies due to new theoretical results, strong demand in several fields including industrial applications, and the development of computing resources [5-7].

Global optimization did not inherit the easiness of the numerical techniques of local optimization. Indeed, the latter use for the most part descent directions, which makes it possible to converge naturally to a local minimum point[8-12].

Global optimization avoids staying at such a point as they must be escaped. This is why many approaches have been used in the attempt to solve problems. They mostly consist of finding a state of minimum and to stop only if it is the best (the global optimum). One of the most widely used methods is the Difference of Convex functions Algorithm , which is a descent method without a linear search (greatly favored by large dimensions). This method is applied with great success to numerous nonconvex optimization problems that rise in diverse fields of applied sciences such as: transportation logistics, telecommunications, finance, data mining, robotics,...

In this work we propose a generalization of method which is based on the decomposition of the function to be minimized into a difference of convex functions and the application of the difference of convex functions algorithm[13-16].

From a good initial point, the difference of convex functions algorithm furnishes

a global minimum, we propose an approach to find a good initial point. For that we minimize the average of approximations of the function from the vertices of the box. This strategy has the advantage of giving generally a minimum to be located in the attraction zone of the global minimum searched. We apply the difference of convex functions algorithm from the minimum found and we arrive certainly to the global minimum searched.

2 Problem formulation

We consider the optimization difference of convex functions problem (DC) as follows:
$(P) \Longleftrightarrow \min\{f(x) = g(x) - h(x), x \in B \subset R^n\}$
$B = \{x = (x_1, x_2...x_n) \in R^n : a_i \leq x_i \leq b_i, i = 1, ..., n\}$ with:
a_i and b_i Constants in R.
f :$R^n \longrightarrow R$ nonconvex of class $\mathcal{C}^2$.
g :$R^n \longrightarrow R$ convex
h :$R^n \longrightarrow R$ convex
We want to solve the problem (P) by applying difference of convex functions algorithm (DCA) to the minimum of the average of approximations of the function from the vertices of the box B.

2.1 The principle of difference of convex functions algorithm

Note that DCA works only with DC components g and h [5,6,14].
At the k-th iteration of DCA, h is replaced by its affine minorant
$h_k(x) = h(x^k) + \langle x - x^k, y^k\rangle$ in the neighborhood of x^k.
Knowing that h is a convex function and B is a box subset of R^n, we have therefore $h(x) \geq h_k(x), \forall x \in B$. As a result, $g(x) - [h(x^k) + \langle x - x^k, y^k\rangle] \geq g(x) - h(x), \forall x \in B$.
That is to say, $g(x) - [h(x^k) + \langle x - x^k, y^k\rangle]$ is a majorant function of function f (x).
Indeed, the surface of f^k can be imagined as a bowl being placed directly above the surface of f; Moreover, the two surfaces ar touching at point $(x^k, f(x^k))$.

2.2 Difference of convex functions algorithm

Initial Step: x^0 given, k=0.
Step 1: We search $y^k \in \partial h(x^k)$.
Step 2: We determine $x^{k+1} \in \partial g^*(y^k)$.
Step 3: If the stop conditions are satisfied then DCA is terminated; Otherwise k=k+1 and we repeat the Step 1.

2.3 The principle of the proposed method

From a good initial point the DCA furnishes a global minimizer[1,2,4].
In the case of a minimizing a nonconvex funtion of classe $\mathcal{C}^2$ defined in R^n, The minimum found starting from a vertices of the box B will generally be different from that found starting from another vertices of the box B.
We propose to find a good initial point.Instead we want to minimize the average of approximations to f from the vertices $s_1, s_2, s_3, ..., s_{2^n}$ of the box B as follows:

$min\frac{1}{2}(\sum_{i=1}^{2^n} f(x; s_i)$ with:
$f(x; s_i) = g(x) - \nabla h(s_i)(x - s_i) - h(s_i)$
This strategy has the advantage of providing in general a minimum to be located in the attraction zone of the minimum global searched.

2.4 Properties of the proposed method

1- The MDCA constructs an initial point x^0 which is the solution of a convex problem, and a sequence $\{x^k\}$ generated by DCA, then it converges to an optimal solution of our problem.
3- It can be seen that MDCA has the option to skip certain neighborhoods of local minima, then arrives at a neighborhood of the global solution. we can understand that the performance of MDCA is the position of the initial point.

3 The algorithm of the proposed method

Step 0: $x^0 = argmin_{x\in B}\frac{1}{2^n}(\sum_{i=1}^{2^n} f(x; s_i)$.
Step 1: Compute x^{k+1} by DCA.
Step 2: If the stop conditions are satisfied then algorithm is terminated; Otherwise k=k+1 and we repeat the Step 1.

4 The application of the proposed method

Example 1.

$$(P) \Leftrightarrow \begin{cases} f(x,y) = y^2 + x - x^2 + y \longrightarrow Min \\ x \in [-3,+3], y \in [-3,+3] \end{cases}$$

$h(x,y) = x^2 - y$
$g(x,y) = y^2 + x$
$\nabla h(x,y) = (2x,-1)$
We aim to minimize the average of the approximations of f from s_1, s_2, s_3 and s_4, with:
$s_1 = (-3,-3), s_2 = (3,-3), s_3 = (3,3), s_4 = (-3,3)$.

$f(x; s_i) = g(x) - \nabla h(s_i)(X - s_i) - h(s_i), i = 1, 2, 3, 4.$
$f(x; s_1) = y^2 + 7x + y + 9.$
$f(x; s_2) = y^2 - 5x + y + 9.$
$f(x; s_3) = y^2 - 5x + y + 9.$
$f(x; s_4) = y^2 + 7x + y + 9.$
Step 0:
Solve the problem convex $P\prime$ following:

$$(P\prime) \Longleftrightarrow \min \frac{1}{4}(f(x, s_1) + f(x, s_2) + f(x, s_3) + f(x, s_4)).$$

$(P\prime) \Longleftrightarrow \min(y^2 + y + x + 9).$
$x = (-3, -\frac{1}{2})$. is the solution of $(P\prime)$
Step 1:
Application of DCA from $x = (-3, -\frac{1}{2})$.
$x = (-3, -\frac{1}{2})$ is the optimal solution (global minimum) of Problem (P).

Remark 1. Using the proposed method (MDCA) we did not choose a starting point, instead we want to minimize the average of the approximations of f from the vertices of the box which will provide a minimum located in the attraction zone of global minimum.
DCA is applied from this minimum, which give the global minimum searched.

5 Conclusions:

This paper proposes the generalization of the MDCA method on an R^n box, using the DCA (difference of convex Algorithm) which concernes a particular class of optimization problems, namely: nonconvex problems.
The strategy of minimizing the average followed by the standard application of DCA has led to the production of the global minimum of function f, while the standard function fminbnd of MATLAB found a non global, local minimum. It now remains to test other examples to better evaluate the pertinence of this strategy, reinforcing the importance of DCA in solving nonconvex problems.

References

1. Leslous, F., Ouanes, M., Marthon, P.: Improving the Robustness of Difference of Convex Algorithm in the Research of a Global Optimum of a nonconvex Differentiable Function Defined on a Bounded Closed Interval. J. Applied Mathematical Sciences. 8, 1, 1-12 (2014)
2. Le Thi, H.A., N.V. Vinh, N.V., Phan Din, T.: A Combined DCA and New Cutting Plane Techniques for Globally Solving Mixed Linear Programming, SIAM Conference on Optimization, Stockholm (2005)
3. Horst, R., Hoang T.: Global Optimization : Deterministic Approches, 2nd revised edition, Springer, Berlin (1993)

4. Nguyen, C.N., Le Thi, H.A., Phan Din. T.: A Branch and Bound Algorithm Based on DC Programming and DCA for Strategic Capacity Planning in supply Chain Design for a New Market Opportnity. Operations rechearch proceedings (2006)
5. Strekalovsky, S.S.: Global Optimality Conditions in Nonconvex Optimization. J. Optim. Theory and Applications. 173, 770-792. (2017)
6. Messine, F., Lagouanelle, J.L.: Enclosure Methods for Multivariate Differentiable Functions and Application to Global Optimization. J. Universal Computer Science. 4, 589-603 (1998)
7. Ratschek, H., Rokne, J.G.: New Computers Methods for Global Optimization, Wiley, New-York (1988)
8. Vinko, T., Lagouanelle, J.L., Csendes, T.: A New Inclusion Function for Optimization: K ite- The One Dimensional Case. J. Global Optimization (2004)
9. Hamzacebi, C., Kutay, F.: A heuristic approach for finding the global minimum: Adaptive random search technique. J. Applied Mathematics and Computation. 173, 1323-1333 (2006)
10. Horst, R., Pardalos, P.M., Romeijn, H.E.: Handbook of global optimization, Springer Publisher, Kluwer (1995)
11. Huang, Z., Miao, X., Wang, P.:A revised cut-peak function method for box constrained continuous global optimization. J. Applied Mathematics and Computation. 194, 224-233 (2007)
12. Liberti, L., Maculan, N.: Global optimization: from theory to implementation, Springer Publisher, (2006)
13. Liu, C.X., Chi, G.: structural optimization using matlab. Engineering Systems and Design Lab, KAIST (2006)
14. Yong, W., Guangla, Z., Xin, Z., Wanquan, L., Louis, C.: The Non-Convex Sparse problem With Nonnegative Constraint for Signal Reconstruction. J. Optim. Theory Appl. 170, 1009-1025 (2016)
15. Nocedal, J., Wright, S.J.: Numerical optimization. Springer Series in Operations Research, Springer, (1999)
16. Ratschek, H., Rokne, J.: New computer methods for global optimization. Ellis Horwood Limited Market Cross House, Cooper Street, Chichester, West Sussex, PO19 1EB, England (1988)

Optimization and Sensitivity Analysis of Queue with Vacations*

Baya Takhedmit[1], Sofiane Ouazine[2], and Karim Abbas[3]

[1] University of Bouira, Faculty of Science and Applied Sciences, Department of Mathematics, 10000-Bouira, Algeria
bayatakh@gmail.com
[2] University of Bejaia, Faculty of Exact Sciences, Department of Mathematics, 06000-Bejaia, Algeria
wazinesofi@gmail.com
[3] University of Bejaia, Faculty of Exact Sciences, Department of Operational Research, 06000-Bejaia, Algeria
karabbas2003@yahoo.fr

Abstract. Queues are very suitable models for modeling many real situations.
The main purpose of the current paper is to use Taylor series expansion approach for analyzing of the GI/M/1/N queue with multiple working vacations, with an emphasis on perturbation analysis. The discussion of some problems for optimization of the queueing model.

Keywords: GI/M/1/N Queue with Vacations· Taylor Series Expansion· Numerical Approximation· Optimization.

1 Introduction

This paper devotes to the performance analysis via functional approximations of the GI/M/1/N queue with multiple working vacation policy. Specifically, denoting the working vacation rate, we seek to compute the stationary distribution of the queue-length process embedded at pre-arrival epochs. Furthermore, in performance analysis one is not only interested in evaluating the system for certain set of parameters but also in the sensitivity of the performance with respect to the parameter.

2 The GI/M/1/N queue with multiple working vacation policy

2.1 Mathematical Model

We consider the $GI/M/1/N$ queue with multiple working vacations policy, where N is the number of buffers in the queue including the one who is in service.

* Research Unit LaMOS.

Assume that customer arrivals occur at discrete-time instants τ_k, where $\tau_0 = 0$. In other words, customers arrive at the system according to a renewal process with inter arrival time distribution $G(t)$ and mean $\frac{1}{\lambda}$. The service time T_s of each server is assumed to be distributed exponentially with service rate μ. Its density function is given by $s(t) = \mu e^{-\mu t}, t \geq 0$.

The service times during a normal service period, the service times during a working vacation period and the working vacation times are exponentially distributed with rate μ, η and λ, respectively.

We introduce the probabilities a_k, b_k and c_k, in order to define the transition matrix of the model.

1. a_k is the conditional probability that k customers have been served during an inter-arrival time when service period is going on: $a_k = \int_0^{+\infty} \frac{(\mu t)^k}{k!} e^{-\mu t} dG(t)$.

2. b_k explains the probability that k services are completed during an inter-arrival time in the working vacation period: $b_k = \int_0^{+\infty} \frac{(\eta t)^k}{k!} e^{-(\gamma+\eta)t} dG(t)$.

3. c_k represents the probability that k customers have been served during an inter-arrival time given that working vacation terminates and service period is going on: $c_k = \int_0^{+\infty} \sum_{j=0}^{k} \left\{ \int_0^t \gamma e^{-\gamma x} \frac{(\eta x)^j}{j!} e^{-\eta x} \times \frac{[\mu(t-x)]^{k-j}}{(k-j)!} e^{-\mu(t-x)} dx \right\} dG(t)$.

The transition matrix of the embedded Markov chain can be written in the partitioned-block form as (see [1])

$$P = \begin{pmatrix} B_{00} & A_{01} & 0 & \dots & 0 \\ B_1 & A_1 & A_0 & \dots & 0 \\ B_2 & A_2 & A_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{N-1} & A_{N-1} & A_{N-2} & \dots & A_0 \end{pmatrix}_{(2N+1)\times(2N+1)}$$

Where $B_{00} = 1 - a_0 - b_0$, $A_{01} = (b_0, c_0)$

$$A_k = \begin{pmatrix} b_k & c_k \\ 0 & a_k \end{pmatrix}, \quad B_k = \begin{pmatrix} 1 - \sum_{i=0}^{k} (b_i + c_i) \\ 1 - \sum_{i=0}^{k} a_i \end{pmatrix}, \quad 1 \leq k \leq N-1$$

3 Taylor series expansion approach

This section presents Taylor series expansion for Markov chains with finite state space. We provide a recursive computation scheme for higher order derivatives of stationary distribution π_η with respect to parameter of interest η. The $n-order$ derivative of the stationary distribution, π_η of an ergodic finite-state Markov chain with respect to η is given by

$$\pi_{\eta}^{(n)} = \pi_{\eta} K_{\eta}(n);$$

where

$$K_{\eta}(n) = \sum_{\substack{1 \leq m \leq n; \\ 1 \leq l_k \leq n; \\ l_1 + \cdots + l_m = n}} \frac{n!}{l_1! \cdots l_m!} \prod_{k=1}^{m} \left(P_{\eta}^{(l_k)} D_{\eta} \right),$$

the matrix $D_{\eta} = (I - P_{\eta} + PI_{\eta})^{(-1)} - P_{\eta}$ exists and it is called the deviation matrix. The recursive Form of the derivatives

$$\pi_{\eta+\Delta} = \sum_{n=0}^{k} \frac{\Delta^n}{n!} \pi_{\eta}^{(n)} + r_{\eta}(k, \Delta); \;\; \Delta > 0.$$

4 Optimization analysis

This section focuses to define criterion to a problem when the service times during a working vacation period change in some interval. The criterion is given by

$$min_{\eta} \pi_{\eta} f \tag{1}$$

for some cost function f. The above optimization problem will be resolved by using an iterative algorithm, such as the a steepest decent algorithm.

$$\eta_{n+1} = \eta_n - \frac{\pi_{\eta} f}{(\pi_{\eta} f)'}.$$

Let us define the cost elements in the following:

C_1: cost per unit time for each customer present in the system; C_2: cost per unit time for service during a normal service period; C_3: cost per unit time for service during a working vacation period; C_4: fixed cost for every lost customer when the system is blocked.
Based on the definitions of each cost element and the corresponding measures of system performance, the total expected cost function per unit time is given by

$$\pi_{\eta} f = C_1 L + C_2 P_{bs}^{1} + C_3 P_{wv}^{0} + C_4 P_b,$$

where

L is the mean queue length, P_{bs}^{1} is the probability that the server is on service period, P_{wv}^{0} is the probability that the server is on working vacation and P_b is the blocking probability.

We can rewrite the objective function as follow
$\pi_\eta f = C_1(V_L \times \pi_\eta) + C_2(V_{P^1_{bs}} \times \pi_\eta) + C_3(V_{P^0_{wv}} \times \pi_\eta) + C_4(V_{P_b} \times \pi_\eta)$,
where:
$V_L = (0, 1, 1, 2, 2, ..., N, N)$, $V_{P^1_{bs}} = (0, 0, 1, 0, 1, ..., 0, 1)$, $V_{P^0_{wv}} = (1, 1, 0, 1, 0, ..., 1, 0)$ and $V_{P_b} = (0, 0, 0, 0, 0, ..., 1, 1)$.

For computation purpose, we let $N = 3, \lambda = 1, \mu = 4, \eta = 1, C_1 = 1, C_2 = 2, C_3 = 8.5, C_4 = 1$ and $\eta \in [0, 1]$. We first depict the total expected cost function per unit time as a function of η in Figure 1. We let the stopping tolerance $\varepsilon = 10^{-4}$ and fix the initial point $\eta_0 = 0$. So, After 8 iterations the minimum expected operating cost per unit time converges to the solution $\eta^* = 0.5739797238398$ with a value 8.292485188301.

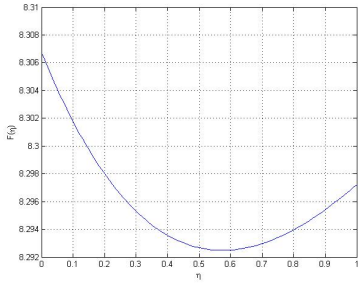


Fig. 1. The curve of cost function with the change of working vacation parameter η.

5 Conclusion

In this paper we have used the Taylor series expansion approach to investigate the numerical evaluation of stationary distribution of the the $GI/M/1/N$ queue with multiple working vacations policy, where we have explored particularly technique to optimization of considered queueing model.

References

1. Banik, A.D., Gupta, U.C., Pathak, S.S.: On the GI/M/1/N queue with multiple working vacations - analytic analysis and computation. Appl. Math. Mode. **31**, 1701–1710 (2007)
2. Ouazine, S., Abbas, K., Heidergott, B.: The Taylor Series Expansions for Performance Functions of Queues: Sensitivity Analysis. Analytical and Stochastic Modeling Techniques and Applications. **7984** , 1–11 (2013)
3. Tian N., Zhang Z.G., Vacation Queueing Models: *Theory and Applications*. Springer-Verlag, New York (2006)

Unscented Kalman Filter Observer

DAID Assia[1,2], AIDENE Mohamed[1] et BUSVELLE Eric[2]

[1] Laboratoire de Conception et Conduite des Systèmes de Production,
Université Mouloud MAMMERI,Tizi-Ouzou,Algérie
aidene@ummto.dz
[2] Laboratoire d'Informatique et des Systèmes,LIS UMR 7020,
Université de TOULON,FRANCE
busvelle@univ-tln.fr, assia.daid@yahoo.fr

Résumé: Dans cet article, nous étudions les propriétés du filtre UKF ("unscented Kalman filter") en tant qu'observateur non linéaire. Sous des hypothèses d'observabilité, le filtre de Kalman étendu (EKF) est un observateur exponentiel sous réserve d'être écrit dans une forme canonique d'observabilité et sous sa forme grand-gain. On montre que contrairement à EKF, UKF n'est pas un observateur à convergence exponentielle. On propose une modification d'UKF qui est une meilleure candidate en tant qu'observateur. Finalement, les propriétés étudiées dans l'article sont illustrées sur un exemple de géolocalisation d'un navire.
Mots clés: Observateurs non-linéaires. Grand gain. Unscented Kalman filter. High-gain Unscented Kalman filter.

1 Introduction

Le filtre de Kalman étendu (EKF, "extended Kalman filter") est trés utilisé par les ingénieurs pour estimer l'état d'un systéme à partir de mesures fournies par des capteurs [10]. Le modéle non-linéaire du systéme est établi d'aprés les connaissances physiques du systéme, de même que la relation entre les variables d'état internes et les mesures.

$$\begin{cases} \frac{dx(t)}{dt} = f(x(t), t) \\ y(t) = h(x(t), t) \end{cases} \tag{1}$$

Usuellement, $x(t) \in \mathbb{R}^n$ représente l'état du systéme et $y(t) \in \mathbb{R}^p$ représente les p mesures.

2 La transformation Unscented

Dans cette section, nous rappellons les bases de la transformation "unscented" et les équations de UKF [7, 8]. Nous nous plaçons dans $\mathbb{R}^n$.

1. Choisir $2n+1$ σ-points:

$$X = [m \cdots m] + \sqrt{c}\left[0 \ \sqrt{P} - \sqrt{P}\right] \tag{2}$$

X est la matrice des σ-points, $c = \alpha^2(n+k)$, avec $k \geq 0$, $\alpha \in (0,1]$.
c, k et α sont des paramétres de réglage. La matrice P est une matrice définie positive. Elle peut donc se décomposer sous la forme de Cholesky $P = BB'$ et on note $B = \sqrt{P}$.

2. Calculer les poids associées aux σ-points:
$W_m = (W_m^0, W_m^1, \ldots W_m^{2n})^t$ où

$$W_m^{(0)} = \frac{\lambda}{n+\lambda};$$
$$W_m^{(i)} = \frac{1}{2(n+\lambda)}, \quad i = 1, \ldots 2n;$$

et $W_c = (W_c^0, W_c^1, \ldots W_c^{2n})^t$ où

$$W_c^{(0)} = \frac{\lambda}{n+\lambda} + (1 - \alpha^2 + \beta);$$
$$W_c^{(i)} = \frac{1}{2(n+\lambda)}, \quad i = 1, \ldots 2n;$$

λ est un paramètre scalaire défini par $\lambda = c - n$.

3. Transformer chaque σ-point par la transformation non linéaire g,

La moyenne et la covariance de $g(X)$ sont données par:

$$E[g(X)] \approx m = g(X)W_m = \sum_{i=0}^{2n} W_m^i g(X_i);$$
$$\mathrm{Cov}(g(X)) \approx \sum_{i=0}^{2n} W_m^i (g(X_i) - m)(g(X_i) - m)^t;$$

On a noté X_i la $i^{\text{ième}}$ colonne de X, et lorsque g est appliquée à la matrice X, $g(X)$ représente la matrice $[g(X_0) \cdots g(X_{2n})]$.

La matrice W est définie comme suit

$$W = \left(I - \left[W_m \cdots \ W_m\right]\right) \times \mathrm{diag}\left(W_c^0 \cdots \ W_c^{2n}\right) \times \left(I - \left[W_m \cdots \ W_m\right]\right)^t \tag{3}$$

2.1 Algorithme de UKF dans le cas continu

Les équations correspondant à UKF dans le cas continu pour le systéme (1) sont données par

$$\begin{cases} K(t) = X(t)Wh'(X(t),t) \\ \dfrac{dm(t)}{dt} = f(X(t),t)W_m + K(t)\left(z(t) - h(X(t),t)W_m\right) \\ \dfrac{dP(t)}{dt} = X(t)Wf'(X(t),t) + f(X(t),t)WX'(t) + Q(t) - K(t)R(t)K'(t) \end{cases} \tag{4}$$

Dans cet algorithme Q et R sont des matrices de covariance de l'état et de bruit de mesure respectivement, elles sont symmétriques définies positives. Dans le cas déterministe, ces deux matrices seront considérées comme des paramètres de réglage et $z(t) = \frac{dy(t)}{dt}$.

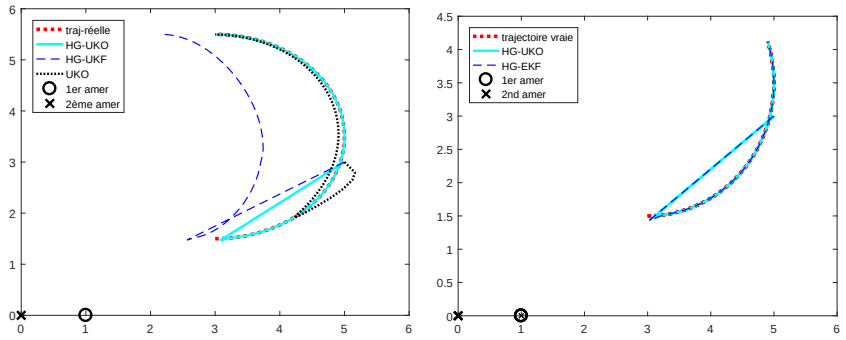
3 Algorithme HG-UKO dans le cas continu

Les équations correspondant à HG-UKO dans le cas continu pour un systéme écrit sous forme canonique d'observabilité [5]sont données par

$$\begin{cases} \dfrac{dm}{dt} = Am + b(m,t) + PC^t R^{-1}(y(t) - Cm) \\ \dfrac{dP}{dt} = XW(AX + b(X,t))' + (AX + b(X,t))WX' + Q^\theta - XWX'C'R^{-1}CXWX' \end{cases} \tag{5}$$

où la ligne i, colonne j de Q^θ est égale à $Q^\theta_{i,j} = \theta^{i+j+1}Q_{i,j}$

4 Comparaison des performances des différents observateurs



(a) Trajectoire réelle et estimée avec uko,hguko et hgukf

(b) Trajectoire réelle et estimée avec hguko et hgekf

Fig. 1: Trajectoire dans l'espace physique

La fig.1a montre la performance des trois observateurs, ce qui permet de vérifier que la convergence du HG-UKO peut être trés rapide par rapport aux deux autres. Nous avons ensuite comparé HG-UKO avec HG-EKF (fig.1b),on constate que les résultats sont similaires ce qui est trés intéressant car on retrouve les bonnes performances du HG-UKO(en tant qu'observateur).

5 Conclusion

Parmi les filtres/observateurs non linéaires présentés, seuls HG-EKF et HG-UKO convergent en simulation, alors que HG-UKF n'est pas un observateur. L'avantage du HG-UKO est une relative simplicité d'écriture puisqu'il n'est pas nécessaire de calculer les Jacobiennes du système. Cette propriété est particulièrement intéressante lorsque le système est écrit dans sa forme canonique.

References

1. N. Boizot, E. Busvelle, and J.-P. Gauthier, An adaptive high-gain observer for nonlinear systems,. *Automatica,* vol. 46, no. 9, pp. 1483–1488, Sep. 2010.
2. Nicolas Boizot, Adaptative high-gain extended Kalman filter and applications, *Ph. D. Université de Bourgogne, Université du Luxembourg*, 2010
3. Nicolas Boizot and Eric Busvelle, Adaptive-gain observers and applications, in *Nonlinear Observers and Applications*, Springer Berlin Heidelberg, 2007, pp. 71–114.
4. M. Doumiati, A. Victorino, A. Charara, and D. Lechner, Unscented Kalman filter for real-time vehicle lateral tire forces and sideslip angle estimation, *IEEE Intelligent Vehicles Symposium*, Jun. 2009.
5. J.-P. Gauthier and I. A. K. Kupka, *Deterministic Observation Theory and Applications,* Cambridge University Press, 2001.
6. Simon J. Julier and Jeffrey K. Uhlmann, New extension of the Kalman filter to nonlinear systems, *Signal processing, sensor fusion, and target recognition* VI. Vol. 3068. International Society for Optics and Photonics, 1997
7. S. Särkkä, On Unscented Kalman Filtering for State Estimation of Continuous-Time Nonlinear Systems, *IEEE Transactions on Automatic Control,* vol. 52, no. 9, Sep. 2007, pp. 1631–1641
8. S. Särkkä, Bayesian Filtering and Smoothing, *Cambridge University Press*, 2009.
9. E. A. Wan and R. Van Der Merwe The unscented Kalman filter for nonlinear estimation. In Adaptive Systems for Signal Processing, *Communications, and Control Symposium 2000* AS-SPCC, 2000, pp. 153–158
10. Z. Yacine, Observateurs pour l'Estimation de la Dynamique Latérale du véhicule et Application à la Détection de Situations Critiques, *Thèse de Doctorat*, Université Mouloud Mammeri, Tizi Ouzou, 2016.

Reformulation de la Requête Web par l'algorithme FireFly

Meriem Zeboudj et Khaled Belkadi

LAMOSI, Département d'informatique, Université des sciences et de la technologie d'Oran (USTO-MB), Oran, Algerie

meriem.zeboudj@univ-usto.dz , khaled.belkadi@univ-usto.dz

Résumé. Un internaute soumet de nombreuses requêtes sur les moteurs de recherche Web afin de récupérer de meilleurs résultats. Si ces requêtes n'expriment pas ces besoins ou bien ces objectifs, ceci implique que, certaines informations ne sont pas formulées, et ce qui nécessite la reformulation de ces requêtes. Dans ce papier, une approche d'optimisation bio-inspirée basée sur l'algorithme des lucioles (FireFly Algorithm) a été employée pour formuler la requête en fournissant une nouvelle suggestion. Expérimentalement, nous étudions la performance de la méthode proposée en la comparant avec d'autres techniques d'optimisations telles que les essaims particulaires (PSO).

Mots clés. Optimisation, Algorithme des lucioles, Suggestion de la requête, Reformulation de la requête, Recherche d'Information.

1 Introduction

Le concept de la reformulation est étudié par de nombreux chercheurs, il est devenu parmi l'un des principales notions dans le domaine de la recherche d'information et a fait l'objet de beaucoup de travaux. Ces travaux apportent des solutions aux utilisateurs selon leurs besoins d'information. L'idée d'amélioration du système de recherche web en modifiant les requêtes a été étudiée dans [7]. Certaines techniques d'améliorations consistent à rajouter à ces requêtes des termes issus de ressources linguistiques existantes comme le WordNet [9, 11], ou bien de ressources construites à partir des collections [8]. D'autres techniques [6, 12] s'appuient sur l'hypothèse que les documents mieux classés (les k premiers documents) sont considérés comme étant pertinents. De plus, les journaux de requêtes ont été largement utilisés dans les différentes approches de Fouille de données Web « Web Mining » [4, 10], afin d'améliorer l'efficacité des moteurs de recherche et leurs utilisabilités.

Notre approche de reformulation de requêtes proposée, est construite essentiellement sur l'algorithme des lucioles (FireFly Algorithm, FF) en s'inspirant de nouvelles suggestions. Cet algorithme s'applique sur le graphe d'association des termes dont chaque requête est considérée comme une luciole, et la reformulation est suggérée en fonction de termes dans les chemins optimaux, et en utilisant une base de données lexicale « WordNet » afin d'ajouter tout le contexte possible qui ne figure pas dans le chemin optimal.

L'article est organisé comme suit: dans la section 1, nous commençons par une introduction. Dans la section 2, la stratégie de reformulation de la requête basée sur l'algorithme FireFly proposée est expliquée. Enfin, la section 3. illustrant la performance de l'approche proposée avec d'autres approches d'optimisation.

2 Algorithme FireFly pour la reformulation des requêtes

À l'arrivée de la requête de l'utilisateur au moteur de recherche Web, la première étape consiste à récupérer les documents pertinents (document, URL, etc.) en s'appuyant sur l'API de Google [2]. Ensuite, on extrait les termes (Metakeywords) de chaque résultat récupéré à cette requête (de façon explicite).

Une phase de prétraitement est utilisée par la suite, dont laquelle on supprime tous les StopWord (mots inutiles). Après, les termes utiles seront annotés avec des informations sur les parties du discours (genre de mots: noms, verbes, infinitifs et participes) et des informations de lemmatisation à l'aide de l'outil treeTagger [1].

Sur la base des termes extraits et annoté (on prend seulement les noms ou bien les adjectifs), ces termes sont insérés dans un arbre de recherche ternaire (TST-Ternary Search Tree). Dans l'étape suivante, l'association entre les documents est définie en fonction de la similarité et un graphe des termes est créé. Chaque traversée du graphe (basée sur la valeur de similarité) est considérée comme un chemin de l'algorithme des lucioles. FireFly converge vers les top-k chemins optimaux et les mots-clés présents dans ces chemins sont extraits. De plus, WordNet [3] sera interrogé afin d'extraire les hyponymes/hyperonymes de ces mots-clés qui seront suggérés pour l'utilisateur par la suite.

Le diagramme schématique de la reformulation de requête proposée est représenté dans la fig. 1.

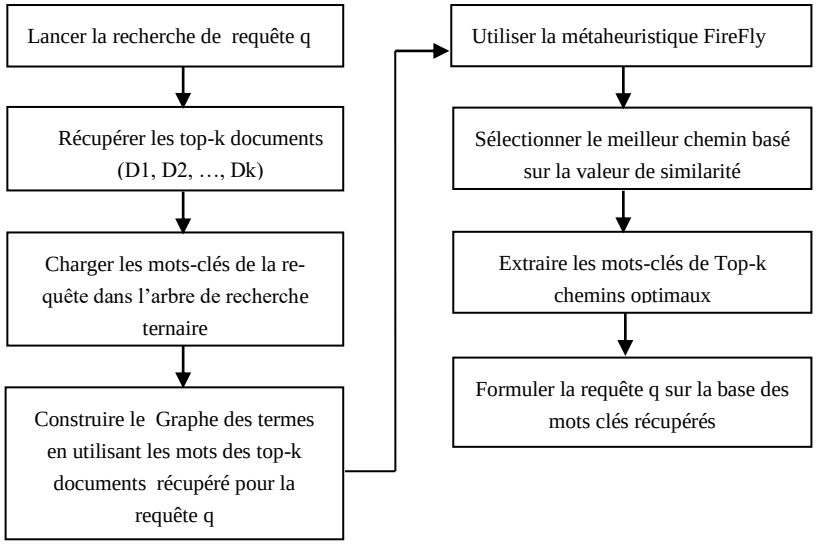


Fig. 1. Diagramme schématique de la reformulation de requête proposé.

2.1 Construction du TST et Graphe des Termes

Le TST est préférée par rapport à d'autre type de structures, comme son efficacité sur la correction des fautes d'orthographe à partir de requêtes des utilisateurs et aussi en terme de calcul d'espace de stockage. À la fin de chaque mot clé dans cet arbre, un graphe de terme sera associé, en utilisant l'Algorithme 1.

Algorithme 1. Construction graphe des termes (q_i, D_n) et création de l'association du graphe des termes (D_i, D_j, W_{ij})

Entrée: les mots-clés extraits de documents d_t

Sortie: structure graphe G_t

Pour chaque document récupéré d_i ∈ D de chaque requête q

- Extraire les termes de d_i
  - Créer un nœud pour chaque terme
  - Créer des arêtes entre ces nœuds pour être un graphe complet en se basant sur la valeur de similarité,
- Calculer la similarité entre (D_i, D_j),

Fin Pour

Pour chaque terme d_t

Si le terme d_t présent dans plusieurs documents

- Créer un arrêt reliant d_t aux termes présents dans les documents

Fin Si

Fin Pour

//Calcul la valeur de similarité

Entrée: Mots-clés (k_i, k_j) des deux documents D_i et D_j

Sortie: valeur de similarité entre D_i et D_j

Pour chaque mot clé dans le document D_i

-Charger le k_i dans une carte M

Pour chaque mot clé dans le document D_j

Si k_j présents dans M **alors**

Incrémenter la valeur du poids

Attribuer le poids sur le bord entre D_i et D_j

Fin Si

Fin Pour

Fin Pour

2.2 Application de l'algorithme FireFly

Notre modèle est construit avec FireFly [5] qui prend la forme d'une structure de graphe où les nœuds racines sont les termes de la requête et les arcs pointent vers l'éventuel raffinement de la requête, et le poids sur les arrêts code l'importance de l'association entre les nœuds.

Le processus de l'algorithme FireFly débute avec l'initialisation de la population des lucioles et chaque luciole dans une population désigne une solution possible. Dans notre cas, la population initiale contient des mots clés liés à l'information ainsi que l'évaluation de ces mots clés. Cette évaluation est mesurée comme la fréquence d'un mot clé dans les documents résultants de la requête d'utilisateur. Elle est considérée comme un poids de phéromone dans le graphe. A la fin de chaque mot clé d'une requête, une liste d'adjacence est créée dans laquelle les mots clés récupérés de la re-

quête initiale sont entrée dans l'ordre de fréquence le plus élevé avec l'association entre les termes.

3 Analyse des résultats

Les critères de mesure des performances les plus connues sont le rappel et la précision. Le rappel mesure la proportion des documents pertinents renvoyés par le système parmi tous ceux qui sont pertinents pour la requête, et la précision mesure la proportion des documents pertinents parmi l'ensemble de ceux renvoyés par le système.

Le tableau 1 illustre des calculs de précision et de rappel pour les 20 premiers documents renvoyés par notre approche FireFly et PSO (respectivement) pour cinq requêtes différentes en utilisant plusieurs collections.

Table 1. Tableau comparatif pour FireFly et PSO.

	FireFly		**PSO**	
Requêtes	Rappel	Précision	Rappel	Précision
R1	0,25	1,00	0,30	1,00
R2	0,70	0,75	0,25	0,50
R3	0,50	0,67	0,50	0,33
R4	0,80	1,00	0,50	0,40
R5	0,80	0,88	0,50	0,30

Selon les résultats obtenus dans tableau comparatif on remarque une amélioration au niveau de la précision avec FireFly par rapport à la PSO, par exemple pour la requête cinq (R5) l'algorithme du FireFLy a donné une bonne précision (0,88) par rapport à la PSO (0,3) et aussi pour le rappel FireFly a donné 0,5 par contre la PSO a donnée 0,3 respectivement.

4 Conclusion

Le Système de Recherche d'Information fournit parfois des résultats qui ne conviennent pas et qui ne satisfont pas ses utilisateurs. Donc, afin de s'approcher de ce dernier le plus possible à la pertinence, une étape de reformulation de la requête est souvent utilisée. Nous avons étudié la performance de notre méthode par rapport à d'autres techniques d'optimisations telle que l'optimisation par les essaims particulaires (PSO), et l'algorithme du FireFly Algorithme est plus efficace pour la reformulation des requêtes.

Références

1. TreeTagger - a part-of-speech tagger for many languages, http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ , consulté le 05/08/2017.
2. Custom Search JSON API | Custom Search | Google Developers, https://developers.google.com/custom-search/json-api/v1/overview.
3. Mark A. Finlayson, "MIT Java Wordnet Interface (JWI)", User's Guide Version 2.4.x October 29, 2015.
4. Z. Kunpeng, W. Yoke and H. Geok, "Wavelet analysis of sensor signals for tool condition monitoring: A review and some new results", In International Journal of Machine Tools & Manufacture 49, p. 537, 2009.

5. X.S. Yang, “Nature-Inspired Metaheuristic Algorithms”, In Luniver Press, UK. 2008.
6. B. Billerbeck and J. Zobel, “Efficient query expansion with auxiliary data structures”, In Information Systems, 31(7):573-584, November 2006, DOI: 10.1016/j.is.2005.11.002.
7. Efthimiadis, E.N., “Query Expansion”, In Williams Martha, E. (ed.) Annual Review of Information Systems and Technology, vol. 31, pp. 121–187. Information Today, 1996.
8. L. Gan and H. Hong, “Improving Query Expansion for Information Retrieval Using Wikipedia”, In International Journal of Database Theory and Application, Vol.8, No.3, pp.27-40 , 2015.
9. H. Fang, “A Re-examination of Query Expansion Using Lexical Resources,” Computational Linguistics, pages 139–147, 2008.
10. R. Baeza-Yates, C.Hurtado and M. Mendoza, “Query Recommendation Using Query Logs in Search Engines”, Current Trends in Database Technology - EDBT 2004 Workshops, Lecture Notes in Computer Science, Volume 3268, pp. 588-596, 2005.
11. J. Zhang, B. Deng, and X. Li, “Concept Based Query Expansion Using WordNet”, 2009 International e-Conference on Advanced Science and Technology, pages 52–55, March 2009.
12. H. Imran and A. Sharan, “A Framework for Automatic Query Expansion”, In WISM 2010, LNCS 6318, pp. 386–393, 2010.

Expérimentation d'un nouvel algorithme pour le calcul de l'intersection entre listes triées sur GPU

Manseur Faiza[1], Zekri Lougmiri[1], Senouci Mohammed[1]

[1] Université d'Oran 1 Ahmed Ben Bella, Oran, Algérie
faiza_inf@hotmail.fr, lougmiri@yahoo.fr, msenouci@yahoo.fr

Résumé. Les moteurs de recherche répondent à des milliers de requêtes par seconde et demandent d'informations sur des milliards de pages web. La taille des données et les coûts d'interrogation se développent à un rythme exponentiel. Pour gérer la charge du travail, beaucoup d'algorithmes de compression et d'intersection sont proposés pour optimiser les comparaisons effectuées et les temps de calcul. Ces algorithmes sont fondés sur l'utilisation des structures de données efficaces ou des techniques de traitement améliorées comme les processeurs graphiques GPU ou les processeurs multi-cœurs. Dans cet article, nous proposons l'algorithme GTWJ (Graphical Test With Jumps) pour le calcul de l'intersection entre des ensembles ordonnés en utilisant une nouvelle structure de données et en exploitant les processeurs graphiques GPU pour profiter du parallélisme qu'ils offrent. L'idée de GTWJ est de regrouper les données dans un tableau en y assignant une clé qui permet de sauter des paquets de données si les clés comparées entre les tableaux ne correspondent pas. Des expérimentations sont réalisées sur des tableaux simulés en faisant varier la taille des tableaux ainsi que la clé de séquencement. Les résultats montrent que l'approche permet de réduire le temps de comparaison ainsi que le nombre de tests réalisés.
Mots clés: Intersection, Listes triées, Séquences, GTWJ, SVS, Naïf, GPU.

1 Introduction

Le calcul de l'intersection, d'une part entre les documents pour créer l'index et d'une autre part entre les requêtes et l'index, est une opération clé dans de nombreuses tâches de traitement de requêtes pour la recherche d'information. Dans ce contexte, beaucoup d'algorithmes sont proposés et améliorés. Ces travaux tentent de réduire le temps de calcul de l'intersection et le nombre de comparaisons, soit en accélérant les temps de calcul séquentiel en utilisant des structures de données bien adaptées, soit en exploitant le parallélisme qu'offrent les processeurs CPU multi-cœurs et les processeurs graphiques GPU. On peut facilement intégrer des cartes GPU dans les machines actuelles. Ces GPU offrent un parallélisme performant. Ainsi, nous proposons un nouvel algorithme GTWJ pour le calcul de l'intersection entre listes triées. Notre premier objectif est de réduire le nombre de tests en évitant de comparer des parties qui ne seront pas forcément partagées entre les listes en entrées. Pour atteindre cet objectif, nous modifions la structure des listes à comparer de façon à pouvoir éviter ces tests inutiles. Notre deuxième objectif est l'exploitation du parallélisme qu'offre la

carte graphique pour accélérer les temps de calcul. Nous avons implémenté notre solution avec le langage CUDA. Ce langage possède un jeu d'instruction complet et permet d'exploiter des milliers de processus (threads) simultanément. Notre solution est comparée aux solutions [1-3].

2 Algorithme Graphical Test With Jump (GTWJ)

La solution que nous proposons vise à minimiser le nombre de comparaisons effectuées autant que possible. Notre idée est de créer une nouvelle structure à partir des ensembles qu'on souhaite croiser. Dans cette structure on va diviser les tableaux initiaux en fragments appelés séquences. Chaque séquence est précédée par deux champs servants d'identifiants à celle-ci.

- **SeqID:** est l'identifiant de la séquence. C'est le quotient de la division entière de la valeur actuelle du tableau sur la valeur de séquencement.
- **Nombre de cases (NSeq):** C'est le nombre d'éléments contenus dans une séquence. Ces champs aideront à prévoir le nombre de cases à sauter pour atteindre la prochaine séquence, d'où le gain en nombre de comparaisons entre éléments des deux tableaux comparés.

Les séquences doivent être créées d'une manière bien étudiée pour profiter des avantages. La valeur de séquencement est obtenue d'une manière expérimentale et varie selon la distribution des valeurs, c'est à dire la valeur minimale et la valeur maximale dans le tableau. Chaque séquence contient toutes les valeurs comprises entre (SeqID * valeur de séquencement) et ((SeqID+1) * valeur de séquencement). A titre d'exemple, si la valeur de séquencement est égale à 100, on obtiendra des séquences 0,100, 200, 300 etc. La fig.1 présente un exemple avec une valeur de séquencement égale à 100.

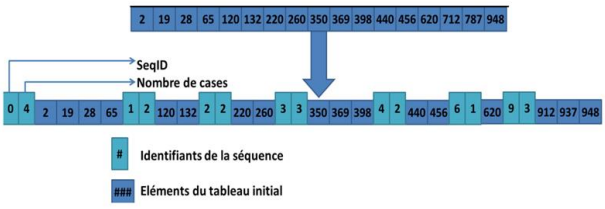


Fig. 1. Exemple de création de séquences pour un tableau d'entiers.

L'algorithme GTWJ prend en entrée deux listes triées transformées en séquences avec une valeur de séquencement bien spécifiée et retourne les éléments communs entre elles. Les éléments de l'ensemble d'intersection retournés sont eux même triés. L'algorithme GTWJ traite les séquences comme étant des éléments d'un tableau trié et procède à une recherche comme celle de l'algorithme Naïf [1] entre les séquences de chaque tableau et entre les éléments des séquences possédants le même SeqID. Nous avons implémenté l'algorithme sur l'architecture CUDA. Nous avons affecté

pour chaque séquence un thread qui va comparer les SeqID et chercher les éléments communs entre les deux séquences de même SeqID (voir fig.2). Chaque fois que les deux SeqID sont égaux, il passe à chercher l'intersection entre ces deux séquences. Si la valeur de SeqID du premier tableau est inférieure à celle du deuxième, alors il passe à la séquence suivante dans le deuxième ensemble et la compare avec la séquence actuelle, sinon il se termine. L'algorithme se termine lorsqu'il n'y a plus de séquences dans l'un des deux ensembles. L'algorithme GTWJ est présenté par le pseudo code suivant pour deux ensembles. Le résultat est enregistré dans C:

Pseudo Code Algorithme GTWJ

```
Intersection_GTWJ(entier[] SeqA, entier[] SeqB, entier[] C ){
Entier i = ThreadIdx.x ; Entier j =0, cp =0;
tant que (j < Taille_SeqB) {
   si (SeqID_Ai == SeqID_Bj) alors{
      tant que (i < NSeqA_i && j < NSeqB_j){
      si (SeqA[i] == SeqB[j]) alors{ i++; j++;C[cp++] = SeqA[i];}
      sinon si (SeqA[i] < SeqB[j]) alors i++;
              sinon j++; } }
      sinon si (SeqID_Ai > SeqID_Bj) alors j = j+ NSeqB_j;
              sinon sortir;}
renvoi C;}
```

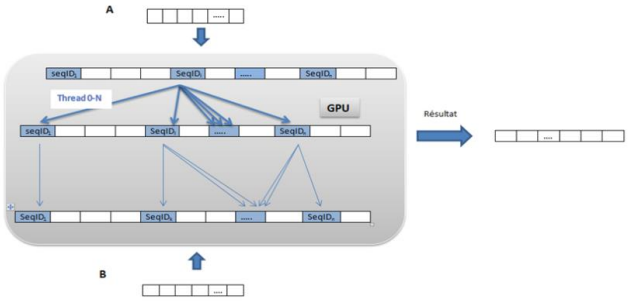


Fig. 2. GTWJ (prétraitement et calcul sur GPU)

Pour deux tableaux A et B, l'algorithme Naïf [1] exécute (tailleA+tailleB) opérations; de part la proposition de la composition des tableaux GTWJ exécuterait (tailleA+tailleB-NS) où NS est le nombre de cases évitées. Dans les expérimentations, nous avons pris les valeurs de séquencement 100, 1000, 5000 et 10000 pour GTWJ et nous avons comparé notre solution avec les algorithmes Naïf[1], SvS[2] et Inoue[3]. Les comparaisons ont été faites sur des tableaux de tailles égales d'une part, et sur des tableaux de tailles différentes d'une part. L'objectif étant d'étudier l'impact de la taille sur le comportement des algorithmes. Fig.3 donne les différents résultats en injectant des tableaux avec des tailles différentes de 1000 à 50000. Nous avons pris la taille du bloc s=4 pour Inoue [3]. Cette valeur lui produit les meilleurs résultats. Fig.3.a montre que SVS a consommé plus de temps que les autres algorithmes. Ceci est dû à la logique avec laquelle il fonctionne. Il est meilleur dans le cas où l'on veut localiser un

très petit tableau dans un autre très grand. Naïf a consommé un temps élevé vu qu'il scanne toutes les valeurs. L'inconvénient de l'algorithme Inoue réside dans l'ensemble d'instructions de test qu'il ajoute dans le corps de Naïf pour éviter les branchements. GTWJ est plus performant vu qu'il saute des parties dans les tableaux à la différence du reste des algorithmes. Fig.3.b montre que pour la valeur 100 les résultats de GTWJ sont les meilleurs alors que les comparaisons augmentent en fonction de la séquence.

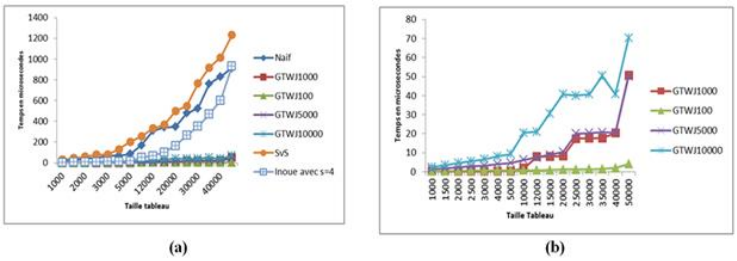


Fig. 3. Temps d'exécution pour des tableaux de tailles égales

Dans la deuxième série d'expérimentation, nous avons fixé la taille du second tableau à 50000 et nous avons varié celle du premier de 1000 à 10000. Nous avons aussi fixé la valeur de séquencement de GTWJ à 100. Le résultat de cette expérience est illustré sur Fig.4.a. On constate que SVS a consommé plus de temps comparativement aux autres algorithmes. Naïf consomme un temps élevé comparativement à Inoue. GTWJ100 a été plus rapide que les autres du fait qu'il arrive à sauter beaucoup de cases grâce à la structure proposée. Fig.4.b montre que les algorithmes SVS, Naïf et Inoue ont touché à toutes les cases des deux tableaux, ce qui augmente leurs temps d'exécution, alors que GTWJ100 a fait un nombre moindre de comparaisons induisant ainsi un temps d'exécution minimal.

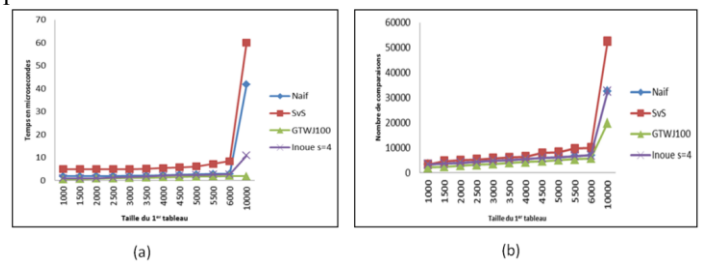


Fig. 4. Temps d'exécution et nombre de comparaisons en variant la taille du premier tableau

Référence

1. Hwang, F.K., Lin, S. Optimal merging of 2 elements with n elements. Acta Informatica. 1971
2. Demaine, E, D., Lopez-Ortiz, A, J. Ian Munro. Experiments on Adaptive set intersections for text retrieval systems. In Proceedings of the 3rd Workshop ALENEX, 91–104, 2001.
3. Inoue, Hi, Ohara, M, Taura, K,. Faster Set Intersection with SIMD instructions .by Reducing Branch Mispredictions. Proceedings of the VLDB Endowment, Vol. 8, No. 3.

Une approche hybride pour l'optimisation de la sélection des services Web composites basée sur les critères non fonctionnels.

Mohammed MERZOUG[1], Amine BRIKCI-NIGASSA[2], Amina BEKKOUCHE[3], Fethallah HADJILA[4], Abdelhak ETCHIALI[5]

[1,2,3,4,5] Département d'Informatique, Faculté des Sciences
Université Aboubakr Belkaid, Tlemcen, Algérie
[1,2,5]{mohammed.merzoug, amine.brikcinigassa, abdelhak.etchiali]@univ-tlemcen.dz,
[3,4] {ami_bekkouche, f_hadjila}@mail.univ-tlemcen.dz

Résumé. La phase de sélection est une étape clé dans le fonctionnement des architectures orientées services (SOA) à base de services Web. Devant la multiplication des services Web disponibles pour e ectuer des tâches similaires, la prise en compte des critères non fonctionnels, à savoir la qualité de service (QoS), est essentielle dans la sélection e cace de ces composants. Pour cela, nous proposons ici une approche hybride basée sur la combinaison de deux métaheuristiques. La première est issue des systèmes immunitaires arti ciels ; il s'agit de l'algorithme de sélection clonale. Nous avons complété cette métaheuristique par un deuxième algorithme inspiré de celui de l'optimisation des essaims de particules (PSO). Cette hybridation, expérimentée sur une base de test, nous a permis d'obtenir des résultats satisfaisants avec des taux d'optimalité nettement plus élevés que ceux obtenus par la sélection clonale seule, avec des performances bien plus supérieures en matière de temps d'exécution.

Mots-clés: Sélection clonale, Optimisation des essaims de particules, Service Web, Qualité de service.

1 Introduction

La sélection des instances de *services Web* basée sur les critères non fonctionnels constituant la qualité de service (*QoS*) représente un problème NP-di cile. La résolution d'un tel problème fait classiquement appel à des algorithmes issus du domaine de l'optimisation. Notre contribution consiste à proposer une solution hybride adaptant les approches de deux métaheuristiques ayant fait leurs preuves : *l'algorithme de sélection clonale* (SC) et *l'optimisation par essaim particulaire* (PSO). La première nous permet grâce à une succession de *clonages* et de *mutations*

de parvenir rapidement à une solution quasi-optimale. La seconde donne l'avantage de retrouver des solutions plus conformes (en terme de *tness*). Pour combiner leurs avantages, nous développons une méta-heuristique hybride notée SC-PSO.

2 Approche proposée

Grosso modo, l'idée de l'algorithme SC-PSO est de diviser la population en deux sous-groupes de même taille et chaque algorithme (SC ou PSO) travaillera sur son propre sous-groupe. A la n de l'itération, la meilleure solution trouvée par PSO est injectée dans le sous groupe de SC, et tout en supprimant la mauvaise solution du sous groupe de SC. Cette tâche est répétée jusqu'à la satisfaction du critère d'arrêt.

Algorithme 1 : Algorithme Hybride SC-PSO

Entrée(s): 1. la requête **R**= <n=taille-composition, BorneQoS1, BorneQoS2, BorneQoS3, BorneQoS4, BorneQoS5> avec :

- *taille-composition* : représente le nombre de tâches abstraites requises par l'utilisateur.
- *BorneQoS1...BorneQoS5* : représentent les exigences globales de l'utilisateur imposées sur les valeurs de QoS agrégées.

2. Une base de services **B** segmentée en n classes (chaque service est caractérisé par R valeurs de QoS).

Sortie(s): une composition maximisant f (notée c^*)

1: Initialiser les cellules B (anticorps) de la population avec des valeurs aléatoires $Pop = \{a_1^0, a_2^0, ..., a_n^0\}$ avec $a_i = (Ins_1, Ins_2, \ldots Ins_n)$. ▷ $(Ins_k \in \{1, 2, \ldots .nombre\_instances\})$

2: Initialiser le compteur d'itérations : $t = 1$

3: $Pop1 = \{a_1^0, a_2^0, ..., a_{|Pop|/2}^0\}$, $Pop2 = \{a_{|Pop|/2+1}^0, ..a_{|Pop|}^0\}$

4: $XGbest = Argmax(f(a_j^0))$, $a_j \in Pop1$

5: **tant que** $t \leq T\_Max$ **faire**

6: Pour chaque $a_i^t \in Pop1$: $a_i^{(t)} = modification\_aléatoire(a_i^{(t)})$, $a_i^{(t+1)} = modification\_sociale(a_i^{(t)}, XGbest)$, $sif(a_i^{(t+1)}) > f(XGbest) alors\ XGbest = a_i^{(t+1)}$

7: $Pop1 = \{a_1^{(t+1)} ... a_{|Pop|/2}^{(t+1)}\}$

8: $Worst\_B\_cell = Argmin(f(a_i^t))$, $a_i^t \in Pop2$

9: $Pop2 = Pop2 - \{Worst\_B\_cell\}$

10: $Pop2 = Pop2 \cup \{XGbest\}$

11: Calculer l'ensemble *clones_set* pour chaque $a_i^{(t)} \in Pop2$, selon le taux de clonage Tc tel que $Tc(a_i^{(t)}) = f(a_i^{(t)})/f(c*)$

12: Faire *l'hypermutation* pour chaque cellule B de *clones_set* et calculer sa nouvelle affinité $f(a_i^{\prime(t)})$ ▷ (variation des cellules clonées)

13: $Union = Pop2 \cup clones\_set$.

14: Trier $Union$ selon l'ordre décroissant des affinités.

15: Mettre à jour la population (création de $a^(t+1)_i$), en retenant les M premières cellules B de $Union$. ▷ (M est la taille de Pop2)

16: Créer L cellules B de façon aléatoire ($L = M * insertion\_rate$). ▷ (génération aléatoire de cellules B).

17: Remplacer les L dernières cellules B de $Pop2$ (en termes d'affinité) par les cellules de l'étape précédente. ▷ (remplacement des mauvaises cellules B par des cellules aléatoires).

18: $c = Argmax(f(a_i^{(t+1)})), a_i^{(t+1)} \in Pop2\,; sif(c) > f(c^*)\ alors\ c^* = c\,;$

19: t=t+1

20: **fin tant que**

21: **retourner** La meilleure cellule B de la population c^*.

3 Expérimentation

La base de test est constituée d'un corpus généré de façon aléatoire. La requête de l'utilisateur est constituée de 5 nombres qui représentent les contraintes globales. Chaque tâche abstraite possède plusieurs instances de services similaires d'un point de vue fonctionnel et di érentes selon le point de vue QoS. Le nombre d'instances varie entre 10 et 100 services. À chaque instance nous associons un vecteur de 5 paramètres de QoS (temps d'exécution, coût, abilité, disponibilité et réputation).

La gure 1 résume les performances d'optimalité de notre contribution ainsi que l'algorithme génétique, l'algorithme d'abeilles et Harmony Search [1]. Nous constatons clairement que l'algorithme SC-PSO est capable d'atteindre l'optimum global dès l'itération 500. De même la gure 2 montre la supériorité de SC-PSO (en termes de temps) par rapport aux 4 approches restantes.

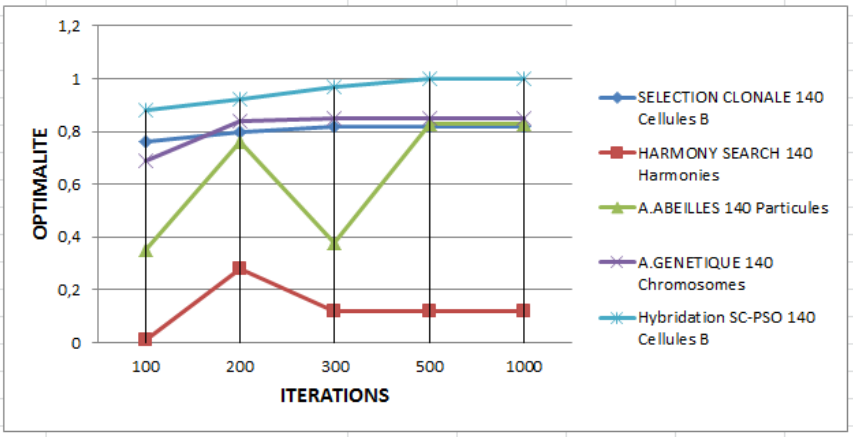


Figure 1 : Le taux d'optimalité obtenu pour les approches de sélection

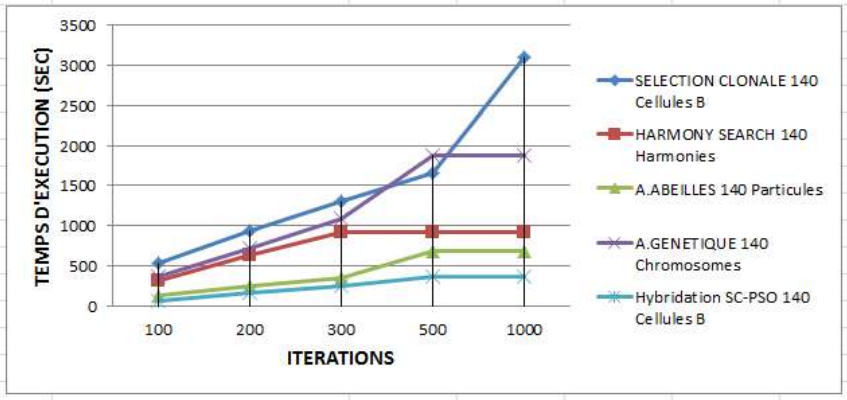


Figure 2 : Le temps d'exécution associé aux approches de sélection

4 Conclusion et perspectives

Dans ce travail, nous avons proposé une approche hybride (SC-PSO) basée sur l'algorithme de sélection clonale que nous avons adapté au problème de la sélection des services Web pour mieux trouver les solutions quasi-optimales (grâce à l'élargissement de l'espace de recherche par les opérateurs de changement aléatoire de services) et sur l'optimisation par essaim particulaire afin de combiner les avantages des deux algorithmes.

En perspective à ce travail, nous pouvons citer les pistes suivantes : la prise en compte de l'incertitude et la variabilité de la QoS pendant la sélection des services Web. Le déploiement des services Web sur les plateformes de cloud computing est de plus en plus répandu, par conséquent le problème de sélection des services Web doit être adapté à ce contexte.

Références

[1] Bekkouche, A., Benslimane, S. M., Huchard, M., Tibermacine, C., Hadjila, F., and Merzoug, M. (2017). Qos-aware optimal and automated semantic web service composition with user's constraints. Service oriented computing & applications, pp. 1-19.

[2] De Castro, L. N. and Von Zuben, F. J. (2002). Learning and optimization using the clonal selection principle. IEEE transactions on evolutionary computation, 6(3):239-251.

[3] Eberhart, R. and Kennedy, J. (1995). A new optimizer using particle swarm theory. In Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on, pages 39–43. IEEE.

Segmentation d'Images par la Méthode des K-moyennes basée sur la Multirésolution et les contraintes spatiales

M. Y. Benzian[1] and N. Benamrane[2]

[1] Département d'Informatique, Université Abou Bekr Belkaid de Tlemcen, Tlemcen, Algérie
[2] Laboratoire SIMPA, Université des Sciences et de la Technologie d'Oran, Oran, Algérie
benzian@mail.univ-tlemcen.dz, nacera.benamrane@univ-usto.dz

Abstract. Ucet article propose une nouvelle approche de segmentation d'images par les K-moyennes appliquée sur des images multirésolution et utilisant des contraintes spatiales. La classification par les k-moyennes est exécutée d'abord séparément sur plusieurs niveaux de résolution d'images à partir de l'image d'origine jusqu'au niveau de résolution bas. Ensuite, le résultat de classification de chaque pixel p dans l'image d'origine est modifié en fonction du résultat de classification k des images à basse résolution et du taux de présence de k au voisinage spatial du pixel p. L'analyse d'images à basse résolution permet d'obtenir une classification grossière qui signifie qu'un pixel est affecté à la classe à laquelle appartient la majorité des pixels de son voisinage. L'objectif de cette approche est de minimiser les erreurs de classification en fonction de la répartition spatiale des classes au voisinage de chaque pixel dans le but d'obtenir des régions plus homogènes et éliminer les zones bruitées dans l'image.

Keywords: Segmentation, K-moyennes, Multirésolution, Bruit Gaussien, Contraintes Spatiales, Classification.

1 Introduction

La méthode des k-moyennes est très utilisée en classification et segmentation d'images, du fait de sa simplicité et rapidité de mise en œuvre [5]. De nombreuses approches de segmentation utilisant la méthode des k-moyennes ont été proposées : certaines ont intégré l'analyse multirésolution [5], d'autres ont combiné les contraintes spatiales ou de voisinage avec les k-moyennes [2], l'approche dynamique et incrémentale [1], les algorithmes génétiques [4] ou la méthode soustractive [3].

2 Approche proposée

Nous avons utilisé dans notre approche plusieurs niveaux de résolution d'images. La limite maximale du niveau de résolution a été fixée à trois. Le calcul des images à basse résolution a été effectué avec le filtre de lissage de Bartlett (3 x 3) qui permet d'avoir une image à basse échelle dont la hauteur et la largeur sont divisées par deux.

$$h_{bartlett}^{3x3} \stackrel{def}{=} \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \quad (1)$$

L'algorithme classique des K-moyennes s'exécute d'abord indépendamment sur les différents niveaux de résolution d'images. Chaque image à basse échelle possède une grille de valeurs de pixels correspondant aux indices de classes auxquelles chaque pixel a été affecté. L'image à échelle 1/2 (resp. 1/4) contient par exemple une matrice de valeurs de classes de taille m/2 (resp. m/4) lignes par n/2 (resp. n/4) colonnes.

Pour obtenir une matrice de classes de taille égale à l'image d'origine (m lignes par n colonnes) à partir de chaque image à basse échelle, un calcul est effectué par propagation pyramidale à partir de la taille de l'image à basse résolution jusqu'à celle de l'image d'origine (fig. 1). Pour chaque pixel à niveau de résolution supérieur, on calcule sa classe à partir du résultat à basse résolution selon la moyenne des classes de son voisinage, par convolution avec les coefficients du filtre de Bartlett:

Etant donné (x, y) la position d'un pixel P dans l'image à basse échelle et la matrice de classe d'appartenance ***Kmeans***, le calcul de la classe d'appartenance du pixel P dans l'image à haute échelle ***KmeansHR*** dans sa nouvelle position (2.x, 2.y) est:

$$KmeansHR(2..x,2.y) = \sum_{i=-1}^{1}\sum_{j=-1}^{1} Kmeans(x+i, y+j)Bartlett(i+1, j+1) \Big/ 16 \quad (2)$$

Ensuite, nous affectons aux positions impaires horizontales et verticales de l'image à haute résolution la classe à laquelle appartiennent la majorité des classes de ses voisins directs calculés précédemment:

$$KmeansHR(2.x+1,2.y) = \begin{array}{c} class\ majeure(KmeansHR(2.x,2.y), \\ KmeansHR(2.x+1,2.y-1), KmeansHR(2.x+2,2.y)) \end{array} \quad (3)$$

$$KmeansHR(2.x,2.y+1) = \begin{array}{c} class\ majeure(KmeansHR(2.x,2.y), \\ KmeansHR(2.x-1,2y+1), KmeansHR(2x,2y+2)) \end{array} \quad (4)$$

$$KmeansHR(2.x+1,2.y+1) = class\ majeure(KmeansHR(2.x+i,2.y+j)), -1 \le i, j \le 1, i \ne j \quad (5)$$

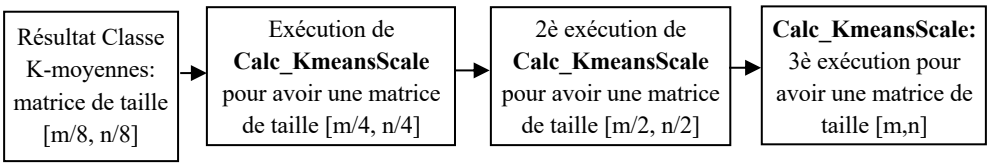


Fig. 1. Processus de propagation pyramidale de l'image à plus basse résolution (1/8^{e}) jusqu'à l'image à plus haute résolution, i.e. l'image d'origine

La dernière étape permet de changer la classe d'appartenance du pixel P de l'image d'origine en fonction des valeurs des classes de P dans les images à basse résolution, en prenant en compte la contrainte spatiale (le taux de présence de la nouvelle classe dans les pixels voisins de P dans l'image d'origine doit être supérieur ou égal à un certain seuil S). Les différents niveaux de résolution d'images sont numérotés du numéro (1) de l'image d'origine jusqu'au no (4) (image à plus basse résolution 1/8^{e}).

L'algorithme général de notre approche est récapitulé comme suit:

1) _echelle = _echelle_limite (plus bas niveau), n=_echelle
2) Exécution des K-moyennes sur les différentes résolutions de l'image séparément
3) **Pour** i **allant** de 2 à n **Faire**
4) Transformation pyramidale de la matrice de classe K du niveau de résolution i **vers** le niveau de résolution de l'image d'origine 5) **Fin Pour**
6) Modification de la classe k des pixels de l'image d'origine en fonction des classes résultats des images à basse résolution et des contraintes spatiales.

3 Résultats Expérimentaux

Notre approche a été testée sur des images médicales présentant un bruit gaussien: image CT du foie avec une lésion tumorale (fig. 2a), image du cerveau avec 3 classes: matières grise et blanche et liquide cérébrospinal (fig. 2b) et image du cerveau présentant une méningite (fig. 2c). Le niveau de résolution a été fixé à 3 pour les 2 1ères images et à 2 pour la dernière. La taille de la fenêtre de voisinage pour les contraintes spatiales est (3 x 3), le seuil S=0.3 (30%). Une comparaison entre notre approche et les K-moyennes standard (fig. 3)[5] donne de meilleurs résultats avec les K-moyennes multirésolution (fig. 4). Pour la validation des résultats, une comparaison est faite avec une classification manuelle réalisée par un expert médical. Le tableau 1 donne un taux d'erreur de classification plus réduit pour notre approche multirésolution.

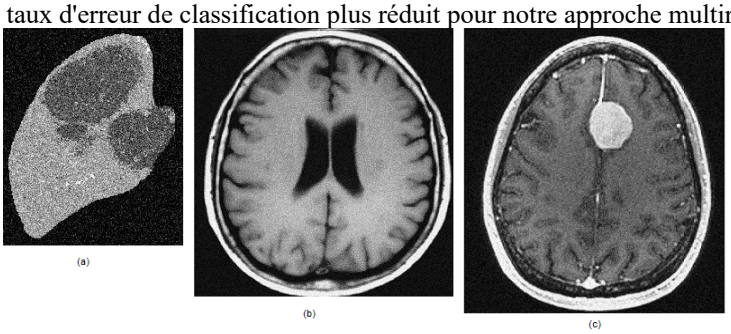


Fig. 2. Images d'origine avec bruit gaussien: foie (a), cerveau (b) cerveau avec méningite (c).

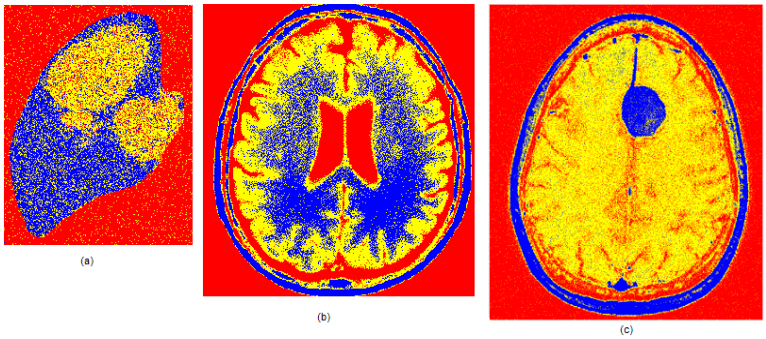


Fig. 3. Résultat de segmentation avec les k-moyennes standard.

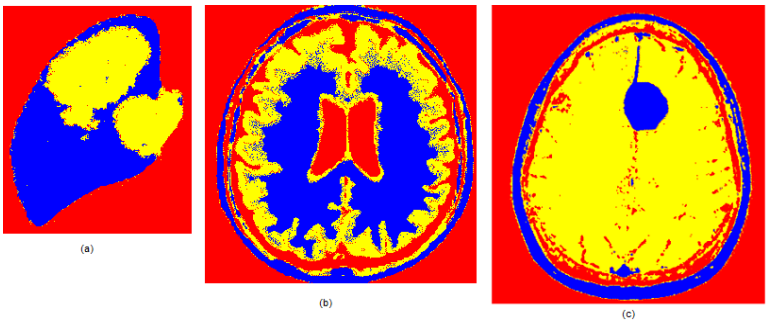


Fig. 4. Résultat de segmentation avec notre approche des k-moyennes multirésolution.

Table 1. Validation des résultats entre la segmentation manuelle et comparaions avec les K-moyennes classiques et les K-moyennes multirésolution.

Type d'image	Nombre total de pixels	Nombre de classes	Taux d'erreur K-moyennes classique	Taux d'erreur K-moyennes multirésolution
Foie CT	49600	3	19.92 %	6.38 %
Cerveau	107520	3	22.33 %	15.84 %
Méningite	460800	3	12.56 %	4.09 %

4 Conclusion

Nous avons présenté dans ce papier une nouvelle approche de segmentation d'images par les K-moyennes basée sur l'analyse multirésolution et les contraintes spatiales. L'approche multirésolution permet de réduire les fausses classes présentes dans les images et éliminer ainsi le bruit présent dans l'image. Les résultats de notre approche testés sur des images médicales sont satisfaisants et le taux d'erreur de classification a été considérablement réduit avec notre approche.

References

1. Aaron, B., Tamir, D.E., Rishe, N., Kandel, A.: Dynamic incremental K-means clustering, In: Intern. Conf. on Computational Science and Computat. Intelligence, Las Vegas (2014)
2. Chandhok, C., Chaturvedi, S., Khurshid, A..A.: An approach to image segmentation using Kmeans clustering algorithm, Intern. Jrnl of Information Technology (IJIT) 1(1), (2012)
3. Dhanachandra, N., Manglem, K., Chanu, Y.J.: Image segmentation using K-means clustering algorithm and subtractive clustering algorithm", IMCIP-2015, Procedia Computer Science 54, pp.764 – 771, (2015)
4. Li, C., Chiao, R.: Multiresolution genetic clustering algorithm for texture segmentation, Image and Vision Computing 21, 955–966 (2003)
5. Lloyd, S.P.: Least squares quantization in PCM », IEEE Transactions on Information Theory 28 (2), pp 129–137 (1982)

Vers une modélisation et implémentation de la sécurité du Dossier Médical Informatisé
(Cas d'une organisation de santé algérienne)

Belaidi Asma[1] and Abderrahim Mohammed El Amine[2]

[1] Laboratoire de Génie Biomédicale, Université de Tlemcen, Algérie
[2] Laboratoire de Génie Biomédicale, Université de Tlemcen, Algérie
belaidi.asma13@gmail.com, mea_abderrahim@mail.univ-tlemcen.dz

Résumé. La sécurité des données représente un aspect très important dans les systèmes d'information. Elle implique deux principes à savoir l'authentification qui concerne la preuve de l'identité et l'autorisation (contrôle d'accès) qui impose ce qu'il est permis et interdit de faire.

Ce travail de recherche est centré principalement sur le contrôle d'accès aux données du dossier médical dans le contexte d'une organisation de santé algérienne. En effet, nous avons proposé dans un premier temps un modèle pour la gestion du dossier médical, ce modèle, et à notre connaissance, n'a jamais été proposé auparavant d'après la littérature disponible en l'état actuelle. Ensuite, dans un second temps, nous avons proposé un modèle pour le contrôle d'accès à ce dossier en se basant sur le modèle Or-BAC (Organization Based Acces Control). La validation de cette politique de sécurité (modèle) à l'aide de la logique de premier ordre nous a permis un passage sûr vers une spécification implémentable et en conséquence le développement d'un ensemble d'outils simples et efficaces pour la prise en charge de cet aspect.

Mots-clés : Dossier Médical Informatisé, Contrôle d'Accès, Or-BAC, Prolog.

1 Introduction

Le Dossier Médical (DM) peut être défini comme un dossier qui contient toutes les informations concernant les épisodes de soins d'un patient, par exemple des informations cliniques, biologique, thérapeutique, etc. L'informatisation de ce DM permet de stocker, rechercher et manipuler l'information saisie lors des consultations des patients, il sert également à partager et échanger les données médicales. Le DM peut être structuré en quatre parties :

1) Les données administratives,
2) Les données concernant les différentes mesures chez le patient,
3) Les données concourant à la coordination, qualité, continuité des soins et prévention,
4) Les données concernant l'espace d'expression du titulaire.

Sur la base de ces données nous avons proposé dans le cadre d'un projet de master [1] un modèle en XML pour le DM. Ce modèle est disponible sous la forme d'une DTD (Document Type Definition) XML.

La sécurisation d'un tel DM constitue un enjeu essentiel pour instaurer un climat de confiance qui encourage le partage des données médicales.

Pour répondre à cette problématique nous avons développé un modèle de Contrôle d'Accès (CA) propre au DM. Ce modèle doit imposer ce qui est permit, ce qui est interdit et ce qui est obligé.

2 Modèles de contrôles d'accès

Un modèle de CA est défini par les éléments (Sujet, Contrôle, Action, Objet) et un ensemble de règles, où un sujet peut avoir une permission afin de réaliser une action sur un ou plusieurs objets. Il existe dans la littérature plusieurs modèles de CA nous citons: Le modèle discrétionnaire DAC (Discretionary Access Control) [2], le modèle de contrôle d'accès obligatoire MAC (Mandatory Access Control) [3], le modèle de contrôle d'accès basé sur les rôles RBAC (Role-Based Access Control) [4] [5], le modèle de contrôle d'accès à base de tâches TBAC (Task Based Access Control) [6], le modèle de contrôle d'accès par équipes TMAC (Team-based Access Control) [4], le modèle de contrôle d'accès à base d'organisation Or-BAC (Organization Based Acces Control) [6] [7] [8] [9].

Ces modèles peuvent être caractérisés par un ensemble de propriété, comme par exemple, l'intégrité, la confidentialité, la facilité de mise à jour, etc. Selon [9], après une étude comparative entre ces modèles, Or-BAC est le modèle le plus complet et le plus approprié dans le domaine médical. Ce modèle permet de spécifier les permissions, les interdictions, les obligations, les recommandations et il gère la notion de contexte qui n'est pas pris en charge dans les autres modèles.

Le modèle Or-BAC a été proposé pour la première fois en 2003 par A. Abou El Kalam, et all [7]. Il est centré sur le concept d'organisation [10] où une organisation peut être vue comme un groupe organisé de sujets, chacun joue un rôle spécifique et tous ses autres concepts sont définis par rapport à cette organisation. A partir des relations ternaires (habilite, utilise et considère), le modèle Or-BAC définit les relations qui existent entre les entités du niveau concret (sujets, objets, et actions) et du niveau abstrait (rôles, vues et activités) [7].

3 Proposition d'un modèle de CA pour le DM

Sur la base du modèle Or-BAC, nous avons proposé un modèle de CA pour le DM. Ce modèle reprend les concepts d'organisation, rôle, activité, vue et contexte. Nous allons décrire ces concepts.

- **Organisation**

Une organisation peut être vue comme un groupe organisé d'entités actives [7]. Dans notre cas nous avons une seule entité appelée « organisation de santé ».

- **Sujets et Rôles**

Dans notre proposition, les sujets sont des personnes. Pour les rôles, nous avons identifié plusieurs rôles : Professeur, médecin, etc.

- **Objets et Vues**

Dans notre modèle les objets sont l'ensemble des données du DM. Pour l'entité vue c'est un ensemble d'objets qui satisfait une propriété commune, par exemple

la vue " dossiers administratifs" correspond à l'ensemble des informations administratives des patients.

- **Actions et activités**
  L'entité action englobe principalement les actions informatiques comme " lire", "écrire", etc. L'entité activité représente un ensemble d'actions qui ont un objectif commun.
- **Contexte**
  Dans notre modèle nous avons identifié plusieurs contextes : Urgence, Temporel, Spatial, etc.
- **Spécification de la politique de Sécurité**
  La politique de sécurité réglemente les accès au système à travers des permissions. Par exemple : L'organisation de santé accorde au spécialiste la permission de modifier la vue rencontre dans le contexte d'urgence. Cette règle de sécurité est exprimée comme suit : *Permission(organisation de santé , spécialiste, modifier, rencontre, urgence).*

4 Formalisation du modèle de CA proposé

Plusieurs méthodes formelles permettent d'exprimer des politiques de CA. Elles se distinguent principalement par leur formalisme ou leur approche. La spécification en utilisant ces méthodes formelles, permettent de découvrir les inévitables erreurs produites lors du développement d'un logiciel le plus tôt possible sur le lieu et dans le temps même de leur production.

Pour ce faire la logique du premier ordre nous offre un cadre formelle idéal pour la description et la validation du modèle de CA que nous avons proposé. Dans ce qui suit nous allons décrire cette formalisation.

- **Les prédicats**

Les rôles, les vues et les contextes peuvent être formalisé en utilisant des prédicats. Par exemple : Pour les rôles nous avons : role(professeur), role(infirmier),etc.

- **Formalisation des règles de CA :**

Prenons par exemple la règle suivante :Le professeur (rôle) a le droit de consulter, envoyer, modifier et ajouter pour la vue « rencontre » dans le contexte urgence. Cette règle est exprimée de la façon suivante : $\forall v$ *if* $(v$=*professeur)*

$$\textit{then consulter}(\textit{professeur,rencontre,urgence}) \wedge \textit{ajouter}(\textit{professeur,rencontre,urgence}) \wedge \textit{modifier}(\textit{professeur,rencontre,urgence}) \wedge \textit{envoyer}(\textit{professeur,rencontre,urgence})\ \textit{end}$$

Pour la prise en compte des règles, nous avons donc élaboré un ensemble de règles, comme par exemple : consulter(professeur,r,s3):-role(professeur),contexte(s3).

Pour l'implémentation complète de cette spécification, nous avons utilisé le langage Prolog.

5 Conclusion

L'accessibilité aux DM dans les systèmes de santé est un aspect très important. Ce travail de recherche, porte sur la protection des données médicales et plus spécifiquement sur l'implémentation d'un modèle de CA pour le DM. Pour ce faire, nous avons proposé un modèle de DM sous la forme d'une DTD XML, ensuite et sur la base du modèle Or-BAC nous avons élaboré une spécification de la politique de CA propre à ce DM. L'utilisation du langage Prolog qui est un système à base de règles et qui repose sur la logique du premier ordre nous a permis de valider nos règles et par conséquent le modèle de CA proposé. Le déploiement du modèle de la politique de sécurité proposé dans un système d'information hospitalier réel est au centre de nos préoccupations actuelles.

References

1. Benouadah Ali, Guendoussi Nourelhouda, Belaidi Asma, Abderrahim Med El Amine; 'Conception et réalisation d'une application pour la gestion du dossier médical personnel (Etude de cas : CHU Algérien)'; Master en Génie Biomédicale option Informatique Biomédicale ; Université Abou Bekr Belkaid Tlemcen, Faculté de Technologie, Septembre 2017.
2. A.Haddad ; 'Modélisation et vérification de politique de sécurité' ; mémoire de master ; Université Joseph Fourier Genève ; 2005.
3. A. Abou El Kalam ;' Politiques de sécurité pour les systèmes d'informations médicales' ; l'Institut National Polytechnique de Toulouse.
4. D.Ferraiolo,R. Kuhn;'Role-Based Access Controls';In: 15th National Computer Security Conference; 1992.
5. A. Abou El Kalam ;'Modèles et politiques de sécurité pour les domaines de la santé et des affaires sociales' ; thèse ; l'Institut National Polytechnique de Toulouse ; 2003.
6. M. Cheaito ; 'Un cadre de spécification et de déploiement de politiques d'autorisation' ; thèse ; Université Toulouse III - Paul Sabatier ; 2012.
7. A. Abou El Kalam, R. El Baida, P. Balbiani, S. Benferhat, F. Cuppens, Y. Deswarte, A. Miège, C. Saurel et G. Trouessin; 'Un modèle de contrôle d'accès basé sur les organisations' ; 2003.
8. C. Coma, N. Cuppens-Boulahia, F.Cuppens ;'Analyse et modélisation de contrôles d'accès au système GED' ; GET/ENST Bretagne, France.
9. S.Medjdoub ; 'Modèle de contrôle d'accès pour XML : « Application à la protection des données personnelles' ;Thèse, Université de Versailles-Saint Quentin en Yvelines France ; 2005.
10. Amine Baïna ; 'Controle d'accès pour les grandes infrastructures critiques. Application au réseau d'énergie électrique' ; Thèse de doctorat en système Informatique critique ; INSA de Toulouse ; 2009.

The impact of clustering method in filter methods results

Nadjla Elong[1] and Sidi Ahmed Rahal[2]

[1] University of Sciences and Technology of Oran, Oran, Algeria
nadjla.elong@univ-usto.dz
[2] University of Sciences and Technology of Oran, Oran, Algeria
Rahalsa2001@yahoo.fr

Abstract. Filter methods are simple methods for feature selection process, but their major problem is that they don't eliminate redundant and similar features. The purpose of the paper is to study the impact of clustering on filter methods' results, by applying a well-known Clustering Algorithm, Hierarchical Agglomerative Clustering HAC, it helps to detect similarities between features in the medical datasets, then we can choose the most relevant features regarding the list of ranked features.

1 Introduction

Feature selection is a crucial step in the process of data mining and data processing due to the data growth in many fields such as: medical field, e-commerce, social media and many other fields of research that became a challenging problem for the researchers. This amount of data contains irrelevant, redundant and noisy data; thus, it stimulates a need to reduce the number of features by eliminating redundant and irrelevant data. Moreover, it will increase the classification accuracy of learning algorithms. Hence, feature selection methods are among the most popular techniques for dimensionality reduction. Feature selection approaches aim to select a small subset of features that minimize redundancy and maximize relevance to the target such as the class labels in classification[1]. Feature selection methods are presented in literature in three models: (1)Filter methods, (2) Wrapper methods and (3)Embedded methods. Filter methods are the simplest one to interpret and to implement, because their process is independent from the learning algorithm. On the other hand, wrapper methods use learning algorithm to evaluate the selected subset, which increase the computational time of feature selection process. In this article we propose a new approach for feature selection which is divided into two steps: (1)Filtering the set of features, to identify the relevant ones, then (2) use a clustering method to identify the redundant. The aim of this study is to show how clustering method can enhance filter methods results. This article is organized as follows: Section 2 Short description of filter methods, In section 3, describes the hierarchical agglomerative clustering method. In section 4,we highlight the proposed approach and we give the experimental result. Section 5 represents the conclusion.

2 Filter methods

Filter methods are feature ranking techniques that evaluate the relevance of features by looking at the intrinsic properties of the data independent of the classification algorithm[2]. A suitable ranking criterion is used to score the variables and a threshold is used to remove the variable below the threshold[2].
The general process of filter-based feature selection method [3] is given in figure (Fig. 1) below:

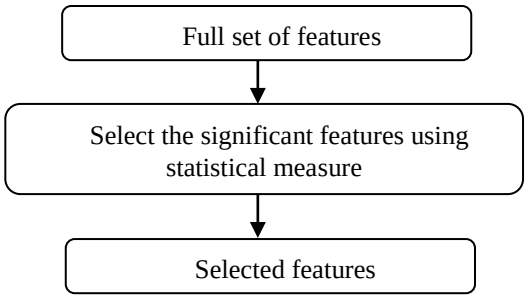


Fig. 1. Feature selection with filter approach

The filter approach (Fig. 1) selects the features without the influence of any supervised learning algorithm. Hence, it works for any classification algorithm and achieves more generality with less computational complexity than the other methods. Therefore, it is suitable for high-dimensional space [3].

3 Hierarchical Agglomerative Clustering method

Hierarchical agglomerative clustering -HAC- constructs K clusters from a larger number of smaller clusters by recursively merging the two clusters that are closest together. it starts with N clusters, each containing one case x^j. Each merge reduces the number of clusters by one [4].
The general agglomerative clustering can be summarized by the following procedure[5]:

1. Start with singleton clusters, calculate the proximity matrix for the clusters.
2. Search the minimal distance :

$$D(C_i, C_j) = \min_{\substack{1 \le m,l \le N \\ m \ne l}} D(C_m, C_l) \qquad (1)$$

where D(∗,∗) is the distance function, and combine cluster C_i and C_j to form a new cluster.

3. Update the proximity matrix by computing the distances between the new cluster and the other clusters.

4. Repeat steps 2) –3) until all objects are in the same cluster.

A fundamental step in HAC is to determine a similarity (dissimilarity) measure to identify the degree of similarity (dissimilarity) between two objects or sets.

4 Proposed approach

Our aim in this research is improving filter methods results by using a clustering method to determine similarities between features. To do so, we proposed a two-step process, which combines filter methods with clustering algorithm.
In the first step we used filter methods presented in section 2, Figure 2 represents a summarized process of this proposed approach:

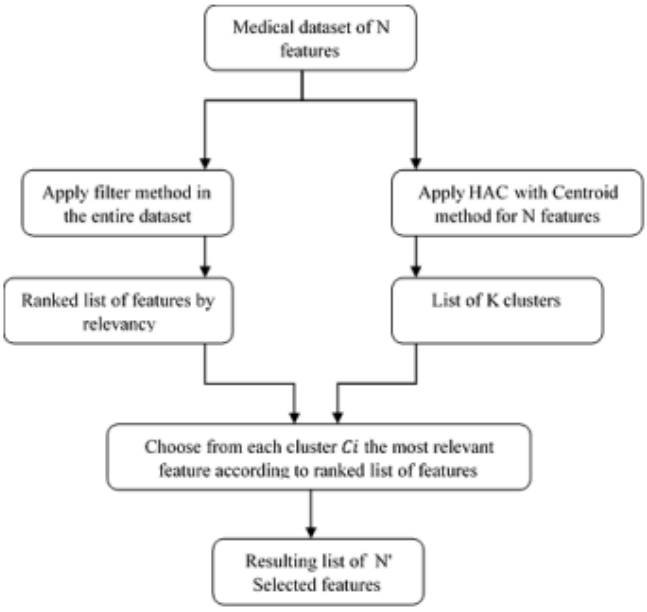


Fig. 2. General process of proposed approach

As described in figure 2, the proposed approach is divided into two steps: First step, we apply filter methods in the given medical datasets. From this step, we obtain ranked list of N features from the most relevant to the less relevant features. Second step, we apply HAC algorithm to cluster the features of the medical datasets. The result of this step is a list of K clusters, each cluster must contain the most similar features to each other.
Then, from the obtained results of these two steps, we can get the final list of N' selected features by choosing from each resulting cluster from step 2 the most relevant feature according to the ranked list of features obtained from step 1.
In the following part, we give a concrete example about the proposed approach. We applied these two steps in WDBC dataset, this dataset is obtained from UCI machine learning repository. WDBC is a dataset of 30 features, 569 instances and 2 classes.

a) The list of attributes in WDBC dataset is: 1)Radius, 2)Texture, 3)Perimeter1, 4)Area1, 5)Smoothness1, 6)Compactness1 7) Concavity1, 8) Concave_points1, 9)Symmetry1, 10)Fractal_dimension1, 11),Radius2 12)Texture2, 13) Perimeter2, 14) Area2, 15) Smoothness2, 16) Compactness2, 17) Concavity2, 18) Concave_points2, 19) Symmetry, 20) Fractal_dimension2, 21)Radius3, 22)Texture3, 23)Perimeter3, 24)Area3, 25)Smoothness3, 26)Compactness3, 27)Concavity3, 28)Concave_points3, 29)Symmetry3, 30)Fractal_dimension3.

b) We apply the first step in this list of features, this means that we apply **Filter method** with I**nformation Gain.** As a result, we obtain a ranked list of features ,from the most relevant to the less relevant one: 1)Perimeter3, 2)Area3, 3)Radius3, 4)Concave_points3, 5)concave_points1, 6)perimeter1, 7)Area1, 8)Radius1, 9)Concavity1, 10)Area2, 11)Concavity3, 12)Radius2, 13)Perimeter2, 14)Compactness3, 15)Compactness1, 16)Concavity2, 17)Concave_points2, 18)Texture3, 19)Texture1, 20)Symmetry3, 21)Compactness2, 22)Smoothness3, 23)Smoothness1, 24)Symmetry1, 25)Fractal_dimension3, 26) Fractal_dimension2, 27) Fractal_dimension1, 28)Symmetry2, 29)Texture2, 30)Smoothness2.

c) As we applied filter method on the attributes listed in (a), we apply also clustering method (HAC) on this list to obtain a list of clusters. We used in this step SPSS. The resulting list is obtained by cutting the Dendrogram in level 20. The results are shown in the following: C1={1,3,21,23,2,22,29,30}; C2={15,19,12,5,25,9,10}; C3={11,13,14,4,24,7,8}; C4={6,26,27,28}; C5={16,20,17,18}.

d) Now, based on the two resulting lists from (b) and (c), we can choose the most relevant attribute from each cluster, C1={1,3,21,23,2,22,29,30}}; C2={15,19,12,5,25,9,10}; C3={11,13,14,4,24,7,8}; C4={6,26,27,28}; C5={16,20,17,18}.

e) Finally, the resulting list of selected features is: {23,25,24,28,17}.

To validate this results, we used "Naïve Bayes" classifier and we compare the accuracy with resulting list of features by applying only filter method. The accuracy obtained by using filter method is 94,3761% and with filter method+HAC algorithm we obtained 95,2548%. From the obtained results, we observe that the proposed approach can enhance classification accuracy.

5 Conclusion

In this study, we proposed new hybrid feature selection approach based on filter methods and HAC algorithm aiming to improve filter methods results.

Experimental results show that this approach can improve the classification accuracy.

We aim in next studies to validate this approach for more datasets and with more than one method for HAC algorithm.

References

[1] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data Classification: Algorithms and Applications*, p. 37, 2014.

[2] M. W. Mwadulo, "A Review on Feature Selection Methods For Classification Tasks," *International Journal of Computer Applications Technology and Research*, vol. 5, pp. 395-402, 2016.

[3] D. Asir, S. Appavu, and E. Jebamalar, "Literature Review on Feature Selection Methods for High-Dimensional Data," *International Journal of Computer Applications*, vol. 136, pp. 9-17, 2016.

[4] M. Meila and D. Heckerman, "An experimental comparison of several clustering and initialization methods," *arXiv preprint arXiv:1301.7401*, 2013.

[5] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, pp. 645-678, 2005.

Index des auteurs