Modélisation et optimisation multi-objectif pour l'extraction interactive de connaissances

***Equipe encadrante** : Samir Loudni & Nicolas Beldiceanu (TASC)*
Thomas Yeung & Haddou benderbal hichem (SLP)

Description du projet de thèse : Objectifs, contexte, bibliographie, perspectives

Objectif et contexte scientifique :
Les algorithmes d'optimisation sont étroitement liés au contexte de données. Ces dernières années, l'interaction entre les sciences des données et l'optimisation combinatoire s'est rapidement développée [1,2]. Modéliser des problèmes de fouille de données en problèmes d'optimisation permet d'extraire des connaissances pertinentes [3]. Cette démarche est une façon de limiter le nombre de motifs et de se focaliser sur les meilleurs motifs. Elle offre également la possibilité de considérer **différents objectifs** que l'on souhaite optimiser simultanément. Cette modélisation multi-objectifs a plusieurs atouts, dont la possibilité d'adjoindre, aux critères de qualité classiques de la fouille de données, des critères qualitatifs métiers, ou encore de traiter de grandes volumétries de données. En particulier, le nombre de variables décrivant les données (caractéristiques décrivant des clients, par exemple) peut être important. Les applications sont nombreuses et se trouvent dans différents domaines. Dans le contexte de la distribution par exemple, ces approches permettent notamment d'aider au profilage des clients en identifiant des sous-groupes de clients partageant des caractéristiques communes.

En classification non supervisée (i.e. clustering), l'objectif global est de bien séparer les données différentes en regroupant les données semblables. Intrinsèquement, le problème est au moins biobjectif (maximiser l'interdissimilarité des groupes tout en maximisant leur intrasimilarité). Pour contrecarrer cette difficulté, la plupart des méthodes d'optimisation utilisées pour le clustering sont souvent des combinaisons de plusieurs objectifs, se ramenant ainsi un problème d'optimisation monocritère [4].

La prise en compte de la dimension multicritère en fouille de données est une piste très prometteuse qui mérite d'être explorée. Dans cette thèse, nous souhaitons en particulier étudier l'apport des approches multiobjectifs de type Pareto. Dès lors, une étape préalable importante consiste à déterminer l'ensemble des solutions efficaces (encore appelées non-dominées ou Pareto optimale). Déterminer l'ensemble des solutions efficaces permet d'une part de mieux appréhender les arbitrages à effectuer entre les différents critères, d'autre part d'identifier le sous-ensemble des solutions parmi lesquelles il convient à l'utilisateur de sélectionner une solution de meilleur compromis.
Cependant, le nombre de solutions efficaces peut croître exponentiellement avec la taille de certaines instances du problème. De plus, il ne s'avère pas pertinent en pratique d'énumérer exhaustivement l'ensemble efficace. C'est le cas par exemple des solutions extrêmes qui ne correspondent pas forcément à des solutions pouvant intéresser l'utilisateur. Une autre façon de définir une préférence globale en fonction des critères consiste à définir un préordre entre ces critères pour comparer les vecteurs de valeurs objectives, pour sélectionner des solutions qui maximisent ce préordre [5].

Récemment, la programmation par contraintes (PPC) a ouvert de nouvelles voies en procurant un cadre déclaratif pour l'extraction de motifs en fouille de données. Plusieurs travaux récents portant sur la fouille d'itemsets [6,7] et le clustering [8,9] ont montré l'intérêt et les apports de la PPC pour modéliser et résoudre de tels problèmes.

L'objectif de cette thèse est de **développer de nouvelles méthodes multicritères** basées sur la PPC, qui intègrent les préférences utilisateur dans le processus d'extraction de connaissances. Pour mener à bien ce projet, nous avons identifié deux axes de travail :

Axe 1 – Modélisation PPC de la dimension multicritère en fouille de donnée
Cet axe portera sur l'étude et la mise en œuvre des fonctions d'agrégation des préférences pour déterminer des solutions de bon compromis. Nous proposons d'aborder cette nouvelle problématique sous l'angle de la programmation par contraintes (PPC)[1]. Nous nous intéressons plus particulièrement à des fonctions d'agrégation collectives, par exemple la *moyenne ordonnée pondérée* (OWA) [10] ou encore *l'intégrale de Choquet* [11]. Le travail consistera à proposer des modélisations efficaces de ces fonctions d'agrégation collectives sous forme de contraintes « métiers », appelées aussi *contraintes globales.*

[1] La PPC offre un cadre unificateur pour modéliser, étudier et résoudre de nombreux problèmes combinatoires. Il a l'avantage de pouvoir représenter les propriétés que doit satisfaire une solution, sous la forme de variables et de contraintes.

Axe 2 – Conception et mise en œuvre de nouveaux algorithmes de recherche intégrant la dimension multicritère
L'objectif est cet axe est de proposer de nouveaux algorithmes permettant d'explorer efficacement l'espace des solutions Pareto ou un raffinement de ces solutions, dans le cas de fonctions d'agrégation collectives. Les techniques issues du monde de la RO pourraient être envisagées pour la partie exploration. Se pose alors la question de comment hybrider efficacement les techniques de "search" de la RO avec les propagateurs PPC dédiés au filtrage.
Un problème majeur envisagé est le passage à l'échelle. Dans la pratique, l'énumération exhaustive de l'ensemble de solutions efficaces n'est pas pertinente. Il est souvent bien plus utile d'en fournir une "bonne" représentation de taille réduite, quitte à explorer exhaustivement, dans une seconde phase, certaines zones d'intérêt. Comme axe de recherche, nous proposons de développer de nouvelles méthodes hybrides combinant recherche locale multiobjectifs et programmation par contraintes [12,13].
D'un point de vue validation, le clustering sous contraintes, où l'aspect multicritère est omniprésent, constitue une cible idéale pour valider les différentes approches qui seront développées dans cette thèse.

Impact attendu:
Le sujet se positionne sur une thématique émergente et prometteuse à l'interface entre l'intelligence artificielle, la recherche opérationnelle et l'apprentissage, thématique faisant partie des priorités nationales (cf. bilan à mi-parcours restitué en novembre 2017, de la mission IA, engagée par le gouvernement) avec comme domaine d'application l'industrie 4.0.
Les récents succès des techniques d'optimisation et d'apprentissage pour la fouille de données ont suscité un regain d'intérêt pour ce type d'approches, notamment par la création récente de **l'EURO working group on Data Science meets Optimization**. Cette thèse s'inscrit dans cette voie de recherche très prometteuse autour des techniques d'optimisation pour l'extraction et la découverte de connaissances. Elle permettra de fédérer les diverses compétences des deux équipes TASC et SLP.
Dans le contexte de l'industrie du futur (Industry 4.0), l'environnement de production devient de plus en plus intelligent et la production doit en tirer parti. Ceci est associé à l'évolution rapide des technologies d'acquisition de données générant une grande quantité (et une variété importante) de données. En conséquence, les techniques développées dans cette thèse seront un composant essentiel pour l'extraction et la classification de données dans un contexte d'optimisation de processus de fabrication allant de la conception de produits, de la génération de plans de production, jusqu'aux stratégies de maintenance et de la logistique.

Connaissances requises :
Connaissances en optimisation discrète (CP, RO), programmation Java pour intégrer les aspect contraintes dans le solveur CHOCO.

Références
[1] C. Dhaenens, L. Jourdan: Metaheuristics for data mining - Survey and opportunities for big data. 4OR 17(2): 115-139 (2019).
[2] A. G. C. Pacheco, R. A. Krohling: Aggregation of neural classifiers using Choquet integral with respect to a fuzzy measure. Neurocomputing 292: 151-164 (2018).
[3] A. Ouali, S. Loudni, Y. Lebbah, P. Boizumault, A. Zimmermann, L. Loukil: Efficiently Finding Conceptual Clustering Models with Integer Linear Programming. IJCAI 2016: 647-654
[4] J. Handl and J. Knowles. Exploiting the trade-off - the benefits of multiple objectives in data clustering. In Evolutionary Multi-Criterion Optimization (EMO 2005), volume LNCS 3410, pages 547–560. Springer-Verlag, 2005.
[5] S. Bouveret, M. Lemaître. Computing leximin-optimal solutions in constraint networks. Artificial Intelligence, 173(2): 343–364, 2009.
[6] T. Guns, S. Nijssen, L. De Raedt. Itemset mining : a constraint programming perspective. Artif. Intell., 175(12-13) :1951–1983, 2011.
[7] M.-B. Belaid, C. Bessière, and N. Lazaar. Constraint programming for association rules. In SDM, pages 127–135, 2019.
[8] Thi-Bich-Hanh Dao, Khanh-Chuong Duong, and C. Vrain. Constrained clustering by constraint programming. Artif. Intell., 244:70–94, 2017.
[9] M. Chabert, *C. Solnon. A Global Constraint for the Exact Cover Problem: Application to Conceptual Clustering. JAIR, 67:509-547 (2020).*
[10] R.R. Yager. On Ordered Weighted Averaging aggregation operators in multi-criteria decision making.IEEE Transactions on systems, Man, and Cybernetics,18(1):183–190, 1988.
[9] U. Höhle , Integration with respect to fuzzy measures, Proceedings of the IFAC Symposium on Theory and Applications of Digital Control, 1982, pp. 35–37.
[11] J. Rissanen. Modeling by shortest data description. Automatica, 14(5) :465–471, 1978.
[12] P. Schaus, R. Hartert. M*ulti-Objective Large Neighborhood Search. CP 2013: 611-627.*
[13] R. Hartert, P. Schaus. A Support-Based Algorithm for the Bi-Objective Pareto Constraint. AAAI 2014: 2674-2679.