Fouille de motifs sans seuil par optimisation multi-objectifs : Application aux règles d'association

Charles Vernerey*, Samir Loudni*, Noureddine Aribi** Yahia Lebbah**

* MT Atlantique, 44307 Nantes, France
{samir.loudni, charles.vernerey}@imt-atlantique.fr
** Université Oran1, Lab. LITIO, 31000 Oran, Algeria
{lebbah.yahia, aribi.noureddine}@univ-oran1.dz

Résumé. Cet article propose un nouveau modèle pour extraire les motifs Pareto dominants à l'aide de la programmation par contraintes. Notre modèle exploite le principe de la représentation condensée pour réduire l'espace de recherche. Nous démontrons que notre approche peut être utilisée pour découvrir des règles d'association intéressantes pour l'utilisateur sans avoir à fixer de seuil. Des expérimentations menées sur un jeu de données génomique ont démontré l'intérêt de cette approche pour l'extraction de règles d'association.

1 Introduction

L'extraction de motifs est une tâche importante en fouille de données, l'objectif étant d'extraire des motifs qui peuvent être interprétés par des experts du domaine ou utilisés comme descripteurs dans d'autres tâches comme par exemple la classification. Depuis la publication de l'article précurseur Agrawal and Srikant [1994], deux problèmes ont limité l'usage de cette approche : 1) comment fixer des seuils qui sont nécessaires dans plusieurs contraintes et 2) comment traiter des résultats qui contiennent parfois des millions de motifs. Utiliser l'approche top-k présente un inconvénient majeur du fait qu'il est difficile de choisir la valeur de k. Faire un post traitement des résultats à l'aide des représentations condensées [Pasquier *et al.*, 1999] ou la fouille d'ensemble de motifs [De Raedt and Zimmermann, 2007] ne fait que repousser le problème.

Depuis une dizaine d'années plusieurs approches ont été proposées pour découvrir des interactions plus complexes entre les motifs. Un exemple d'interaction est l'optimisation multi-objectifs (MO) où plusieurs objectifs (souvent antagonistes) doivent être optimisés simultanément . Peu d'approches ont été proposées concernant la découverte de motifs et la MO. Ghosh and Nath [2004] ont proposé une approche MO où des algorithmes génétiques ont été utilisés. van Leeuwen and Ukkonen [2013] ont proposé un algorithme pour la fouille de sous-groupes skylines. La notion de motif skyline a été exploitée dans [Ugarte *et al.*, 2017] pour extraire des motifs de haut niveau par rapport à plusieurs mesures.

Cet article s'inscrit dans la voie qui vise à connecter la programmation par contraintes (PPC) à la fouille de motifs [Guns *et al.*, 2011]. Nous proposons un nouveau modèle PPC compact et flexible pour découvrir des motifs Pareto optimaux (a.k.a. skypatterns) par rapport

à un ensemble de mesures. Notre modèle utilise le principe de représentations condensées pour réduire l'effort de fouille. Nous proposons une nouvelle contrainte globale, ADEQUATECLOSURE, pour assurer la contrainte de fermeture par rapport à plusieurs mesures. Nous montrons comment les skypatterns peuvent être utilisés pour dériver des règles d'association non redondantes de haute qualité sans avoir à fixer des seuils. Enfin, nous montrons l'intérêt de notre approche dans un cas pratique où nous cherchons des associations intéressantes sur des données génomiques. Le présent article est un résumé de l'article publié dans la conférence IJCAI 22 [Vernerey *et al.*, 2022a].

2 Préliminaires

2.1 Fouille de motifs et de règles

Soit $\mathcal{I} = \{1, ..., n\}$ un ensemble de n *items*, un motif P est un sous-ensemble non vide de $\mathcal{I}$. Le langage des motifs correpond à $\mathcal{L}_\mathcal{I} = 2^\mathcal{I} \backslash \emptyset$. Un jeu de données transactionnel $\mathcal{D}$ est un ensemble de transactions, où chaque *transaction* $t \subseteq \mathcal{I}$; $\mathcal{T} = \{1, ..., m\}$ est un ensemble de m indices de *transaction*. Un motif P *apparaît* dans une transaction t, ssi $P \subseteq t$. La *couverture* de P dans $\mathcal{D}$ est l'ensemble des transactions dans lesquelles il apparaît : $\boldsymbol{t}(P) = \{t \in \mathcal{D} \mid P \subseteq t\}$. Le *support* de P dans $\mathcal{D}$ est la taille de sa couverture : $sup(P) = |\boldsymbol{t}(P)|$. Un motif P est dit *fréquent* dans $\mathcal{D}$ si $sup(P) \geq \theta$, où θ est un seuil minimal fixé par l'utilisateur. Etant donné $T \subseteq \mathcal{D}$, $\boldsymbol{i}(T)$ est l'ensemble d'items qui sont communs à toutes les transactions de T : $\boldsymbol{i}(T) = \{i \in \mathcal{I} \mid \forall t \in T, i \in t\}$. La fermeture d'un motif P par rapport à un ensemble de mesures M [Soulet and Crémilleux, 2008] , noté $clos_M(P)$, est l'ensemble d'items tel que $clos_M(P) = \{i \in \mathcal{I} \,|\, \forall m \in M, m(P \cup \{i\}) = m(P)\}$. P est clos par rapport à M ssi $clos_M(P) = P$.

Fouille de règles. Une règle d'association est une implication $r : X \Rightarrow Y$ où X et Y sont des motifs tel que $X \cap Y = \emptyset$ et $Y \neq \emptyset$. X est appelé antécédent de la règle et Y conséquence. Le support de la règle est donné par $sup(r) = sup(X \cup Y)$. La confiance de la règle indique la probabilité qu'elle soit vraie dans la base, i.e. $conf(r) = \frac{sup(r)}{sup(X)}$. Etant donné une confiance minimale c et un support minimum θ, l'objectif est de découvrir toutes les règles r tel que $conf(r) \geq c$ et $sup(r) \geq \theta$. Le lift d'une règle r est défini par $lift(r) = \frac{conf(r) \times |\mathcal{D}|}{sup(Y)}$. Pour réduire le nombre de règles, Bastide *et al.* [2000] ont proposé la notion de *minimal non-redundant rules* (MNR). Une règle d'association $r : X \Rightarrow Y$ est une MNR ssi : (1) $sup(r) \geq \theta$ et $conf(r) \geq c$; (2) $X \cup Y$ est clos par rapport au support; et (3) X est un générateur. Un motif X est un générateur s'il n'a pas de sous-ensemble avec la même fréquence.

2.2 Optimisation Multi-Objectifs (MO)

Un problème MO $\mathcal{P}$ consiste en un ensemble de m fonctions objectifs $f_i : \mathbb{R}^n \to \mathbb{R}$ pour $i = 1..m$ et un ensemble discret $\mathcal{X}$ de solutions réalisables. Pour simplifier, nous supposons que les fonctions doivent être maximisées simultanément. On note $F : \mathbb{R}^n \to \mathbb{R}^m$, la fonction qui associe chaque solution réalisable $x \in \mathcal{X}$ au vecteur objectif correspondant $F(x) = (f_1(x), \ldots, f_m(x))$. Nous notons $\mathcal{Y}$ l'image de $\mathcal{X}$ dans l'espace objectif, s.t. $\mathcal{Y} = \{y | y = F(x),\ x \in \mathcal{X}\}$. Comparer les solutions dans $\mathcal{X}$ revient à les comparer dans l'espace

$\mathcal{Y}$. Si le décideur ne spécifie pas de préférence, on compare généralement l'ensemble des solutions à l'aide de la dominance Pareto. Soit $y, y' \in \mathcal{Y}$ deux solutions de $\mathcal{P}$. On dit que y domine y', noté $y \succ y'$, ssi : $\forall i \in [1..m] : y_i \geq y'_i$ et $\exists j \in [1..m] : y_j > y'_j$. Une solution $y^* \in \mathcal{Y}$ est Pareto optimale ssi il n'existe aucune solution $y \in \mathcal{Y}$ qui domine y^*, i.e. $\nexists y \in \mathcal{Y} : y \succ y^*$. L'ensemble des solutions Pareto optimales est appelé le front de Pareto. Une archive $\mathcal{A} \subseteq \mathcal{Y}$ est un ensemble de solutions tel qu'il n'existe aucune solution dans $\mathcal{A}$ qui domine une autre solution dans $\mathcal{A} : \nexists y, y' \in \mathcal{A} : y \succ y'$.

Notre tâche de fouille est la découverte de toutes les solutions Pareto optimales par rapport à un ensemble de mesures M. Pour un motif P donné, chaque variable du vecteur objectif $(obj_1, ..., obj_{|M|})$ représente la valeur $m(P)$ d'une mesure $m \in M$. Un motif associé au front de Pareto est appelé *skypattern*. Le problème de la fouille de skypatterns consiste à trouver l'ensemble de skypatterns $\mathcal{S}ky$ par rapport à un ensemble de mesures M. L'inconvénient majeur est que le nombre de skypatterns est exponentiel dans le pire des cas (i.e. égal au nombre total de motifs $|\mathcal{S}ky| = 2^{|\mathcal{I}|} - 1$). Pour réduire l'espace de recherche, Soulet *et al.* [2011] ont proposé la notion de *maximal skylineability*. Ce dernier consiste à trouver un ensemble de mesures M' tel que chaque skypattern par rapport à M est soit clos par rapport à M' ou est un sous-ensemble d'un pattern clos par rapport à M'. En d'autres termes, l'ensemble des skypatterns par rapport à M et clos par rapport à M' forme une représentation condensée de l'ensemble de tous les skypatterns. Par conséquent, l'espace de recherche se réduit aux motifs clos par rapport à M'. Dans cet article, nous n'expliquons pas comment obtenir M' à partir de M. Pour plus de détails, nous invitons le lecteur à consulter [Soulet *et al.*, 2011].

3 Un modèle PPC pour la fouille de skypatterns

La PPC est au coeur des approches génériques pour la fouille de motifs. Avec le développement récent des contraintes globales avec des algorithmes de filtrage efficaces, la PPC est devenue compétitive pour résoudre des problèmes de fouille de données [Belaid *et al.*, 2019; Schaus *et al.*, 2017]. Nous présentons un nouveau modèle PPC, appelé CLOSEDSKY, qui tire parti de contraintes globales afin d'obtenir un modèle efficace et compact pour la fouille de skypatterns. Pour construire ce modèle, nous avons besoin d'un jeu de données $\mathcal{D}$, un ensemble de mesures M et une archive $\mathcal{A}$, qui est dynamiquement mise à jour chaque fois que nous trouvons une nouvelle solution, étant donné que $\mathcal{A}$ ne doit pas contenir de solution dominée. Le modèle $\text{CLOSEDSKY}_{\mathcal{D},M,\mathcal{A}}(x, obj)$ est donné de la façon suivante :

$$\begin{cases} \text{PARETO}_{\mathcal{A}}(obj) & (1) \\ \text{ADEQUATECLOSURE}_{\mathcal{D},M'}(x) & (2) \\ \text{MEASURES}_{\mathcal{D},M}(x, obj) & (3) \end{cases}$$

(a) Variables. Notre modèle a : (i) n *variables Booléennes* x, où x_i représente la présence de l'item $i \in \mathcal{I}$ dans le motif. (ii) *variables objectifs* : obj est un vecteur de variables entières, s.t. obj_i représente la valeur d'une mesure $m \in M$.

(b) Contraintes. Notre modèle exploite un ensemble de contraintes qui ont des algorithmes de filtrage efficaces.

- **Pareto.** La contrainte (1) est une contrainte d'optimisation globale qui permet d'extraire les skypatterns sans avoir à ajouter des contraintes dynamiques comme dans Ugarte *et al.* [2017]. Formellement, $\text{PARETO}_{\mathcal{A}}(obj) \equiv \bigwedge_{y \in \mathcal{A}} \bigvee_{i=1..|M|} obj_i > y_i$. Elle impose que le prochain

vecteur objectif $obj = (obj_1, ..., obj_{|M|})$ n'est pas dominé par une solution de l'archive $\mathcal{A}$, i.e. $\nexists y \in \mathcal{A} : y \succ obj$. La contrainte est détaillée dans Schaus and Hartert [2013].
- **AdequateClosure.** Nous introduisons la contrainte (2) comme une nouvelle contrainte globale pour assurer que x est clos par rapport à un ensemble de mesures M'. Cela permet de découvrir des représentations condensées sans avoir à utiliser des contraintes réifiées. M' est calculé automatiquement tel que M est maximalement M'-skylineable. Les règles de filtrage de cette contrainte sont données dans la section 3.1.
- **Measures.** La contrainte (3) est utilisée afin de lier chaque variable objectif obj_i à une mesure $m \in M$ tel que $obj_i = m(x)$.

3.1 La contrainte globale ADEQUATECLOSURE

Dans cette section, nous introduisons une nouvelle contrainte globale ADEQUATECLOSURE pour la fouille de représentations condensées de motifs par rapport à un ensemble de mesures M'. Cela est possible grâce à l'opérateur de fermeture $clos_{M'}$ qui est adéquat pour un ensemble de mesures. Contrairement à Ugarte *et al.* [2017], notre contrainte globale ne nécessite pas de contraintes réifiées ou de variables additionnelles. Toutes les preuves sont données dans l'annexe supplémentaire du papier [Vernerey *et al.*, 2022b]. Nous utilisons les notations : $x^+ = \{i \in \mathcal{I} | dom(x_i) = \{1\}\}$, $x^- = \{i \in \mathcal{I} | dom(x_i) = \{0\}\}$, $x^* = \mathcal{I} \setminus \{x^+ \cup x^-\}$, où $dom(x_i)$ représente l'ensemble des valeurs autorisées pour la variable x_i.

Définition 1 (ADEQUATECLOSURE) *Soit x un vecteur de variables Booléennes, $\mathcal{D}$ un jeu de données transactionnel et M' un ensemble de mesures. La contrainte* ADEQUATECLOSURE $_{\mathcal{D},M'}(x)$ *est respectée ssi $clos_{M'}(x^+) = x^+$.*

Nous définissons à présent l'opérateur d'inclusion de fermeture $\texttt{cl}_{inc}$ exploité par les règles de filtrage de notre contrainte globale, pour la fouille de représentations condensées par rapport à un ensemble de mesures M'.

Définition 2 (Inclusion de fermeture) *Soit x une affectation partielle de variables $\{x_1, \ldots, x_{|\mathcal{I}|}\}$, M' un ensemble de mesures et i un item.* $\texttt{cl}_{inc}(x^+, i, M')$ *ssi* $i \in clos_{M'}(x^+)$.

La définition 2 fournit une condition nécessaire et suffisante pour la propriété de représentation condensée par rapport à un ensemble de mesures M', lorsqu'on étend le motif x^+ avec un item libre (appartenant à x^*). En d'autres termes, $\texttt{cl}_{inc}(x^+, i, M') \Leftrightarrow i \in clos_{M'}(x^+)$.

Le lemme 1 caractérise une affectation partielle cohérente par rapport à la contrainte ADEQUATECLOSURE, c'est à dire une affectation partielle qui peut être étendue à une solution qui respecte cette contrainte.

Lemme 1 (Affectation partielle cohérente) *Soit x^+ une affectation partielle des variables dans $\{x_1, \ldots, x_{|\mathcal{I}|}\}$ et M' un ensemble de mesures. x^+ est une affectation partielle cohérente ssi $\nexists j \in x^-$ s.t.* $\texttt{cl}_{inc}(x^+, j, M')$ *est vérifiée.*

Le propagateur que nous proposons pour ADEQUATECLOSURE est basé sur deux règles de filtrage données dans la proposition 1.

Proposition 1 (Règles de filtrage) *Etant donné une affectation partielle cohérente x, un ensemble M' de mesures, pour tout $i \in x^*$:* $(\mathbf{R_1})$ *Si* $\texttt{cl}_{inc}(x^+, i, M') \Rightarrow 0 \notin dom(x_i)$. $(\mathbf{R_2})$ *Si* $\exists\, j \in x^-$ *s.t.* $\texttt{cl}_{inc}(x^+ \cup \{i\}, j, M') \Rightarrow 1 \notin dom(x_i)$.

La première règle filtre la valeur 0 de $dom(x_i)$ si $\{i\}$ est une inclusion de fermeture x^+ (Def. 2). Cela permet d'ajouter tous les items $i \in x^*$ qui sont nécessaires pour étendre x^+ à un motif clos par rapport à M'. La seconde règle filtre la valeur 1 de $dom(x_i)$ si $\texttt{cl}_{inc}(x^+ \cup \{i\}, j, M')$ est respectée où $j \in x^-$. Cette règle vérifie si $x^+ \cup \{i\}$ ne peut pas être étendu à un motif clos par rapport à M' sans ajouter j (lemme 1).

4 Extraction de MNRs à l'aide des skypatterns

Dans la fouille de règles d'association, fixer à priori les valeurs de seuils appropriées pour c et θ (qui sont nécessaires dans les contraintes de support et de confiance) est généralement difficile. Dans cette section, nous montrons comment les skypatterns peuvent être utilisés afin d'extraire des MNRs de bonne qualité **sans avoir à spécifier de seuil**. Le processus complet est divisé en deux étapes : (i) *Générer les skypatterns* : les skypatterns sont générés à l'aide du modèle CLOSEDSKY détaillé dans la section 3, avec l'ensemble de mesures $M_r = \{sup, area, aconf\}$; (ii) *Générer les MNRs* : générer toutes les règles MNRs à partir de la collection de skypatterns représentatifs obtenus à la première étape à l'aide d'un nouveau modèle PPC.

Soit $P = X \cup Y$ un skypattern et $r : X \Rightarrow Y$ une règle. Nous avons $conf(r) \geq aconf(P)$ et $sup(r) = sup(P)$. Comme nous maximisons à la fois les mesures $aconf$ et sup lors de la génération des skypatterns, toutes les règles produites à partir de P satisfont la première condition des MNRs (i.e. $sup(r) \geq \theta$ et $conf(r) \geq c$) avec les seuils implicites $c = aconf(P)$ et $\theta = sup(P)$. Par conséquent, la génération de MNRs est **threshold-free**. De plus, notre approche permet de découvrir des règles avec un support bas mais une confiance élevée (i.e. règles *rares*).

Soit $clos_{sup}(P)$ la fermeture de P par rapport à $\{sup\}$ et $clos^-_{sup}(P) = clos_{sup}(P) \setminus P$. Pour respecter la seconde condition des MNRs, nous imposons que $X \cup Y = clos_{sup}(P)$. Cependant, comme $aconf(clos_{sup}(P)) \leq aconf(P)$, nous ajoutons la contrainte $clos^-_{sup}(P) \subseteq Y$ pour garantir que $conf(r) \geq aconf(P)$. Enfin (3^{eme} condition), nous imposons que X est un générateur à l'aide de la contrainte globale GENERATOR introduite par Belaid *et al.* [2019]. La proposition 2 résume ce résultat important. Le détail de notre modèle PPC se trouve dans [Vernerey *et al.*, 2022a].

Proposition 2 *Soit P un skypattern, $r : X \Rightarrow Y$ une règle tel que $X \cup Y = P$, et X un générateur. Soit $r' : X \Rightarrow Y'$ une règle s.t. $X \cup Y' = clos_{sup}(P) \wedge Y' = Y \cup clos^-(P)$. r' est une MNR avec $sup(r) = sup(r') \wedge conf(r') = conf(r) \geq aconf(P)$.*

5 Expérimentation

Nous rapportons un cas pratique avec une base de données génomique. Les autres expériences sur les jeux de données UCI se trouvent dans notre article [Vernerey *et al.*, 2022a]. Notre approche est implémentée à l'aide du solveur `Choco` [Prud'homme *et al.*, 2016] version 4.10.8, une librairie Java pour la PPC, le code étant disponible en ligne [Vernerey *et al.*, 2022b]. Nous avons implémenté une variante relâchée de CLOSEDSKY (notée CLOSEDSKY-WC) en désactivant la règle ($\mathbf{R_2}$), ce qui permet de réduire sa complexité théorique. Toutefois,

la qualité de filtrage diminue par rapport à CLOSEDSKY. Tout d'abord, nous avons comparé les différentes approches pour extraire les skypatterns. Comme baseline, nous avons retenu le modèle PPC CP+SKY [Ugarte *et al.*, 2017] implémenté en `Choco` et l'implémentation en C++ de la méthode spécialisée AETHERIS [Soulet *et al.*, 2011]. Ensuite, nous avons effectué une évaluation pour l'extraction de MNRs, où nous nous sommes comparés avec le modèle PPC CP4MNR [Belaid *et al.*, 2019] et la méthode spécialisée ECLAT-Z [Szathmary *et al.*, 2008]. Pour toutes les expériences, nous avons spécifié une limite de temps d'une heure.

L'objectif est d'étudier l'apport des skypatterns afin de découvrir des relations entre des expressions de gènes et des informations biologiques dans une base de données biologique. Les expériences ont été menées sur le jeu de données Eisen [1] qui contient des expressions de 2465 Yeast genes pour 79 conditions biologiques. Chaque gène a été annoté avec des IDs GO qui associent des termes dans la Yeast Gene Ontology, les PubMed IDs représentent les associations avec des papiers de recherche, les IDs des KEGG pathways dans lesquels il est impliqué, les annotations phénotypes et les noms des gènes de transcriptions régulatrices. Toutes les annotations ont été transformées en données Booléennes, qui indiquent si une annotation donnée est reliée ou non à un gène donné. Le jeu de données obtenu comporte 2465 transactions représentant des gènes et 9634 items représentant des expressions et des annotations.

Fouille de skypatterns. CLOSEDSKY-WC obtient la meilleure performance : il prend 372 secondes pour terminer l'extraction, ce qui produit 13 skypatterns, alors que les deux autres approches CP+SKY et AETHERIS dépassent la limite de temps. Notons aussi que CLOSEDSKY dépasse la limite de temps d'une heure.

Fouille de MNRs. Nous rapportons les résultats comparatifs de SKY4MNR vs CP4MNR et ECLAT-Z. Nous avons aussi comparé ces approches à GENMINER [Martinez *et al.*, 2008], un outil spécialisé pour la fouille de MNRs dans des jeux de données génomique. Il utilise l'algorithme NORDI pour discrétiser les valeurs continues et l'algorithme CLOSE [Pasquier *et al.*, 1999] pour l'extraction de MNRs. SKY4MNR est threshold-free ; pour les autres approches, nous considérons les mêmes seuils θ et c utilisés dans GENMINER, i.e. $\theta = 0.3\%$ (au moins 7 gènes) et $c = 50\%$. GENMINER arrive à trouver plus de 1.33×10^6 MNRs dans la limite de temps de 2 heures, tandis que CP4MNR extrait 364119 MNRs en 364 secondes. SKY4MNR extrait 63 MNRs en 375 seconds. Enfin, ECLAT-Z dépasse la limite de temps autorisée.

Analyse des règles. La table 1 montre trois formes de règles extraites par SKY4MNR. Les règles de la forme *annotations* $\Rightarrow$ *expressions* (règles 5-7) signifie qu'un groupe de gènes associé avec un ensemble spécifique d'annotations a une probabilité importante d'être sur ou sous-exprimé. Les règles de la forme *expressions* $\Rightarrow$ *annotations* signifient que quand un groupe de gènes est sur ou sous-exprimé, ces gènes ont les annotations associées correspondantes (règles 8-10). Par exemple, la règle 8 met en valeur un groupe de gènes associé à une structure ribosomique de type (path :03010) qui sont sous-exprimées après un choc thermique et une expérience de sporulation [Carmona-Saez *et al.*, 2006]. Enfin, les règles 1-4 révèlent des liens possibles entre des annotations de différentes sources comme le lien entre les termes KEGG pathways et Gene Ontology.

1. `i3s.unice.fr/~pasquier/web/`

Rule	Antecedent	Consequent	Supp.(#)	Conf. (%)
1	pmid :14576278	go :0005737, go :0005739	430	100
2	pr :FHL1	go :0005737, go :0005840, go :0006412	105	79
3	path :00970	go :0005737, go :0016874, go :0006412, pmid :1108023	32	100
4	go :0005739	go :0005737	532	100
5	go :0005737, go :0006412	heat3↓	109	38
6	path :03010	heat3↓	97	74
7	path :03010	heat3↓, ndt80-1↓	66	50
8	heat3↓, ndt80-1↓	path :03010	66	90
9	heat3↓	go :0005737, go :0006412	109	83
10	heat3↓	go :0005737, go :0005840, go :0006412, pr :FHL1	85	64

TAB. 1 – Exemples de MNRs générés par SKY4MNR. heat3 et ndt80 font référence aux points temporels de l'activation thermique et des expériences de sporulation, respectivement. ↓ désigne une sous-expression. Les préfixes `go`, `path`, `pmid`, `pr` identifient les termes GO, les KEGG pathways, les identifiants PubMed et les noms des transcriptions régulatrices, respectivement.

6 Conclusions

Nous avons proposé un nouveau modèle PPC compact et flexible pour extraire efficacement des skypatterns par rapport à un ensemble de mesures. Notre modèle exploite le principe de représentation condensée afin de réduire l'espace de recherche. Nous avons introduit une nouvelle contrainte globale pour garantir la fermeture d'un motif sur plusieurs mesures données. Nous avons montré comment les skypatterns pouvaient être utilisés afin d'extraire un ensemble réduit mais intéressant de MNRs sans avoir à spécifier de seuil. Une étude empirique menée sur des jeux de données UCI et de gènes ont démontré l'efficacité de notre approche pour extraire efficacement des skypatterns et des règles d'association comparée à des approches spécialisées et basées sur la PPC.

Références

R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th VLDB*, pages 487–499, San Francisco, CA, USA, 1994.

Yves Bastide, Nicolas Pasquier, Rafik Taouil, Gerd Stumme, and Lofti Lakhal. Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets. In *Computational Logic — CL 2000*, pages 972–986. Springer, 2000.

Mohamed-Bachir Belaid, Christian Bessiere, and Nadjib Lazaar. Constraint Programming for Association Rules. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*, pages 127–135, 2019.

Pedro Carmona-Saez, Monica Chagoyen, Andres Rodríguez, Oswaldo Trelles, Jose María Carazo, and Alberto Pascual-Montano. Integrated analysis of gene expression by association rules discovery. *BMC Bioinform.*, 7 :54, 2006.

Luc De Raedt and Albrecht Zimmermann. Constraint-based pattern set mining. In *7th SIAM SDM*, pages 237–248. SIAM, 2007.

Ashish Ghosh and Bhabesh Nath. Multi-objective rule mining using genetic algorithms. *Inf. Sci.*, 36(1-3) :123–133, 2004.

Tias Guns, Siegfried Nijssen, and Luc De Raedt. Itemset mining : A constraint programming perspective. *Artificial Intelligence*, 175(12) :1951–1983, 2011.

Ricardo Martinez, Nicolas Pasquier, and Claude Pasquier. GenMiner : mining non-redundant association rules from integrated gene expression data and annotations. *Bioinformatics*, 24(22) :2643–2644, 2008.

Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lofti Lakhal. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th ICDT*, pages 398–416, 1999.

C. Prud'homme, J-G. Fages, and X. Lorca. Choco Solver Documentation, 2016.

Pierre Schaus and Renaud Hartert. Multi-Objective Large Neighborhood Search. In *Proceedings of CP 2013*, 2013.

P. Schaus, J. Raoul Aoga, and T. Guns. Coversize : A global constraint for frequency-based itemset mining. In *Proceedings of the 23rd CP 2017*, pages 529–546, 2017.

Arnaud Soulet and Bruno Crémilleux. Adequate condensed representations of patterns. *Data Min. Knowl. Discov.*, 17(1) :94–110, 2008.

A. Soulet, C. Raïssi, M. Plantevit, and B. Crémilleux. Mining dominant patterns in the sky. In *Proceedings of the ICDM 2011*, pages 655–664. IEEE Computer Society, 2011.

L. Szathmary, P. Valtchev, A. Napoli, and R. Godin. An Efficient Hybrid Algorithm for Mining Frequent Closures and Generators. pages 47–58, 2008.

Willy Ugarte, Patrice Boizumault, Bruno Crémilleux, Alban Lepailleur, Samir Loudni, Marc Plantevit, Chedy Raïssi, and Arnaud Soulet. Skypattern mining : From pattern condensed representations to dynamic constraint satisfaction problems. *Artif. Intell.*, 244 :48–69, 2017.

Matthijs van Leeuwen and Antti Ukkonen. Discovering skylines of subgroup sets. In *ECML PKDD 2013*, pages 272–287, 2013.

C. Vernerey, S. Loudni, N. Aribi, and Y. Lebbah. Threshold-free pattern mining meets multi-objective optimization. In *Proceedings of IJCAI 2022*, pages 1880–1886, 2022.

C. Vernerey, S. Loudni, Y. Lebbah, and N. Aribi. Code et matériel supplémentaire. `https://gitlab.com/chaver/data-mining`, 2022. Accessed : 2022-05-16.

Summary

This paper investigates a Multi-objective Optimization approach where several functions need to be optimized at the same time. We introduce a new model for efficiently mining Pareto optimal patterns with constraint programming. Our model exploits condensed pattern representations to reduce the mining effort. We design a new global constraint for ensuring the closedness over a set of measures. We show how our approach can derive high-quality non redundant association rules without the use of thresholds whose added-value is studied on a case study related to the analysis of genes expression data.