Découverte des soft-skypatterns avec une approche PPC

Willy Ugarte*, Patrice Boizumault*
Samir Loudni*, Bruno Crémilleux*, Alban Lepailleur**

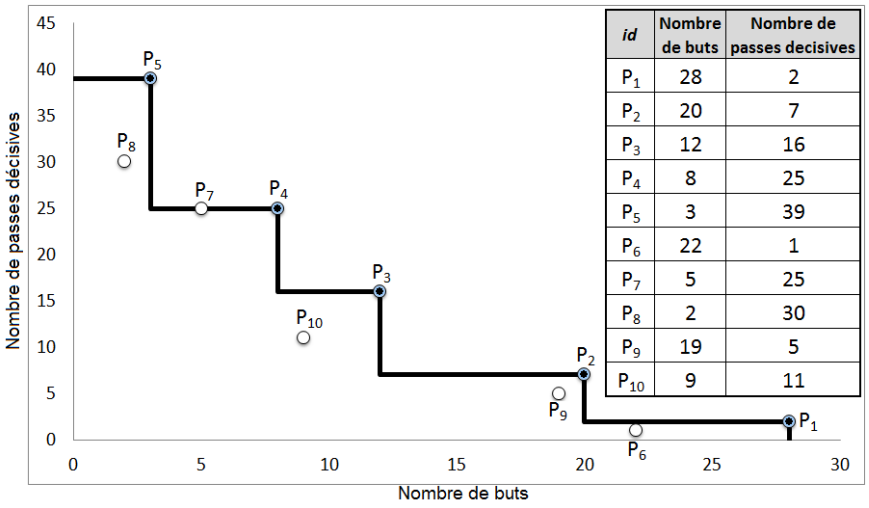
* GREYC (CNRS UMR 6072)-Université de Caen Basse-Normandie
Campus Côte de Nacre, Bd du Maréchal Juin BP 5186 - 14032 Caen CEDEX - France
** CERMN (UPRES EA 4258 - FR CNRS 3038 INC3M)
Université de Caen Basse-Normandie - Bd Becquerel, 14032 CAEN Cedex -France
{prénom.nom}@unicaen.fr

Résumé. Les skypatterns sont des motifs traduisant des préférences de l'utilisateur selon une relation de dominance. Dans cet article, nous introduisons la notion de souplesse dans la problématique des skypatterns et nous montrons comment celle-ci permet de découvrir des motifs intéressants qui seraient manqués autrement. Nous proposons une méthode efficace d'extraction de skypatterns ainsi que de *soft*-skypatterns, méthode fondée sur la programmation par contraintes. La pertinence de notre approche est illustrée à travers une étude de cas en chémoinformatique pour la découverte de toxicophores.

1 Introduction

La découverte de motifs est une tâche centrale en fouille de données et est utilisée avec succès dans un grand nombre d'applications. Une limite bien connue des processus de fouille de données est la production d'un grand nombre de motifs qu'il n'est pas possible d'examiner manuellement et parmi lesquels l'information utile est diluée. L'extraction de motifs sous contraintes permet de cibler l'information recherchée selon les centres d'intérêt de l'utilisateur. Un prolongement récent de cette voie de recherche est la prise en compte de l'intérêt d'un motif en fonction des autres motifs extraits, afin de produire des ensembles de motifs qui satisfont des propriétés sur l'ensemble des motifs considérés conjointement (Raedt et Zimmermann, 2007; Khiari et al., 2010). Notre travail se situe dans cette lignée et porte sur la notion de requêtes *skylines* (Börzsönyi et al., 2001). Notre originalité est d'introduire la souplesse dans la relation de dominance caractérisant les skylines dans le contexte de la fouille de données et de montrer l'apport de la Programmation Par Contraintes (PPC) pour cela.

La notion de skylines a été récemment étendue à la fouille de données pour extraire des *motifs skylines* (appelés *skypatterns*) (Soulet et al., 2011). Les skypatterns traduisent les préférences d'un utilisateur selon une relation de *dominance*. Dans un espace multidimensionnel où chaque dimension définit une préférence, un point p_1 domine un autre point p_2 ssi p_1 est meilleur ou égal à p_2 sur toutes les dimensions, et est strictement meilleur sur au moins une dimension. Par exemple, un utilisateur peut préférer les motifs ayant une fréquence peu élevée, une petite taille et une confiance élevée. Dans ce cas, un motif p_1 domine un autre motif p_2 ssi : $freq(p_1) \leq freq(p_2) \wedge taille(p_1) \leq taille(p_2) \wedge confiance(p_1) \geq confiance(p_2)$, où au moins une



id	Nombre de buts	Nombre de passes decisives
P_1	28	2
P_2	20	7
P_3	12	16
P_4	8	25
P_5	3	39
P_6	22	1
P_7	5	25
P_8	2	30
P_9	19	5
P_{10}	9	11

FIG. 1 – *Exemple de points skylines dans un espace à deux dimensions.*

de ces inégalités est stricte. Les *skypatterns* sont intéressants à double titre : ils permettent de s'affranchir de la notion de seuil sur les mesures et la relation de dominance exprime une forme d'intérêt global sur toutes les dimensions qui est facilement compréhensible par l'utilisateur. Néanmoins, le cadre actuel est rigide et des motifs pouvant en réalité s'avérer intéressants (les motifs dominés proches des skypatterns) ne sont pas proposés. L'exemple suivant (portant sur des points *skylines*) illustre cette situation.

Exemple. L'entraîneur d'une équipe de football souhaite recruter un ou plusieurs joueurs pour la prochaine saison. Chaque joueur sur le marché est caractérisé par le nombre de buts qu'il a marqués et le nombre de passes décisives qu'il a effectuées durant la saison actuelle (cf. figure 1). Seuls les joueurs p_1, p_2, p_3, p_4 et p_5 sont des *skylines*, les autres joueurs (i.e. p_6, p_7, p_8, p_9 et p_{10}) étant dominés par au moins un joueur skyline. Néanmoins, ces joueurs *non-skylines* peuvent s'avérer intéressants pour un recrutement :

- pour le recrutement d'un attaquant, l'entraîneur privilégierait le nombre de buts marqués. Outre les joueurs p_1 et p_2, les joueurs *non-skylines* p_6 et p_9 sont des candidats intéressants.
- pour le recrutement d'un milieu de terrain offensif, l'entraîneur privilégierait le nombre de passes décisives. Outre les joueurs p_4 et p_5, les joueurs *non-skylines* p_7 et p_8 deviennent intéressants.
- pour le recrutement d'un joueur polyvalent, l'entraîneur privilégierait plutôt le compromis entre le nombre de buts marqués et le nombre de passes décisives réalisées. Outre les joueurs p_3 et p_4, le joueur *non-skyline* p_{10} est alors un candidat intéressant.

Par ailleurs, les joueurs *skylines* sont très recherchés et ils sont coûteux parce qu'ils sont fortement sollicités : leurs salaires pourraient être hors de portée du budget du club. Les joueurs *non-skylines* proches des joueurs skylines peuvent ainsi devenir d'un grand intérêt pour l'entraîneur. Nous verrons comment de tels joueurs sont découverts grâce à l'introduction de la souplesse dans la relation de dominance.

Contributions. Nous proposons une approche nouvelle et efficace pour l'extraction de *skypat-*

terns (durs et souples) en utilisant le cadre PPC. Nous montrons comment l'extraction des (soft-)*skypatterns* peut être modélisée et résolue avec des techniques PPC. Un tel choix présente deux avantages majeurs. Tout d'abord, nous sommes en mesure d'améliorer l'étape d'extraction grâce à l'ajout de contraintes postées dynamiquement, ces contraintes étant déduites de l'ensemble des motifs candidats déjà extraits. De plus, l'aspect déclaratif du cadre PPC permet de gérer de manière unifiée plusieurs types de relaxation des skypatterns. Finalement, la pertinence de l'approche est mise en évidence par une étude de cas en chémoinformatique pour la découverte de toxicophores.

La section 2 de cet article introduit les concepts de base. La section 3 présente le problème de l'extraction de soft-skypatterns. Notre méthode d'extraction des skypatterns (durs et souples) fondée sur la PPC est détaillée en section 4. La section 5 présente un état de l'art synthétique. Les résultats de nos expérimentations sur la découverte de toxicophores sont donnés en section 6.

2 Extraction des skypatterns

2.1 Extraction de motifs sous contraintes

Soit $\mathcal{I}$ un ensemble de littéraux distincts appelés *items*. Un itemset (ou motif) est un sous-ensemble non nul de $\mathcal{I}$. La langage d'itemsets correspond à $\mathcal{L} = 2^{\mathcal{I}} \backslash \emptyset$. Une base de données transactionnelle est un multi-ensemble de motifs de $\mathcal{L}$. Une entrée de la base de données est appelée transaction et est un élément de $\mathcal{L}$. La table 1 (gauche) présente un ensemble de données transactionnelles $\mathcal{D}$ où chaque transaction t_i rassemble des articles décrits par des items notés A,...,F. L'exemple classique est une base de données de supermarchés dans laquelle chaque transaction correspond à un client et chaque item de la transaction à un produit acheté par ce client. Un prix est associé à chaque produit (cf. table 1, à droite).

La fouille de motifs sous contraintes a pour objectif d'extraire l'ensemble des motifs de $\mathcal{L}$ qui satisfont une requête (i.e., une conjonction et/ou disjonction de contraintes). La contrainte de fréquence est l'exemple le plus classique, elle conduit à extraire les motifs X_i dont le nombre d'occurrences dans $\mathcal{D}$ dépasse un seuil minimal min_{fr} fixé par l'utilisateur : $\mathit{freq}(X_i) \geq min_{fr}$. D'autres mesures quantifient l'intérêt des motifs recherchés comme par exemple la *taille* (nombre d'items composant un motif), le prix moyen (*prixMoy* : moyenne des prix associés aux items d'un motif), ou encore l'*aire* ($aire(X_i) = freq(X_i) \times taille(X_i)$). Dans de nombreuses applications, on souhaite caractériser des contrastes entre sous-ensembles de transactions, tels que par exemple en chémoinformatique des molécules toxiques versus non toxiques. A cet effet, le taux de croissance est une mesure de contraste très courante (Novak et al., 2009) que nous utiliserons en section 6. Soit $\mathcal{D}$ une base de données partitionnée en deux sous-ensembles $\mathcal{D}_1$ et

Transaction	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F

Items	A	B	C	D	E	F
Prix	30	40	10	40	70	55

TAB. 1 – *Exemple de contexte transactionnel $\mathcal{D}$.*

$\mathcal{D}_2$. Le taux de croissance d'un motif X_i de $\mathcal{D}_2$ vers $\mathcal{D}_1$ est : $m_{gr_1}(X_i) = \frac{|\mathcal{D}_2| \times freq(X_i, \mathcal{D}_1)}{|\mathcal{D}_1| \times freq(X_i, \mathcal{D}_2)}$. Par ailleurs, la communauté porte maintenant une grande attention aux *ensembles* de motifs (Raedt et Zimmermann, 2007; Khiari et al., 2010; Guns et al., 2011) où l'intérêt d'un motif dépend des autres motifs extraits. La conception de classifieurs, les top-k motifs, ou encore les skypatterns se situent dans cette lignée.

2.2 Notion de skypattern

Cette section présente la problématique de l'extraction des skypatterns. Nous commençons par définir la notion de dominance.

DÉFINITION 1 (Dominance) : *Étant donné un ensemble de mesures $M \subseteq \mathcal{M}$, un motif X_i domine un motif X_j par rapport à M, noté $X_i \succ_M X_j$, ssi $\forall m \in M$, $m(X_i) \geq m(X_j)$ et $\exists\, m \in M$ tel que $m(X_i) > m(X_j)$.*

Considérons l'exemple de la table 1 et supposons que M=$\{freq,\ aire\}$. Le motif BCD domine le motif BC car $freq(BCD)$=$freq(BC)$=5 et $aire(BCD)$>$aire(BC)$. Pour M=$\{freq, taille, prixMoy\}$, le motif BDE domine BCE car $freq(BDE)$=$freq(BCE) =$ 4, $taille(BDE)$=$taille(BCE) = 3$ et $prixMoy(BDE)$>$prixMoy(BCE)$. Étant donné un ensemble de mesures M, si un motif est dominé par un autre par rapport aux mesures de M, il est considéré comme non-intéressant et ne doit pas être dans le résultat final. Cette idée est au cœur de la notion de skypatterns.

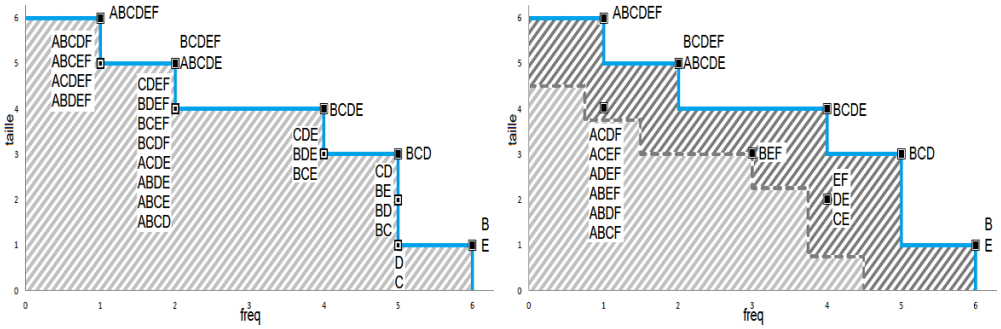
DÉFINITION 2 (Opérateur skypattern) : *Étant donnés un ensemble de motifs $P \subseteq \mathcal{L}$ et un ensemble de mesures $M \subseteq \mathcal{M}$, un skypattern de P par rapport à M est un motif non-dominé dans P par rapport à M. L'opérateur skypattern $Sky(P, M)$ renvoie tous les skypatterns de P par rapport à M : $Sky(P, M) = \{X_i \in P \mid \nexists X_j \in P, X_j \succ_M X_i\}$*

Le problème de l'extraction des skypatterns consiste à évaluer la requête $Sky(\mathcal{L}, M)$. Ainsi, pour le jeu de données de la table 1, $Sky(\mathcal{L}$,$\{freq$,$size\})$=$\{ABCDEF, BCDEF, ABCDE, BCDE, BCD, B, E\}$. Les skypatterns sont représentés à la figure 2(a). L'aire hachurée à la figure 2(a), nommée *zone interdite*, est la zone où il ne peut pas y avoir de skypatterns. La *zone de dominance* est au-dessus de la ligne bleu, cette dernière est appelée *bordure de l'aire de dominance*. Elle marque la frontière entre ces deux zones.

L'extraction des skypatterns est un problème difficile en raison du nombre très élevé de motifs candidats (i.e. $|\mathcal{L}|$) et une énumération naïve de $\mathcal{L}$ n'est pas réaliste. Par exemple, avec 1000 items une approche naïve aurait besoin de calculer $(2^{1000} - 1) \times |M|$ valeurs de mesures, et ensuite de comparer les 2^{1000} motifs entre eux. Dans (Soulet et al., 2011), les auteurs proposent une méthode efficace qui tire partie des relations théoriques entre les représentations condensées de motifs et les skypatterns, rendant ainsi le processus d'extraction faisable lorsque la représentation condensée des motifs peut être extraite. Toutefois, cette méthode est limitée à la version dure de la relation de dominance.

3 Extraction des soft-skypatterns

Cette section montre comment introduire de la souplesse dans la problématique des skypatterns. L'idée consiste à relâcher la relation de dominance afin de capturer les motifs prometteurs de la zone interdite. À cette fin, nous définissons deux types de soft-skypatterns : les



(a) edge-skypatterns. (b) δ-skypatterns avec $\delta = 0.25$.

FIG. 2 – *Soft-skypatterns extraits de l'exemple de la table 1.*

edge-skypatterns qui appartiennent à la bordure de l'aire de dominance (voir la section 3.1) et les δ-*skypatterns* qui sont proches de cette bordure (voir la section 3.2).

3.1 Edge-skypattern

Comme pour les skypatterns, les *edge-skypatterns* sont définis selon une relation de dominance et un opérateur Sky. Ces deux notions sont reformulées comme suit :

DÉFINITION 3 (Dominance stricte) : *Étant donné un ensemble de mesures $M \subseteq \mathcal{M}$, un motif X_i domine strictement un motif X_j par rapport à M, noté $X_i \gg_M X_j$, ssi $\forall m \in M, m(X_i) > m(X_j)$.*

DÉFINITION 4 (Opérateur edge-skypattern) : *Étant donnés un ensemble de motifs $P \subseteq \mathcal{L}$ et un ensemble de mesures $M \subseteq \mathcal{M}$, un edge-skypattern de P par rapport à M, est un motif non-strictement dominé dans P par rapport à M. L'opérateur $Edge\text{-}Sky(P, M)$ retourne tous les edge-skypatterns de P par rapport à M : $Edge\text{-}Sky(P, M) = \{X_i \in P \mid \nexists X_j \in P : X_j \gg_M X_i\}$*

Le problème d'extraire les edge-skypatterns consiste à évaluer la requête $Edge\text{-}Sky(\mathcal{L}, M)$. Notons que $Sky(P, M) \subseteq Edge\text{-}Sky(P, M)$. La figure 2(a) illustre les 28 = 7+(4+8+3+4+2) edge-skypatterns extraits de l'exemple de la table 1. On remarque que tous les edge-skypatterns appartiennent à la bordure de l'aire de dominance, 7 d'entre eux sont des skypatterns durs.

3.2 δ-skypattern

Comme illustré par l'exemple donné en introduction, l'utilisateur peut être intéressé par des skypatterns exprimant un compromis entre les mesures. Les δ-skypatterns expriment un tel compromis.

DÉFINITION 5 (δ-dominance) : *Étant donné un ensemble de mesures $M \subseteq \mathcal{M}$, un motif X_i δ-domine un motif X_j par rapport à M, noté par $X_i \succ^{\delta}_{M} X_j$, ssi $\forall m \in M\ (1 - \delta) \times m(X_i) \geq m(X_j)$ et $\exists\, m \in M$ tel que $m(X_i) > m(X_j)$.*

DÉFINITION 6 (δ-skypattern) : *Étant donné un ensemble de motifs $P \subseteq \mathcal{L}$ et un ensemble de mesures $M \subseteq \mathcal{M}$, un δ-skypattern de P par rapport à M est un motif non-dominé dans P par rapport à M. L'opérateur $\delta\text{-}Sky(P, M)$ retourne tous les δ-skypatterns de P par rapport à M : $\delta\text{-}Sky(P, M) = \{X_i \in P \mid \nexists X_j \in P : X_j \succ^{\delta}_{M} X_i\}$*

Le problème d'extraction des δ-skypatterns consiste à évaluer la requête δ-$Sky(P, M)$. Notons que $Sky(P, M) \subseteq \delta$-$Sky(P, M)$ (l'égalité a lieu pour δ=0) et que $Edge$-$Sky(P, M) \subseteq \delta$-$Sky(P, M)$. En appliquant l'opérateur δ-$Sky(P, M)$ sur notre exemple (cf. table 1), avec δ=0.25, nous obtenons 10 = 6+1+3 nouveaux motifs (cf. figure 2(b)), auxquels il faut ajouter les 28 edge-skypatterns (cf. section 3.1).

4 Mise en œuvre de l'extraction à l'aide de la PPC

L'extraction des skypatterns (durs ou souples) s'effectue en deux étapes. La première consiste à accroître la zone interdite par extensions successives. Pour cela, on construit une suite de motifs X_1, X_2, ..., X_n où chaque X_{i+1} améliore X_i selon au moins une des mesures, étendant ainsi la zone interdite courante. Le processus s'arrête lorsque la zone interdite ne peut plus être étendue. Soit $Cand$=$\{X_1, X_2, ..., X_n\}$ l'ensemble des points ainsi obtenus. $Cand$ constitue un sur-ensemble des skypatterns recherchés, qu'il suffit de filtrer dans une seconde étape. La mise en œuvre s'effectue de manière simple et déclarative grâce aux CSP dynamiques. L'implantation a été réalisée en `Gecode` en étendant l'extracteur de motifs (basé CSP) développé par (Khiari et al., 2010).

4.1 Exemple introductif

L'exemple de la figure 3 illustre la construction de la zone interdite pour deux mesures m_1 et m_2. Soit X_1 le premier motif extrait avec <$m_1(X_1)$=2, $m_2(X_1)$=2> ; il ne peut y avoir de skypattern dans l'aire délimitée par X_1 (zone interdite en vert à la figure 3). On recherche alors un motif X qui améliore l'ensemble des éléments de $Cand$, ici X_1, selon au moins une des mesures (i.e. tel que $(m_1(X)\geq 2)\vee(m_2(X)\geq 2)$). Soit X_2<3, 4> un tel motif (X_2 améliore X_1 pour m_1 et m_2), alors la zone bleue devient interdite. On recherche alors un motif X qui améliore X_2 selon au moins une des mesures. Soit X_3<5, 4> un tel motif (X_3 améliore X_2 pour m_1), alors la zone rouge devient interdite. On recherche alors un motif X qui améliore X_3 selon au moins une des mesures. Soit X_4<6, 3> un tel motif, alors la zone jaune devient interdite. À nouveau, on cherche un motif X qui améliore l'ensemble des éléments de $Cand$. Soit X_5<4, 5> un tel motif. Supposons désormais qu'il n'existe aucun motif améliorant X_5 pour au moins une des

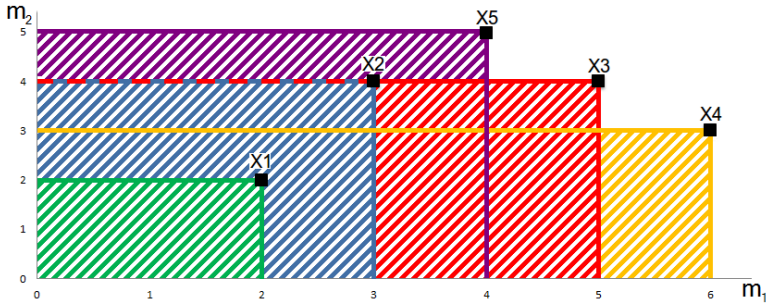


FIG. 3 – *Exemple illustratif :* $Cand$=$\{X_1, X_2, X_3, X_4, X_5\}$.

deux mesures. L'ensemble $Cand$ des motifs ainsi obtenus contient les motifs X_1, X_2, X_3, X_4 et X_5 et constitue un sur-ensemble des skypatterns. $Cand$ est alors filtré afin de ne retenir que les motifs qui sont des skypatterns.

4.2 Extraction des skypatterns à l'aide des CSP dynamiques

Un problème de satisfaction de contraintes (CSP) est défini par un triplet $(\mathcal{X}, \mathcal{D}, \mathcal{C})$ où $\mathcal{X}$ est un ensemble de variables, $\mathcal{D}$ est l'ensemble de leurs domaines finis, et $\mathcal{C}$ est un ensemble de contraintes portant sur $\mathcal{X}$. Pour l'extraction de motifs en fouille de données (Khiari et al., 2010), $\mathcal{X}$ est l'ensemble des motifs inconnus, chacun d'eux ayant $\mathcal{L}{=}2^{\mathcal{I}}\setminus\emptyset$ comme domaine. $\mathcal{C}$ peut contenir des contraintes arithmétiques sur les mesures, des contraintes ensemblistes sur les motifs, ainsi que des propriétés spécifiques que doivent vérifier certains motifs (par exemple, être un motif fermé pour une ou plusieurs mesures).

Un CSP dynamique (DCSP) est une séquence de CSP, où chaque CSP résulte de changements apportés au précédent (Verfaillie et Jussien, 2005). Ces changements peuvent affecter les variables (ajout/suppression), les domaines (ajout/suppression de valeurs), les contraintes (ajout/suppression). Dans notre approche, variables et domaines restent identiques. Seules sont ajoutées de nouvelles contraintes permettant d'agrandir (dynamiquement) la zone interdite comme l'illustre l'exemple de la figure 3. Considérons le DCSP P_1, P_2, …, P_n où chaque $P_i = (\{X\}, \mathcal{L}, q_i(X))$ avec :

- $q_1(X) = closed_M(X)$.
- $q_{i+1}(X) = q_i(X) \bigwedge(\bigvee_{m\in M} m(X) \geq m(X_i))$ où X_i est solution de la requête $q_i(X)$.

La contrainte $closed_M(X)$, qui impose que X soit un motif fermé pour les mesures de M, permet de réduire le nombre de motifs redondants. En effet, les skypatterns fermés par rapport à M constituent une représentation condensée exacte de l'ensemble des skypatterns (Soulet et al., 2011).

$Cand$ est l'ensemble des solutions successives obtenues en résolvant le DCSP $P_1, P_2, ..., P_n$. Chaque fois qu'un nouveau motif X_i (solution de la requête $q_i(X)$) est extrait, on recherche alors un nouveau motif X permettant d'agrandir la zone interdite dans au moins une de ses mesures (requête $q_{i+1}(X)$). Le processus s'arrête pour la valeur n_0 telle que la requête $q_{n_0+1}(X)$ ne possède aucune solution (on ne peut plus agrandir la zone interdite). $Cand$ est alors filtré afin de ne conserver que les motifs qui sont des skypatterns.

4.3 Extraction des soft skypatterns à l'aide des CSP dynamiques

L'extraction des δ-skypatterns s'effectue de manière analogue aux skypatterns. Le seul changement concerne les contraintes disjonctives ajoutées dynamiquement. Considérons le DCSP $P'_1, P'_2, ..., P'_n$ où chaque $P'_i = (\{X\}, \mathcal{L}, q'_i(X))$ avec :

- $q'_1(X) = closed_M(X)$.
- $q'_{i+1}(X) = q'_i(X) \bigwedge(\bigvee_{m\in M} m(X) \geq (1-\delta) \times m(X_i))$ où X_i est solution de $q'_i(X)$

Comme précédemment, $Cand$ est l'ensemble des solutions successives obtenues en résolvant le DCSP $P'_1, P'_2, ..., P'_n$. Chaque fois qu'un nouveau motif X_i est extrait, on recherche alors un nouveau motif X permettant d'agrandir la zone interdite. Le processus s'arrête lorsqu'on ne peut plus agrandir cette zone et $Cand$ est alors filtré pour ne contenir que les motifs qui sont des δ-skypatterns.

5 État de l'art

Le calcul des skylines dans les bases de données s'apparente à la recherche d'un vecteur maximal en géométrie algorithmique (Matousek, 1991), au calcul de la frontière Pareto (Kung et al., 1975) et à la recherche d'une solution optimale en optimisation multi-critères (Steuer, 1992). (Jin et al., 2004) ont introduit la notion de thick skylines dans les bases de données. Un thick skyline est soit un skyline P, soit un point P' dominé par un skyline P et tel que P' soit proche de P (leur distance est inférieure à un seuil ϵ). La notion de δ-skypattern que nous proposons est une extension des thick skylines dans le cadre de l'extraction de motifs.

La notion de skypattern, ainsi que leur calcul à partir de représentations condensées selon différentes mesures, a été introduite par (Soulet et al., 2011). (Gavanelli, 2002) a proposé une première méthode fondée sur la PPC pour déterminer la frontière Pareto. Cette technique est proche de notre méthode de résolution mais elle est développée uniquement pour les skylines durs portant sur les n-uplets de la base.

6 Expérimentations

La toxicologie est la science qui étudie les effets toxiques des substances chimiques sur les organismes vivants. Un défi majeur en chémoinformatique est d'identifier les caractéristiques des structures chimiques liées à une activité toxique donnée (e.g. $CL50$ [1].). Les fragments [2] moléculaires responsables des propriétés toxiques d'une substance chimique sont appelés *toxicophores* et leur découverte est bien souvent à la base des modèles de prédiction en (éco)toxicité (Verhaar et al., 1992; Benigni et Bossa, 2011). L'objectif de la présente étude menée en collaboration avec le laboratoire CERMN [3], est d'étudier l'apport de soft-skypatterns pour la découverte semi-automatique de toxicophores.

6.1 Protocole expérimental

Nous avons utilisé un jeu de données issu du site de l'ECHA [4]. Pour évaluer le potentiel toxique d'une substance, des indicateurs de toxicité également appelés *endpoints* permettent de quantifier leur impact sur la santé humaine et sur l'environnement. Nous avons utilisé l'indicateur quantitatif de toxicité vis à vis des organismes aquatiques $CL50$. En fonction de la valeur de cet indicateur, les molécules sont réparties dans les trois catégories suivantes : $H400$ très toxique ($CL50 \leq 1$ mg/L), $H401$ toxique (1 mg/L $< CL50 \leq 10$ mg/L), et $H402$ nocif (10 mg/L $< CL50 \leq 100$ mg/L). Dans cette étude, nous nous concentrons uniquement sur les classes $H400$ et $H402$ afin de maximiser le contraste entre classes. Le jeu de données $\mathcal{D}$ utilisé contient 567 molécules, 372 de la classe $H400$ et 195 de la classe $H402$. Les molécules sont représentées en utilisant 1450 sous-graphes fréquents fermés, initialement extraits de $\mathcal{D}$ [5] avec un seuil de fréquence de 1%.

1. Concentration nécessaire d'une substance pour causer la mort de 50% d'une population (poisson, daphnie ou algue) dans des conditions expérimentales précises.
2. Un fragment désigne une partie connectée d'une structure chimique contenant au moins une liaison chimique.
3. Centre d'Etudes et de Recherche sur le Médicament de Normandie UPRES EA 4258 - FR CNRS 3038 INC3M
4. European Chemicals Agency : `http://echa.europa.eu/`
5. Une substance chimique Ch contient un élément A si Ch supporte A, et A est un sous-graphe fréquent de $\mathcal{D}$.

	Skypattern		Edge-skypattern		δ-skypattern			
					δ			
					0.1		0.2	
$\mathcal{M}$	$\lvert Cand \rvert$	# of sol.	$\lvert Cand \rvert$	# of sol.	$\lvert Cand \rvert$	# of sol.	$\lvert Cand \rvert$	# of sol.
{émergence, fréquence}	2259	8	2259	24	19468	25	21710	80
	28m :42s				32m :22s		44m :50s	
{émergence, aromaticité}	6522	5	6522	76	16235	181	18543	1027
	25m :59s				38m :02s		2h :23m :04s	
{fréquence, aromaticité}	10954	2	10954	72	27836	181	30583	1011
	26m :33s				35m :59s		2h :30m :32s	
{émergence, fréquence, aromaticité}	23887	21	23887	144	32322	223	33744	1055
	40m :30s				1h :4m :27s		4h :35m :27s	

TAB. 2 – *Analyse de la performance de l'extraction de soft-skypattern.*

Afin de découvrir des toxicophores candidats, nous combinons des mesures de contraste très utilisées en fouille (Novak et al., 2009) et caractérisant les motifs du point de vue de leur présence dans les données telles que le taux de croissance et la fréquence avec des mesures exprimant des connaissances chimiques telles que l'aromaticité d'une molécule qui est un indicateur connu de la toxicité. Notre méthode offre un moyen naturel de combiner simultanément dans un même cadre ces mesures provenant de différentes origines. Nous présentons maintenant ces mesures.
Taux de croissance. Cette mesure permet de caractériser une molécule d'une classe (toxique) par rapport à une autre classe (non-toxique). Une substance chimique qui possède dans sa structure des fragments moléculaires d'un motif avec une valeur élevée du taux de croissance est particulièrement susceptible d'être toxique.
Fréquence. Les motifs de faible fréquence sont souvent dus à des artefacts dans les données et constituent du bruit. Afin d'assurer une représentativité de l'information extraite, nous imposons une contrainte de fréquence minimale.
Aromaticité. L'intérêt de cette mesure est qu'elle véhicule une hypothèse toxicophore : plus la valeur d'aromaticité est forte, plus la molécule possédant ces fragments moléculaires tend à être toxique puisque ses métabolites peuvent mener à des espèces réactives pouvant interagir de manière néfaste avec des biomacromolécules. L'aromaticité d'un motif est calculée en utilisant la fonction $mean = \frac{min+max}{2}$ des valeurs d'aromaticité de ses fragments moléculaires.
Enfin, la redondance est réduite à l'aide des **skypatterns fermés** qui sont une représentation exacte condensée de l'ensemble des skypatterns pour ces trois mesures (cf. section 4.2).

Nous avons réalisé plusieurs expérimentations avec différentes combinaisons de ces trois mesures. Pour le paramètre δ, nous avons considéré deux valeurs : 10% et 20%. Une analyse qualitative des soft-skypatterns extraits effectuée par les chimistes permet l'identification de toxicophores connus. Toutes nos expérimentations ont été réalisées sur un processeur Intel core i3 à 2,13 GHz ayant 4 Go de RAM. Nous analysons d'abord les performances de notre approche sur un plan quantitatif. Ensuite, nous procédons à une analyse qualitative de résultats.

6.2 Analyse des performances et de la qualité des skypatterns extraits

La table 2 compare les performances des trois opérateurs aussi bien en termes de nombre de skypatterns extraits qu'en temps de calcul, pour différentes combinaisons de mesures.

L'augmentation du nombre de mesures conduit à un plus grand nombre de soft-skypatterns extraits. En effet, un motif domine rarement tous les autres motifs sur l'ensemble des mesures. Néanmoins, dans nos expérimentations, ce nombre reste raisonnablement réduit, au plus, il y a un maximum de 1052 δ-skypatterns. Par ailleurs, les temps de calcul indiquent que notre approche

est efficace (moins de 1 heure), même avec l'augmentation du nombre de mesures (sauf pour $\delta = 0, 2$, où le nombre de δ-skypatterns et le temps de calcul augmentent sensiblement). De ces résultats, nous dressons le bilan suivant.

Tout d'abord, en utilisant le taux de croissance et la fréquence, seulement huit skypatterns ont été détectés mais trois toxicophores bien connus ont été retrouvés. Deux d'entre eux sont des *composés aromatiques* : le chlorobenzène (code smiles [6] : {`Clc`}) et le phénol {`c1(ccccc1)O`}). La contamination de l'eau et des sols par les produits chimiques aromatiques est en effet très répandue puisqu'on les retrouve dans de nombreux produits industriels tels que des pesticides, des solvants, des colorants et même des explosifs. Beaucoup de ces produits peuvent s'accumuler dans la chaîne alimentaire et ont un effet néfaste aussi bien sur les plantes que sur les animaux et l'homme. Le troisième toxicophore mis en évidence correspond au motif *organophosphate* ({`OP`, `OP=S`}) qui figure également dans de nombreux pesticides. Concernant les soft skypatterns, aucune information supplémentaire n'est venue enrichir notre liste de toxicophores dans ce premier cas. On peut cependant souligner une énumération plus précise des cycles aromatiques substitués par un atome de chlore (ex. {`Clc(ccc)c`, `Clcccc`}) et des organophosphates (ex. {`OP(=S)O)`, `COP(=S)O`}).

Ensuite, en considérant le taux de croissance et l'aromaticité, ou la fréquence et l'aromaticité, les résultats sont relativement similaires. Bien que la liste des toxicophores détectés grâce aux skypatterns soit limitée en comparaison de celle obtenue précédemment (taux de croissance et fréquence), l'extraction des soft-skypatterns permet d'identifier de façon beaucoup plus exhaustive des cycles aromatiques de différentes natures. En effet, cette nature peut varier en fonction i) de la présence/absence d'hétéroatomes (`N`, `S`, `O`), ii) du nombre de cycles, iii) de la présence/absence de substituants. Les edge-skypatterns permettent ainsi d'extraire des *composés aromatiques azotés* (ex. indole {`ncc`, `c1ccccc1`}, benzoimidazole {`ncnc`, `c1ccccc1`}), des *composés aromatiques soufrés* (ex. benzothiophene {`scc`, `c1ccccc1`}), des *composés aromatiques oxygénés* (ex. benzofurane {`coc`, `c1ccccc1`}) et des *composés aromatiquess hydrocarbonés* (ex. naphthalène {`c1ccc2ccccc2c1`}). Les δ-skypatterns viennent compléter cette liste avec de nouveaux systèmes *polycycliques* (ex. biphényle {`c1ccccc1c2ccccc2`}). Il est également important de noter que dans ce cas, les δ-skypatterns permettent de détecter un autre type de toxicophore très néfaste pour les organismes aquatiques, à savoir les *amines aromatiques* (ex. aniline {`c1(ccccc1)N`}).

Enfin, les meilleurs résultats ont été obtenus en considérant simultanément trois mesures (taux de croissance, fréquence et aromaticité). En effet le phénol, le chlorobenzène et le motif organophosphate sont détectés dès l'énumération des skypatterns. De plus, une information concernant les cycles aromatiques azotés est également extraite. Les edge et δ-skypatterns permettent ensuite de retrouver l'ensemble des cycles aromatiques plus "exotiques" discutés précédemment. De plus, les edge-skypatterns permettent d'identifier de façon beaucoup plus efficace les organophosphates (ex. {`COP(=S)O`, `O(P(OC)=S)C`, `O(CC)P=S`}).

La figure 4 montre la répartition des skypatterns (durs et souples) pour les trois mesures considérées. Les skypatterns durs sont situés dans différentes régions (cf. les patterns p_1 et p_2 et ceux inclus dans les quatre ellipses e_1, e_2, e_3 et e_4). Le motif p_1 correspond à un cycle aromatique substitué par un atome de chlore, alors que p_2 désigne un motif *organophosphate*. Les autres skypatterns inclus dans e_1, e_2, e_3 et e_4 correspondent respectivement aux *cycles aromatiques azotés* (ex. {`nc`}), aux *cycles aromatiques alkylés* (ex. {`cC`}), au chlorobenzène et au phénol.

6. Code smiles est une notation en ligne pour décrire la structure des molécules chimiques : `http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html`

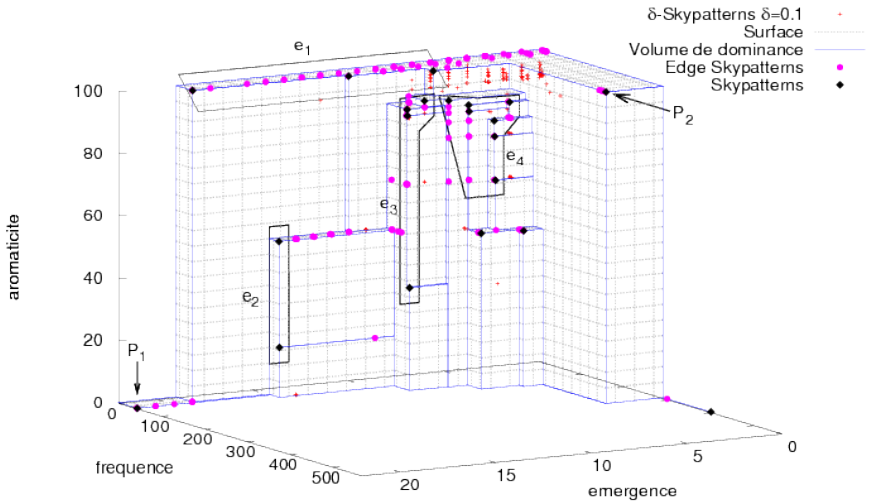


FIG. 4 – *Répartition des (soft-)skypatterns pour M={émergence,fréquence,aromaticité}.*

Concernant les soft-skypatterns, la plupart des edge-skypatterns sont situés sur la bordure du volume de dominance correspondants aux motifs inclus dans e_1. Ces motifs complètent la liste des cycles aromatiques qui n'ont pas été trouvés lors de l'extraction des skypatterns durs, tels que les *cycles aromatiques soufrés* (ex. {cs}) et le biphényle. Enfin, pour les δ-skypatterns, les plus informatifs sont ceux qui se trouvent autour des motifs appartenant à e_1 (ex. naphtalène et aniline).

7 Conclusion

Dans cet article, nous avons proposé d'introduire de la souplesse dans la problématique des skypatterns. Nous avons montré comment celle-ci permet de découvrir des motifs intéressants qui seraient manqués autrement. En s'appuyant sur la PPC, nous avons proposé une méthode efficace pour l'extraction des skypatterns et des soft-skypatterns. L'apport et l'efficacité de notre approche sont mises en évidence par une étude de cas en chémoinformatique portant sur la découverte de toxicophores. Les résultats expérimentaux montrent l'intérêt d'utiliser les soft-skypatterns afin d'obtenir de nouvelles connaissances en chimie. Dans le futur, nous voulons étudier l'introduction de la souplesse sur de nouvelles tâches comme le clustering, l'apport des soft-skypatterns pour la recommandation et d'étendre notre approche aux skycubes.

Remerciements. Ce travail a été soutenu par l'Agence Nationale de la Recherche, référence ANR-10-BLA-0214. Nous voulons aussi remercier Bertrand Cuissart, Guillaume Poezevara et Ronan Bureau pour les discussions stimulantes et fructueuses à ce sujet.

Références

Benigni, R. et C. Bossa (2011). Mechanisms of chemical carcinogenicity and mutagenicity : A review with implications for predictive toxicology. *Chemical Revue 111*, 2507–2536.

Börzsönyi, S., D. Kossmann, et K. Stocker (2001). The skyline operator. In *Proceedings of the 17th International Conference on Data Engineering*, pp. 421–430. IEEE Computer Society.

Gavanelli, M. (2002). An algorithm for multi-criteria optimization in csps. In F. van Harmelen (Ed.), *ECAI*, pp. 136–140. IOS Press.

Guns, T., S. Nijssen, et L. D. Raedt (2011). Itemset mining : A constraint programming perspective. *Artif. Intell. 175*(12-13), 1951–1983.

Jin, W., J. Han, et M. Ester (2004). Mining thick skylines over large databases. In *PKDD'04*, pp. 255–266.

Khiari, M., P. Boizumault, et B. Crémilleux (2010). Constraint programming for mining n-ary patterns. In *CP'10*, Volume 6308 of *LNCS*, pp. 552–567. Springer.

Kung, H. T., F. Luccio, et F. P. Preparata (1975). On finding the maxima of a set of vectors. *Journal of ACM 22*(4), 469–476.

Matousek, J. (1991). Computing dominances in e. *Inf. Process. Lett. 38*(5), 277–278.

Novak, P. K., N. Lavrač, et G. I. Webb (2009). Supervised descriptive rule discovery : A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research 10*, 377–403.

Raedt, L. D. et A. Zimmermann (2007). Constraint-based pattern set mining. In *Proceedings of the Seventh SIAM International Conference on Data Mining*, pp. 237–248.

Soulet, A., C. Raïssi, M. Plantevit, et B. Crémilleux (2011). Mining dominant patterns in the sky. In *11th IEEE Int. Conf. on Data Mining series (ICDM 2011)*, pp. 655–664.

Steuer, R. E. (1992). *Multiple Criteria Optimization : Theory, Computation and Application*. Radio e Svyaz, Moscow, 504 pp. (in Russian).

Verfaillie, G. et N. Jussien (2005). Constraint solving in uncertain and dynamic environments : A survey. *Constraints 10*(3), 253–281.

Verhaar, H. J. M., C. J. van Leeuwen, et J. L. M. Hermens (1992). Classifying environmental pollutants. 1 : Structure - activity relationships for prediction of aquatic toxicity. *Chemosphere 25*, 471–491.

Summary

Within the pattern mining area, skypatterns enable to express a user-preference point of view according to the domination relation. In this paper, we deal with the introduction of softness in the skypattern mining problem. We show how softness can provide convenient patterns that would be missed otherwise. Then, thanks to Constraint Programming, we propose an efficient method to mine skypatterns as well as soft ones. Finally, we show the relevance of our approach through a case study in chemoinformatics for discovering toxicophores.