Calcul du cube de skypatterns

Willy Ugarte[1] **Patrice Boizumault**[1] **Samir Loudni**[1] **Bruno Crémilleux**[1]

[1] GREYC (CNRS UMR 6072),
Université de Caen Basse-Normandie, 14032 CAEN
{prénom.nom}@unicaen.fr

Abstract

In [5], we introduce skypattern cubes and propose an efficient bottom-up approach to compute them. Our approach relies on derivation rules collecting skypatterns of a parent node from its child nodes without any dominance test. Non-derivable skypatterns are computed on the fly thanks to Dynamic CSP. The bottom-up principle enables to provide a concise representation of the cube based on skypattern equivalence classes without any supplementary effort. Experiments on mutagenicity datasets show the effectiveness of our proposal.

1. Introduction

La notion de requêtes *skyline* [1] a été récemment intégrée dans la découverte de motifs pour extraire des motifs appelés *skypatterns* [4, 6]. Les skypatterns sont des motifs basés sur la notion de Pareto-dominance pour lesquels aucune mesure ne peut être améliorée sans en dégrader au moins une autre. De tels motifs sont intéressants car ils n'obligent pas à fixer de seuil sur les mesures et possèdent un très fort intérêt global.

Dans la pratique, l'utilisateur ne connait pas a priori le rôle exact de chaque mesure, et ne peut déterminer à l'avance le sous-ensemble le plus approprié de mesures. De façon similaire au cube de skylines [3], l'utilisateur aimerait disposer du *cube de skypatterns*. Chaque élément du cube est un nœud qui associe, à un sous-ensemble des mesures, son ensemble de skypatterns. De plus, l'utilisateur peut facilement repérer les sous-ensembles de mesures ayant le même ensemble de skypatterns (qui forment une classe d'équivalence).

2. Contexte et définitions

Soit $\mathcal{I}$ un ensemble de littéraux appelés *items*. Un motif est un sous-ensemble non-vide de $\mathcal{I}$. Le langage d'itemsets correspond à $\mathcal{L}_\mathcal{I} = 2^\mathcal{I}\backslash\emptyset$. Un jeu de données est un multiset de motifs appelées transactions. La Figure 1a représente un jeu de données $\boldsymbol{r}$ où chaque transaction t_i est décrite par les items notés $A, \ldots, F$.

Exemple 1. Pour le jeu de données de Fig. 1a, on a $\texttt{freq}(BC)$=5, $\texttt{area}(BC)$=10 et $\texttt{mean}(BCD.price)$=25, avec les mesures suivantes définies par :

- $\texttt{freq}(\texttt{x}) = |\{t \in \boldsymbol{r} \mid \texttt{x} \subseteq t\}|$.
- $\texttt{area}(\texttt{x}) = \texttt{freq}(\texttt{x}) \times \texttt{taille}(\texttt{x})$ où $\texttt{taille}(\texttt{x})=|\texttt{x}|$.
- $\texttt{min}(\texttt{x}.att)$ (resp. $\texttt{max}(\texttt{x}.att)$) est la plus petite (resp. grande) valeur de x pour l'attribut *att*.
- $\texttt{mean}(\texttt{x}) = (\texttt{min}(\texttt{x}.att) + \texttt{max}(\texttt{x}.att))/2$.

Les skypatterns permettent d'exprimer une préférence de l'utilisateur via une relation de dominance [4].

Définition 1 (Dominance Pareto). Soit M un ensemble de mesures, un motif $\texttt{x}_i$ domine un autre motif $\texttt{x}_j$ sur M (noté $\texttt{x}_i \succ_M \texttt{x}_j$), ssi $\forall m{\in}M, m(\texttt{x}_i){\geq}m(\texttt{x}_j)$ et $\exists m \in M, m(\texttt{x}_i) > m(\texttt{x}_j)$.

Définition 2 (Skypattern et opérateur skypattern). Soit M un ensemble de mesures, un skypattern sur M est un motif non-dominé. L'opérateur skypattern est $Sky(M) = \{\texttt{x}_i \in \mathcal{L}_\mathcal{I} \mid \nexists \texttt{x}_j \in \mathcal{L}_\mathcal{I}, \texttt{x}_j \succ_M \texttt{x}_i\}$

Exemple 2. Pour $M = \{\texttt{freq}, \texttt{area}\}$, on a $Sky(M) = \{BCDE, BCD, B, E\}$ (cf Figure 1b).

Soit M un ensemble de mesures, deux motifs $\texttt{x}_i$ et $\texttt{x}_j$ sont *indistincts* sur M (noté $\texttt{x}_i{=_M}\texttt{x}_j$) ssi $\forall m{\in}M$, $m(\texttt{x}_i){=}m(\texttt{x}_j)$. $\texttt{x}_i$ et $\texttt{x}_j$ sont *incomparables* sur M (noté $\texttt{x}_i \prec\succ_M \texttt{x}_j$) ssi $(\texttt{x}_i \not\succ_M \texttt{x}_j)$ et $(\texttt{x}_j \not\succ_M \texttt{x}_i)$ et $(\texttt{x}_i{\neq_M}\texttt{x}_j)$.

Définition 3 (Skypattern incomparable). Un motif $\texttt{x} \in Sky(M)$ est incomparable sur M ssi $\forall \texttt{x}_i{\in}Sky(M)$ tel que $\texttt{x}_i \neq \texttt{x}, \texttt{x}_i \prec\succ_M \texttt{x}$.

Définition 4 (Skypattern indistinct). Un motif $\texttt{x}{\in}Sky(M)$ est indistinct sur M ssi $\exists \texttt{x}_i{\in}Sky(M)$ tel que $(\texttt{x}_i \neq \texttt{x}) \wedge (\texttt{x}_i =_M \texttt{x})$.

On peut regrouper les skypatterns indistincts.

Définition 5 (Groupe de skypatterns indistincts (GSI)). $S{\subseteq}Sky(M)$ est un GSI ssi $|S|{\geq}2$ et $\forall \texttt{x}_i, \texttt{x}_j{\in}S, (\texttt{x}_i{=_M}\texttt{x}_j){\wedge}\forall \texttt{x}_i{\in}S, \forall \texttt{x}_j{\in}Sky(M)\backslash S, (\texttt{x}_i{\prec\succ_M}\texttt{x}_j)$.

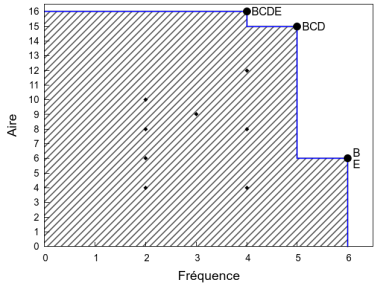
Exemple 3. Pour $M{=}\{\texttt{freq}, \texttt{area}\}$, $BCDE$ et BCD sont incomparables. B et E (indistincts) forment un GSI.

Définition 6 (Cube de skypatterns pour M). $SkyCube(M){=}\{(M_u, Sky(M_u)) \mid M_u \subseteq M, M_u \neq \emptyset\}$

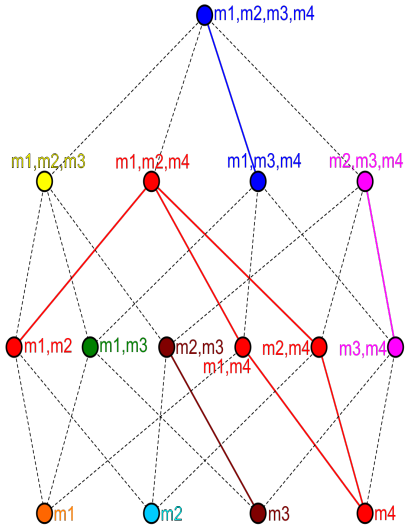
Trans.	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F

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

(a) Jeu de données $\boldsymbol{r}$.



(b) Skypatterns pour M={`freq`, `area`}.



(c) Treillis associé à M.

Sous-ensemble de M	Ensemble de skypatterns
$\{m_1, m_2, m_3, m_4\}$	{BCDE, BCD, BDE, EF, BE, E}
$\{m_1, m_2, m_3\}$	{BCDE, BCD, BE, E}
$\{m_1, m_2, m_4\}$	{E}
$\{m_1, m_3, m_4\}$	{BCDE, BCD, BDE, EF, BE, E}
$\{m_2, m_3, m_4\}$	{BCDE, BDE, EF, E}
$\{m_1, m_2\}$	{E}
$\{m_1, m_3\}$	{BCDE, BCD, B, E}
$\{m_1, m_4\}$	{E}
$\{m_2, m_3\}$	{BCDE}
$\{m_2, m_4\}$	{E}
$\{m_3, m_4\}$	{BCDE, BDE, EF, E}
$\{m_1\}$	{B, E}
$\{m_2\}$	{ABCDEF, ABCEF, ABDEF, ABEF, ABCDE, ABCE, ABDE, ABE, ACDEF, ACEF, ACDE, ACE, ADEF, ADE, AEF, AE, BCDEF, BCEF, CDEF, CEF, BCDE, BCE, CDE, CE, BDEF, DEF, BDE, DE, BEF, EF, BE, E}
$\{m_3\}$	{BCDE}
$\{m_4\}$	{E}

(d) Cube de skypatterns pour M.

FIGURE 1 – $M = \{m_1 : \texttt{freq}, m_2 : \texttt{max}, m_3 : \texttt{area}, m_4 : \texttt{mean}\}$.

Exemple 4. La Figure 1c représente le treillis associé à M. La Figure 1d associe à chaque sous-ensemble non-vide de M son ensemble de skypatterns.

3. Règles de dérivation et calcul du cube

Deux règles de dérivation permettent construire de manière ascendante le cube de skypatterns, : l'une pour les skypatterns incomparables (cf. le théorème 1) et l'autre pour les GSI (cf. le théorème 2).

Théorème 1 (*Règle pour les incomparables*). Soit $M_u \subseteq M$, si $\mathtt{x}$ est un skypattern incomparable pour M_u, alors $\forall m \in M \setminus M_u$, $\mathtt{x} \in Sky(M_u \cup \{m\})$.

Théorème 2 (*Règle pour les GSI*). Soient $M_u \subseteq M$ et S un GSI pour M_u. $\forall m \in M \setminus M_u$, $\forall \mathtt{x} \in S$ t.q. $m(\mathtt{x}) = \max_{\mathtt{x}' \in S} \{m(\mathtt{x}')\}$, $\mathtt{x} \in Sky(M_u \cup \{m\})$.

Mais, ces deux règles ne permettent pas toujours de déterminer tous les skypatterns d'un nœud père à l'aide des skypatterns de ses fils.

Exemple 5. Pour M_u={m_1, m_3}, les skypatterns dérivés sont : B, E et $BCDE$ (le motif $BCDE$ est incomparable alors que les motifs B et E sont indistincts). Mais, les deux règles ne permettent pas de déduire que $BCD \in Sky(M_u)$.

Les CSP Dynamiques permettent de calculer à la volée les skypatterns manquants (non-dérivables). Considérons la séquence $P_1, \ldots, P_n$ de CSP où chaque $P_i = (\{\mathtt{x}\}, \mathcal{L}_\mathcal{I}, \mathtt{q}_i(\mathtt{x}))$ et :
- $\mathtt{q}_1(\mathtt{x}) = (B \not\succ_{M_u} \mathtt{x}) \wedge (E \not\succ_{M_u} \mathtt{x}) \wedge (BCDE \not\succ_{M_u} \mathtt{x})$.
- $\mathtt{q}_{i+1}(\mathtt{x}) = \mathtt{q}_i(\mathtt{x}) \wedge (s_i \not\succ_{M_u} \mathtt{x})$ où s_i est la première solution à la requête $\mathtt{q}_i(\mathtt{x})$.

Chaque requête $\mathtt{q}_i(\mathtt{x})$ permet d'agrandir la zone de non-dominance jusqu'à ce qu'elle soit totalement déterminée, i.e. $\exists n$ t.q. $\mathtt{q}_n(\mathtt{x})$ n'a pas de solution. Mais, tous les s_i ainsi obtenus ne sont pas forcément des skypatterns pour M_u. Une étape de post-traitement doit être effectuée afin de les déterminer [6].

4. Expérimentations

Nous avons mené une évaluation expérimentale sur différents jeux de données de l'*UCI* et sur un jeu de données réel Mutagénicité [2], problème majeur dans l'évaluation des risques des substances chimiques fourni par le CERMN (`www.cermn.unicaen.fr`). Les résultats obtenus montrent l'efficacité de notre proposition et la qualité de nos règles de dérivation.

Conclusion. Nous avons conçu une méthode bottom-up efficace pour calculer le cube de skypatterns. Les expérimentations menées montrent l'intérêt de notre proposition. La navigation à travers le cube est une perspective très prometteuse.

Références

[1] S. Börzsönyi, D. Kossmann, and K. Stocker. The Skyline Operator. In *ICDE*, pages 421–430, 2001.

[2] K. Hansen and al. Benchmark data set for in silico prediction of ames mutagenicity. *Journal of Chemical Information and Modeling*, 49(9) :2077–2081, 2009.

[3] C. Raïssi, J. Pei, and T. Kister. Computing Closed Skycubes. *PVLDB*, 3(1) :838–847, 2010.

[4] A. Soulet, C. Raïssi, M. Plantevit, and B. Crémilleux. Mining Dominant Patterns in the Sky. In *ICDM*, 2011.

[5] W. Ugarte, P. Boizumault, S. Loudni, and B. Crémilleux. Computing Skypattern Cubes. In *ECAI*, 2014.

[6] W. Ugarte, P. Boizumault, S. Loudni, B. Crémilleux, and A. Lepailleur. Mining (Soft-) Skypatterns using Dynamic CSP. In *CPAIOR*, pages 71–87, 2014.