Enumération des itemsets rares minimaux à partir des bases de données transactionnelles

Amel Hidouri*, Badran Raddaoui**
Said Jabbour***

*ICUBE (UMR CNRS 7357) INSA Strasbourg, France
amel.hidouri@insa-strasbourg.fr
**SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France
badran.raddaoui@telecom-sudparis.eu
*** CRIL & CNRS, Universite d'Artois, Lens, France
jabbour@cril.fr

1 Résumé

L'énumération des itemsets rares minimaux est une tâche bien connue en fouille de données, avec de nombreuses applications. Ce travail présente une nouvelle approche pour la fouille des itemsets rares minimaux. Tout d'abord, nous introduisons une généralisation des itemsets rares minimaux appelée motifs k-rare minimaux qui est défini comme un motif rare qui ne devient fréquent qu'après suppression d'au moins k de ces items. Nous présentons ensuite un cadre basé sur la SAT pour découvrir efficacement ces motifs k-rare minimaux à partir de bases de transactions. Pour le passage à l'échelle, nous utilisons une approche de décomposition. Enfin, pour évaluer l'efficacité de notre approche, nous menons une analyse expérimentale en utilisant divers bases transactionelles tout en la comparant à des algorithmes spécialisés et des algorithmes basés sur la programmation par contraintes .

2 Définition du Problème et encodage vers SAT

Dans ce travail nous considérons Ω un ensemble d'items. Un motif est un sous-ensemble de Ω i.e., $X \subseteq \Omega$ et une table de transactions $\mathcal{D}$ est un ensemble de transactions définies sur Ω. La couverture d'un motif est l'ensemble des transactions qui le contiennent i.e., $\texttt{Cover}(X, \mathcal{D}) = \{T \in \mathcal{D} \mid X \subseteq T\}$ et le support est la cardinalité de sa couverture i.e., $\texttt{Supp}(X, \mathcal{D}) = |\texttt{Cover}(X, \mathcal{D})|$. Pour un seuil de support λ donné, un itemset X est dit fréquent si $\texttt{Supp}(X, \mathcal{D}) \geq \lambda$. Dans le cas contraire, i.e.,$\texttt{Supp}(X, \mathcal{D}) < \lambda$, il est dit *rare*. Un motif rare X est dit *minimal* si tout sous-ensemble strict de X est fréquent et il est dit *k-rare minimal* si (i) $\forall\, Y \subset X$ s.t. $|Y| \leq k-1$, $X \setminus Y$ est rare, et (ii) $\forall Y \subset X$ s.t. $|Y| \geq k$, $X \setminus Y$ est fréquent.

Plusieurs approaches ont été proposées pour l'extraction des motifs rares minimaux Szathmary et al. (2007); Belaid et al. (2019). Nous présentons un encodage des motifs rares minimaux classiques (1-MRI) avant de donner une généralisation aux k-MRI.
Pour notre modèle basé sur SAT, nous allons introduire un ensmeble de varaibles propositionnelles où pour chaque item $a \in \Omega$ on associe une variable propositionnelle x_a où x_a est *vrai*

si et seulement si a appartient au motif rare candidat. De manière similaire, pour chaque transaction T_i dans $\mathcal{D}$ on associe deux variables propositionnelles p_i et q_i où p_i (resp. q_i) est *vrai* si et seulement si la transaction T_i contient le motif rare (resp. le motif rare à l'exception d'un seul item). Maintenant, afin d'établir une correspondance entre l'ensemble des 1-MRIs dans la base de transctions $\mathcal{D}$ et les modèles de la formule propositionnelle correspondante, nous introduisons les contraintes logiques comme indiqué dans la Figure 1.

$$\bigwedge_{T_i \in \mathcal{D}} (p_i \leftrightarrow (\bigwedge_{a \in \Omega \setminus T_i} \neg x_a)) \quad (1) \qquad \sum_{T_i \in \mathcal{D}} p_i < \lambda \quad (2)$$

$$\bigwedge_{T_i \in \mathcal{D}} (q_i \leftrightarrow (\sum_{a \in \Omega \setminus T_i} x_a = 1)) \quad (3) \qquad \bigwedge_{a \in \Omega} (x_a \rightarrow (\sum_{\substack{T_i \in \mathcal{D} \\ a \in T_i}} p_i + \sum_{\substack{T_i \in \mathcal{D} \\ a \notin T_i}} q_i \geq \lambda)) \quad (4)$$

FIG. 1 – Encodage basé sur SAT pour la fouille des MRIs

Pour énumérer l'ensemble de tous les k-MRI dans $\mathcal{D}$, nous procédons de manière similaire. Pour chaque item a on associe une variable x_a pour représenter chaque item a. Toutefois, pour chaque transaction on associe un ensemble de variables $p_{i,0}, p_{i,1}, \ldots, p_{i,k}$ pour identifier si la transaction contient tout l'itemset sauf j pour tout $0 \leq j \leq k$.

$$\bigwedge_{j=0}^{k} \bigwedge_{T_i \in \mathcal{D}} (p_{i,j} \leftrightarrow (\sum_{a \in \Omega \setminus T_i} x_a = j)) \quad (5) \qquad \bigwedge_{\substack{X \subseteq \Omega \\ |X| = k-1}} (\bigwedge_{a \in X} x_a \rightarrow (\sum_{j=0}^{k-1} (\sum_{\substack{T_i \in \mathcal{D} \\ |(\Omega \setminus T_i) \cap X| = j}} p_{i,j} < \lambda))) \quad (6)$$

$$\bigwedge_{\substack{X \subseteq \Omega \\ |X| = k}} (\bigwedge_{a \in X} x_a \rightarrow (\sum_{j=0}^{k} (\sum_{\substack{T_i \in \mathcal{D} \\ |(\Omega \setminus T_i) \cap X| = j}} p_{i,j} \geq \lambda))) \quad (7)$$

La formule propositionnelle $\Phi_{\mathcal{D}}^{k,\lambda} = (5) \wedge (6) \wedge (7)$ encode le problème d'extraraction des k-MRI dans $\mathcal{D}$. Dès lors l'énumération des modèles de $\Phi_{\mathcal{D}}^{k,\lambda}$ permet de d'obtenir tous les k-MRIs. Par conséquent, un solveur SAT dédié à l'énumération de modèles peut être utilisé. Les expérimetations sont disponibles sur le papier original Hidouri et al. (2023) où pour le passage à l'échelle, la décomposition a été utilisée pour accélérer l'énumération de ces motifs.

Références

Belaid, M.-B., C. Bessiere, et N. Lazaar (2019). Constraint programming for mining borders of frequent itemsets. In *IJCAI*, pp. 1064–1070.

Hidouri, A., B. Raddaoui, et S. Jabbour (2023). Targeting minimal rare itemsets from transaction databases. In *Thirty-Second International Joint Conference on Artificial Intelligence {IJCAI-23}*, pp. 2114–2121.

Szathmary, L., A. Napoli, et P. Valtchev (2007). Towards rare itemset mining. In *ICTAI*, pp. 305–312.