Extraction de motifs fermés dans des relations n-aires bruitées

Loïc Cerf, Jérémy Besson, Jean-François Boulicaut

Université de Lyon, CNRS
INSA-Lyon, LIRIS UMR5205, F-69621 Villeurbanne, France
Prénom.Nom@liris.cnrs.fr

Résumé. L'extraction de motifs fermés dans des relations binaires a été très étudiée. Cependant, de nombreuses relations intéressantes sont n-aires avec $n > 2$ et bruitées (nécessité d'une tolérance aux exceptions). Récemment, ces deux problèmes ont été traités indépendamment. Nous introduisons notre proposition pour combiner de telles fonctionnalités au sein d'un même algorithme.

1 Introduction

La fouille de relations binaires a été très étudiée via notamment les usages multiples des ensembles fermés fréquents. Cependant, il est courant que les données à traiter se représentent dans des relations n-aires avec $n \geq 3$ et il semble donc naturel de vouloir étendre le calcul de motifs fermés dans ce contexte (Ji et al., 2006; Jaschke et al., 2006; Cerf et al., 2008b). Dans le cas des relations binaires (calcul de 2-ensembles fermés ou concepts formels selon (Ganter et al., 2005)), nous savons que le nombre et la qualité des motifs extraits sont déjà problématiques. De nombreuses raisons (e.g., une erreur de mesure) peuvent mener à l'absence d'un couple dans la relation et un « véritable » motif donne lieu à plusieurs motifs fermés distincts et plus petits : quand la quantité de bruit augmente, le nombre de motifs fermés explose et leur pertinence se dégrade. Cette situation empire dramatiquement lorsque l'arité de la relation à fouiller augmente. Nous introduisons ici un algorithme de calcul de tous les motifs fermés ayant un nombre borné d'exceptions par élément (de n'importe quel attribut) sur n'importe quelle relation n-aire. Cet article est une version courte de (Cerf et al., 2008a).

2 Notion de ET-n-ensemble fermé

Soit $D^1, \dots, D^n$ les domaines de n attributs. Soit $\mathcal{R}$ une relation n-aire sur ces attributs, i.e., $\mathcal{R} \subseteq D^1 \times \cdots \times D^n$. Appelons X un motif $\langle X^1, \dots, X^n \rangle \in 2^{D^1} \times \cdots \times 2^{D^n}$. $\forall i = 1 \dots n, \forall e \in D^i$, l'hyper-plan de X sur e, noté $H(X, e)$, est $\langle X^1, \dots, \{e\}, \dots, X^n \rangle$. $0(X)$ est le nombre de n-uplets de X qui sont absents de $\mathcal{R}$, i.e., $|(X^1 \times \cdots \times X^n) \setminus \mathcal{R}|$. Un n-ensemble fermé de $\mathcal{R}$ désigne l'extension naturelle de la notion de concept formel aux relations n-aires quand $n > 2$. Un n-ensemble fermé satisfait deux contraintes, la contrainte dite de connexion et celle dite de fermeture. $X = \langle X^1, \dots, X^n \rangle$ vérifie la contrainte de connexion notée $\mathcal{C}_{\text{cx}}$ ssi $\forall i = 1 \dots n, \forall e \in X^i, 0(H(X, e)) = 0$. X est dit fermé, i.e. satisfait la contrainte