M2IA - Semestre 2

Université d'Artois
Faculté des Sciences Jean Perrin

AIC – Algorithmique pour l'Inférence et Contraintes

Fiche de TP numéro 3

La fouille des itemsets fréquents et des règles d'association sont certaines des tâches standards en fouille de données les plus utilisées dans divers domaines d'application. Elles visent à découvrir des relations d'association fortes ou des implications entre les éléments dans une base de données transactionnelle sous certains seuils spécifiés.

Ces problèmes sont définis de façon formelle comme suit.

Soit Ω un univers d'items (ou de symboles). Un itemset classique X sur Ω est défini comme un sous-ensemble de Ω, c'est-à-dire $X \subseteq \Omega$. Il peut être vu comme une conjonction d'éléments. On note 2^{Ω} l'ensemble de tous les itemsets sur Ω. Une base de données de transactions $\mathcal{D}$ est définie comme un ensemble fini de paires $\{(1, T_1), \ldots, (m, T_m)\}$ telle que $T_i \in 2^{\Omega}$, $\forall 1 \leq i \leq m$.

— La couverture d'un itemset est l'ensemble des transactions contenant cet itemset i.e.,

$$Cover(X, \mathcal{D}) = \{(i, T_i) \in \mathcal{D},\ X \subseteq T_i\}$$

— Le support d'un itemset est la cardinalité de la couverture :'

$$Supp(X, \mathcal{D}) = |Cover(X, \mathcal{D})|$$

— Pour un support minimum α, un itemset est dit **fréquent** si $Supp(X) \geq \alpha$.

— Un itemset est dit **fermé** si pour tout $X \subset Y$, $Supp(Y) < Supp(X)$

Le problème de la fouille des itemsets fréquents (fermés) ((C)FIM) est le problème qui consiste à énumérer tous les itemsets fréquents (fermés) :

$$S = \{X \in \Omega,\ Supp(X, \mathcal{D}) \geq \alpha\}$$

Soit une base de données de transactions D, une règle d'association (AR) est une expression de la forme $X \rightarrow Y$ où X et Y sont deux itemsets disjoints appelés respectivement antécédent et conséquence de la règle [Rakesh et al., 1993]. l'intérêt d'une AR est calculé à l'aide des deux mesures :

— Le support de $X \rightarrow Y$ dans $\mathcal{D}$, défini comme suit et qui détermine la fréquence de la règle dans D :

$$Supp(X \rightarrow Y, \mathcal{D}) = \frac{Supp(X \cup Y, \mathcal{D})}{|\mathcal{D}|}$$

— La **confidence** de $X \rightarrow Y$ dans $\mathcal{D}$ est défini comme suit :

$$Conf(X \rightarrow Y, \mathcal{D}) = \frac{Supp(X \cup Y, \mathcal{D})}{Sup(X, \mathcal{D})}$$

Elle représente la probabilité conditionnelle de l'occurrence du conséquent compte tenu de l'antécédent.

— Une règle d'association $X \rightarrow Y$ est **fermée** si $X \cup Y$ est fermé.

— **Le problème de la fouille des règles d'association** est le problème qui est définit étant donnée une table de transactions $\mathcal{D}$ et deux entiers α et β de calculer les règles d'association de support supérieur à α et de confidence au moins égale à β.

$$S_{AR} = \{X \rightarrow Y,\ Supp(X \rightarrow Y, \mathcal{D}) \geq \alpha \text{ and } Conf(X \rightarrow Y, \mathcal{D}) \geq \beta\}$$

Example

Considérons la table de transactions $\mathcal{D}$ de la figure 1. Nous avons :
— $Cover(\{f, g\}, \mathcal{D}) = \{T_7, T_8, T_9\}$
— $Supp(\{f, g\}, \mathcal{D}) = 3$
— $\{f, g\}$ est un itemset fermé alors que $\{e, h\}$ ne l'est pas.

tid	**Itemset**								
t_1		b	c	d					
t_2	a		c	d					
t_3	a	b		d	e				
t_4	a	b	c		e	f			
t_5	a		c			f		h	
t_6				d	e		g	h	i
t_7					e	f	g	h	i
t_8					e	f	g		i
t_9						f	g	h	

TABLE 1 – A Transaction Database $\mathcal{D}$

Exercice 1 : Proposer un encodage en logique propositionnelle du problème de la fouille des itemsets fréquents.

Exercice 2 : Écrire un programme qui prend en paramètres une table de transactions et un entier α et qui encode le problème de l'énumeration des itemsets fréquents en logique propositionnelle. Ajouter à votre programme une fonction qui permet d'énumérer toutes les règles en faisant un appel itératif à un solveur SAT.

Exercice 3 : Proposer une extension de l'encodage des itemsets fréquents vers les règles d'association.

Exercice 4 : Écrire un parseur qui prend une table de transactions et qui encode le problème de la fouille des règles d'association de support α et de confidence β en logique propositionnelle. Ajouter également à votre programme une fonction qui permet d'énumération toutes les règles en faisant un appel itératif à un solveur SAT.

Encodage des itemsets fréquents en logique propostionnelle

— Contrainte de couverture
  — $|\Omega| = n$ et $|\mathcal{D}| = m$
— Contrainte de support

$$\Phi^{cov} = \bigwedge_{i=1}^{n} (\neg q_i \leftrightarrow \bigvee_{a \in \Omega \setminus T_i} p_a) \tag{1}$$

— Contrainte de support

$$\Phi^{freq} = \sum_{i=1}^{m} q_i \geq \alpha \tag{2}$$

— Contrainte de fermeture

$$\Phi^{clos} = \bigwedge_{a \in \Omega} ((\bigvee_{(i,T_i) \in \mathcal{D},\ a \notin T_i} q_i) \vee p_a) \tag{3}$$

La formule qui encode le problème FIM est la suivante. Il existe une bijection entre les modèles de $\Phi(\mathcal{D}, \alpha)$ et les itemsets fréquents de $\mathcal{D}$.

$$\Phi(\mathcal{D}, \alpha) = \Phi^{cov} \wedge \Phi^{freq} \wedge \Phi^{clos}$$