Sous-bases k-faibles pour des règles d'association valides au sens de la confiance

Jean Diatta, Régis Girard

IREMIA, Université La Réunion
15, Avenue Réné Cassin- 97715-St Denis, FRANCE
{ jean.diatta, rgirard}@univ-reunion.fr

Résumé. Nous introduisons la notion de sous-base k-faible pour les règles d'association valides au sens de la confiance. Ces sous-bases k-faibles sont caractérisées en termes d'opérateurs de fermeture correspondant à des familles de Moore k-faiblement hiérarchiques.

1 Introduction

L'un des problèmes majeurs rencontrés dans la fouille des règles d'association valides au sens de la confiance est le nombre souvent très élevé de ces règles. Plusieurs solutions à ce problème ont été proposées ou considérées dans la littérature. Parmi ces solutions figurent les bases, c'est-à-dire, des familles génératrices minimales (Zaki et Ogihara, 1998; Pasquier et al., 1999). La plupart de ces bases se caractérisent en terme d'un opérateur de fermeture de Galois sur l'ensemble des motifs du contexte considéré. Or, cet opérateur de fermeture correspond à une famille de Moore m-faiblement hiérarchique, où $m \geq 2$ est un entier (Diatta, 2004). Plus précisément, les fermés de cet opérateur de fermeture coïncident avec les classes faibles associées à une certaine mesure de dissimilarité m-voies et forment donc, de ce fait, la hiérarchie m-faible associée à cette mesure de dissimilarité.

Dans cet article, nous considérons la caractérisation de ces bases pour les règles d'association, en remplaçant l'opérateur de fermeture de Galois par un opérateur de fermeture correspondant à la hiérarchie k-faible associée à une mesure de dissimilarité k-voies donnée, où $2 \leq k \leq m$. Pour chaque valeur de k, l'ensemble de règles ainsi caractérisé sera appelé sous-base k-faible. Ces sous-bases k-faibles offrent une approximation de l'ensemble des règles valides, relativement à des ensembles d'items (classes k-faibles) ayant un certain degré d'homogénéité exprimé par le biais d'un indice d'isolation. Par ailleurs, la possibilité d'associer une sous-base $(k-)$ faible à une mesure de dissimilarité (k-voies) permet d'intégrer la sémantique de cette mesure de dissimilarité dans le choix des règles à générer.

2 Règles d'association

2.1 Définition générale

Étant donné un contexte binaire $\mathbb{K} = (\mathcal{E}, \mathcal{V})$, où $\mathcal{E}$ désigne un ensemble fini d'entités et $\mathcal{V}$ un ensemble fini de variables booléennes (ou attributs) définies sur $\mathcal{E}$. On appelle *motifs* les

sous-ensembles de $\mathcal{V}$ et on dit qu'une entité possède un attribut "x" si $x(e) = 1$.

Une règle d'association de $(\mathcal{E}, \mathcal{V})$ est un couple (X, Y) de motifs, notée $X \rightarrow Y$, où Y est non vide. X et Y sont respectivement appelés "*prémisse*" et "*conséquent*".

Étant donné un motif X, X' désignera l'ensemble des entités qui possèdent tous les attributs de X, *i.e.*, $X' = \{e \in \mathcal{E} : \forall x \in X, x(e) = 1\}$.

Un contexte binaire $(\mathcal{E}, \mathcal{V})$ contient $2^{|\mathcal{V}|}(2^{|\mathcal{V}|} - 1)$ règles d'association parmi lesquelles beaucoup ne sont pas pertinentes. On utilise des mesures de qualité pour sélectionner uniquement les règles qui vérifient des contraintes données.

2.2 Règles d'association confiance-valides

Une *mesure de qualité* pour les règles d'association d'un contexte $\mathbb{K}$ est une application à valeurs réelles, définie sur l'ensemble des règles d'association de $\mathbb{K}$. Beaucoup de mesures de qualité on été proposées dans la littérature, les plus utilisées d'entre elles étant le support et la confiance (Agrawal et al., 1993).

Le support d'un motif X est la proportion des entités de $\mathcal{E}$ qui possèdent tous les attributs de X ; on le définit par $\mathrm{Supp}(X) = \frac{|X'|}{|\mathcal{E}|}$, où, pour un ensemble fini S, $|S|$ désigne le cardinal de S. Si on note p la mesure de probabilité intuitive définie sur $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$ par $p(E) = \frac{|E|}{|\mathcal{E}|}$ pour $E \subseteq \mathcal{E}$, alors le support de X peut s'écrire en termes de p par : $\mathrm{Supp}(X) = p(X')$. Le support d'une règle d'association $X \rightarrow Y$ est défini par $\mathrm{Supp}(X \rightarrow Y) = \mathrm{Supp}(X \cup Y) = p((X \cup Y)') = p(X' \cap Y')$.

La confiance d'une règle d'association $X \rightarrow Y$ est la proportion des entités qui possèdent tous les attributs de Y, parmi celles qui possèdent tous les attributs de X ; elle est définie par $\mathrm{Conf}(X \rightarrow Y) = \frac{\mathrm{Supp}(X \rightarrow Y)}{\mathrm{Supp}(X)} = \frac{p(X' \cap Y')}{p(X')} = p(Y'|X')$, où $p(Y'|X')$ est la probabilité conditionnelle de Y' sachant X'.

Une règle d'association est confiance-valide si sa confiance est au moins égale à un seuil minimum (de validité) fixé. Une règle est dite confiance-exacte si sa confiance est égale à 1 ; elle est dite confiance-approximative si sa confiance est strictement inférieure à 1.

3 Bases pour les règles confiance-valides

L'un des problèmes majeurs de l'extraction de règles d'association est le nombre très élevé de règles générées. En effet, pour une mesure de qualité donnée μ, l'ensemble des règles d'association μ-valides contient fréquemment de nombreuses règles redondantes, *i.e* qui peuvent être déduites d'autres règles μ-valides. Pour palier ce problème on cherche à calculer une base de l'ensemble des règles valides, c'est à dire un ensemble minimal (au sens de l'inclusion) de règles d'association valides, à partir duquel toute règle valide peut être dérivée. Dans ce papier, nous considérons des bases qui se caratérisent en termes d'opérateurs de fermeture de Galois.

3.1 Opérateurs de fermeture de Galois

Le contexte binaire $\mathbb{K}$ induit une correspondance de Galois entre les ensembles ordonnés $(\mathcal{P}(\mathcal{E}), \subseteq)$ et $(\mathcal{P}(\mathcal{V}), \subseteq)$, définie par les applications $f : X \mapsto \bigcap\limits_{x \in X} \{v \in \mathcal{V} : v(x) = 1\}$ et

$g : Y \mapsto \underset{v\in Y}{\cap}\{x \in \mathcal{E} : v(x) = 1\}$, pour $X \subseteq \mathcal{E}$ et $Y \subseteq \mathcal{V}$ (Barbut et Monjardet, 1970). Par ailleurs, la correspondance de Galois (f, g) induit un opérateur de fermeture $\varphi := f \circ g$ sur $(\mathcal{P}(\mathcal{V}), \subseteq)$ (Birkhoff, 1967). Cet opérateur de fermeture sera dit de Galois, et un motif X sera dit fermé de Galois du contexte $\mathbb{K}$, ou φ-fermé, si $\varphi(X) = X$. Pour un motif X, $\varphi(X)$ sera appelé sa φ-fermeture.

3.2 Bases de Guigues-Duquenne et de Luxenburger

L'ensemble des règles d'association confiance-exactes est un système complet d'implications, *i.e.* il satisfait les axiomes d'inférence de Armstrong (1974). Il en découle que la base de Guigues et Duquenne (1986) pour les systèmes complets d'implications est aussi une base pour les règles d'association confiance-exactes.

La base de Guigues-Duquenne pour les règles d'association confiance-exactes est l'ensemble $\mathrm{GD} = \{X \rightarrow \varphi(X) \setminus X : X \text{ est } \varphi\text{-critique}\}$, où un motif X est φ-critique s'il n'est pas φ-fermé et contient strictement la φ-fermeture de tout motif φ-critique Y tel $Y \subset X$. Pour tout motif X, $\varphi(X)$ est applelé la φ-fermeture de X.

La base de Luxenburger (1991) pour les règles d'association confiance-approximatives est l'ensemble $\mathrm{LB} = \{X \rightarrow Y : X = \varphi(X), Y = \varphi(Y), X \prec Y \text{ et } \mathrm{Conf}(X \rightarrow Y) \geq \alpha\}$, où $X \prec Y$ signifie que $X \subset Y$ et qu'il n'existe aucun ensemble φ-fermé Z tel que $X \subset Z \subset Y$.

4 Sous-bases k-faibles pour les règles confiance-valides

4.1 Opérateurs de fermeture et familles de Moore

Soit E un ensemble. Une *famille de Moore* sur E est une partie $\mathcal{F}$ de l'ensemble $\mathcal{P}(E)$ des parties de E telle que $E \in \mathcal{F}$ et $\mathcal{F}' \subseteq \mathcal{F}$ impliquent $\cap\mathcal{F}' \in \mathcal{F}$. Si $\mathcal{F}$ est finie, donc si E est fini, alors la seconde condition peut être remplacée par : $X, Y \in \mathcal{F}$ implique $X \cap Y \in \mathcal{F}$.

Par ailleurs, étant donnée un famille de Moore $\mathcal{F}$ sur E, l'application $\phi_{\mathcal{F}}$ définie sur $\mathcal{P}(E)$ par $\phi_{\mathcal{F}}(X) = \cap\{Y \in \mathcal{F} : X \subseteq Y\}$ est un *opérateur de fermeture* sur $\mathcal{P}(E)$.

Réciproquement, étant donné un opérateur de fermeture ϕ sur E, la collection $\mathcal{F}_\phi$ de sous-ensembles de E, définie par $\mathcal{F}_\phi = \{X \subseteq E : \phi(X) = X\}$ est une famille de Moore sur E.

4.2 Fermés de Galois et classes faibles

Une mesure de dissimilarité (mutuelle) sur E est une fonction $d_2 : E \times E \rightarrow \mathbb{R}$ vérifiant les conditions $d_2(x, x) = 0$, $d_2(x, y) \geq 0$ et $d_2(x, y) = d_2(y, x)$. Dans ce papier, nous remplaçons les deux premières conditions par la condition moins forte $d_2(x, y) \geq d_2(x, x)$. Une justification de cet affaiblissement peut être trouvée dans (Diatta, 2006).

Un sous-ensemble X de E est une *classe faible* associée à une mesure de dissimilarité d_2 (ou *classe d_2-faible*), si son *indice d'isolation d_2-faible*

$$\mathbf{i}^w_{d_2}(X) := \min_{x,y\in X, z\notin X}\{\max\{d_2(x, z), d_2(y, z)\} - d_2(x, y)\}$$

est strictement positif. En d'autres termes, pour tous x, y dans la classe et tout z extérieur à la classe, au moins l'une des dissimilarités $d_2(x, z)$ et $d_2(y, z)$ est strictement supérieure à la dissimilarité $d_2(x, y)$.

Les mesures de dissimilarité (mutuelle) se généralisent naturellement en mesures de dissimilarité dites multivoies permettant d'évaluer le degré de dissemblance globale entre entités d'un ensemble de plus de deux éléments.

Etant donnés un ensemble S et un entier $k \geq 1$, désignons par $S^*_{\leq k}$ l'ensemble des sous-ensembles non vides de S ayant au plus k éléments. Alors, en adoptant la définition ensembliste proposée dans (Diatta, 2006), une *mesure de dissimilarité k-voies* sur E est une fonction monotone croissante définie sur $E^*_{\leq k}$ à valeurs réelles positives ou nulles, *i.e.*, une fonction $d_k : E^*_{\leq k} \rightarrow \mathbb{R}_+$ telle que $d_k(X) \leq d_k(Y)$ lorsque $X \subseteq Y$.

On notera que les mesures de dissimilarité (mutuelle) correspondent au cas où $k = 2$. Par ailleurs, on parlera de mesure de dissimilarité multivoies pour signifier une mesure de dissimilarité k-voies quelconque, pour $k \geq 2$. De plus, telle qu'observée dans (Bandelt et Dress, 1994), la notion de classe faible s'adapte tout aussi naturellement aux mesures de dissimilarité multivoies.

Ainsi, un sous-ensemble X de E est une classe faible associée à une mesure de dissimilarité k-voies d_k si son indice d'isolation d_k-faible

$$\min_{Y \in X^*_{\leq k}, z \notin X} \{ \max_{Z \in Y^*_{\leq k-1}} d_k(Z \cup \{z\}) - d_k(Y) \}$$

est strictment positif.

Par ailleurs, il a été montré dans (Diatta, 2004) qu'il existe un entier $m \geq 2$ tel qu'un sous-ensemble X de $\mathcal{V}$ est φ-fermé si et seulement si X est une classe faible associée à une certaine mesure de dissimilarité m-voies d_m sur $\mathcal{V}$. Il découle de ce résultat que les motifs φ-fermés forment une famille de Moore m-faiblement hiérarchique appelée la *hiérarchie m-faible* associée à la mesure de dissimilarité d_m. Ainsi, φ est tout simplement l'opérateur de fermeture correspondant à cette famille de Moore au sens de la section 4.1.

4.3 Sous-bases k-faibles

Nous avons vu dans la section 3 qu'aussi bien la base de Guigues-Duquenne que celle de Luxenburger pour les règles d'association confiance-valides se caractérisent en terme de l'opérateur de fermeture φ. Par ailleurs, nous venons de voir dans la section 4.2 que l'opérateur de fermeture φ correspond à une famille de Moore m-faiblement hiérarchique, pour un certain entier $m \geq 2$. Ces deux observations nous conduisent à considérer les analogues respectifs des bases de Guigues-Duquenne et de Luxenburger, en remplaçant l'opérateur de fermeture φ par un opérateur de fermeture correspondant à une famille de Moore k-faiblement hiérarchique quelconque, où $2 \leq k \leq m$. Pour $k = m$, on peut retrouver ces bases sous certaines conditions mais, dans le cas gńéral, les ensembles de règles obtenus ne sont pas des bases au sens algébrique.

Ainsi, nous appellerons *sous-base k-faible de Guigues-Duquenne* pour les règles d'association confiance-valides un ensemble de règles GD_{ϕ_k} défini par :

$$\mathrm{GD}_{\phi_k} = \{X \rightarrow \phi_k(X) \setminus X : X \text{ est } \phi_k\text{-critique}\}$$

où ϕ_k est un opérateur de fermeture correspondant à une famille de Moore k-faiblement hiérarchique.

On notera que les règles de GD_{ϕ_k} ne sont pas nécessairement exactes. En effet, certaines de ces règles peuvent avoir des exceptions pour la simple raison que la φ-fermeture de la prémisse d'une telle règle peut être strictement contenue dans sa ϕ_k-fermeture, *i.e.*, X ϕ_k-critique et $\varphi(X) \subset \phi_k(X)$.

On notera également qu'il a été montré dans (Diatta, 2005) que si $\mathcal{F}$ une famille de Moore k-faiblement hiérarchique sur E, contenant toutes les parties de E d'au plus $k-1$ éléments, alors $X \subset E$ est $\phi_{\mathcal{F}}$-critique si et seulement si $X \notin \mathcal{F}$ et $|X| = k$. Ce résultat généralise en fait son analogue obtenu par Domenach et Leclerc (2004) dans le cas particulier des familles de Moore faiblement hiérarchiques ($k = 2$).

De même, nous appellerons *sous-base k-faible de Luxenburger* pour les règles d'association confiance-valides un ensemble de règles LB_{ϕ_k} défini par :

$$\mathrm{LB}_{\phi_k} = \{X \rightarrow Y : X = \phi_k(X), Y = \phi_k(Y), X \prec Y \text{ et } \mathrm{Conf}(X \rightarrow Y) \geq \alpha\}$$

où ϕ_k est un opérateur de fermeture correspondant à une famille de Moore k-faiblement hiérarchique et où $X \prec Y$ signifie $X \subset Y$ et il n'existe pas de ϕ_k-fermé Z tel que $X \subset Z \subset Y$.

On notera que les règles de LB_{ϕ_k} ne sont pas nécessairement approximatives. En effet, certaines de ces règles peuvent être exactes si par exemple leur conséquent est contenu dans la φ-fermeture de leur prémisse, *i.e.*, $X = \phi_k(X) \prec \phi_k(Y) = Y \subseteq \varphi(X)$.

5 Conclusion et discussion

Nous avons introduit la notion de sous-base k-faible relative à l'opérateur de fermeture correspondant à une famille de Moore k-faiblement hiérarchique. Ces sous-bases k-faibles peuvent permettre de réduire le nombre de règles générées tout en assurant que les règles ainsi générées soient relatives à des ensembles d'items (classes k-faibles) ayant un certain degré d'homogénéité exprimé par le biais d'un indice d'isolation. Le degré d'homogénéité de ces classes peut, dans une certaine mesure, garantir la cohérence du lien entre la prémisse et le conséquent d'une règle. En effet, on sait que des règles générées dans l'approche classique peuvent lier des motifs qui sont en réalité très peu liés voire statistiquement indépendants. Par ailleurs, la possibilité d'associer une sous-base ($k-$) faible à une mesure de dissimilarité (k-voies) permet d'intégrer la sémantique de la mesure de dissimilarité dans le choix des règles à générer. On notera qu'une sous-base k-faible n'est pas nécessairement une famille génératrice, donc pas une base au sens algébrique du terme. Toutefois, cela ne nous semble pas être un handicap. En effet, l'intérêt d'un ensemble de règles générées nous parait être plus dans la pertinence des règles qu'il comporte que dans sa capacité à permettre de reconstruire toutes les règles valides. Cela étant, les idées présentées dans ce travail posent, entre autres, les deux questions suivantes : (a) comment générer efficacement une sous-base k-faible ? (b) quel est le degré d'approximation des bases usuelles par les sous-bases k-faibles ?

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In P. Buneman et S. Jajodia (Eds.), *Proc. of the ACM SIGMOD International Conference on Management of Data*, Volume 22, Washington, pp. 207–216. ACM press.

Armstrong, W. W. (1974). Dependency structures of data base relationships. *Information Processing 74*, 580–583.

Bandelt, H.-J. et A. W. M. Dress (1994). An order theoretic framework for overlapping clustering. *Discrete Mathematics 136*, 21–37.

Barbut, M. et B. Monjardet (1970). *Ordre et classification*. Paris : Hachette.

Birkhoff, G. (1967). *Lattice theory*. 3rd edition, Coll. Publ., XXV. Providence, RI : American Mathematical Society.

Diatta, J. (2004). A relation between the theory of formal concepts and multiway clustering. *Pattern Recognition Letters 25*, 1183–1189.

Diatta, J. (2005). Charactérisation des ensembles critiques d'une famille de Moore finie. In *Rencontres de la Société Francophone de Classification*, Montréal, Canada, pp. 126–129.

Diatta, J. (2006). Description-meet compatible multiway dissimilarities. *Discrete Applied Mathematics 154*, 493–507.

Domenach, F. et B. Leclerc (2004). Closure systems, implicational systems, overhanging relations and the case of hierarchical classification. *Mathematical Social Sciences 47*, 349–366.

Guigues, J. L. et V. Duquenne (1986). Famille non redondante d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences humaines 95*, 5–18.

Luxenburger, M. (1991). Implications partielles dans un contexte. *Math. Inf. Sci. hum. 113*, 35–55.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Closed set based discovery of small covers for association rules. In *Proc. 15emes Journees Bases de Donnees Avancees, BDA*, pp. 361–381.

Zaki, M. J. et M. Ogihara (1998). Theoretical Foundations of Association Rules. In *3rd SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*, pp. 1–8.

Summary

We introduce the notion of a k-weak sub-basis for confidence-valid association rules. These k-weak sub-bases are characterized in terms of closure operators corresponding to k-weakly hierarchical Moore families.