Extraction de Règles de Corrélation Décisionnelles

Alain Casali*, Christian Ernst**

* Laboratoire d'Informatique Fondamentale de Marseille (LIF), CNRS UMR 6166
Aix-Marseille Université, Case 901
163 Avenue de Luminy, 13288 Marseille Cedex 9
casali@lif.univ-mrs.fr
** Ecole des Mines de St Etienne, CMP-Georges Charpak
880 avenue de Mimet, 13541 Gardanne
ernst@emse.fr

Résumé. Dans cet article, nous introduisons deux nouveaux concepts : les règles de corrélation décisionnelles et les vecteurs de contingence. Le premier résulte d'un couplage entre les règles de corrélation et les règles de décision. Il permet de mettre en évidence des liens pertinents entre certains ensembles de motifs d'une relation binaire et les valeurs d'un attribut cible (appartenant à cette même relation) en se basant à la fois sur la mesure du Khi-carré et sur le support des motifs extraits. De par la nature du problème, les algorithmes par niveaux font que l'extraction des résultats a lieu avec des temps de réponse élevés et une occupation mémoire importante. Afin de palier à ces deux inconvénients, nous proposons un algorithme basé sur l'ordre lectique et les vecteurs de contingence.

1 Introduction et Motivation

Un axe majeur de la fouille de données est d'exprimer des liens entre les valeurs d'une relation binaire en des temps de calcul raisonnables. Agrawal et al. (1996) ont introduit les algorithmes par niveaux pour calculer les règles d'association : un lien directionnel $X \rightarrow Y$ basé sur la plateforme support / confiance. En s'appuyant sur les littéraux, Wu et al. (2004) proposent le calcul des règles d'association positives et/ou négatives, afin d'extraire des règles du type $\neg X \rightarrow Y$, ... Brin et al. (1997) extraient des règles de corrélation en utilisant la mesure statistique Khi-carré, usuellement notée χ^2. Cet indicateur est approprié pour plusieurs raisons : (i) il est plus significatif au sens statistique du terme qu'une règle d'association ; (ii) il tient compte de la présence et de l'absence des valeurs ; et (iii) il est non directionnel : il met en évidence des liens existants plus complexes qu'une simple implication. Le problème crucial, lors du calcul de règles de corrélation, provient de l'utilisation mémoire requise par les algorithmes par niveaux. En effet, pour un motif X, le calcul du χ^2 s'appuie sur son tableau de contingence qui contient $2^{|X|}$ cases. Ainsi, pour un niveau i donné, et dans le pire des cas, il faut $4 * C^i_{|\mathcal{R}|} * 2^i$ octets en mémoire. Pour cette raison, Brin et al. (1997) ne calculent que des corrélations entre deux valeurs d'une relation binaire. Etant donné un seuil $MinCor$ donné par l'utilisateur, Grahne et al. (2000) montrent que la contrainte $\chi^2(X) \geq MinCor$ est monotone. En conséquence, l'ensemble des règles obtenues forme un espace convexe, représenté par sa

bordure minimale L. La déduction d'une approximation de la valeur du χ^2, pour un motif appartenant à l'espace convexe, s'effectue à l'aide des valeurs du χ^2 des motifs de L inclus dans ce motif.

Par ailleurs, en micro-électronique, le domaine de l'APC[1] cherche à mettre en évidence des corrélations entre des paramètres liés à la production (produits, équipements, procédés, recettes, étapes, ...), de manière à pouvoir rectifier d'éventuelles dérives de la chaîne de fabrication. Nous cherchons à mettre en évidence des corrélations entre les valeurs d'une relation et celles d'une colonne cible (ici, le gain). C'est pourquoi, nous introduisons le concept de règles de corrélation décisionnelles. Celui-ci est une restriction des règles de corrélation contenant une valeur de l'attribut cible. Pour pouvoir calculer ces règles, (1) nous utilisons l'ordre lectique (Ganter et Wille, 1999) pour parcourir le treillis des parties ; (2) nous proposons le concept de vecteur de contingence : une nouvelle approche des tableaux de contingence ; nous montrons comment construire le vecteur de contingence d'un motif de cardinalité i à partir du vecteur de contingence d'un de ses sous-ensembles de cardinalité $i-1$ (ce qui est impossible avec les tableaux de contingence) ; et (3), nous tirons partie de l'ordre lectique, des vecteurs de contingence et des mécanismes de construction récurrents pour proposer l'algorithme LHS-CHI2. Enfin, nous menons des expérimentations sur des relations fournies par des industriels du domaine, et comparons nos résultats avec une approche par niveaux.

L'article est organisé comme suit : au paragraphe 2, nous rappelons le fondement des règles de corrélation et de l'ordre lectique. La section 3 décrit les concepts utilisés pour l'extraction des règles de corrélation décisionnelles ainsi que notre algorithme. Les évaluations expérimentales sont synthétisées au paragraphe 4. En conclusion, nous présentons le bilan de notre contribution et les perspectives de recherche.

2 Travaux antérieurs

Soit r une relation binaire (ou base de transactions) sur un ensemble de motifs à valeurs dans $\mathcal{R} = \mathcal{I} \cup C$. Dans notre approche, $\mathcal{I}$ représente les valeurs (ou motifs) de la relation binaire servant de critère d'analyse et $\mathcal{C}$ un attribut cible. Afin de simplifier les notations, nous introduisons le treillis de littéraux associé à un motif $X \subseteq \mathcal{R}$. Cet ensemble, noté $\mathbb{P}(X)$, contient l'ensemble des littéraux pouvant être construits à partir de X, et ayant pour cardinalité $|X|$; *i.e.* $\mathbb{P}(X) = \{Y\overline{Z}$ tel que $Y \subseteq X$ et $Z = X \backslash Y\}$. Le calcul de la valeur de la fonction χ^2 pour un motif X s'appuie sur son tableau de contingence dans lequel chaque case correspond à la fréquence du littéral $Y\overline{Z}$ qui lui est associée (nombre de transactions de la relation r contenant Y et n'ayant aucune valeur commune avec Z). Ce calcul s'effectue en deux étapes :

1. pour chaque cellule $Y\overline{Z}$ du tableau de contingence, nous mesurons la fréquence théorique de la partie positive du littéral en cas d'indépendance des 1-motifs (motifs de cardinalité 1) inclus dans Y : $E(Y) = |r| * \prod_{y \in Y} \frac{Supp(y)}{|r|}$

2. nous mesurons l'écart entre le carré du support réel de $Y\overline{Z}$ et son espérance, le tout divisé par l'espérance de $Y\overline{Z}$. Enfin, nous sommons toutes ces valeurs pour obtenir la valeur du χ^2 : $\chi^2(X) = \sum_{Y\overline{Z} \in \mathbb{P}(X)} \frac{(Supp(Y\overline{Z}) - E(Y))^2}{E(Y)}$

[1] Advanced Process Control : Contrôle Avancé des Procédés

Une règle de corrélation est représentée par un motif pour lequel la valeur de la fonction χ^2 est supérieure ou égale à un seuil $MinCor$ fourni à l'avance. Il existe une bijection entre les valeurs des centiles et celles de la distribution du χ^2 pour un seul degré de liberté (Spiegel, 1990), que nous utilisons afin d'établir le taux de corrélation des règles valides.

L'ordre lectique, noté $<_{lec}$, est un ordre total permettant l'énumération de tous les sous-ensembles de $\mathcal{R}$. Supposons que les 1-motifs sont totalement ordonnés et donc comparables deux à deux via un ordre noté $\preceq$. Soit X et $Y \subseteq \mathcal{R}$, alors nous avons : $X <_{lec} Y \Leftrightarrow max_{\preceq}(X \backslash (X \cap Y)) \preceq max_{\preceq}(Y \backslash (X \cap Y))$. La proposition suivante exprime le fait que l'ordre lectique est compatible avec les contraintes anti-monotones. En conséquence, nous pouvons modifier les algorithmes (Ganter et Wille, 1999; Laporte et al., 2002) d'énumération afin qu'ils prennent en compte une conjonction de contraintes anti-monotones.

Proposition 2.1 - Soit $X,\ Y$ deux motifs. Si $X \subset Y$, alors $X <_{lec} Y$ (Ganter et Wille, 1999).

3 Algorithme LHS-Chi2

Les vecteurs de contingence sont une autre représentation des tables de contingence. Contrairement à ces dernières, nous montrons que, pour un motif $X \cup A$ donné ($A \in \mathcal{R} \backslash X$), le calcul de son vecteur de contingence est possible à partir du vecteur de contingence de X et de la liste des identifiants des tuples de la relation contenant A.

Définition 3.1 (Classe d'équivalence associée à un littéral) - Soit $Y\overline{Z}$ un littéral. On note par $[Y\overline{Z}]$ la classe d'équivalence associée à $Y\overline{Z}$. Cette classe contient l'ensemble des identifiants des transactions de la relation qui contiennent Y et qui ne contiennent aucun 1-motif de Z ($i.e.$, $[Y\overline{Z}] = \{i \in Tid(r)$ tel que $Y \subseteq Tid(i)$ et $Z \cap Tid(i) = \emptyset\}$).

La proposition suivante assure qu'un identifiant d'une transaction ne peut appartenir qu'à une unique classe d'équivalence.

Proposition 3.1 - L'union des classes d'équivalence $[Y\overline{Z}]$ du treillis des littéraux associé à un motif $X \subseteq \mathcal{R}$ forme une partition (Laurent et Spyratos, 1988) des identifiants de la relation r.

Définition 3.2 (Vecteur de contingence) - Le vecteur de contingence d'un motif X, noté $VC(X)$, regroupe l'ensemble des classes d'équivalence des littéraux appartenant à $\mathbb{P}(X)$ ordonnées selon l'ordre lectique sur la partie positive des littéraux.

En conséquence de la proposition 3.1, pour un motif X donné, son vecteur de contingence est une représentation exacte de son tableau de contingence. Pour dériver ce tableau à partir de son vecteur de contingence, il suffit de calculer la cardinalité de chacune de ses classes d'équivalence. La proposition 3.2 montre comment calculer le vecteur de contingence du motif $X \cup A$ ($A \notin X$) à partir du vecteur de contingence de X et de la liste des identifiants des tuples de la relation contenant A.

Proposition 3.2 - Soit X un motif et $A \in \mathcal{R} \backslash X$ un 1-motif. Le vecteur de contingence du motif $X \cup A$ peut être calculé à partir des vecteurs de contingence de X et de A comme suit :

$$VC(X \cup A) = (VC(X) \cap [\overline{A}]) \cup (VC(X) \cap [A]) \quad (1)$$

$$\Leftrightarrow VC(X \cup A) = (VC(X) \cap (Tid(r) \backslash Tid(A))) \cup (VC(X) \cap Tid(A)) \quad (2)$$

Alg. 1 Algorithme LHS-CHI2.

Entrée : X et Y deux ensembles de motifs
Sortie : $\{motifs\ Z \subseteq X\ tq\ \chi^2(Z) \geq MinCor\}$
1: **si** $Y = \emptyset$ **et** $\exists c \in \mathcal{C} : c \in X$ **et** $\chi^2(X) \geq MinCor$ **alors**
2: **Afficher** X, $\chi^2(X)$
3: **fin si**
4: $A := max(Y)$
5: $Y := Y \backslash \{A\}$
6: LHS-CHI2(X,Y)
7: $Z := X \cup \{A\}$
8: **si** $\forall z \in Z, \exists W \in BD^+ : \{Z \backslash z\} \subseteq W$ **alors**
9: $VC(Z) := \{\emptyset\}$
10: **pour tout** $Y\overline{Z} \in \mathbb{P}(X)$ selon l'ordre lectique sur la partie positive **faire**
11: $VC(Z) := VC(Z) \cup ([Y\overline{Z}] \cap (Tid(r) \backslash (Tid(A))) \cup ([Y\overline{Z}] \cap Tid(A))$
12: **fin pour**
13: **si** $|Z| \leq MaxCard$ **et** $CtPerc(VC(Z), MinPerc, MinSup)$ **alors**
14: $BD^+ := max_{\subseteq}(BD^+ \cup Z)$
15: LHS-CHI2(Z,Y)
16: **fin si**
17: **fin si**

Nous introduisons le concept de règle de corrélation décisionnelle : une restriction des règles de corrélation, au sens où ne sont gardées que les règles contenant une valeur de l'attribut cible.

Définition 3.3 (Règle de corrélation décisionnelle) - Soit $X \subseteq \mathcal{R}$ un motif, et $MinCor$ un seuil donné. Nous disons que le motif X constitue une règle de corrélation décisionnelle valide si et seulement si : (i) X contient une valeur de l'attribut cible $\mathcal{C}$ et (ii) $\chi^2(X) \geq MinCor$.

L'algorithme Lectic Hybrid Search-Chi2, ou LHS-CHI2, permet d'extraire l'ensemble des règles de corrélation décisionnelles d'une relation r satisfaisant la contrainte de seuil $MinCor$ pour la fonction χ^2. Cet algorithme résulte d'une modification de l'algorithme LS (Laporte et al., 2002) à notre contexte. Cette adaptation permet la prise en compte des vecteurs de contingence ainsi que celle de plusieurs contraintes monotones et anti-monotones afin d'élaguer l'espace de recherche (Grahne et al., 2000). Les contraintes utilisées sont les suivantes : (i) une valeur de l'attribut cible doit être présente parmi les motifs extraits (contrainte monotone, ligne 1 de notre algorithme) ; (ii) la fonction χ^2 étant une fonction croissante, nous imposons une cardinalité maximale aux motifs à examiner, qui, usuellement, n'excède pas la valeur 8 (contrainte anti-monotone, ligne 13) ; (iii) afin d'obtenir des règles ayant une sémantique sur la relation, au moins $MinPerc\%$ des cases du tableau de contingence doivent avoir un support supérieur ou égal à $MinSup$. Cette contrainte s'exprime par le prédicat *CtPerc* qui prend trois paramètres : le vecteur de contingence, $MinPerc$ et $MinSup$ (contrainte anti-monotone, ligne 13). La proposition 2.1 justifie l'intégration de ces contraintes anti-monotones dans notre algorithme. Nous effectuons un élagage utilisant la bordure positive (Mannila et Toivonen, 1997) relative aux deux dernières contraintes afin de réduire l'espace de recherche. Pour ce faire, nous nous assurons que le motif Z, pouvant servir de paramètre au second appel récursif,

a tous ses sous-ensembles directs inclus dans un des éléments de la bordure positive (ligne 8). Par définition, nous avons $VC(\emptyset) = \{Tid(r), \emptyset\}$. La bordure positive est initialisée avec pour valeur l'ensemble $\{\emptyset\}$. Le premier appel récursif à LHS-CHI2 s'effectue avec pour paramètres $X = \emptyset,\ Y = \mathcal{R}$.

4 Evaluations Expérimentales

Les expérimentations ont été menées sur différents fichiers CSV de mesures numériques fournis par STMicroElectronics (STM) et ATMEL (ATM). Ces fichiers comportent entre 800 et 1500 colonnes, environ 300 lignes (pouvant ne pas comporter de valeur), et un ou plusieurs attributs cibles. Nous leur avons appliqué divers prétraitements puis une transformation (discrétisation des valeurs) pour obtenir une base de transactions, en utilisant des méthodes classiques. Les valeurs absentes sont gérées comme un *item* comme les autres, que nous ne cherchons pas à corréler. Ensuite, nous avons comparé les temps de réponse rendus par LHS-CHI2 à ceux obtenus avec un algorithme par niveaux (LEVELWISE) standard, et basés tous deux sur les mêmes contraintes de monotonie et d'anti-monotonie introduites ci-dessus. Les schémas des figures 1 et 2 montrent l'évolution des temps d'exécution des deux méthodes lorsque $MinSup$ varie tandis que $MinPerc$ et $MinCor$ sont fixes (avec agrandissement des résultats pour des sous intervalles de $MinSup$). On constate que LHS-CHI2 fournit des temps de réponse entre 30% et 70% meilleurs que LEVELWISE. Notez enfin que la plupart des jeux d'essais utilisés produisent des règles de corrélation décisionnelles de cardinalité 4.

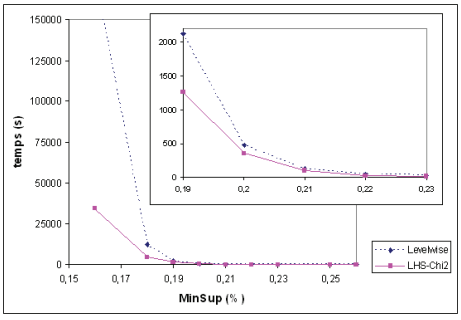

FIG. 1: Durée d'exécution avec $MinPerc = 0.34$, $MinCor = 1.6$ (fichier STM - cible1).

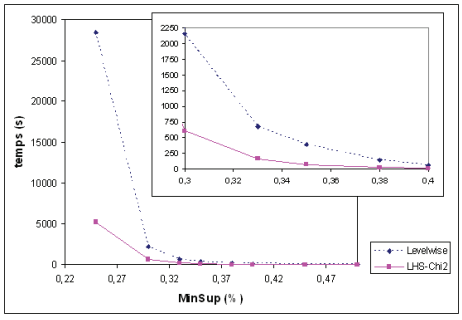

FIG. 2: Durée d'exécution avec $MinPerc = 0.24$, $MinCor = 2.8$ (fichier ATM - cible2).

5 Conclusion

Cet article introduit deux concepts : (i) les règles de corrélation décisionnelles, une restriction des règles de corrélation comportant une valeur d'un attribut cible, et (ii) les vecteurs de contingence, une autre représentation des tables de contingence. Nous avons proposé un algorithme basé sur l'ordre lectique pour parcourir le treillis des parties d'un ensemble de motifs.

Cet algorithme utilise la propriété d'inférence d'un vecteur de contingence d'un motif à partir de celui d'un de ses sous-ensembles direct. Les expérimentations menées montrent que la méthode proposée permet le calcul des règles en des temps inférieurs à ceux d'un algorithme par niveaux. Notre approche a permis de découvrir de nouvelles corrélations entre les paramètres des fichiers fournis : un quart environ des corrélations déterminées par le premier test n'étaient ainsi pas connues de STM, et la quasi-totalité des résultats obtenus ont été validés par nos partenaires manufacturiers. Nous travaillons actuellement (i) à optimiser les étapes de prétraitements avec préservation du contexte, en vue d'obtenir des règles plus significatives ; et (ii) à calculer des règles de corrélation sur des littéraux.

Références

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, et A. I. Verkamo (1996). Fast Discovery of Association Rules. In *Advances in Knowledge Discovery and Data Mining*, pp. 307–328.

Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : generalizing association rules to correlations. In *Proceedings of the International Conference on Management of Data, SIGMOD*, pp. 265–276.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis : Mathematical Foundations*. Springer.

Grahne, G., L. Lakshmanan, et X. Wang (2000). Efficient Mining of Constrained Correlated Sets. In *Proceedings of the 16th International Conference on Data Engineering, ICDE*, pp. 512–524.

Laporte, M., N. Novelli, R. Cicchetti, et L. Lakhal (2002). Computing full and iceberg datacubes using partitions. In *Proceedings of the 13rd International Symposium on Methodologies for Intelligent Systems, ISMIS*, pp. 244–254.

Laurent, D. et N. Spyratos (1988). Partition semantics for incomplete information in relational databases. In *Proceedings of the International Conference on Management of Data, SIGMOD*, pp. 66–73.

Mannila, H. et H. Toivonen (1997). Levelwise Search and Borders of Theories in Knowledge Discovery. In *Data Mining and Knowledge Discovery*, Volume 1(3), pp. 241–258.

Spiegel, M. R. (1990). *Théorie et applications de la statistique*. Schaum.

Wu, X., C. Zhang, et S. Zhang (2004). Efficient mining of both positive and negative association rules. *ACM Trans. Inf. Syst. 22*(3), 381–405.

Summary

In this paper, we introduce two concepts: decision correlation rules and contingency vectors. The first one results from a cross fertilization between correlation and decision rules. It makes it possible to highlight relevant links between sets of patterns of a binary relation and the values of a target attribute on the twofold basis of the Chi-squared measure and on the extracted patterns support. Due to the very nature of the problem, levelwise algorithms only allow extraction of results with high execution times and huge memory occupation. To offset these two problems, we propose an algorithm based on lectical order and contingency vectors.