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