Découverte d'associations quantitatives générales et atypiques

Sylvie Guillaume*, Leïla Nemmiche-Alachaher*, Michel Schneider*

* Laboratoire LIMOS, UMR 6158 CNRS, Université Blaise Pascal
Complexe scientifique des Cézeaux, 63177 AUBIERE Cedex - France
{sylvie.guillaume, nemmiche, michel.schneider}@isima.fr

Résumé. Dans ce papier, nous proposons un nouveau type d'association mettant en jeu au moins une variable quantitative. Il s'agit de rechercher dans une population d'individus, les catégories qui s'écartent significativement du comportement normal de cette population. Plus exactement nous recherchons les catégories d'individus qui sont sur-représentées ou au contraire sous-représentées pour les fortes ou les faibles valeurs de la variable cible. Pour caractériser l'association, nous avons utilisé et étendu une mesure existante, l'intensité d'inclination. Comme toutes les catégories d'individus (*ou associations de variables*) ne sont pas d'égal intérêt pour l'utilisateur, ce type de connaissance permet, dans un premier temps, d'avoir une vision globale de l'association. Dans un deuxième temps il est possible de rechercher les intervalles de valeurs pour lesquels les écarts interviennent. Cette recherche des intervalles s'appuie sur les tableaux de contingence des écarts entre la situation observée et la situation attendue.

1 Introduction

La découverte des règles d'association Agrawal et al. (1993), Mannila et al. (1994) et Agrawal et al. (1996)) repose généralement sur deux mesures de fréquence (*support et confiance*) qui sont peu adaptées dans le cas d'analyses de données quantitatives. Il n'est pas possible de segmenter directement les données et une étape préliminaire de discrétisation est nécessaire.

Srikant et Agrawal (1996) ont proposé une technique basée sur une discrétisation automatique des données en maîtrisant la perte d'information engendrée par cette étape de préparation. Cependant Bay (2001a) et Ludl et Widmer (2000) ont montré qu'une discrétisation selon uniquement la distribution de la variable, sans tenir compte du contexte, peut conduire à des solutions non optimales. Mehta et Parthasarathy (2005) ont donc proposé une discrétisation qui prend en compte la distribution de chacune des variables mises en jeu. Lee et al. (2004) proposent une méthode augmentant la confiance accordée dans les motifs fréquents obtenus après discrétisation des variables quantitatives. Quant à Kuok et al. (1998), Zhang (1999) et Subramanyam et Goswami (2006), ils utilisent la technique des ensembles flous et Miller et Yang (1997) et Tong et al. (2005) utilisent des clusters pour la recherche des règles d'association quantitatives. D'autres auteurs optimisent