Mesure de la qualité des règles d'association par l'intensité d'implication entropique

Julien Blanchard, Pascale Kuntz, Fabrice Guillet, Régis Gras

IRIN – Ecole polytechnique de l'université de Nantes
La Chantrerie – BP 50609 – 44306 Nantes cedex 3
{julien.blanchard, pascale.kuntz, fabrice.guillet, regis.gras}@polytech.univ-nantes.fr

Résumé. Le filtrage de l'information pertinente par des mesures de qualité reste l'une des étapes les plus délicates d'un processus d'extraction de règles d'association. Afin de prendre en compte la taille du jeu de données, contrairement à l'indice traditionnel de confiance, et de souligner le caractère naturellement asymétrique de la notion d'implication, Gras a défini l'intensité d'implication qui mesure l'étonnement statistique des règles découvertes. Cependant, comme de nombreuses autres mesures dans la littérature, cette dernière ne tire pas profit de la contraposée (*non b*) $\Rightarrow$ (*non a*) qui permet pourtant de renforcer l'affirmation de la relation implicative de *a* sur *b*. Nous introduisons ici une extension de l'intensité d'implication qui exploite l'entropie de Shannon pour quantifier les déséquilibres entre exemples et contre-exemples à la fois pour la règle et sa contraposée. Ce nouvel indice est construit de façon à mieux mesurer la notion complexe qu'est la qualité en intégrant simultanément étonnement statistique et qualité inclusive. Des comparaisons numériques avec la confiance et l'indice de Loevinger sont effectuées sur des jeux de données synthétiques et sur des données réelles de domaines variés allant des ressources humaines aux pannes d'ascenseurs. Les distributions statistiques des indices sur les corpus de règles montrent que notre mesure fait preuve d'une forte capacité de filtrage.

Mots-clés : fouille de données, règles d'association, mesures de qualité des règles, étonnement statistique, qualité inclusive, contraposée, entropie de Shannon

1. Introduction

Les règles d'association de la forme $a \Rightarrow b$ sont devenues un concept majeur en fouille de données pour représenter les relations quasi-implicatives entre des variables booléennes (dénommées items). Depuis les premiers travaux d'Agrawal *et al.* (1993), de nombreux algorithmes ont été proposés pour découvrir efficacement de telles connaissances dans de grandes bases de données. Tous engendrent d'énormes quantités de règles, dues à l'explosion combinatoire du nombre de conjonctions d'items traitées. Afin que l'utilisateur-décideur puisse mettre en évidence un ensemble restreint de relations pertinentes et intelligibles pour la prise de décisions, il est nécessaire de mesurer la qualité des règles extraites. Dans l'importante littérature consacrée à l'évaluation de cette notion complexe de qualité, les mesures sont souvent classées en deux catégories : les subjectives (orientées décideur) et les objectives (orientées données). Les mesures subjectives prennent en compte les objectifs du