

Catégorisation des mesures d'intérêt pour l'extraction des connaissances

Sylvie Guillaume^{*,***}, Dhouha Grissa^{**,***,****} et Engelbert Mephu Nguifo^{**,***}

Clermont Université, Université d'Auvergne^{*} et Université Blaise Pascal^{**}, LIMOS,
BP 10448, F-63000 Clermont-Ferrand
CNRS^{***}, UMR 6158, LIMOS, F-63173 AUBIERE
URPAH^{****}, Département d'Informatique, Faculté des Sciences de Tunis, Campus
Universitaire, 1060 Tunis, Tunisie
guillaum@isima.fr, dgrissa@isima.fr, mephu@isima.fr

Résumé. La recherche de règles d'association intéressantes est un domaine de recherche important et actif en fouille de données. Les algorithmes de la famille *Apriori* reposent sur deux mesures pour extraire les règles, le support et la confiance. Bien que ces deux mesures possèdent des vertus algorithmiques accélératrices, elles génèrent un nombre prohibitif de règles dont la plupart sont redondantes et sans intérêt. Il est donc nécessaire de disposer d'autres mesures filtrant les règles inintéressantes. Des travaux ont été réalisés pour dégager les "bonnes" propriétés des mesures d'extraction des règles et ces propriétés ont été évaluées sur 61 mesures. L'objectif de cet article est de dégager des catégories de mesures afin de répondre à une préoccupation des utilisateurs : le choix d'une ou plusieurs mesures lors d'un processus d'extraction des connaissances dans le but d'éliminer les règles valides non pertinentes extraites par le couple (*support*, *confiance*). L'évaluation des propriétés sur les 61 mesures a permis de dégager 9 classes de mesures, classes obtenues grâce à deux techniques : une méthode de la classification ascendante hiérarchique et une version de la méthode de classification non-hiérarchique des *k*-moyennes.

1 Introduction

Les algorithmes d'extraction de règles d'association (Agrawal et Srikant 1994), fondés sur les mesures *support* et *confiance*, ont tendance à générer un nombre important de règles. Ces deux mesures ne sont pas suffisantes pour extraire uniquement les règles réellement intéressantes et ce constat a été mis en évidence dans de nombreux travaux comme par exemple (Sese et Morishita 2002, Carvalho et al. 2005). Une étape supplémentaire d'analyse des règles extraites est donc indispensable et différentes solutions ont été proposées. Une première solution consiste à restituer facilement et de façon synthétique l'information extraite grâce à des techniques de représentation visuelle (Hofmann et Wilhelm 2001, Blanchard et al. 2003). Une seconde voie consiste à réduire le nombre de règles extraites. Certains auteurs (Zaki 2000, Zaman Ashrafi et al., 2004, Ben Yahia et al. 2009) éliminent les règles