Analyse du comportement limite d'indices probabilistes pour une sélection discriminante

Sylvie Guillaume* et Israël-César Lerman**

*Clermont Université, Université d'Auvergne, LIMOS, BP 10448, F-63000 Clermont-Fd
sylvie.guillaume@isima.fr
**Irisa, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex
lerman@irisa.fr

Résumé. Nous étudions ici le comportement de deux types d'indices probabilistes discriminants en présence de données dont le volume va en croissant. À cet égard, un modèle spécifique de croissance de la taille des données et de liaison entre variables est mis en œuvre et celui-ci va permettre de déterminer le comportement limite des différents indices quel que soit le niveau de liaison entre la prémisse et la conclusion de la règle donnée. La clarté des résultats obtenus nous conduit à en chercher l'explication formelle. L'expérimentation a été effectuée avec la base de données UCI *Wages*.

1 Introduction

L'extraction des règles d'association (Agrawal et Srikant 1994) est un domaine largement étudié dans la communauté "*extraction des connaissances*". Une règle d'association est une implication du type $a \rightarrow b$ où a et b sont des conjonctions de variables binaires disjointes. Afin de déterminer les règles intéressantes, deux indices[1] sont communément utilisés : (1) le support $p(a \wedge b)$ (*ou taux de couverture*) qui mesure la fréquence d'apparition de la règle et indique le pourcentage d'individus contenant toutes les variables de la règle, (2) la confiance $p(b/a)$ (*ou probabilité conditionnelle*) qui mesure la force de la règle et évalue le pourcentage d'individus vérifiant la conclusion b parmi ceux qui vérifient la prémisse a. Une règle sera dite valide si les valeurs prises par le couple de mesures sont supérieures à deux seuils fixés par l'utilisateur : le support minimum sup_{min} et la confiance minimum $conf_{min}$. De nombreux auteurs comme par exemple Sese et Morishita (Sese et Morishita 2002) ont montré les faiblesses de ce couple de mesures qui valide des règles qui ne sont pas toujours pertinentes. De nombreux indices ont été proposés dans la littérature pour palier les faiblesses de ces deux mesures. Le lecteur pourra consulter les articles de synthèse (Tan et al. 2002, Lallich et Teytaud 2004, Geng et Hamilton 2007, Feno 2007, Vaillant 2007 et Guillaume et al. 2010) mettant en évidence non seulement les nombreux indices d'intérêt, mais aussi les propriétés de ces mesures d'intérêt, afin d'aider l'utilisateur dans le choix d'une ou plusieurs mesures complémentaires capables d'éliminer les règles valides non pertinentes. Une famille d'indices, celle reposant sur une échelle probabiliste, a montré tout son intérêt puisqu'elle est capable d'éliminer certains types de règles valides inintéressantes. Fleury (Fleury, 1996) et Guillaume (Guillaume, 2000) révèlent les types de règles éliminées. L'indice fondateur de

[1] ou mesures.