Comparaison entre deux indices pour l'évaluation probabiliste discriminante des règles d'association

Israël-César Lerman*, Sylvie Guillaume**

*Irisa - Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cédex
lerman@irisa.fr,
**Clermont Université, Auvergne, LIMOS, BP 10448, F-63000 Clermont-Fd
guillaum@isima.fr

Résumé. L'élaboration d'une échelle de probabilité discriminante pour la comparaison mutuelle entre plusieurs attributs observés sur un échantillon d'objets de "grosse" taille, nécessite une normalisation préalable. L'objet de cet article est l'analyse comparée entre deux approches. La première dérive de l"'Analyse de la Vraisemblance des Liens Relationnels *Normalisée*". La seconde est fondée sur la notion de "Valeur Test" sur un échantillon *virtuel* de taille 100, synthétisant l'échantillon initial.

1 Introduction

Relativement à une base de données où on distingue un ensemble $\mathcal{A}$ d'attributs booléens observés sur un ensemble $\mathcal{O}$ d'entités (objets, individus, ...), le problème fondamental et bien connu en "Fouille des Données" ("Data Mining") est de pouvoir inférer un ensemble significatif et exploitable de règles d'association, on dit encore d'implications, entre attributs. Pour $(a, b) \in \mathcal{A} \times \mathcal{A}$, une règle de la forme $a \rightarrow b$ où a et b sont des attributs singletons, est un cas particulier d'une règle d'association. Intuitivement elle signifie que si a est à $VRAI$ sur un élément de $\mathcal{O}$, alors généralement - mais sans que cela soit un absolu - b est également à $VRAI$ sur cet élément de $\mathcal{O}$. Pour détecter de telles associations orientées, il importe de disposer d'un indice (on dit encore "coefficient" ou "mesure") statistique pertinent d'implication qui permette de dégager des "règles d'association" *intéressantes* ; c'est - à - dire, qui augmentent notre connaissance du réseau des tendances causales entre attributs de $\mathcal{A}$. Dans les travaux de Agrawal et al. (1993) qui ont lancé ces recherches dans le domaine de la "Fouille des Données" le souci d'une *bonne* mesure ne se manifestait pas encore. Il transparaît clairement dans Tan et al. (2002) où différents critères sont considérés pour organiser un ensemble de coefficients définissant des indices d'implication. D'autres études comparatives avec des facettes différentes sont également considérées dans Lallich et Teytaud (2004); Lenca et al. (2004). La mise à contribution de la notion d'indépendance statistique entre attributs intervient dans l'élaboration de nombreux coefficients Lallich et Teytaud (2004); Lenca et al. (2004); Lerman et Azé (2007); Piatetsky-Shapiro (1991); Tan et al. (2002). Historiquement, l'élaboration d'une échelle de probabilité pour éprouver l'existence d'un lien entre *deux* attributs descriptifs a été établie dans l'optique des tests d'hypothèses statistiques. L'adaptation au problème