Critère VT100 de sélection des règles d'association

Alain Morineau*, Ricco Rakotomalala**

*MODULAD, Paris
alain.morineau@modulad.fr
http://www.modulad.fr
**Laboratoire ERIC – Université Lyon 2
ricco.rakotomalala@univ-lyon2.fr
http://eric.univ-lyon2.fr/~ricco

Résumé. L'extraction de règles d'association génère souvent un grand nombre de règles. Pour les classer et les valider, de nombreuses mesures statistiques ont été proposées ; elles permettent de mettre en avant telles ou telles caractéristiques des règles extraites. Elles ont pour point commun d'être fonction croissante du nombre de transactions et aboutissent bien souvent à l'acceptation de toutes les règles lorsque la base de données est de grande taille. Dans cet article, nous proposons une mesure inspirée de la notion de valeur-test. Elle présente comme principale caractéristique d'être insensible à la taille de la base, évitant ainsi l'écueil des règles fallacieusement significatives. Elle permet également de mettre sur un même pied, et donc de les comparer, des règles qui auront été extraites de bases de données différentes. Elle permet enfin de gérer différents seuils de signification des règles. Le comportement de la mesure est détaillé sur un exemple.

1 Introduction

1.1 Les valeurs-tests

Pour faire un test de l'hypothèse nulle H_0, le statisticien calcule une « *probabilité critique* » (ou *p-value*). C'est la probabilité, calculée sous H_0, d'un événement au moins aussi extrême que l'événement observé. De façon intuitive, on comprend que cette probabilité est d'autant plus faible qu'on est loin de l'hypothèse nulle. Si l'événement observé est très improbable sous l'hypothèse nulle, on jugera que les observations sont vraisemblablement régies par un mécanisme *non nul*. Il est donc tentant d'utiliser cette valeur numérique pour évaluer l'écart entre ce qu'on a observé et la situation « sans intérêt » correspondant à ce qu'on aurait observé sous H_0. Dans ce contexte, plus l'évaluation de l'écart est forte (plus la *probabilité critique* est faible), plus ce qu'on a observé est *intéressant* (Gras *et al.*, 2002 ; Lerman et Azé, 2003 ; Lallich et Teytaud, 2004). Dans la pratique, on se rend compte que la p-value est difficile à manipuler ; elle peut atteindre des valeurs très faibles, très peu lisibles ; pire, dans certains cas, elle est inutilisable car on se heurte aux limites de l'approximation