Extraction de « pépites » de connaissance dans les données : une nouvelle approche et une étude de la sensibilité au bruit

Jérôme Azé, Yves Kodratoff

CNRS, Laboratoire de Recherche en Informatique
Bât. 490, Université. Paris-Sud
91405 Orsay Cedex - France
{aze,yk}@lri.fr

Résumé. La plupart des méthodes permettant d'extraire des règles d'association dans les données sont basées sur l'utilisation de mesures et de seuils prédéfinis par l'expert pour optimiser la recherche des règles. Le choix des seuils permettant de séparer les règles intéressantes des règles triviales est difficile, même pour un expert du domaine étudié.

Si l'on considère que les données sont bruitées et que l'extraction d'informations du type « pépites » de connaissance (c'est-à-dire, règles ayant un faible support) peut intéresser l'expert, alors les méthodes classiques sont souvent mises en défaut dans de telles situations.

Nous proposons donc une nouvelle mesure d'extraction des règles d'association, appelée « moindre-contradiction ». Nous montrons que cette mesure (i) permet d'extraire des « pépites » de connaissance dans les données, sans pour autant être submergés par le nombre de règles ayant un faible support, (ii) se comporte légèrement mieux que les mesures classiques étudiées lorsque les données sont bruitées.

Mots Clés : règles d'association, mesures de qualité, bruit.

1 Introduction

La découverte non supervisée de règles d'association dans les bases de données est particulièrement intéressante et de nombreux travaux ont été effectués pour caractériser les motifs cachés dans les bases de données (Agrawal et al. 1993, Agrawal et Srikant 1995, Brin et al. 1997, Lavrač et al. 1999, Sahar 1999, Gras et al. 2001). Les effets du bruit et le comportement de ces algorithmes ont été peu étudiés, bien que les bases de données réelles soient rarement parfaites. Nous pensons qu'il est intéressant d'étudier le comportement des algorithmes d'extractions de règles d'association lorsque les données contiennent entre 1% et 10% de bruit. Il nous semble qu'au delà de 10% de bruit, les données sont trop imparfaites pour pouvoir être correctement analysées.

Ayant fait cette constatation, nous proposons, dans cet article, une mesure d'intérêt qui est moins sensible au bruit que la plupart des mesures classiques (que nous avons étudiées).
Comme nous l'avons déjà montré dans (Azé et Kodratoff 2002a, Azé et Kodratoff 2002b), de nombreuses mesures d'intérêt classiques, comme le support, la confiance