ARQAT : plateforme exploratoire pour la qualité des règles d'association

Xuan-Hiep Huynh*, Fabrice Guillet*, Henri Briand*
*LINA CNRS FRE 2729 - Ecole polytechnique de l'université de Nantes
La Chantrerie, BP 50609, 44306 Nantes Cedex 3, France
{xuan-hiep.huynh, fabrice.guillet, henri.briand}@polytech.univ-nantes.fr

Résumé. Le choix de mesures d'intérêt pour la validation des règles d'association constitue un défi important dans le contexte de l'évaluation de la qualité en fouille de données. Mais, comme l'intérêt dépend à la fois de la structure des données et des buts de l'utilisateur (décideur, analyste), certaines mesures peuvent s'avérer pertinentes dans un contexte donné, et ne plus l'être dans un autre. Dans cet article, nous proposons un outil original ARQAT afin d'étudier le comportement spécifique de 34 mesures d'intérêt dans le contexte d'un jeu de règles, selon une approche résolument exploratoire mettant en avant l'interactivité et les représentations graphiques.

1 Introduction

L'étude et la conception de mesures d'intérêt (MI) adaptées aux règles d'association constitue un important défi pour l'évaluation de la qualité des connaissances en ECD. Les règles d'association (Agrawal et al. 1993) proposent un modèle non supervisé pour la découverte de tendances implicatives dans les données. Malheureusement, en phase de validation, l'utilisateur (expert des données, ou analyste) se trouve confronté à un problème majeur : une grande quantité de règles parmi lesquelles il doit isoler les meilleures en fonction de ses préférences. Une manière de réduire le coût cognitif de cette tâche consiste à le guider à l'aide de mesures d'intérêt adaptées à la fois à ses préférences et à la structure des données étudiées.

Les travaux précurseurs sur les règles d'association (Agrawal et al. 1993) (Agrawal et Srikant 1994) proposent l'utilisation de 2 mesures statistiques : le support et la confiance. Ce couple de mesures dispose de vertus algorithmiques accélératrices, mais n'est pas suffisant pour capter l'intérêt des règles. Afin de compenser cette limite, de nombreuses mesures complémentaires ont été proposées dans la littérature et dissociées en 2 groupes (Freitas 1999) : les mesures objectives et les mesures subjectives. Les mesures subjectives dépendent essentiellement des buts, connaissances, croyances de l'utilisateur qui doivent être préalablement recueillis. Elles sont associées à des algorithmes supervisés ad hoc (Padmanabhan et Tuzhilin 1998) (Liu et al. 1999) permettant de n'extraire que les règles conformes ou au contraire en contradiction avec les croyances de l'utilisateur, et ainsi d'orienter la notion d'intérêt vers la nouveauté (novelty) ou l'inattendu (unexpectedness). Les mesures objectives, quant à elles, sont des mesures statistiques s'appuyant sur la structure des données ou plus exactement la fréquence des combinaisons fréquentes d'attributs (itemsets). De nombreux travaux de synthèse récapitulent et comparent leurs définitions et leurs propriétés (Bayardo et Agrawal 1999) (Hilderman et Hamilton 2001) (Tan et al. 2002) (Tan et al. 2004)