Modèle décisionnel basé sur la qualité des données pour sélectionner les règles d'associations légitimement intéressantes

Laure Berti-Équille
IRISA, Campus Universitaire de Beaulieu,
35042 Rennes, France
berti@irisa.fr

Résumé. Dans cet article nous proposons d'exploiter des mesures décrivant la qualité des données pour définir la qualité des règles d'associations résultant d'un processus de fouille. Nous proposons un modèle décisionnel probabiliste basé sur le coût de la sélection de règles légitimement, potentiellement intéressantes ou inintéressantes si la qualité des données à l'origine de leur calcul est bonne, moyenne ou douteuse. Les expériences sur les données de KDD-CUP-98 montrent que les 10 meilleures règles sélectionnées d'après leurs mesures de support et confiance ne sont intéressantes que dans le cas où la qualité de leurs données est correcte voire améliorée.

1 Introduction

La qualité des règles d'association est généralement évaluée par des mesures d'intérêt (classiquement le support et la confiance) et de nombreuses autres mesures ont été proposées (Tan *et al.*, 2002). Mais, on peut légitimement se demander quel est l'intérêt de telles règles, notées LHS → RHS, si 30 % des données de LHS sont obsolètes, 20% des données de RHS sont imprécises, et 15% des données de LHS proviennent d'une source réputée peu fiable. La thèse défendue dans cet article est que les mesures d'intérêt pour la découverte de règles d'associations ne sont pas autosuffisantes pour représenter effectivement la qualité des règles. Des mesures décrivant la qualité des données à partir desquelles sont calculées les règles doivent être intégrées au processus de découverte, de même que le coût d'une décision de choisir (ou non) ces règles « supposées intéressantes » doit être également considéré. Ceci a motivé donc nos travaux que nous formalisons dans les sections suivantes.

2 Caractérisation de la qualité des règles d'association à partir de la qualité des données d'origine

Soit I un ensemble d'items. Une règle d'association R est une implication de la forme: $LHS \rightarrow RHS$ où $LHS \subseteq I$, $RHS \subseteq I$ et $LHS \cap RHS = \varnothing$. LHS et RHS sont des conjonctions de variables telles que l'extension de LHS est : $g(LHS) = x_1 \wedge x_2 \wedge \ldots \wedge x_n$ et l'extension de Y est $g(RHS) = y_1 \wedge y_2 \wedge \ldots \wedge y_{n'}$. Soit j ($j=1, 2, \ldots, k$) une dimension décrivant un aspect de la qualité des données (*i.e.*, complétude, fraîcheur, précision, cohérence, crédibilité, etc.). Soit $q_j(I_i) \in [min_{ij}, max_{ij}]$ le score de la dimension de qualité j pour le sous-ensemble de données I_i