Sélection des variables informatives pour l'apprentissage supervisé multi-tables

Dhafer Lahbib*,** Marc Boullé*, Dominique Laurent**

*France Télécom R&D - 2, avenue Pierre Marzin, 23300 Lannion
dhafer.lahbib@orange-ftgroup.com
marc.boulle@orange-ftgroup.com
**ETIS-CNRS-Universite de Cergy Pontoise-ENSEA, 95000 Cergy Pontoise
dominique.laurent@u-cergy.fr

Résumé. Dans la fouille de données multi-tables, les données sont représentées sous un format relationnel dans lequel les individus de la table cible sont potentiellement associés à plusieurs enregistrements dans des tables secondaires en relation un-à-plusieurs. La plupart des approches existantes opèrent en transformant la représentation multi-tables, notamment par mise à plat. Par conséquent, on perd la représentation initiale naturellement compacte mais également on risque d'introduire des biais statistiques. Notre approche a pour objectif d'évaluer l'informativité des variables explicatives originelles par rapport à la variable cible dans le contexte des relations un-à-plusieurs. Elle consiste à résumer l'information contenue dans chaque variable par un tuple d'attributs représentant les effectifs des modalités de celle-ci. Des modèles en grilles multivariées sont alors employés pour qualifier l'information apportée conjointement par les nouveaux attributs, ce qui revient à une estimation de densité conditionnelle de la variable cible connaissant la variable explicative en relation un-à-plusieurs. Les premières expérimentations sur des bases de données artificielles et réelles montrent qu'on arrive à identifier les variables explicatives potentiellement pertinentes sur tout le domaine relationnel.

1 Introduction

Tandis que dans les méthodes de fouille de données classiques, les données sont stockées dans une seule table, la *Fouille de données mutli-tables* (en anglais, Multi-Relational Data Mining, MRDM) s'intéresse à l'extraction de connaissances à partir de bases de données relationnelles multi-tables (Knobbe et al., 1999). Typiquement, en MRDM les individus sont contenus dans une table *cible* en relation un-à-plusieurs avec des *tables secondaires*. En apprentissage supervisé, un *attribut cible* devrait être défini au sein de la table cible ce qui correspond à la *variable à expliquer* par analogie au cas mono-table.

Pour prendre en compte les relations un-à-plusieurs, la plupart des méthodes MRDM opèrent en transformant la représentation relationnelle : dans le paradigme de la Programmation Logique Inductive ILP (Džeroski, 1996), les données sont recodées sous la forme de