Un Critère d'évaluation pour la construction de variables à base d'itemsets 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 le contexte de la fouille de données multi-tables, les données sont représentées sous un format relationnel dans lequel les individus de la table cible sont potentiellement liés à plusieurs enregistrements dans des tables secondaires en relation un-à-plusieurs. Dans cet article, nous proposons un Framework basé sur des itemsets pour la construction de variables à partir des tables secondaires. L'informativité de ces nouvelles variables est évaluée dans le cadre de la classification supervisée au moyen d'un critère régularisé qui vise à éviter le sur-apprentissage. Pour ce faire, nous introduisons un espace de modèles basés sur des itemsets dans la table secondaire ainsi qu'une estimation de la densité conditionnelle des variables construites correspondantes. Une distribution a priori est définie sur cet espace de modèles, pour obtenir ainsi un critère sans paramètres permettant d'évaluer la pertinence des variables construites. Des expérimentations préliminaires montrent la pertinence de l'approche.

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, une *variable cible* devrait être définie au sein de la table cible. La nouveauté en MRDM est de considérer les variables se trouvant dans les tables secondaires (*variables secondaires*) pour prédire la classe. Plusieurs solutions ont été proposée dans la littérature, notamment la Programmation Logique Inductive PLI (Džeroski, 1996) qui utilise le formalisme logique ou encore la propositionalisation qui opèrent par mise à plat afin de pouvoir utiliser un classifieur monotable classique (Kramer et al., 2001).

Dans cet article, nous introduisons un espace de modèles basé sur des itemsets de variables secondaires. Ces itemsets permettent de construire de nouvelles variables binaires dans les tables secondaires. Ensuite nous évaluons la pertinence de ces variables pour la tâche de classification supervisée. Afin de prendre en compte le risque de sur-apprentissage, qui augmente