Prétraitement Supervisé des Variables Numériques pour la Fouille de Données 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

Le prétraitement des variables numériques dans le contexte de la fouille de données multi-tables diffère de celui des données classiques individu-variable. La difficulté vient principalement des relations un-à-plusieurs où les individus de la table cible sont potentiellement associés à plusieurs enregistrements dans des tables secondaires. Dans cet article, nous décrivons une méthode de discrétisation des variables numériques situées dans des tables secondaires. Nous proposons un critère qui évalue les discrétisations candidates pour ce type de variables. Nous décrivons un algorithme d'optimisation simple qui permet d'obtenir la meilleure discrétisation en intervalles de fréquence égale pour le critère proposé. L'idée est de projeter dans la table cible l'information contenue dans chaque variable secondaire à l'aide d'un vecteur d'attributs (un attribut par intervalle de discrétisation). Chaque attribut représente le nombre de valeurs de la variable secondaire appartenant à l'intervalle correspondant. Ces attributs d'effectifs sont conjointement partitionnés à l'aide de modèles en grille de données afin d'obtenir une meilleure séparation des valeurs de la classe. Des expérimentations sur des jeux de données réelles et artificielles révèlent que l'approche de discrétisation permet de découvrir des variables secondaires pertinentes.

Summary

In Multi-Relational Data Mining (MRDM), data are represented in a relational form where the individuals of the target table are potentially related to several records in secondary tables in one-to-many relationship. Variable pre-processing (including discretization and feature selection) within this multiple table setting differs from the attribute-value case. Besides the target variable information, one should take into account the relational structure of the database. In this paper, we focus on numerical variables located in a non target table. We propose a criterion that evaluates a given discretization of such variables. The idea is to summarize for each individual the information contained in the secondary variable by a feature tuple (one feature per interval of the considered discretization). Each feature represents the number of values of the

secondary variable ranging in the corresponding interval. These count features are jointly partitioned by means of data grid models in order to obtain the best separation of the class values. We describe a simple optimization algorithm to find the best equal frequency discretization with respect to the proposed criterion. Experiments on a real and artificial data sets reveal that the discretization approach helps one to discover relevant secondary variables.