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.

1 Introduction

La plupart des algorithmes de fouille de données existants se basent sur une représentation des données de la forme attribut-valeur. Dans ce format dit à plat, chaque enregistrement représente un individu et les colonnes représentent les variables décrivant ces individus. Dans les applications réelles, les données présentent souvent une structure interne difficile à exprimer sous un format tabulaire. Cette structure peut être naturellement décrite à l'aide du formalisme relationnel dans lequel chaque objet est relié à un ou plusieurs enregistrements dans d'autres tables (tables secondaires) à l'aide de relations de clés étrangères.

Exemple 1. *Prenons par exemple un problème CRM (Customer Relationship Management). La Figure 1 représente un extrait du schéma relationnel d'une base de données fictive d'une*