Vers une Automatisation de la Construction de Variables pour la Classification Supervisée

Marc Boullé *, Dhafer Lahbib *

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
marc.boulle@orange.com,
http://perso.rd.francetelecom.fr/boulle/

Résumé. Dans cet article, nous proposons un cadre visant à automatiser la construction de variables pour l'apprentissage supervisé, en particulier dans le cadre multi-tables. La connaissance du domaine est spécifiée d'une part en structurant les données en variables, tables et liens entre tables, d'autre part en choisissant des règles de construction de variables. L'espace de construction de variables ainsi défini est potentiellement infini, ce qui pose des problèmes d'exploration combinatoire et de sur-apprentissage. Nous introduisons une distribution de probabilité a priori sur l'espace des variables constructibles, ainsi qu'un algorithme performant de tirage d'échantillons dans cette distribution. Des expérimentations intensives montrent que l'approche est robuste et performante.

1 Introduction

Dans un projet de fouille de données, la phase de préparation des données vise à extraire une table de données pour la phase de modélisation (Pyle, 1999), (Chapman et al., 2000). La préparation des données est non seulement coûteuse en temps d'étude, mais également critique pour la qualité des résultats escomptés. La préparation repose essentiellement sur la recherche d'une représentation pertinente pour le problème à modéliser, recherche qui se base sur des étapes complémentaires de construction et de sélection de variables. La sélection de variables a été largement étudiée dans la littérature (Guyon et al., 2006). Dans ce papier, nous nous focalisons sur l'approche filtre, qui évalue la corrélation entre les variables explicatives et la variable cible indépendamment de la méthode de classification utilisée, et est adaptée à la phase de préparation des données dans le cas d'un grand nombre de variables descriptives.

La construction de variables (Liu et Motoda, 1998) est un sujet nettement moins étudié dans la littérature scientifique, qui représente néanmoins un travail considérable pour l'analyste de données. Celui-ci exploite sa connaissance du domaine pour créer de nouvelles variables potentiellement informatives. En pratique, les données initiales sont souvent issues de bases de données relationnelles et ne sont pas directement exploitables pour la plupart des techniques de classification qui exploitent un format tabulaire attributs-valeurs. La fouille de données relationnelle, en anglais Multi-Relational Data Mining (MRDM), introduit par (Knobbe et al., 1999) vise à exploiter directement le formalisme multi-tables, en transformant la représentation relationnelle. En programmation logique inductive (ILP) (Džeroski et Lavrač, 2001), les données sont recodées sous forme de prédicats logiques. D'autres méthodes, dénommées