A Robust Method for Partitioning the Values of Categorical Attributes

Marc Boullé *

* France Telecom R&D, 2, Avenue Pierre Marzin,
22300 Lannion, France
marc.boulle@francetelecom.com

Résumé. Dans le domaine de l'apprentissage supervisé, les méthodes de groupage des modalités d'un attribut symbolique permettent de construire un nouvel attribut synthétique conservant au maximum la valeur informationnelle de l'attribut initial et diminuant le nombre de modalités. Nous proposons ici une généralisation de l'algorithme de discrétisation Khiops[1] pour le problème du groupage des modalités. L'algorithme proposé permet de contrôler a priori le risque de sur-apprentissage et d'améliorer significativement la robustesse des groupages produits. Cette caractéristique de robustesse a été obtenue en étudiant la statistique des variations du critère du Khi2 lors de regroupements de lignes d'un tableau de contingence et en modélisant le comportement statistique de l'algorithme Khiops. Des expérimentations intensives ont permis de valider cette approche et ont montré que la méthode de groupage Khiops aboutit à des groupages performants, à la fois en terme de qualité prédictive et de faible nombre de groupes.

1. Introduction

While the discretization problem has been studied extensively in the past, the grouping problem has not been explored so deeply in the literature. However, in real data mining datasets, there are many cases where the grouping of values of categorical attributes is a mandatory preprocessing step. The grouping problem consists in partitioning the set of values of a categorical attribute into a finite number of groups. For example, most decision trees exploit a grouping method to handle categorical attributes, in order to increase the number of instances in each node of the tree [Zighed et Rakotomalala, 2000]. Neural nets are based on numerical attributes and often use a 1-to-N binary encoding to preprocess categorical attributes. When the categories are too numerous, this encoding scheme might be replaced by a grouping method. This problem arises in many other classification algorithms, such as bayesian networks, linear regression or logistic regression. Moreover, the grouping is a general-purpose method that is intrinsically useful in the data preparation step of the data mining process [Pyle, 1999].

The grouping methods can be clustered according to the search strategy of the best partition and to the grouping criterion used to evaluate the partitions. The simplest algorithm tries to find the best bipartition with one category against all the others. A more interesting approach consists in searching a bipartition of all categories. The Sequential Forward Selection method derived from [Cestnik *et al.*, 1987] and evaluated by [Berckman, 1995] is a

[1] French patents N° 01 07006 and N° 02 16733