Discretization of Continuous Features by Resampling

Taimur Qureshi*, D.A.Zighed*

*University of Lyon 2 - Lab ERIC
5, Avenue Pierre Mendès France, 69676 Bron Cedex - France
taimur.qureshi, abdelkader.zighed@univ-lyon2.fr

Résumé. Les arbres de décision sont largement utilisés pour générer des classificateurs à partir d'un ensemble de données. Le processus de construction est une partitionnement récursif de l'ensemble d'apprentissage. Dans ce contexte, les attributs continus son discrétisés. Il s'agit alors, pour chaque variable à discrétiser de trouver l'ensemble des points de coupure. Dans ce papier nous montrons que la recherche des ces points de coupure par une méthode de ré-échantillonnage, comme le BOOTSTRAP conduit à des meilleurs résultats. Nous avons testé cette approche avec les méthodes principales de discrétisation comme MDLPC, FUSBIN, FUSINTER, CONTRAST, Chi-Merge et les résultats sont systématiquement meilleurs en utilisant le bootstrap. Nous exposons ces principaux résultats et ouvrons de nouvelles pistes pour la construction d'arbres de décision.

1 Introduction

In the process of knowledge discovery from a raw data set, we first preprocess the data to remove noise and handle missing data fields. Then data transformation, such as the reduction of the number of variables and the *discretization of attributes* defined on a continuous domain, is often performed, which is later provided to a data mining algorithm. One of the most important and complex issues in data mining is related to the transformation process such as discretization which consists of converting numerical data into symbolic or discrete form. Kusiak [9] emphasized that the quality of knowledge discovery from a data set can be enhanced by discretization because many of the knowledge discovery techniques are very sensitive to size of data in terms of complexity. Thus, the choice of discretization technique has important consequences on the induction model used such as CART [2].

In addition, numerical value ranges are not easy enough for evaluation functions to handle in a nominal domain ; for example, the original versions of the popular machine learning algorithms ID3 could be used only for categorical data and Quinlan [11] had to transform continuous ones into discrete values in his C4.5 decision tree learner. Many real-world classification algorithms are hard to solve unless the continuous attributes are discretized. It is hard to determine the intervals for a discretization of numerical attributes that has an infinite number of candidates. A simple discretization procedure divides the range of a continuous variable into equal-width intervals or equal-frequency intervals. Fayyad et al. [6] suggested a class dependent algorithm which reduce the number of attributed values maintaining the relationship between the class and attribute values. Liu et al. [10] classified discretization methods from