Invited Session *Statistical Inference in Data Mining*
COMPSTAT 2006, Roma, Italy, from August 28 to September 1, 2006

Organizer :
Stéphane Lallich, Université de Lyon, Laboratoire ERIC, Lyon 2,

Computing and using the deviance for classification trees

Gilbert Ritschard, Department of Econometrics, University of Geneva, Switzerland
gilbert.ritschard@themes.unige.ch

Summary. The reliability of induced classification trees is most often evaluated by means of the error rate. Whether computed on test data or through cross-validation, this error rate is suited for classification purposes. We claim that it is, however, a partial indicator only of the quality of the knowledge provided by trees and that there is a need for additional indicators. For example, the error rate is not representative of the quality of the description provided. In this paper we focus on this descriptive aspect. We consider the deviance as a goodness-of-fit statistic that attempts to measure how well the tree is at reproducing the conditional distribution of the response variable for each possible profile (rather than the individual response value for each case) and we discuss various statistical tests that can be derived from them. Special attention is devoted to computational aspects.
Key words: Classification tree, Deviance, Goodness-of-fit, Chi-square statistics, BIC.

Self Organizing Maps : Understanding, Measuring and Reducing Variability

Patrick Rousset**,** CEREQ, Marseille
rousset@cereq.fr

Summary. In classification self-organizing maps is used as a generalisation of the K-means method including a neighbourhood organization between clusters. The correspondence between this clusters organization and the input proximity is called the topology preservation. The aim of this paper is to measure, reduce and understand variability of SOM. Considering the property of topology preservation, a local approach of variability (at an individual level) is preferred to a global one. A complementary visualising tool, called Map of Distances between Classes (MDC), is presented to complete this local approach relating variability to the complexity of the data's intrinsic structure. It basically allows the main information to be extracted from a very large matrix of distances between Self-Organizing Maps' classes. In the presence of a complex structure, it enlarges the information set, linking the variability of acceptable representations to the data structure complexity. To reduce variability, a stochastic method based on a bootstrap process aims to increase the reliability of the induced neighbourhood structure. The resulting (robust) map, called R-Map, is more robust relative to the sensitivities of the outputs to the sampling method and to some of the learning options of the SOM' algorithm (initialisation and order of data presentation). This method consists of selecting one map from a group of several solutions resulting from the same self-organizing map algorithm, but obtained with various inputs. The Rmap can be perceived as the map, among the group of solutions, and corresponds to the most common interpretation of the data set structure. When an R-map can be perceived as the representative of a given SOM network, the relevance of the chosen network depends on R-map's ability to adjust the data structure. As an application, a criterion to validate the network size is proposed comparing its ability of adjustment with SOM outputs of a larger network.

Key words: Self-Organizing Maps, Robustness, Reliability, Bootstrap, Neighbourhood, Variability, R-Map.

Statistical inference and data mining: false discoveries control

Stéphane Lallich, Laboratoire ERIC, Université Lyon 2, France, stephane.lallich@univ-lyon2.fr,
Elie Prudhomme, Laboratoire ERIC, Université Lyon 2, France, elie.prudhomme@eric.univ-lyon2.fr
OlivierTeytaud, CNRS, UMR-8623, Équipe TAO INRIA Futurs, LRI, Université Paris Sud, Orsay, France.
Olivier.Teytaud@lri.fr

Summary. Data Mining is characterized by its ability at processing large amounts of data. Among those are the data ”features”- variables or association rules that can be derived from them. Selecting the most interesting features is a classical data mining problem. That selection requires a large number of tests from which arise a number of false discoveries. An original non parametric control method is proposed in this paper. A new criterion, UAFWER, defined as the risk of exceeding a pre-set number of false discoveries, is controlled by BS FD, a bootstrap based algorithm that can be used on one- or two-sided problems. The usefulness of the procedure is illustrated by the selection of differentially interesting association rules on genetic data.
Key words: feature selection, multiple testing, false discoveries, bootstrap.