Some Aspects on Classification, Variable Selection and Categorical Clustering

Gero Szepannek, Uwe Ligges, and Claus Weihs

Department of Statistics, Technische Universität Dortmund, Germany

Abstract. The package `klaR` contains several utilities to handle classification problems, e.g. Friedman's RDA, an interface to `svmlight` (Joachims, 1999) as well as variable selection procedures like the `stepclass` algorithm or Wilk's Λ, a visualization tool for SOMs or several classification performance measures (see Weihs et al., 2006).
This poster presents recent extensions towards classification on minimal variable subspaces for multi class problems by performing class pair wise variable selection (see Szepannek and Weihs, 2006). Examples of situations are presented where this approach may be highly beneficial in terms of misclassification rates.
Furthermore, the k-modes algorithm (Huang, 1998) is implemented allowing to perform a k-means like clustering for categorical data.

Keywords

CLASSIFICATION, VARIABLE SELECTION, CLUSTERING, CATEGORICAL DATA, DATA MINING

References

Huang, Z. (1998): Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3), 283-304.

Joachims, T., (1999): Making large-Scale SVM learning practical, In: B. Schlkopf and C. Burges and A. Smola (ed.) (eds.): *Advances in Kernel Methods - Support Vector Learning*, MIT-Press.

Szepannek, G., Weihs, C. (2006): Variable selection for more than two classes where data are sparse. In: M.Spiliopolou, R.Kruse, C.Borgelt, A.Nürnberger and W.Gaul (eds): *From Data and Information Analysis to Knowledge Engineering*, Springer, 700-707.

Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005): klaR - analyzing German business cycles, In: D. Baier, R. Becker and L. Schmidt-Thieme (Eds.): *Data Analysis and Decision Support*, Springer, Berlin, 335-343.