Adv Data Anal Classif (2009) 3:1–2
DOI 10.1007/s11634-009-0040-0

EDITORIAL

Editorial

Published online: 9 June 2009

This first issue of volume 3 (2009) of the journal *Advances in Data Analysis and Classification* (*ADAC*) contains three papers on clustering (unsupervised learning) and classification (supervised learning), and one paper that deals with a somewhat puzzling fact in the data analysis literature concerning reported classification errors.

- The article by *Jukka Corander, Mats Gyllenberg* and *Timo Koski* is devoted to the reconstruction of an unknown partition of objects characterized by qualitative data vectors. In contrast to Bayesian latent class analysis that deals with mixtures and priors for the unknown parameters (including the class frequencies), the authors describe a partition-based approach where priors are directly defined on the set of admissible partitions of the objects, similar to the papers of Geisser (1966), Bock (1972, 1974) and Van Cutsem (1996). They describe a Bayesian stochastic search algorithm (using parallel MCMC algorithms on the set of partitions) in order to learn the posterior probabilities of partitions. The method is illustrated by two examples from biology, dealing with human populations and DNA strains of bacteria.
- The paper by *Adrien Jamain* and *David J. Hand* is based on a meta-analysis of comparative studies on the performance of classification (discrimination) rules and their error rates that are reported in the literature (45 articles) for various (146) datasets. Within this analysis it was observed that error rates decreased with increasing size of dataset. The authors discuss various possible explanations for this fact related to the way datasets are collected by the research community.
- The third article of this issue, written by *Marc Boullé*, deals with supervised classification of multivariate data points by using data grid models where the (continuous or categorical) variables are compressed by a suitable segmentation (in the case of continuous variables: discretization) of their domains, thereby creating a grid of cells with data-dependent cell occupation frequencies that are the basis for applying a (fixed given) classification rule. The purpose of the paper is the simultaneous optimization of the variable-specific partitions, by considering the observed correlation between data and class memberships (filter methods). Its innovative aspect

resides in the fact that this problem is considered in a Bayesian model selection framework with suitably chosen priors for the grid partitions. Various numerical experiments with different classifiers illustrate the usefulness of the approach that leads, e.g., to a significant improvement of classification accuracy.

- The last article in this issue considers clustering problems for dynamically evolving data streams with smoothly changing data sets and similarity matrices, e.g., when considering the purchase of (old and new) products in a supermarket (cross-selling numbers) over time, or the cross-usage histories of objects (e.g., books) in the Online Public Access Catalog (OPAC) of a library. Since recalculating a classification at each new time point is computationally prohibitive, *Markus Franke* and *Andreas Geyer-Schulz* have developed a method for dynamically clustering the objects. Essentially it proceeds by simulating many restricted random walks (Markov chain) on the set of *edges* of the similarity graph, and summarizing the results in the form of a hierarchy of clusters (RRW clustering) that is dynamically adapted to the new data (dynamic RRW clustering). The paper proves various mathematical properties of the algorithm and concludes by some experiments for the OPAC that demonstrate the efficiency of the method in comparison to a simple reclustering after each change.
- On its last pages this issue contains a Call for Papers for a Special Issue of ADAC devoted to the topic

  Robust Methods for Classification and Data Analysis

  to be published in 2010. Guest editors will be Marco Riani, Andrea Cerioli and Peter J. Rousseeuw. All ADAC readers are kindly encouraged to submit manuscripts for this issue. For details and deadlines see the Call for Papers and the ADAC website http://www.springer.com/11634.

Hans-Hermann Bock (*Aachen*)
Wolfgang Gaul (*Karlsruhe*)
Akinori Okada (*Tokyo*)
Maurizio Vichi (*Rome*)

References

Bock HH (1972) Statistische Modelle und Bayes'sche Verfahren zur Bestimmung einer unbekannten Klassifikation normalverteilter zufälliger Vektoren. Metrika 18:120–132

Bock HH (1974) Automatische Klassifikation. Vandenhoeck & Ruprecht, Göttingen, Chapters 10–13

Geisser S (1966) Predictive discrimination. In: Krishnaiah PR (ed) Multivariate analysis. Academic Press, New York, pp 149–163

Van Cutsem B (1996) Combinatorial structures and structures for classifications. Comput Stat Data Anal 23:165–188