Adv Data Anal Classif (2015) 9:367–369
DOI 10.1007/s11634-015-0224-8

EDITORIAL

Special issue on "New trends on model-based clustering and classification"

Preface by the Guest Editors:
Salvatore Ingrassia, Geoffrey J. McLachlan, Gérard Govaert

Published online: 26 November 2015

This Special Issue of ADAC is devoted to recent developments in Model-Based Clustering and Classification which is an increasingly active area in both theoretical and applied research. This area has attracted the interest of a growing number of researchers due to the high potential of such approaches in applications, and new different topics have been investigated by several authors.

The Call for Papers for this special issue resulted in 22 manuscript submissions; at present, 6 have been accepted for publication and 3 are still under review. The latter ones, if accepted for publication, will be published in a regular issue of ADAC.

This Special Issue contains six papers dealing with quite different topics:

The first article, entitled 'Maximum Likelihood Estimation of Gaussian Mixture Models without Matrix Operations' by Hien D. Nguyen and Geoffrey J. McLachlan, focuses on the EM algorithm for Gaussian mixture models. It is well known that the EM algorithm is slow to converge and thus computationally expensive in some situations. To overcome this problem, the authors introduce the linear regression characterization (LRC) of Gaussian mixture models and propose a minorization-maximization (MM) algorithm which maintains the desirable properties of the EM algorithm without involving matrix operations. The estimators of the LRC parameters are then proved to be consistent and asymptotically normal. Finally, to compare the MM and EM algorithms, a large numerical study is provided under different settings, including also many mixture components and large data situations. The second article concerns again Gaussian mixture models but now the focus is on the initialization of the EM algorithm. The paper 'Improved initialisation of model-based clustering using Gaussian hierarchical partitions', by Luca Scrucca and Adrian E. Raftery, introduces an approach based on the model-based hierarchical agglomerative clustering (MBHAC) to provide initial partitions in the popular mclust R package. In this framework, different recipes are proposed based on data projection through a suitable transformation before applying the MBHAC at the initialization step. Numerical studies based on well-known datasets show that the procedure is computationally convenient and often yields good clustering partitions.

In model-based clustering, an important issue concerns the nature of the detected clustering. This topic is investigated in the article 'Probabilistic assessment of model-based clustering' by Xuwen Zhu and Volodymyr Melnykov. The authors develop an approach for assessing the variability in classifications carried out by the Bayes decision rule. The proposed technique is then applied to the detection of influential observations. In this context, visualization tools called significance and influence plots are also proposed. The former one is devoted to reflecting the significance of each classification through color hues; the latter display helps to summarize the information from influence tables for all observations. The proposed diagnostic methodology is finally studied and illustrated on synthetic and well-known real datasets.

The fourth article, entitled 'Robust model-based clustering via mixtures of skew-t distributions with missing information' by Wan-Lun Wang and Tsung-I Lin concerns multivariate mixture modeling of incomplete data using the skew-t distribution. Indeed, the occurrence of missing data is a ubiquitous problem in many practical problems. The authors present first a computationally flexible EM-type algorithm and then an information-based approach to approximating the asymptotic covariance matrix of the maximum likelihood estimators using the outer product of scores. The methodology is illustrated through a large numerical study based on both simulated data and a real dataset with genuine missing values.

Evaluation of the performance of clustering algorithms is a relevant topic in model-based clustering. In this framework, simulation studies play a fundamental role in computational statistics and tools for data generation are more and more important for the assessment of classification methods. The paper entitled 'Simulating mixtures of multivariate data with fixed cluster overlap in FSDA', by Marco Riani, Andrea Cerioli, Domenico Perrotta and Francesca Torti addresses robust clustering scenarios in the presence of contamination and outliers. The authors extend the capabilities of MixSim, a framework which is useful for evaluating the performance of clustering algorithms, on the basis of measures of agreement between data partitioning and flexible generation methods for data, outliers and noise. With MixSim, the FSDA (flexible statistics for data analysis) integrates in the same environment robust state of the art clustering algorithms and principled routines for their evaluation and calibration.

The last article is entitled 'Latent drop-out based transitions in linear quantile hidden Markov models for longitudinal responses with attrition', by M. Francesca Marino and Marco Alfò, and concerns longitudinal data, which are characterized by the presence of dependence between observations coming from the same individual. The paper takes steps from the literature on hidden Markov models for longitudinal responses in the exponential family. A common feature of this kind of studies is that individuals may leave the study before its designed end; in these cases, we observe variable-length individual sequences that may represent a further challenge, since not all individuals have the same weight in building up the log-likelihood function. The proposed approach is based on hidden Markov models and focuses on monotone missingness which may lead to selection bias and, therefore, to unreliable inference. The proposal is detailed by re-analyzing a well-known dataset on the time progression of CD4 cell counts in HIV seroconverters.

The Editors gratefully acknowledge the assistance of the following experts and colleagues in the process of reviewing the manuscripts that were submitted for this special issue:

Francesco Bartolucci (Italy), Hans-Hermann Bock (Germany), Charles Bouveyron (France), Faicel Chamroukhi (France), Stephane Chrétien (France), Francesca Greselin (Italy), Kevin J. Grimm (USA), Jeanine Houwing-Duistermaat (The Netherlands), Dimitri Karlis (Greece), Victor-Hugo Lachos Davila (Brazil), Alessio Farcomeni (Italy), Luis-Angel Garcia-Escudero (Spain), Tsung-I Lin (Taiwan), Marco Marozzi (Italy), Antonello Maruotti (Italy), Agustin Mayo-Iscar (Spain), Facundo Memoli (USA), Volodymyr Melnykov (USA), Hien Nguyen (Australia), Shu-Kay Ng (Australia), Cécile Proust-Lima (France), Antonio Punzo (Italy), Cinzia Viroli (Italy).

Salvatore Ingrassia (Catania, Italy)
Geoffrey J. McLachlan (Brisbane, Australia)
Gérard Govaert (Compiégne, France)