Editorial for ADAC issue 1 of volume 16 (2022)

Salvatore Ingrassia¹ · Julien Jacques² · Weixin Yao³

Published online: 1 March 2022

© Springer-Verlag GmbH Germany, part of Springer Nature 2022

This is the fifth Special Issue of ADAC dedicated to recent developments in Models and Learning in Clustering and Classification, an area which provides increasingly active research in both theoretical and applied domains and has attracted the interest of a growing number of researchers.

This special issue is divided in two parts due to the large number of papers that have been submitted for publication. The first part contains 8 papers which have been accepted for publication after a blinded peer-reviewed process, dealing with quite different topics. Four papers focus on mixture models. The first three contributions deal with outlier detection with missing data, deep models for mixed data sets, unobserved classes and extra variables in the test set, cluster-weighted models with skewed-distributions; the fourth paper on mixture models presents a more theoretical contribution about consistency of the maximum likelihood estimator under a special finite mixture of two-parameter Gamma distributions. Two subsequent papers concern robust techniques for classification trees for binary data and trimming approaches for functional data. The last paper presents a detailed characterization of many clustering methods and related properties. Below, we provide a short overview on the papers published in this special issue.

The mixture of multivariate contaminated normal (MCN) distributions is popularly used to cluster data with mild outliers. However, this approach requires complete data, which limits its application. The paper titled “Model-based clustering and outlier detection with missing data” by *Hung Tong* and *Cristina Tortora* develops a framework for fitting a mixture of multivariate contaminated normal (MCN) distributions to

✉ Salvatore Ingrassia
s.ingrassia@unict.it

Julien Jacques
Julien.Jacques@univ-lyon2.fr

Weixin Yao
weixiny@ucr.edu

¹ Department of Economics and Business University of Catania Corso Italia, 55 – 95129 Catania, Italy

² Laboratoire ERIC Bureau K067, 5 av. Mendès France, 69676 BRON Cedex, France

³ Department of Statistics, University of California Riverside, Riverside, CA 92521, USA

incomplete data sets. An expectation-conditional maximization algorithm is used for parameter estimation. The new model-based clustering method can be applied to data sets with outliers and missing data. The clustering performance of the proposed method is compared with some existing robust clustering methods in simulation studies and a real data application.

The paper entitled “Mixed Deep Gaussian Mixture Model: A clustering model for mixed datasets” by *Robin Fuchs, Denys Pommeret, Cinzia Viroli* introduces a multilayer architecture model-based clustering method called Mixed Deep Gaussian Mixture Model (MDGMM) which performs clustering via subsequent dimensionally reduced latent spaces in a very flexible way. Moreover, the model provides continuous low-dimensional representations of the data that allows to visualize mixed datasets. Parameter estimation plays a fundamental role in the paper and a suitable computational procedure is also introduced because some of the terms of the expected log-likelihood contain intractable expectations and a Monte Carlo extension of the EM is then required. Identifiability results for the model are also established. The approach is illustrated on the ground of many datasets concerning both discrete and mixed-type data.

The paper “Unobserved Classes and Extra Variables in High-dimensional Discriminant Analysis” by *Michael Fop, Pierre-Alexandre Mattei, Charles Bouveyron and Thomas Brendan Murphy* deals with the problem of classification when the test set contains new classes, not observed in the training phase, and new features, not measures on the training set. The authors propose to solve this transfer learning problem in a model-based way, considering a Gaussian mixture model that takes into account the evolution between the training and test phase. An inductive estimation strategy is proposed, starting by estimating the training model and then using an EM algorithm for inferring the test model with additional features and potentially novel classes. The ability of the proposed framework to deal with complex situations is illustrated through simulation and onto an artificial example build from a Mid-infrared spectra classification problem. An R package DAMDA implements the proposed model.

The paper “Multivariate Cluster Weighted Models Using Skewed Distributions” by *Michael P.B. Gallaughier, Salvatore Daniele Tomarchio, Paul D. McNicholas, and Antonio Punzo* introduces a family of 24 novel multivariate cluster weighted models (CWM), which allows both the covariates and response variables to be modelled using skewed distributions, or the normal distribution. The cluster weighted model extends the finite mixture of regression (FMR) to include modelling of the covariates and hence offers more flexibility compared to FMR. The expectation–maximization algorithm is applied for the parameter estimation. Both simulated and real data are used to evaluate the performance of the proposed models.

The paper “Strong consistency of the MLE under two-parameter Gamma mixture models with a structural scale parameter” by *Mingxing He and Jiahua Chen* studies the consistency of the maximum likelihood estimator under a special finite mixture of two-parameter Gamma distributions. Such mixture model could be useful for clustering observations with positives values. When the consistency of the shape parameter estimator is already known, this paper establishes the consistency of the scale parameter. An application to salary data illustrates the use of such model.

A quite different topic is presented in the paper entitled “Robust Optimal Classification Trees under Noisy Labels” by *Victor Blanco, Alberto Japón, Justo Puerto* that concerns a novel supervised method for binary classification. The method profits both from the ideas of Support Vector Machines and Optimal Classification Trees to build classification rules. It is based on two main elements: the splitting rules for the classification trees are designed to maximize the separation margin between classes applying the paradigm of Support Vector Machines; and some of the labels of the training sample are allowed to be changed during the construction of the tree trying to detect the label noise. The proposal is illustrated based on a large numerical study involving different classification tree-based methods on many popular real-life datasets.

The paper “Robust Clustering of Functional Directional Data” by *Pedro César Alvarez Esteban and Luis Angel García Escudero* introduces a robust approach for clustering functional directional data. The robustness is achieved by allowing to discard a proportion of data based on the techniques of “impartial trimming”, which uses the dataset itself to perform adaptive trimming. A feasible algorithm is proposed for its computation. Some theoretical properties of the algorithm are also provided. A “time warping” approach is further introduced to address misalignment problems within clusters and to detect typical “templates” which are useful to describe the detected clusters. The proposed methodology is finally applied to cluster aircraft trajectories to illustrate the practical interest of the approach.

The paper “An empirical comparison and characterization of nine popular clustering methods” by Christian Hennig compares nine clustering methods based on popular cluster validation indexes on 42 real data sets. This study gives a detailed characterization of these clustering methods and of the clustering properties which could be expected from the use of each of these methods. This study can be helpful to choose a clustering method in practice, according to what we want to discover in the data.

Acknowledgements 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: Margareta Ackerman (USA), Daniel Ahfock (Australia), Julien Ah-Pine (France), Andreas Alfons (Netherlands), Silvia Bacci (Italy), Zsuzsa Bakk (Netherlands), Michel Broniatowski (France), Ryan Browne (Canada), Gabriela Ciuperca (France), Pietro Coretto (Italy), Marco Corneli (France), Roberto Di Mari (Italy), Jean-Baptiste Durand (France), John Dziak (USA), Brian Franczak (Canada), Luis Angel García-Escudero (Spain), Michael Fop (Italy), Claire Gormley (Ireland), David Hitchcock (USA), Tsung-I Lin (Taiwan), Dimitris Karlis (Greece), Agustin Mayo-Isca (Spain), Geoffrey McLachlan (Australia), Igor Melnykov (USA), Colin Pawlowski (USA), Monia Ranalli (Italy), Shuchismita Sarkar (USA), Cristina Tortora (USA), Cinzia Viroli (Italy), Adalbert Wilhelm (Germany), Darren Wraith (Australia), Ping-Feng Xu (China), Jie Yang (USA), Ying Daisy Zhuo (USA).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.