Advances in Data Analysis and Classification (2024) 18:539–543
https://doi.org/10.1007/s11634-024-00605-6

EDITORIAL

# Special issue on "New methodologies in clustering and classification for complex and/or big data"

**Paula Brito[1] · Andrea Cerioli[2] · Luis Angel García-Escudero[3] · Gilbert Saporta[4]**

Accepted: 30 August 2024 / Published online: 4 September 2024


This Special Issue of ADAC is dedicated to recent developments in methodologies for clustering and classification, with a particular focus on the analysis of complex and big data. It is associated to the conference.

> *Classification and Data Science in the Digital Age: The 17th conference of the International Federation of Classification Societies*.

which took place in Porto, Portugal, on 19–23 July 2022.

The main goal of the Special Issue is to collect high-quality research papers in different areas related to modern statistical approaches for the analysis of complex and/or big data. We received 29 submissions for this Special Issue, 11 of which are published in this volume. The topics covered by the different papers span over diverse sub-areas of the general field of clustering and classification for complex (big) data. They also present both methodological results and challenging applications.

The first three contributions deal with mixture models and model-based clustering in challenging situations, such as those arising from time-dependent data, functional observations and multivariate heavy-tailed distributions. The three subsequent papers are instead concerned with methodological issues related to depth, robustness and anomaly detection in different contexts of interest for modern statistical applications. Then, the next two papers cover issues at the interface between statistics

✉ Andrea Cerioli
andrea.cerioli@unipr.it

[1] Faculdade de Economia, Universidade do Porto, Rua Dr. Roberto Frias, Porto 4200-464, Portugal

[2] Dipartimento di Scienze Economiche e Aziendali, Università di Parma, Via Kennedy 6, Parma 43125, Italy

[3] Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Paseo de Belén s/n, Valladolid, Spain

[4] Conservatoire National des Arts et Métiers, 292 rue Saint Martin, Paris 75003, France

and machine learning, considering the use of neural networks for classification of point patterns and neighbourhood solutions to imbalance problems with multi-label data. The ninth and tenth contributions are instead more in the wake of multivariate statistics, although they address challenges arising from modern non-standard data sources, such as multimode networks and compositions. The final paper is more on the applied side and deals with interesting classification problems of audio data collected from music recordings. Below, we provide a short overview of the papers published in this Special Issue.

The first paper, titled "Finite mixture of hidden Markov models for tensor-variate time series data" and authored by *Abdullah Asilkalkan*, *Xuwen Zhu* and *Shuchismita Sarkar*, deals with challenges arising from the need to model data with higher dimensions, such as those displayed in a tensor-variate framework where each observation is considered as a three-dimensional object. In particular, the authors develop a finite mixture of hidden Markov models for tensor-variate time series data and develop the corresponding EM algorithm for parameter estimation. They further show the classification performance of the proposed model through simulation studies and a real-life application.

The second contribution, by *Cristina Anton* and *Iain Smith*, is titled "Model-based clustering of functional data via mixtures of $t$ distributions". Its aim is to propose a procedure for clustering multivariate functional data with outliers through a mixture of multivariate $t$ distributions. For this purpose, the authors first define a family of latent mixture models following the approach used for the parsimonious models considered in the literature and also considering different options for constraining or not the degrees of freedom of the multivariate $t$ distributions to be equal across the mixture components. They then develop an appropriate EM algorithm for parameter estimation and inference, whose properties are studied from a theoretical and a computational point of view. The proposed approach is compared to potential competitors in the literature on simulated functional data with outliers and on real-world functional data.

The third paper in the Special Issue is "Parsimony and parameter estimation for mixtures of multivariate leptokurtic-normal distributions", by *Ryan P. Browne*, *Luca Bagnato* and *Antonio Punzo*. This work deals with a particular class of mixture models which have been recently introduced in the literature for the purpose of clustering multivariate data that originate from elliptical heavy-tailed distributions. An advantage of such models is that their parameters can be directly related to the moments of practical interest. The main proposal of this work is the derivation of two estimation procedures for the mixtures under consideration. The first one is based on the majorization-minimization algorithm, while the second is based on a fixed point approximation. Moreover, parsimonious forms of the considered mixtures are considered, and the suggested estimation procedures are used to fit them both in simulated and real data sets.

The subsequent paper moves to a topic related to robustness aspects in the analysis of complex multivariate data. It is titled "Theory of angular depth for classification of directional data" and it is written by *Stanislav Nagy*, *Houyem Demni*, *Davide Buttarazzi*, and *Giovanni C. Porzio*. Its goal is to investigate the potential of using depths in the problem of nonparametric supervised classification of directional data, that is

classification of data that naturally live on the unit sphere of a Euclidean space. This work addresses the problem mainly from a theoretical point of view, and its final goal is to offer guidelines on which angular depth functions should be adopted in classifying directional data. For this purpose, a set of desirable properties of an angular depth is put forward and the theoretical consequences of such properties are investigated. The most widely adopted angular depth functions are then extensively compared and contrasted with respect to the proposed properties. Simulated and real data are finally exploited to showcase the main implications of the discussed theoretical results, with an emphasis on potentials and limits of the often disregarded angular halfspace depth.

The fifth contribution, titled "Robust and sparse logistic regression" and written by *Dries Cornilly*, *Lise Tubex*, *Stefan Van Aelst* and *Tim Verdonck*, deals with robust statistical analysis of high-dimensional data through logistic regression. This work proposes a robust and sparse logistic regression estimator, where robustness is achieved by means of the $\gamma$-divergence. An elastic net penalty ensures sparsity in the regression coefficients such that the model is more stable and interpretable. Some theoretical properties of the proposed estimator are shown. In particular, the authors prove that its influence function is bounded. The good performance of the proposed estimator is then illustrated through a simulation study and in an empirical application that deals with classifying the type of fuel used by cars.

The sixth paper discusses novelty detection in a classification framework from a Bayesian perspective. It is titled "Variational inference for semiparametric Bayesian novelty detection in large datasets" and is authored by *Luca Benedetti*, *Eric Boniardi*, *Leonardo Chiani*, *Jacopo Ghirri*, *Marta Mastropietro*, *Andrea Cappozzo* and *Francesco Denti*. Novelty detection is meant as a method which aims to classify the instances of an unlabeled test set while allowing for the presence of previously unseen classes. This work focuses on a two-stage Bayesian semiparametric novelty detector recently introduced in the literature. Its main contribution is the proposal to resort to a variational Bayes approach for estimation, thus providing an efficient algorithm for posterior approximation which scales up to large high-dimensional datasets. A significant gain in efficiency and excellent classification performance of the proposed algorithm are demonstrated through extensive simulation studies. A novelty detection analysis is then performed on a large collection of satellite imaging spectra, to search for novel soil types.

With the seventh paper we move towards the interface between statistics and machine learning. Its title is "Neural networks with functional inputs for multi-class supervised classification of replicated point patterns" and it is written by *Kateřina Pawlasová*, *Iva Karafiátová* and *Jiří Dvořák*. Its main goal is to show the possibility of solving the supervised multi-class classification task for spatial point patterns via functional neural networks. To predict the class membership for a newly observed point pattern, it is suggested to compute an empirical estimate of a selected functional characteristic. Then, this estimated function is considered to be a functional variable entering the network. In a simulation study, the authors show that the neural network approach outperforms the kernel regression classifier which is considered as a benchmark method in the point pattern setting. They also analyse a real dataset of point patterns of intramembranous particles and illustrate the practical applicability of the proposed method.

The subsequent work, "Natural-neighborhood based, label-specific undersampling for imbalanced, multi-label data" by *Payel Sadhukhan* and *Sarbani Palit*, considers the practically relevant problem of imbalance in multi-label classification problems. Specifically, it suggests an undersampling scheme that uses the principles of the natural nearest neighbourhood and follows a paradigm of label-specific undersampling. In the proposed framework a single natural neighbour search is sufficient to identify all the label-specific overlaps. Natural neighbour information is also used to find the key lattices of the majority class. An empirical study, involving twelve real-world multi-label datasets, seven competing methods and four evaluating metrics, is conducted to show the performance of the proposed method.

The ninth contribution is titled "An analytic strategy for data processing of multimode networks" and is written by *Vincenzo Giuseppe Genova*, *Giuseppe Giordano*, *Giancarlo Ragozini* and *Maria Prosperina Vitale*. Its goal is to discuss an analytic strategy for simplifying multipartite networks in which different sets of nodes are linked. By considering the connection of multimode networks and hypergraphs as theoretical concepts, a three-step procedure is introduced to simplify, normalise, and filter network data structures. A model-based approach is then introduced for the derived bipartite weighted networks in order to extract statistically significant links. The usefulness of the strategy is demonstrated in handling two application fields, such as intranational student mobility in higher education and research collaboration in European framework programs.

The tenth paper is written by *Michael Greenacre* and titled "The chiPower transformation: a valid alternative to logratio transformations in compositional data analysis". It supports an alternative approach to analysing compositional data which does not rely on logratio transformations. The advantages of the suggested method are that it allows data zeros, without having to substitute them, and that it combines the standardisation inherent in the chi-square distance in correspondence analysis with the essential elements of the Box-Cox power transformation. It is argued that the new approach can be especially helpful in the area of high-dimensional data and in supervised learning contexts.

The Special Issue ends with the paper "Liszt's Étude S.136 no.1: audio data analysis of two different piano recordings", authored by *Matteo Farnè*. This work discusses the main signal processing tools of Music Information Retrieval from audio data and applies them to two recordings of the same piece, with the aim of uncovering the macro-formal structure and comparing the interpretative styles of the two performers. This goal is achieved through a segmentation based on the degree of novelty, in the sense of spectral dissimilarity, calculated frame-by-frame via the cosine distance. The metrical, temporal and timbrical features of the two executions are compared by Music Information Retrieval tools. The proposed analysis thus represents a case study able to show the potentialities of Music Information Retrieval from audio data in supporting traditional music score analyses and in providing objective information for statistically founded musical execution analyses.

Finally, we conclude our short description of this volume by gratefully acknowledging the assistance of many experts and colleagues in the process of reviewing the manuscripts that were submitted for the Special Issue. Its production would not have

been possible without their collaboration, which has greatly improved the quality of all the published papers.