Advances in Data Analysis and Classification (2024) 18:1–4
https://doi.org/10.1007/s11634-024-00584-8

EDITORIAL

Special issue on "advances in models and learning for clustering and classification"

Luis-Angel García-Escudero[1] · Salvatore Ingrassia[2] · T. Brendan Murphy[3]

Received: 12 February 2024 / Accepted: 12 February 2024 / Published online: 27 February 2024

This is the sixth Special Issue of ADAC dedicated to recent developments in Models and Learning in Clustering and Classification presenting recent results in both methodological and applied areas.

The first five contributions present topics in model-based clustering: the first two contributions deal with new approaches in mixtures of regressions and the subsequent two papers concern models for skew data for both vector and matrix-variate data; finally, the fifth paper focuses on modelling of spherical data.

Quite different domains are approached in the next two papers that concern robust co-clustering procedures and a clustering model designed for skew-symmetric data, respectively. Thus, from a the data type point of view, we can state that this special issue highlights recent developments in the analysis of skewed data.

The last two papers concern fraud detection, which is a very relevant domain recent years; in particular, fraud detection in daily credit card transactions and insurance are considered, respectively. Below, we provide a short overview on the papers published in this special issue.

The first paper, titled "Semiparametric Mixture of Linear Regressions with Nonparametric Gaussian Scale Mixture Errors", authored by *Sangkon Oh* and *Byungtae Seo*, suggests the utilization of nonparametric Gaussian scale mixture distributions for the component errors within a mixture of regressions framework. This approach aims to enhance robustness against outliers and reduce efficiency loss caused by misspecification of the distributions for the component errors. To achieve that greater flexibility, the procedure requires the nonparametric maximum likelihood estimation of some mixing distributions. The authors introduce a feasible algorithm for implementing this methodology along with a theoretical result on identifiability. The efficacy of the

✉ Salvatore Ingrassia
salvatore.ingrassia@unict.it

1 Department of Statistics and Operational Research and IMUVA, University of Valladolid, Facultad de Ciencias-Paseo Belen s/n, 47011 Valladolid, Spain

2 Department of Economics and Business, University of Catania, Corso Italia 55, 95128 Catania, Italy

3 School of Mathematics and Statistics, University College Dublin, Dublin 4, Ireland

proposed methodology is illustrated through simulation studies and the analysis of two real datasets.

Mixtures of regressions are taken into account also in the second paper entitled "Flexible mixture regression with the generalized hyperbolic distribution" by *Nam-Hwui Kim* and *Ryan P. Browne* introduces mixture of regressions but here the error components are modeled according to generalized hyperbolic distributions, which are highly flexible models that include several robust distributions such as the hyperbolic, normal-inverse Gaussian, variance-gamma and t distributions. Model indentifiability is analyzed and computational issues about parameter estimation via the EM algorithm according to the maximum likelihood approach are also presented. Many numerical studies based on both simulate and real data are presented. The paper is finally enriched by a quite large number of references.

An asymmetric generalization of the multivariate shifted exponential normal distribution has been introduced in the third paper entitled "Model-based clustering using a new multivariate skew distribution" by *Salvatore Daniele Tomarchio, Luca Bagnato, and Antonio Punzo*. The flexibility of this distribution enables effective handling of non-normal deviations, such as skewness and heavy tails, commonly present in the data. Furthermore, the distribution can be employed within a mixture modeling context, and an EM algorithm is provided for the estimating the parameters of the corresponding mixture. A simulation study comparing the performance of the proposal with several alternative models is presented. In a real-world data example, the application of this methodology distinguishes between two different types of co-movements among cryptocurrencies.

Mixture models for matrix-variate data matrix are becoming more and more popular in the last years. In this perspective, the fourth paper of this special issue entitled "Contaminated transformation matrix mixture modeling for skewed data groups with heavy tails and scatter" by *Xuwen Zhu, Yana Melnykov* and *Angelina S Kolomoytseva* deals with skewed matrix-variate data and extends to matrix-variate data previous results about a contaminated transformation mixture model capable of fitting multivariate data that can be skewed and heavy-tailed simultaneously. Finally, the proposal is assessed on the ground of numerical studies based on both simulated data and a real data concerning COVID-19.

The fifth paper "Mixture Modeling with Normalizing Flows for Spherical Density Estimation" by *Tin Lok James Ng* and *Andrew Zammit-Mangion* deals with the problem of modelling of spherical data. The authors propose a mixture-of-normalizing-flows model to offer a flexible modelling framework for spherical data. The proposed model offers an alternative modelling framework for spherical data to existing non-parametric, parametric and mixture modelling approaches. The proposed model is estimated using an EM-like algorithm and is demonstrated on two datasets where it achieves excellent modelling performance.

The next paper by *Edoardo Fibbi, Domenico Perrotta, Francesca Torti, Stefan Van Aelst,* and *Tim Verdonck* is titled "Co-clustering contaminated data: a robust model-based approach". Co-clustering, also known as block clustering, two-way partitioning, or biclustering, aims to simultaneously cluster the rows and columns of a data matrix. The authors highlight the importance of robustifying co-clustering procedures in response to the potentially adverse effects of outliers. With that in mind, they suggest

a robustification of the Latent Block Model by implementing an impartial trimming of rows and columns. Latent Block Models effectively tackle co-clustering problems within the framework of finite mixture models. The paper demonstrates the achieved robustness through simulations and presents results from five real data applications, including its application to a new dataset on customs declarations.

Paper number seven in this special issue focuses on investigating exchanges between objects. In their paper titled "A Between-Cluster Approach for Clustering Skew-Symmetric Data," *Donatella Vicari* and *Cinzia Di Nuzzo* analyze these exchanges using a clustering model designed for this particular type of data. This novel model aims to elucidate the between-cluster effects of skew-symmetries, reflecting the imbalances in the observed exchanges. A notable advantage of this approach is that it provides a visual representation of the exchanges that is not complex to interpret by relying on clustering results. The authors present an algorithm for implementation based on Alternating Least Squares. The procedure is illustrated using a brand-switching data.

The last two papers focus on fraud detection. The number of daily credit card transactions is ever and ever growing: the e-commerce market expansion and the recent constraints for theCovid-19 pandemic have significantly increased the use of electronic payments. The paper entitled "The role of diversity and ensemble learning in credit card fraud detection" by *Gian Marco Paldino Bertrand Lebichot, Yann-Aël Le Borgne, Wissam Siblini, Frédéric Oblé, Giacomo Boracchi, Gianluca Bontempi* concerns learning a strategy that relies on diversity based ensemble learning and allows to preserve past concepts and reuse them for a faster adaptation to changes. The effectiveness of our proposed learning strategy is assessed on extracts of two real datasets from two European countries, containing more than 30 M and 50 M transactions, provided by our industrial partner, Worldline, a leading company in the field.

In the same field, the paper "Claims Fraud Detection with Uncertain Labels" by *Félix Arthur Vandervorst*, *Wouter Verbeke* and *Tim Verdonck* concerns insurance fraud. This is a challenging problem and a high priority for insurers because, unlike their customers, there is no initially knowledge on the true fraud nature of an insurance contract, or a claim. In this framework, imprecisely observed labels can be represented in the Dempster-Shafer theory of belief functions. The paper shows how partial information from the historical investigations can add valuable and learnable information for the fraud detection system, improving its performances. Moreover, belief function theory provides a flexible mathematical framework for concept drift detection and cost sensitive learning. Finally, an application to a real-world motor insurance claim fraud based on a dataset composed of half a million motor insurance claims from a major insurance company, spanning over 8 years, is presented.

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.

Olcay Arslan, Gianluca Bontempi, Luis M. Castro Cepero, Andrea Cerioli, Sebastian Ciobanu, Pietro Coretto, Alessio Farcomeni, John Kent, Paul D. McNicholas, Keefe Murphy, Dominik Olszewski, Domenico Perrotta, Gabriele Soffritti, Salvatore Daniele Tomarchio, Cristina Tortora, Tim Verdonck, Cinzia Viroli, Darren Wraith, Hiroshi Yadohisa, Camila Borelli Zeller, Xuwen Zhu.