Advances in Data Analysis and Classification (2021) 15:1–4
https://doi.org/10.1007/s11634-021-00438-7

EDITORIAL

Editorial for ADAC issue 1 of volume 15 (2021)

Maurizio Vichi[1] · Andrea Cerioli[2] · Hans Kestler[3] · Akinori Okada[4] · Claus Weihs[5]

Published online: 27 March 2021

This issue 1 of volume 15 (2021) of the journal *Advances in Data Analysis and Classification (ADAC)* contains 10 articles that deal with interval forecasts, Boolean matrix factorization, robust spatial autoregressive scalar-on-function regression, Combined *k*-means and DBSCAN algorithm, Regularized Spectral Data Embedding, cost-sensitive constrained Lasso, semi-supervised support vector machine with asymmetric squared loss, Kappa Coefficients for Dichotomous-Nominal Classifications, Clustering Discrete-Valued Time Series, Simultaneous Dimension Reduction and Clustering.

Xin Zhao, Stuart Barber, Charles C. Taylor and Zoka Milan contribute the first paper of this ADAC issue with the title "Interval forecasts based on regression trees for streaming data". They have proposed two tree-based methods to deal with forecasting in a streaming data context, focusing on forecast intervals rather than point forecasts. The methodology is accomplished by updating the forecast interval by using root mean square prediction error calculated from the most recent batch of data, so as to update the model for future prediction. The interval is not necessarily symmetric, being initially based on quantiles but then it is adapted to ensure that the interval forecast will always include the point forecast.

In the second paper "From-below Boolean matrix factorization algorithm based on MDL" written by *Tatiana Makhalova* and *Martin Trnecka*, an Minimum-Description-Length-based (MDL) from-below factorization algorithm, which utilizes formal concept analysis, has been proposed. The method produces a small subset of formal concepts having the lowest information loss rate. The proposed algorithm does not compute the whole set of formal concepts but identifies factor sets that have better overall characteristics than factor sets computed by the existing Boolean

✉ Maurizio Vichi
maurizio.vichi@uniroma1.it

[1] Sapienza University of Rome, Rome, Italy

[2] University of Parma, Parma, Italy

[3] Ulm University, Ulm, Germany

[4] Rikkyo (St. Paul's) University, Tokyo, Japan

[5] Universität Dortmund, Dortmund, Germany

matrix factorization algorithms. The generated factor sets are small, contain few single-attribute factors, and have high coverage with a low overlapping rate.

The third article entitled "A robust spatial autoregressive scalar-on-function regression with t-distribution" is written by *Tingting Huang*, *Gilbert Saporta*, *Huiwen Wang*, and *Shanshan Wang*. Authors propose a robust spatial autoregressive scalar-on-function regression that incorporates a spatial lagged term into the classical scalar-on-function regression (SoFR) to accommodate the spatial dependence and allow for a thick-tailed noise term. An estimation method based on basis expansion and EM algorithm is developed to obtain the estimators of the spatial autoregressive parameter and the slope function. Authors include a simulation study to demonstrate the consistency of the proposed estimators. In particular, the new model performs better than the SoFR when the spatial correlation is present, and the SSoFR when the error term has thick tails. The new proposed methodology constitutes a practical statistical tool for modelling the spatial dependent data with functional covariates and scalar response that complements the widely popular spatial autoregressive scalar-on-function regression with normality assumption.

The next article on "A combination of k-means and DBSCAN algorithm for solving the multiple generalized circle detection problem" is written by *Rudolf Scitovski* and *Kristian Sabo*. The authors wish to solve the problem of identifying rod-shaped particles (e.g. bacilliform bacterium) by considering the multiple generalized circle detection problem. This is a complex global optimization problem that cannot be directly and efficiently solved by using one of the well-known global optimization methods. This paper proposes a modification of the well-known k-means algorithm for generalized circles as cluster-centers. The algebraic distance between a point and the generalized circle is used in the paper, but if several outliers can be expected among the data, LAD-distance should be used.

In the fifth paper, written by *Lazhar Labiod* and *Mohamed Nadif* on "Efficient regularized spectral data embedding", a regularized dimensionality reduction technique is proposed based on clustering. The work develops a simultaneous learning approach for data embedding and clustering that reinforces the relationships between these two tasks. It is suitable for high-dimensional data, where working with a low-dimensional space can often be useful for partitioning purposes. The proposal is based on an algorithm that alternates iteratively between embedding and clustering. It relies on a matrix decomposition technique for learning a spectral data embedding, a cluster membership matrix, and a rotation matrix that closely maps out the continuous spectral embedding, in order to obtain a good clustering solution. The suggested approach is compared with some traditional clustering methods through numerical experiments showing that it usually yields better embedding approximation and clustering accuracy.

The sixth article is written by *Rafael Blanquero*, *Emilio Carrizosa*, *Pepa Ramírez-Cobo* and *M. Remedios Sillero-Denamiel* on "A cost-sensitive constrained Lasso". The authors of this work propose a novel version of the Lasso in which quadratic performance constraints are added to Lasso-based objective functions. The goal is to set threshold values that bound the prediction errors in specific groups of interest to the researcher, instead of optimizing the overall prediction error. As a result, a constrained sparse regression model is defined by

a nonlinear optimization problem. The theoretical properties of the new methods are studied in detail in the paper. These properties include the existence of a unique optimal solution to the stated problem, the behavior to the limit of the solution, and some results about consistency. The empirical behavior of the suggested estimators is also investigated through a simulation study and the analysis of relevant data sets from biomedical and sociological contexts. It is shown that the proposed cost-sensitive constrained Lasso can have a direct application in heterogeneous samples where data are collected from distinct sources.

In the next paper entitled “A novel semi-supervised support vector machine with asymmetric squared loss”, *Huimin Pei*, *Qiang Lin*, *Liran Yang* and *Ping Zhong* present a semi-supervised SVM with asymmetric squared loss based on the expectile distance. The main advantage of the proposed method is that of being less sensitive to noise-corrupted data than the existing Laplacian SVM. The latter is based on the semi-supervised manifold regularization learning framework and is known to perform better than the standard SVM, especially when the supervised information is insufficient, but is also sensitive to the noise around the decision boundary. The goal of the paper is achieved by the addition of manifold regularization, which has the ability to encode the geometric information embedded in the unlabeled data, and by the fact that the expectile distance is stable to noise-corrupted data. A simple and efficient functional iterative method is then adopted to solve the involved optimization problems. The convergence of this iterative method is proved both theoretically and numerically. Experimental results on a number of benchmark datasets show that the proposed method compares well with several popular supervised and semi-supervised learning algorithms.

In the eighth article, by *Matthijs J Warrens* with the title "Kappa coefficients for dichotomous-nominal classifications", the author introduces a family of similarity coefficients that can deal with an absence category such as "no disorder". Cohen’s kappa coefficient may not be appropriate here, as a disagreement between classifiers on a ’presence’ category and the ’absence’ category may be much more severe than disagreement on two ’presence’ categories, for example, in a clinical context. Consequently, in the extension presented, different agreement types are distinguished: agreement on the ’presence’ categories, disagreement among the ’presence’ categories, and disagreement between all ’presence’ categories and the single ’absence’ category. Properties of this family of coefficients are derived. If the ’absence’ category is not used, the coefficients are identical to Cohen’s kappa.

In the next paper, "Clustering discrete valued time series", *Tyler Roick*, *Dimitris Karlis*, and *Paul McNicholas* introduce a model-based approach for clustering discrete valued time series data. Model parameters are estimated using the EM algorithm. Their model-based technique, focusing on the time series’ discreteness, was applied to both simulated and real data to illustrate its clustering capabilities. Model selection was made using the BIC, and for the simulated data, performance assessment was carried out using the adjusted Rand index. They could show that the technique performed very well under various simulated scenarios, and for the real data, they were able to find relevant results.

Finally, in the 10th paper, with the title "Simultaneous dimension reduction and clustering via the NMF-EM algorithm", the authors *Léna Care* and *Pierre Alquier* propose a new model that can be regarded as an adaption of the mixture of factor analyser (MFA) to combinations of distributions with nonnegative parameters. Here the decomposition into Gaussian factors in MFA is replaced by a nonnegative matrix factorisation (NMF). The authors show the utility of their approach in a simulation study and an application to real-world ticketing data.