Editorial for ADAC issue 2 of volume 15 (2021)

Editorial for ADAC issue 2 of volume 15 (2021)

This issue 2 of volume 15 (2021) of the journal *Advances in Data Analysis and Classification (ADAC)* contains 10 articles that deal with fuzzy set-valued data, fuzzy cluster, feature selection, fuzzy clustering for financial time series, random forest, hierarchical clustering, functional data, visualisation scheme for high-dimensional fold-change data, mixtures of factor analyzers, clustering of modal-valued symbolic data.

Beatriz Sinova, Stefan Van Aelst and Pedro Terán contribute the first paper of this ADAC issue with the title "M-estimators and trimmed means: from Hilbert-valued to fuzzy set-valued data". In the statistical literature several robust approaches have been proposed to measure location with data associated to a random experiment. M-estimators and trimmed means have been studied to handle Hilbert-valued data. In this paper it has been proven, that the behaviour of both alternatives is more robust than for the Aumann-type mean. Fuzzy M-estimators of location are a more robust approach than fuzzy trimmed means when the trimming proportion is less than 0.5. Only fuzzy trimmed means are always scale equivariant, even if both estimators share the translation equivariance, symmetry with respect to symmetrically distributed random fuzzy set values and strong consistency. The real-life example has empirically shown the robustness of fuzzy M-estimators of location and fuzzy trimmed means as an alternative to the Aumann-type mean.

Authors suggest, as future research lines, to use other tools from robust statistics, such as the influence function, so that realistic parametric families of distributions on the class of all fuzzy sets-valued data should be proposed first. In addition, scale equivariant fuzzy M-estimators could be defined by means of robust scale measures for fuzzy set-valued data. Last but not least, hypothesis testing procedures related to these measures could be established.

In the second paper entitled, "Isotonic boosting classification rules", written by *David Conde, Miguel A. Fernández, Cristina Rueda and Bonifacio Salvador*, authors give the definition of novel rules developed for binary and multiclass classification problems. In fact, in many real classification problems a monotone relation between some predictors and the classes may be assumed when higher (or lower) values of those predictors are related to higher levels of the response. New boosting algorithms, based on Logit-Boost, that incorporate the isotonicity information, yielding more accurate and easily interpretable rules are proposed. These algorithms are based on theoretical developments that consider isotonic regression. Authors show the good performance of these procedures not only on simulations, but also on real data sets coming from two very different contexts, namely cancer diagnostic and failure of induction motors. In the first case the new rules reduce the error rates between 33% and 66%. In the second case, that deals with the diagnostic of induction motors, the training sample fulfils the expected monotone relationships and the error rates are quite low. Also, in this case the new rules manage to reduce the error rates significantly.

The third article is written by *Giovanni De Luca and Paola Zuccolotto* on "Regime dependent interconnectedness among fuzzy clusters of financial time series". Authors analyse the relationships among clusters of assets identified according to the lower tail dependence. So, assets belonging to the same cluster show a high lower tail dependence, while this type of dependence is low with respect to the assets of other clusters. Clusters are identified by means of a fuzzy cluster analysis algorithm and the tail dependence coefficients are estimated using the Joe-Clayton copula function. The 75th percentile within clusters is used as a measure of each cluster's overall tail dependence. Using a Granger causality approach, in order to determine whether the pattern of a cluster can be predicted based on the past values of the others, the interdependence structure of the clusters' tail dependence dynamics has been analysed. The hypothesis of a possible regime switching dynamics in tail

dependence is also investigated by means of a Threshold Vector Auto-Regressive model and the results are compared to those obtained with a linear autoregression.

The proposed procedure is described with reference to a case study dealing with the assets composing Eurostoxx 50, but it can be viewed as the proposal of a general method, that can be relevantly applied to whatever set of asset returns time series.

The next paper is entitled “Active Learning of Constraints for Weighted Feature Selection”, and it is written by *Samah Hijazi, Denis Hamad, Mariam Kalakech, Ali Kalakech*. Recently, researchers were interested in using pairwise constraints, a cheaper kind of supervision information that does not need to reveal the class labels of data points. These were suggested for feature selection to enhance the performance of clustering algorithms. However, in most current methods, pairwise constraints are provided passively and generated randomly over multiple algorithmic runs by which the results are averaged. This leads to the need of a large number of constraints that might be redundant, unnecessary, and under some circumstances even inimical to the algorithm's performance. Therefore, authors suggest a framework for actively selecting and then propagating constraints for feature selection by using graph Laplacian that is defined on the similarity matrix. They highlighted three main characteristics of their methodology: first, the use of the margin-based feature selection algorithm that utilizes constraints; second, the process of active selecting these constraints, which represents their core contribution; third, the augmentation of supervision information through propagating these constraints. Authors assume that when a small perturbation of the similarity value between a data couple leads to a more well-separated cluster indicator based on the second eigenvector of the graph Laplacian, this couple is definitely expected to be a pairwise query of higher and more significant impact. Constraints propagation on the other side ensures increasing supervision information while decreasing the cost of human-labour. Experimental results validated the proposal in comparison to other known feature selection methods and proved to be prominent.

In the fifth paper, written by *Thiago Salles*, *Leonardo Rocha* and *Marcos Gonçalves* on “A bias-variance analysis of state-of-the-art random forest text classifiers”, the authors perform a detailed comparison of several random-forest classifiers in view of the bias-variance decomposition of error rate. This analysis allows the authors to shed light on the main causes of the observed improvements enjoyed by the best performing variants of random-forest classifiers, for which significant reductions in variance couple with stability in bias. Furthermore, the analysis also suggests new promising directions for further enhancements in random-forest learners.

After an overview of the bias-variance decomposition for classification, the paper describes a strategy for estimating the bias and variance factors obtained through an error-rate decomposition framework suitable for the analysis of text data. Accuracy results for six random-forest learners are then obtained through an empirical study of data sets dealing both with topic characterization and sentiment analysis, two major tasks in text classification.

The sixth article is written by *Kadri Umbleja*, *Manabu Ichino* and *Hiroyuki Yaguchi* on “Hierarchical conceptual clustering based on quantile method for identifying microscopic details in distributional data”. The authors of this work propose an algorithm for clustering distributional data from a “microscopic” perspective by using quantile values. The suggested “microscopic” approach is intended to take into account the underlying properties of the distribution, as opposed to the more classical “macroscopic” approach where only limited characteristics of data are considered. The goal is reached by measuring the dissimilarity between two objects at multiple points defined through quantiles.

The goal of the proposed algorithm, having multiple points for comparison, is to identify similarities in small sections of the distribution under study while producing more adequate hierarchical concepts. The authors show that the algorithm has a monotonicity property and produces

more adequate conceptual clusters in experiments. Furthermore, it allows the user to compare different types of symbolic data easily.

In the next paper entitled "Robust archetypoids for anomaly detection in big functional data", *Guillermo Vinue* and *Irene Epifanio* suggest the use of robust functional archetypoids, combined with an adjusted boxplot for skewed distributions, to identify functional outliers. Robustness is achieved by means of M-estimators with a suitable loss function, while the adjusted boxplot provides the necessary cut-off for outlier labelling. The proposed method compares well with several state-of-the-art techniques for functional outlier detection in a controlled study.

Furthermore, the authors present a new scalable archetypoid algorithm that can be used to analyse large data sets in reasonable time. This method is applied to two large time series of data, where outlying curves are present. The reduction in computational time allowed by the new scalable algorithm is discussed and a new R package, that includes the algorithms used in the paper, is also introduced.

In "A perceptually optimised bivariate visualisation scheme for high dimensional fold change data" *André Müller, Ludwig Lausser, Adalbert Wilhelm, Timo Ropinski, Matthias Platzer, Heiko Neumann and Hans A. Kestler* describe a method for visualising ratios together with their absolute values. This is often important, as the visualisation of pure ratios might be deceptive if there is no reference to the absolute numbers that are used to generate them. This can lead to strong misinterpretations especially in the field of life-sciences where often ratios from high-throughput experiments like micro-arrays or RNA-Seq are analysed and visualised.

The visualisation scheme consists of two parts: a uniform colour scale and a patch grid representation. This uniform colour scale was derived from sub-sampling a CIE-LUV or CIE-LAB colour space to generate a perceptually uniform representation. It was shown that bivariate colour scales encoding two dimensions are difficult to read, therefore the authors proposed perceptually separable visual dimensions encoding ratio as colour and patch size as absolute values. This new visualisation scheme was then investigated with a number of experiments by twenty-five observers. It was shown that this subsampling approach together with the patch visualisation was superior to representations based on standard RGB colour spaces.

The next article "Mixtures of factor analyses with scale mixtures of fundamental skew normal distributions" by *Sharon Lee, Tsung-I Lin and Geoffrey McLachlan* deals with mixtures of factor analyzers for modelling high dimensional data. They present a novel generalisation of the mixture of factor analyzers model based on a general skew distributional form that defines the class of SMCFUSN (scale mixtures of canonical fundamental skew normal) distributions. The proposed model provides a tool for the flexible modelling of data exhibiting non-normal features including multimodality, skewness, and heavy-tailedness. This class of mixture of skew factor analyzers formally embeds other existing skew factor analyzers including MSNFA, MSTFA, and CFUSHFA models. They perform pattern parameter estimation by an EM-type algorithm and show the usefulness and the potential of the method on for real-world datasets.

The final paper of this edition on "Clustering of modal valued symbolic data" by *Nataša Kejžar, Simona Korenjak-Černe and Vladimir Batagelj* deals with the extension of two popular clustering methods, both based on representatives, to the clustering of symbolic objects. These descriptions can encompass more than only a single value. A special type of these symbolic objects consists of descriptions with frequency or probability distributions. In this way a simultaneous analysis of both single value variables and variables with richer descriptions is possible. The authors adapt two classical clustering methods, namely a generalisation of *K*-means clustering and Wards hierarchical clustering for analysing the symbolic objects. This is achieved by the use of weights for each symbolic object and affects the similarity measure between the symbolic object and the representative for that cluster. The extensions of these algorithms are derived and applied to synthetic and real-world

datasets. The authors argue that their method allows the cluster representatives to preserve the structure of the data while additionally preventing the clustering process from favouring only one value of a variable generating more interpretable results.

Maurizio Vichi (Rome)
Andrea Cerioli (Parma)
Hans Kestler (Ulm)
Akinori Okada (Tokyo)
Claus Weihs (Dortmund)