Special issue on “Science of big data: theory, methods and applications”

Preface by the Guest Editors

Hans A. Kestler¹ · Paul D. McNicholas² · Adalbert F. X. Wilhelm³

Published online: 1 November 2018

© Springer-Verlag GmbH Germany, part of Springer Nature 2018

This special issue of ADAC is devoted to a broad range of innovative research developments in tackling the challenges big data pose for statistical classification and data analysis. As one of the most common statistical learning techniques, classification requires flexible approaches and algorithms to deal with the challenges posed by big data to compress and extract the essential underlying information.

The Call for Papers for this special issue resulted in 24 manuscript submissions, of which eight have been accepted for publication. Main topical foci of the selected manuscripts are ensemble techniques and high-dimensional data.

The first article, entitled ‘Ensemble of a subset of kNN classifiers’ by Asma Gul, Aris Perperoglou, Zardad Khan, Osama Mahmoud, Miftahuddin Miftahuddin, Werner Adler and Berthold Lausen proposes to ensemble a specially selected subset of k-nearest neighbours (kNN) classifiers in a two-stage process. In the first stage, these classifiers are selected according to their performance using out-of-sample accuracy. In the second stage, they are then combined sequentially, in the order of performance. The collective performance is then assessed via a validation set. In a benchmark study, that augments the data sets with uninformative features, the proposed ensemble method compares favourably with state-of-the-art procedures, such as kNN, bagged kNN, random kNN, the multiple feature subset method, random forests and support vector machines.

In the second article, entitled ‘Understanding non-linear modeling of measurement invariance in heterogeneous populations’, Deana Desa addresses the challenge of measurement invariance in large-scale international surveys, in particular, com-

✉ Adalbert F. X. Wilhelm
a.wilhelm@jacobs-university.de

¹ Institute of Medical Systems Biology, Ulm University, 89069 Ulm, Germany

² Department of Mathematics and Statistics, McMaster University, Hamilton, ON L8S 4L8, Canada

³ Department of Psychology and Methods, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany

mon in international comparative educational research. By examining the non-linear modelling of ordered categorical variables within multiple-group confirmatory factor analysis, measurement invariance across countries was empirically investigated combining two steps. First, a separated confirmatory factor analysis was performed to model the complex structure of the relationships within each country. Second, a categorical multiple-group confirmatory factor analysis was applied to examine full measurement invariance between the countries. The chosen approach supported all three kinds of invariance—pattern, metric and scalar invariance—for the latent factor structure under investigation.

The next paper in this section by Daniel Horn, Aydın Demircioğlu, Bernd Bischl, Tobias Glasmachers and Claus Weihs entitled “A Comparative Study on Large-Scale Kernelized Support Vector Machines” compares approximate support vector machine solvers for big data on twelve benchmark data sets with different accuracy/run-time trade-offs using Pareto-fronts. Surprisingly the standard LIBSVM with subsampling is a strong baseline in these settings. Additionally, some solvers systematically outperform others, which aids in providing concrete recommendations for use.

The paper ‘A computationally fast variable importance test for random forests for high-dimensional data’ by Silke Janitza, Ender Celik and Anne-Laure Boulesteix proposes a fast heuristic to test variable importance measures for classification and ranking candidate predictors. Ranking predictors according to some variable importance measure is a commonly used approach in feature selection and model building. The determination of a cutoff is a challenging question, and some hypothesis tests have been proposed for this task. Because existing approaches require the repeated computation of random forests, they are computationally inefficient in high-dimensional settings. The proposed procedure uses permutation variable importance and is especially useful in high-dimensional scenarios in which many dimensions are of low information. Comparative tests using simulation studies show that the new approach is computationally as well as statistically superior.

The next paper entitled ‘Rank-based classifiers for extremely high-dimensional gene expression data’ and authored by Ludwig Lauser, Florian Schmidt, Lyn-Rouven Schirra, Adalbert F. X. Wilhelm and Hans A. Kestler looks at classification tasks in a small n , large p situation. The authors propose to rank transform the real-valued gene expression profiles to construct invariant classifiers that are less affected by noisy data. Based on a large-scale cross-validation experiment on numerous data sets and nineteen different classification models, they conclude that classifiers largely benefit from the rank transformation. Their results show that rank transformations are particularly effective for random forests and support vector machines.

The sixth paper by Afef Ben Brahim and Mohamed Limam titled ‘Ensemble feature selection for high dimensional data: a new method and a comparative study’ deals with the generation of diverse feature subsets via different feature relevance criteria for the generation of robust feature sets. They use a two-step approach consisting of ensemble generation and subsequent ensemble aggregation. The method is then tested on seven gene expression data sets and it turns out that homogeneous ensembles formed with unstable base learners are superior to heterogeneous ensembles.

Paper number seven in this special issue is concerned with fitting random forests in high-dimensional covariate spaces. In their paper ‘An efficient random forests algo-

rithm for high dimensional data classification', Qiang Wang, Thanh-Tung Nguyen, Joshua Z. Huang and Thuy Thi Nguyen propose to modify the standard random forest algorithm by subdividing the set of predictors into subsets of informative and non-informative variables and by imposing different weights on these subsets. Using this approach maintains the diversity and randomness of the forest, while reducing the computational complexity. The approach has been evaluated on real world data sets from gene and image classification. The reported results indicate the aptness of the approach to significantly reduce prediction error in comparison to other random forest implementations.

The final paper in this special issue by Ravi Sankar Sangam and Hari Om entitled 'Equi-Clustream: a framework for clustering time evolving mixed data' deals with clustering of mixed type time-evolving data. A framework is introduced which includes hybrid drifting concept detection algorithm, a hybrid data labeling algorithm, and a visualization approach analyses the relationship between the clusters at different timestamps. The efficacy of the proposed framework is shown by experiments on synthetic and real-world datasets and also compared to other approaches.

This special issue would not have been possible without the support and contributions of the experts and colleagues reviewing the manuscripts. As Guest Editors we gratefully acknowledge the valuable assessment, evaluations, and critical remarks by them: Daniel Baier, Simona Balbi, Bernd Bischl, Hans Hermann Bock, Krisztian Buza, Claudio Conversano, Vincenzo Esposito-Vinzi, Holger Fröhlich, Alexander Groß, Bettina Grün, Christian Hennig, Iulian Ilies, Markus Kaechele, Johann Michael Kraus, Berthold Lausen, Ludwig Lausser, Geoffrey McLachlan, Günther Palm, Giuseppe Rizzo, Lyn-Rouven Schirra, Florian Schmid, Matthias Schmid, Rainer Schuler, Friedhelm Schwenker, Eric Sträng, Marieke Timmerman, Jacobo Toran, Claus Weihs, Maurizio Vichi, Gunnar Völkel.