Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-020-00431-6

EDITORIAL

Special issue on “Learning in data science: theory, methods and applications”—preface by the guest editors

Daniel Baier[1] · Berthold Lausen[2] · Angela Montanari[3] · Ute Schmid[4]

Recently, the interplay of disciplines involved in data science, most notably statistics and computer science has intensified. Impressive advances in statistical, deep, and machine learning (both supervised and unsupervised) have been achieved by developing and applying more and more complex methods for data, data stream, text, or image processing. They are now further developed and used in many fields of applications like, e.g., engineering, finance, genomics, industrial automation, industry 4.0, marketing, personalised medicine or health care, systems biology.

Many of these learning methods are about to make their way into real-world usage now with—in many cases and for good reasons—strict legal requirements. This has renewed the demand that methods should not only be accurate but allow the practitioner to obtain important insights about the learning process and results at hand. So, ensuring interpretability is of central importance in application domains where trust into the system is essential for its acceptance and where malfunctioning may result in legal liability.

In this special issue, we therefore solicit contributions that describe and apply new developments in the field of deep, machine, and statistical learning, discuss and evaluate the interpretability of various types of learning methods and/or their assessment of uncertainty, use dimension reduction and representations which are understandable

✉ Daniel Baier
daniel.baier@uni-bayreuth.de

Berthold Lausen
blausen@essex.ac.uk

Angela Montanari
angela.montanari@unibo.it

Ute Schmid
ute.schmid@uni-bamberg.de

[1] Faculty of Law, Business and Economics, University of Bayreuth, 95447 Bayreuth, Germany

[2] Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK

[3] Department of Statistical Sciences, University of Bologna, 40126 Bologna, Italy

[4] Faculty of Information Systems and Applied Computer Science, University of Bamberg, 96047 Bamberg, Germany

Published online: 18 November 2020

by experts and even non-experts, introduce methods that can be inspected, verified, and possibly also modified by non-expert users, offer explanations or visualizations of their decisions, and develop methods for interpretable learning in complex domains.

The Call for Papers for this Special Issue resulted in 31 manuscript submissions, of which 13 have been accepted for publication. They have been allocated by the Guest Editors into the following five sections on

- statistical learning (three articles),
- interpretable machine learning and explainability (three articles),
- dimension reduction and visualization (two articles),
- applications in biostatistics and bioinformatics (three articles),
- applications in marketing (two articles).

The special issues starts with a section on **statistical learning** with three articles. The first one, entitled "Data generation for composite-based structural equation modeling methods" by Rainer Schlittgen, Marko Sarstedt, and Christian M. Ringle, examines wide-spread malpractices when the efficacy of estimators for composite-based structural equation modeling (SEM) is studied. New composite model-based data generation procedures for prespecified model parameters in the structural model and the measurements models are proposed and their superiority is demonstrated. The findings will assist researchers in their composite-based SEM simulation studies.

The second article, entitled "Mixtures of Dirichlet-multinomial distributions for supervised and unsupervised classification of short text data" by Laura Anderlucci and Cinzia Viroli, focuses on topic detection in short textual data. A new approach is proposed that extends the well-known mixture of unigrams approach by including a Dirichlet prior to the compound mixtures of Dirichlet-multinomial distributions which is preferable for sparse data. Real empirical problems are used to demonstrate the usefulness of the new approach.

A last article in this section, entitled "On the use of quantile regression to deal with heterogeneity: the case of multi-block data" by Cristina Davino, Rosaria Romano, and Domenico Vistocco, proposes a quantile regression based strategy to assess heterogeneity in a multi-block type data structure. Specifically the article deals with a particular structure where all the blocks of variables are observed on the same units and a structure of relations is assumed between the different blocks. In an application, the new approach is applied to a framework where consumers are grouped according to their similarities in the dependence structure.

The section on **interpretable machine learning and explainability** contains three articles. The first one, entitled "Editable machine learning models? A rule-based framework for user studies of explainability" by Stanislav Vojíř and Tomáš Kliegr, deals with new approaches to improve the comprehensibility of machine learning models. Instead of using acceptance measurement questionnaires or surveys, an extension of the EasyMiner system for generating classification and explorative models based on association rules is developed and applied. The presented web-based rule editing software allows the user to perform common editing actions such as modify rule (add or remove attribute), delete rule, create new rule, or reorder rules. To observe the effect of a particular edit on predictive performance, the user can validate the rule list against a selected dataset using a scoring procedure.

The second article in this section, entitled "A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C" by Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou, discusses the well-known comprehensibility and transparency problems of predictive systems when high-dimensional behavioral and textual data are the basis: Linear as well as non-linear models contain thousands of parameters that are difficult to keep track of. Instead, the authors align the recently proposed Linear Interpretable Model-agnostic Explainer (LIME) and SHapley Additive ExPlanations (SHAP) with the notion of counterfactual explanations and empirically compare the effectiveness and efficiency of these novel algorithms against a model-agnostic heuristic search algorithm for Searching EviDence Counterfactuals (SEDC) using 13 behavioral and textual data sets. The results are promising.

The section on interpretable machine learning and explainability closes with a third article, entitled "A process framework for inducing and explaining datalog theories" by Mark Gromowski, Michael Siebers, and Ute Schmid. The authors develop a general process framework for logic-rule-based classifiers facilitating mutual exchange between system and user. The framework constitutes a guideline for how a system can apply inductive logic programming in order to provide comprehensive explanations for classification choices and empowering users to evaluate and correct the system's decisions. It also includes users' corrections being integrated into the system's core logic rules via retraining in order to increase the overall performance of the human-computer system.

The section on **dimension reduction and visualization** consists of two articles. The first one, entitled "The ultrametric correlation matrix for modelling hierarchical latent concepts" by Carlo Cavicchia, Maurizio Vichi, Giorgia Zaccaria, discusses a new approach to derive tree-like structures of latent concepts from correlation matrices. Algorithms are proposed that detect consistent latent concepts and their relationships starting from an observed correlation matrix. Applications to simulated data and to the well-known Bechtoldt data set on intellectual abilities (memory, verbal, words, space, number, reasoning) demonstrate the usefulness of the new approach.

The second article in this section, entitled "SEM-tree hybrid models in the preferences analysis of the members of Polish households" by Adam Sagan and Mariusz Lapczyński, also develops tree-like structures for latent concepts. SEM-tree hybrid models, that combine the confirmatory structural equation models (SEM) with exploratory and predictive classification and regression trees (CART), are applied to empirical data on resources allocation of 1020 Polish households members. The new model and application develops new insights: The choice of an altruistic (patriarchal person) strategy (taking care of the "common good" with a strong control function of the family head) strongly encourages the members of the household to submit the value of money ("struggle for existence"), even at the expense of free time and loss of one's reputation.

The section on **applications in biostatistics and bioinformatics** consists of three articles. The first one, entitled "Chained correlations for feature selection" by Ludwig Lausser, Robin Szekely, and Hans A. Kestler proposes a new approach to extend the data sources for supervised learning algorithms designed in molecular pheno- or genotyping. Samples of (related) foreign classes are incorporated in the training, which changes the magnitude of available data. Intermediate foreign classes are detected by evaluating the correlation between class labels and features for each pair of original and foreign categories. Interestingly, this approach does not require direct comparisons of the initial diagnostic groups and therefore, might be suitable for settings with restricted data access.

The second article in this section, entitled “Adapted single-cell consensus clustering (adaSC3)” by Cornelia Fuetterer, Thomas Augustin, and Christiane Fuchs, extends single-cell consensus clustering by Kiselev et al. in so far that the linear principal components analysis is replaced by diffusion maps, a non-linear method that takes the transition of single cells into account. The performance of the new approach is compared to alternative approaches in terms of accuracy on simulated and real data sets. The results are promising.

The last article in the section on applications in biostatistics and bioinformatics, entitled “Automatic gait classification patterns in spastic hemiplegia” by Ana Isabel Aguilera Faraco and Alberto Subero, explores different techniques for the selection of attributes in clinical gait analysis to get the best classification scores. Comparison of these results to a qualitative assessment from physicians shows a surprising high success rate. The approach has been integrated into an end-user application in order to support medical decision-making.

The Special Issue closes with a section on **applications in marketing** with two articles. The first one, entitled, “Predicting brand confusion in imagery markets based on deep learning of visual advertisement content” by Atsuho Nakayama and Daniel Baier demonstrates how convolutional neural networks (CNNs) with pre-trained network parameters (VGG16, ImageNet challenge) can be used to predict the uniqueness of brand positionings. The ability of the CNN to predict the advertised brand (with masked slogans and logos) based on advertisements with visual content (e.g. photos, video clips) is used as a proxy for this uniqueness. An application to the German Pils beer market shows that the new approach extends the traditional approach with confusion experiments and samples of consumers in so far that now ads across a multitude of channels (e.g., TV, cinema, newspaper, search engines, social networks, photo-, video-sharing platforms) can be analyzed with a high accuracy.

The section and the Special Issue closes with an article, entitled “The GNG neural network in analyzing consumer behaviour patterns: empirical research on a purchasing behaviour processes realized by the elderly consumers” by Kamila Migdał-Najman, Krzysztof Najman, and Sylwia Badowska. The purchasing behaviour of students from Slovenia, the Czech Republic and Poland in the years 2017–2018 is analyzed using a Growing Neural Gas (GNG) approach. The main purpose was to learn and predict purchases of smartphones and of other innovative new products and the underlying heterogeneity across these three countries. The derived market segmentation appears to be very understandable and helpful for marketeers.

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 and the Editors of this journal: Hans Hermann Bock, Ines Brusch, Michael Brusch, Andrea Cerioli, Reinhold Decker, Johannes Fürnkranz, Claas Christian Germelmann, Wolfgang Gaul, Andreas Geyer-Schulz, Dominik Heider, Christian Hennig,

Carmela Iorio, Krzysztof Jajuga, Hans A. Kestler, Ludwig Lausser, Hermann Locarek-Junge, Karsten Lübke, Eneldo Loza Mencía, Ulrich Müller-Funck, Atsuho Nakayama, Akinori Okada, Francesco Palumbo, Józef Pociecha, Dirk van den Poel, Johannes Rabold, Alexandra Rese, Roberto Rocci, Lars Schmidt-Thieme, Ingo Schmitt, Roberta Siciliano, Winfried Steiner, Alfred Ultsch, Maurizio Vichi, Claus Weihs, Adalbert Wilhelm, and Herbert Woratschek.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.