mathematics

MDPI

Editorial

Special Issue "Statistical Data Modeling and Machine Learning with Applications"

Snezhana Gocheva-Ilieva

Department of Mathematical Analysis, University of Plovdiv Paisii Hilendarski, 4000 Plovdiv, Bulgaria; snow@uni-plovdiv.bg

Give Us Data to Predict Your Future!

The modeling and processing of empirical data is one of the main subjects and goals of statistics. Nowadays, with the development of computer science, the extraction of useful and often hidden information and patterns from data sets of different volumes and complex data sets in warehouses has been added to these goals. New and powerful statistical techniques with machine learning (ML) and data mining paradigms have been developed. To one degree or another, all of these techniques and algorithms originate from a rigorous mathematical basis, including probability theory and mathematical statistics, operational research, mathematical analysis, numerical methods, etc. Popular ML methods, such as artificial neural networks (ANN), support vector machines (SVM), decision trees, random forest (RF), among others, have generated models that can be considered as straightforward applications of optimization theory and statistical estimation. The wide arsenal of classical statistical approaches combined with powerful ML techniques allow many challenging and practical problems to be solved.

This Special Issue belongs to the section "Mathematics and Computer Science". Its aim is to establish a brief collection of carefully selected papers presenting new and original methods, data analyses, case studies, comparative studies, and other research on the topic of statistical data modeling and ML as well as their applications. Particular attention is given, but is not limited, to theories and applications in diverse areas such as computer science, medicine, engineering, banking, education, sociology, economics, among others.

This Special Issue begins with a contribution to computer science and educational data mining (EDM) by Gocheva et al. [1]. A new framework based on three ML regression methods for predicting student achievement in mathematics through the statistical processing of empirical data is proposed. In the first stage, classification and regression trees (CART) as well as CART ensembles and bagging models are built and evaluated. The importance of predictors for student success is determined. In the next stage, the predicted values of the best models from the first stage are combined through stacking with the multivariate adaptive regression splines (MARS) method. It is shown that the combined MARS models are superior to the single models for all of the statistical indicators.

An engineering application to predict dam inflow levels by means of time series modeling using a long short-term memory (LSTM) network is developed in the work of Tran et al. [2]. A robust statistical criterion called the "correlation threshold" for partial autocorrelation and cross-correlation functions is introduced to the appropriate input predictors and the number of their time lag variables. A wavelet transformation and a hyper-parameter optimization determined by K-fold cross-validation were also applied to improve the overall performance of the LSTM. The resulting streamflow predictions are shown to be more accurate than those of ANNs, recurrent NNs, support vector regression, and multilayer perceptron (MLP).

The application of data modeling in medicine is presented in the work of Mwata-Velu et al. [3]. The problem of information processing in the interactions between brain signals and controllable machines, which requires instantaneous brain data decoding,

Citation: Gocheva-Ilieva, S. Special Issue "Statistical Data Modeling and Machine Learning with Applications". *Mathematics* **2021**, *9*, 2997. https://doi.org/10.3390/math9232997

Received: 16 November 2021
Accepted: 18 November 2021
Published: 23 November 2021

is considered. The authors have developed a real-time embedded brain–computer interface (BCI) system for signal recognition, and this system is applied in the locomotion of a hexapod robot. A hybrid convolutional neural network (CNN)–LSTM model is implemented to process and classify motor imagery electroencephalogram signals that have been captured by specialized sensors. By using stratified 10-fold cross-validation, the average model accuracy is determined to be about 85%.

The next article in this Special Issue models the hedonistic prices of the housing market in Catalonia. The authors Torres-Pruñonosa et al. [4] analyze a large amount of data provided by two banks. Regression models with ANN, quantile regression (QR), and semi-logarithmic regression (SLR) were obtained and studied. A comparison of the results showed that QR models are less accurate than other models, and therefore, the QR method cannot be recommended. The ANN and SLR models demonstrated similar statistics, with SLR showing the best performance in the case of smaller data volumes. The study offers recommendations to Spanish banks, encourage the use of ANN or SLR for real estate appraisal.

Interesting results have been reported in [5] by Cerquides and his co-authors. This study contributes to the crowdsourcing paradigm based on "the wisdom of crowd". In particular, this article deals with methods for assessing the reliability and accuracy of public data that have been collected and analyzed by volunteers—members of different citizen science communities. An abstract probabilistic consensus model is proposed, which summarizes other label aggregation models. Practical evidence supporting this approach is demonstrated by assessing the accuracy of crowdsourced graphical data on the earthquake in Albania during 2019, difficult to process statistically by means of classical and ML statistical techniques.

The work of Grané and Sow-Barry [6] explores a hybrid approach that consists of two classical multivariate techniques—multidimensional scaling (MDS) and the k-prototype clustering algorithm—to visualize the profiles of individuals from weighted and mixed big data. Initially, a small random working sample (up to 10%) was extracted from the whole dataset. This sample was clustered using the k-prototype clustering algorithm with Gower's distance, the individuals were labeled trough weighted MDS, and the MDS configuration was determined. Then, the rest of the individuals were projected onto the MDS configuration by the Gower's interpolation. The algorithm was implemented to classify and visualize large survey data depicting the health and socioeconomic living conditions of European citizens.

The Special Issue continues with the work of Shao et al. [7], who focuses on the multi-category classification problem. A framework and ML algorithm based on data-adaptive kernel SVM for binary imbalanced data have been proposed, including a new method for constructing the data-dependent kernel for a multi-class setting. The improved performance of the method is mainly due to the more robust decay rate and flexibility of the constructed kernel functions. This ensures the optimal adaptation of the data-dependent kernel to the studied data. Tests were performed with different datasets, both artificial and real. An application of the new method is conducted in the medical field for the classification of a multi-class cancer imaging dataset. The simulations illustrate that the model results are significantly better compared to six standard classifiers for all of the statistical indicators.

In paper [8], Zhou et al. develop two new second-order optimization methods, namely the damped Newton stochastic gradient descent (DN-SGD) method and the stochastic gradient descent-damped Newton (SGD-DN) method, which were designed to train deep neural networks (DNNs). A mathematical expression for calculating the Hessian matrices of the last layer parameters and the penultimate layer of the loss function is derived, and it is proven that the Hessian matrices are positive and semidefinite. Furthermore, the DN-SGD and SGD-DN algorithms are implemented to iterate the parameters of the last layer using a variational damped Newton method. In this way, faster algorithm convergence is achieved, and the cost of the calculations is reduced. Numerical experiments have been

performed, which demonstrate the higher efficiency of the proposed methods for solving various regression and classification problems with data from the housing market, finance, and other areas.

A system for classifying breast cancer subtypes using the cascade deep forest method is presented in the work of El-Nabawy et al. [9]. The cascade deep forest model incorporates both DNNs and ensemble models such as RF. The model learns the class distribution features by assembling decision tree-based forests while supervising different types of input features. The classification is employed on various omics METABRIC sub-datasets that contain preprocessed and integrated data profiles of clinical datasets, gene expression datasets, and two types of datasets obtained through statistical feature engineering for different classes of dataset configurations. A model accuracy of up to 83% for 5 subtypes and of up to 78% for 10 subtypes was reported. It is shown that with the developed system, the time needed to obtain the results is shorted compared to the previous results in the field.

The Special Issue concludes with an article by Chen et al. [10], which proposes a new algorithm for the non-linear clustering of categorical data. The algorithm includes a self-expressive kernel density estimation scheme and a probability-based non-linear feature-weighted similarity measure. A non-linear optimization method in kernel subspace is implemented in the developed self-expressive kernel subspace clustering algorithm with embedded feature selection. Experiments were performed that highlighted the better performance of the algorithm compared to classical clustering algorithms used for categorical data with non-linear relationships.

In conclusion, the resulting palette of methods, algorithms, and applications for statistical modeling and ML presented in this Special Issue is expected to contribute to the further development of research in this area. We also believe that the new knowledge acquired here as well as the applied results are attractive and useful for young scientists, doctoral students, and researchers from various scientific specialties.

Funding: This research received no external funding.

Acknowledgments: The research activity of the Guest Editor of this Special Issue has been conducted in the framework and has been partially supported by the MES (Grant No. D01-387/18.12.2020) for NCDSC, part of the Bulgarian National Roadmap on RIs.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Gocheva-Ilieva, S.; Kulina, H.; Ivanov, A. Assessment of Students' Achievements and Competencies in Mathematics Using CART and CART Ensembles and Bagging with Combined Model Improvement by MARS. *Mathematics* **2021**, *9*, 62. [CrossRef]
2. Tran, T.D.; Tran, V.N.; Kim, J. Improving the Accuracy of Dam Inflow Predictions Using a Long Short-Term Memory Network Coupled with Wavelet Transform and Predictor Selection. *Mathematics* **2021**, *9*, 551. [CrossRef]
3. Mwata-Velu, T.; Ruiz-Pinales, J.; Rostro-Gonzalez, H.; Ibarra-Manzano, M.A.; Cruz-Duarte, J.M.; Avina-Cervantes, J.G. Motor Imagery Classification Based on a Recurrent-Convolutional Architecture to Control a Hexapod Robot. *Mathematics* **2021**, *9*, 606. [CrossRef]
4. Torres-Pruñonosa, J.; García-Estévez, P.; Prado-Román, C. Artificial Neural Network, Quantile and Semi-Log Regression Modelling of Mass Appraisal in Housing. *Mathematics* **2021**, *9*, 783. [CrossRef]
5. Cerquides, J.; Mülâyim, M.O.; Hernández-González, J.; Shankar, A.R.; Fernandez-Marquez, J.L. A Conceptual Probabilistic Framework for Annotation Aggregation of Citizen Science Data. *Mathematics* **2021**, *9*, 875. [CrossRef]
6. Grané, A.; Sow-Barry, A.A. Visualizing Profiles of Large Datasets of Weighted and Mixed Data. *Mathematics* **2021**, *9*, 891. [CrossRef]
7. Shao, J.; Liu, X.; He, W. Kernel Based Data-Adaptive Support Vector Machines for Multi-Class Classification. *Mathematics* **2021**, *9*, 936. [CrossRef]
8. Zhou, J.; Wei, W.; Zhang, R.; Zheng, Z. Damped Newton Stochastic Gradient Descent Method for Neural Networks Training. *Mathematics* **2021**, *9*, 1533. [CrossRef]
9. El-Nabawy, A.; Belal, N.A.; El-Bendary, N. A Cascade Deep Forest Model for Breast Cancer Subtype Classification Using Multi-Omics Data. *Mathematics* **2021**, *9*, 1574. [CrossRef]
10. Chen, H.; Xu, K.; Chen, L.; Jiang, Q. Self-Expressive Kernel Subspace Clustering Algorithm for Categorical Data with Embedded Feature Selection. *Mathematics* **2021**, *9*, 1680. [CrossRef]