Overcoming “Big Data” Barriers in Machine Learning Techniques for the Real-Life Applications

Czarnowski, I., Jędrzejowicz, P., Chao, K-M. & Yildirim, T.

DOI 10.1155/2018/1234390
ISSN 1076-2787
ESSN 1099-0526

Publisher: Hindawi

Hindawi
Complexity
Volume 2018, Article ID 1234390, 3 pages
https://doi.org/10.1155/2018/1234390

Editorial

Overcoming "Big Data" Barriers in Machine Learning Techniques for the Real-Life Applications

Ireneusz Czarnowski,[1] **Piotr Jedrzejowicz**,[1] **Kuo-Ming Chao**,[2] **and Tülay Yildirim**[3]

[1] *Department of Information Systems, Gdynia Maritime University, Morska 83, 81-225 Gdynia, Poland*
[2] *Coventry University, Priory Street, Coventry CV1 5FB, UK*
[3] *Department of Electronics and Communication Eng., Yildiz Technical University, Davutpasa Campus, 34220 Esenler/Istanbul, Turkey*

Correspondence should be addressed to Ireneusz Czarnowski; irek@am.gdynia.pl

Received 12 November 2018; Accepted 5 December 2018; Published 19 December 2018

1. Introduction

Data analysis, regardless of whether the data are expected to explain a quantitative (as in regression) or categorical (as in classification) models, often requires overcoming various barriers. They include unbalanced datasets, faulty measurement results, and incomplete data.

A special group of barriers is typical for what nowadays is referred to as "Big Data." The term is used to characterize problems where the available datasets are too large to easily deal with traditional machine learning tools and approaches. It is now generally accepted that dealing with huge and complex data sets poses many processing challenges and opens a range of research and technological problems and calls for new approaches.

Big Data is often characterized by the well-known 5V properties:

(i) Volume: typically a huge amount of data,

(ii) Velocity: a speed at which data are generated including their dynamics and evolution in time,

(iii) Variety: involving multiple, heterogeneous, and complex data representations,

(iv) Veracity: the uncertainty of data and lack of its quality assurance,

(v) Value: a potential business value that big data analysis could offer.

Big Data environment is often a distributed one with a distributed data sources. These sources can be heterogeneous, differing in various respects including storage technologies and representation methods.

Challenges of the Big Data not only involve a need to overcome the 5V properties but also include a need to develop techniques for data capturing, transforming, integrating, and modelling. Yet other important issues are concerned with privacy, security, governance, and ethical aspects of the Big Data analysis.

Current advances in dealing with the Big Data problems, albeit in many cases spectacular, are far from being satisfactory for the real-life applications. This becomes especially true in numerous domains where machine learning tasks are crucial to obtaining knowledge of different processes and properties in areas such as bioinformatics, text mining, or security. Unfortunately, the majority of the current algorithms become ineffective when the problem becomes very large since underlying combinatorial optimization problems are, as a rule, computationally difficult. There exists a variety of methods and tools which are excellent at solving small and medium size machine learning tasks but become unsatisfactory when dealing with the large ones.

Current hot topics in the quest to improve the effectiveness of the machine learning techniques include a search for compact knowledge representation methods and better tools for knowledge discovery and integration. Machine learning may also profit from integrating collective intelligence techniques, applying evolutionary and bioinspired techniques, and exploring further deep and extreme learning techniques.

2. Contributions Included in the Special Issue

The purpose of this special issue is to publish some of the current research results advancing different techniques for dealing with large and complex problems. The issue consists of fourteen papers covering some novel methods and techniques as well as their applications. The selected of them are extended versions of papers presented at the IEEE INISTA Conference in 2017 year.

The paper of I. Czarnowski and P. Jędrzejowicz proposes an approach to data reduction for learning from Big Data sets by integrating stacking, rotation, and agent population learning techniques. The paper shows that combining the proposed techniques can improve the performance of the classifier learning from large and complex datasets. The approach is based on the classifier ensemble paradigm where stacking ensembles have been produced using the rotation-based techniques, guaranteeing their heterogeneity. To reduce the dimensionality of the data, data reduction in an instance and feature dimensions has been applied.

Dimensionality reduction has been also discussed in the paper of J. Lee and D.-W. Kim. The authors consider a multilabel classification problem. Multilabel classification is a variant of multiclass classification, where multiple labels may be assigned to each instance; i.e., each instance corresponds to multiple class labels. The approach is based on dimensionality reduction by feature selection. The approach is based on analysis of the information content and remedies the computational burden by discarding the labels that are unimportant to feature importance scores.

Semisupervised learning is a class of the machine learning tasks where both labeled and unlabeled data are used to induce a learning model. The paper of E. Protopapadakis et al. deals with the problem of instance selection and training set preprocessing. Several approaches to instance selection based on sampling are discussed and compared. An extensive experimental evaluation of the considered approaches is included in the paper.

In their paper, J. Jedrzejowicz and P. Jedrzejowicz consider an approach to the data stream mining. The problem has been solved using the incremental Gene Expression Programming classifier with metagenes and data reduction. As it has been shown, the proposed concept of metagenes assured increasing the classification accuracy while data reduction allowed controlling computation time. The advantage of the proposed approach is also allowing work with the data stream that results from implementation of simple drift detection mechanisms. The proposed approach offers also scalability through the possibility to adjust computation times to the user needs at the expense of the classification accuracy.

Z. H. Kilimci and S. Akyokus focus on the text classification problem. The authors propose the ensemble learning and deep learning approaches to enhance the text classification performance. The ensemble of base classifiers proposed for solving the considered classification problem includes traditional machine learning algorithms such as naïve Bayes, support vector machine and random forest, and a deep learning based Conventional Network classifier. The different document representations and different ensemble approaches on eight different datasets have been evaluated. Finally, it has been shown that using heterogeneous ensembles together with deep learning methods and word embedding enhances text classification performance.

A. Nowak-Brzezińska deals with the problem of knowledge management and proposes a new approach for the rule management mechanisms in the decision support systems. The approach is based on hierarchically organized rule-base structure. Such a structure is produced based on the clustering approach. Making use of the similarity approach the proposed algorithm tries to discover new facts (new knowledge) from rules and facts already known. The computational experiment involves an analysis of the impact of the proposed methods on the efficiency of a decision support system with hierarchical knowledge representation.

B. Huynh and B. Vo focus on the problem of mining erasable itemsets. Mining erasable itemsets is a class a frequent pattern mining problem. In general, the problem of mining erasable itemsets belongs to the NP-hard class and the existing algorithms for mining erasable itemsets have high computational complexity. Computational experiments results show that the proposed approach ensures quite reasonable and competitive results as compared with earlier approaches.

Y. Guo et al. deal with the problem of complex power system status monitoring and evaluation. In the paper, a special Big Data platform, used as an analytical tool, is presented as discussed. Based on the case study, authors show how to improve the decision-making process in power systems.

The paper written by P. Caşcaval focuses on the problem of modeling and evaluating the reliability of the complex networks. In general, the problem of computing complex network reliability belongs to NP-hard class problems. The paper contributes by proposing a novel approach to network reliability evaluation. The proposed method reduces the computation time for large networks to a great extent, compared with an exact method as well with other known methods.

In his paper, J. P. Paplinski investigates the stability of the Bat Algorithm. The analysis is based on the assumption that the considered algorithm can be treated as a stochastic discrete-time system which allows using the Lyapunov stability theory for analyzing the behavior of the algorithm. The computational experiment proves the correctness of the approach.

The paper of C. A. Martin et al. is dedicated to the problem of classifying comments that tourists publish online. The paper discusses the case study, where Convolutional Neural Networks and Long Short-term Memory Networks are used in the process of decision making with respect to the quality of the service improvements.

A. Wosiak and D. Zakrzewska discuss the real problem of detection and diagnosis of heart diseases. One approach for preparing a model of heart diseases for medical diagnostic is based on clustering. The authors propose a new approach based on combining unsupervised feature selection and clustering. The proposed approach has been validated using the real-life datasets of cardiovascular cases. The experiment results show the advantage of the approach as compared to other approaches based on feature selection but without the clustering for supporting the statistical inference.

In the paper written by J. Jakubik and H. Kwaśnicka, the music data are analyzed using machine learning methods. SVM is used as the classification tool, but before the adequate data representation has to be prepared using the Recurrent Neural Network. The computational experiment results show that the proposed hybrid machine learning tool is competitive as compared with other approaches.

The paper written by D. Świetlik deals with simulation of natural brain processes, including three typical processes, i.e., learning, memory, and forgetting. The processes are simulated based on the model of the CA1 region of the hippocampus. A possibility of the hardware implementation of the pyramidal cells of the CA1 region of brain hippocampus is also discussed. The problem considered is an example of the problem where different signals influence the brain processes. Their analysis can be useful from the point of view of the medical diagnostics as well from the point of view of extracting knowledge important for preparing and improving artificial models and algorithms applied for brain data analyses.

3. Conclusions

The editors believe that the special issue has been an important and timely initiative. The editors hope that the presented research results will be of value to the scientific community working in the field of Big Data, data science, machine learning, analysis of complex data, data mining, knowledge discovery, and project management. Presented results are also addressed for other researchers who are currently or will be in the future implementing different data analysis tools trying to solve the real-life problems.

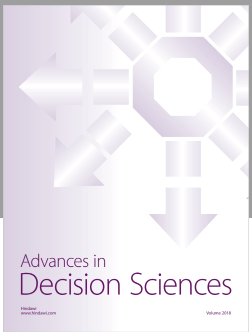
We would like to take this opportunity to thank all the authors for their valuable contributions. The submitted papers have been reviewed by at least two referees. We wish to thank all peer reviewers whose invaluable work, suggestions, and detailed feedback have helped to improve the quality of the papers included in the special issue. Special thanks are due to Sergio Gómez and Vincent Labatut, who supported the editors in their work.

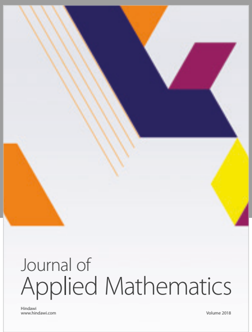
Conflicts of Interest

The editors declare that they have no conflicts of interest regarding the publication of this Special Issue.

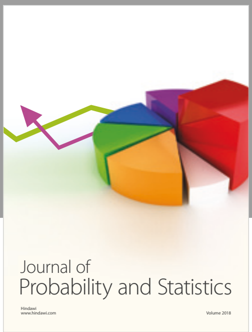
Ireneusz Czarnowski
Piotr Jedrzejowicz
Kuo-Ming Chao
Tülay Yildirim

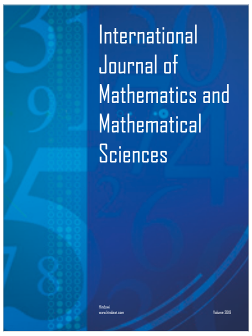
Advances in
Operations Research


Advances in
Decision Sciences

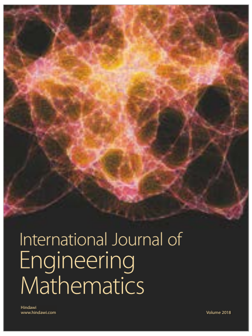

Journal of
Applied Mathematics

The Scientific
World Journal

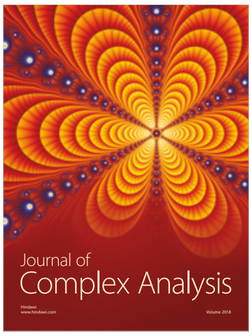

Journal of
Probability and Statistics

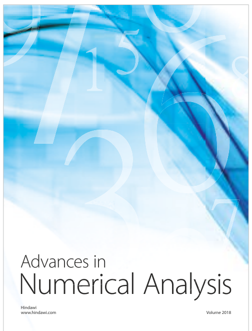

International
Journal of
Mathematics and
Mathematical
Sciences

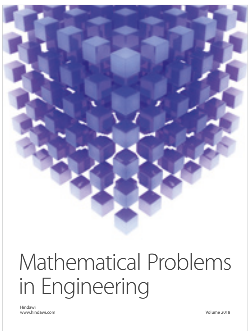
Journal of
Optimization

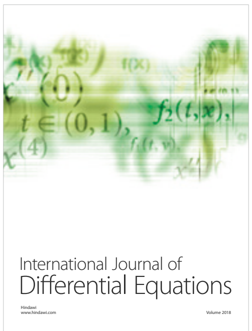

International Journal of
Engineering
Mathematics

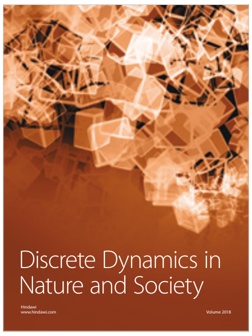
International Journal of
Analysis

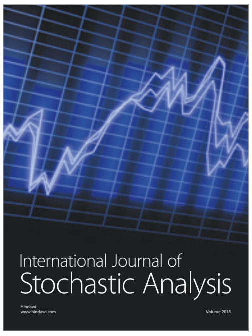

Journal of
Complex Analysis


Advances in
Numerical Analysis


Mathematical Problems
in Engineering

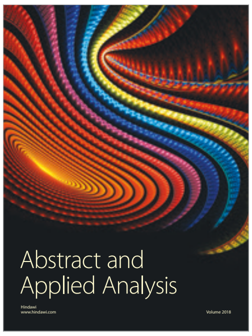

International Journal of
Differential Equations

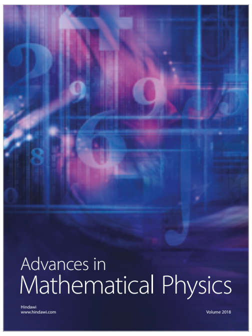

Discrete Dynamics in
Nature and Society


International Journal of
Stochastic Analysis

Journal of
Mathematics

Journal of
Function Spaces


Abstract and
Applied Analysis


Advances in
Mathematical Physics