PREFACE

Special section on
Exploring Complex and Big Data

Machine learning and data mining have made a tremendous progress in the last decades and have become important sub-fields of computer science. Nevertheless, many of the current approaches assume that the data being processed are static and well-structured, which turns out to be too restrictive, as many modern applications and technologies produce complex, poorly structured, and dynamic data. This is particularly evident in the rapidly developing field of processing and analysing big data. New challenges in processing big data stem not only from a vast volume of data, but also from their heterogeneity, complexity, multi-dimensionality, dynamics, changes with time and uncertainty. Therefore, it is necessary to focus more on the *properties* of such data, rather than solely on data volumes.

The property of data that is particularly in the focus of this special section is *complexity*, which originates both from the need for getting richer and more precise descriptions of the real world, and from the advancement of new technologies of data acquisition. Processing and exploring such data pose new challenges for researchers and call for a variety of new approaches. This observation was one of the motivations behind the big data seminar organized by the guest editors of this special section at the Poznań University of Technology in April 2016.[1] Given the success of the seminar and an invitation from Professor Józef Korbicz, we decided to prepare this special section of the *International Journal of Applied Mathematics and Computer Science*, devoted to the challenges in big data exploration, with a particular emphasis on those originating from data complexity.

The best presenters of the big data seminar and some other researchers in the field were invited to contribute to this special section. Two rounds of rigorous reviews and revisions led us to the final selection of six papers, briefly summarized below.

The opening paper, entitled *Exploring complex and big data*, by J. Stefanowski, K. Krawiec, and R. Wrembel, discusses selected challenges of complex and big data management, as well as analysis methods. The authors survey different views on big data and focus on their properties. Then, the state-of-the-art architectures for storing, managing, and pre-processing techniques are discussed, accompanied with the highlighting of open research problems. A separate part of the paper is devoted to new requirements for data mining algorithms in the context of two recently emerged fields: mining concept-drifting data streams and applying deep learning to complex data representations.

The second paper, entitled *Methods for mining co-location patterns with extended spatial objects*, by R. Bembenik, W. Jóźwicki, and G. Protaziuk, addresses the problems of spatial data mining. The paper focuses on *extended objects*, i.e., those that cannot be represented by single points, and on *co-location patterns*, i.e., sets of spatial objects that satisfy certain constraints on their features and neighborhoods. The authors compare the DEOSP algorithm proposed in their earlier work with three different implementations of a reference method (EXCOM), and discuss the differences between these and other approaches. This conceptual comparison is accompanied by empirical evaluation with respect to the efficiency and quality of detected patterns, which allows the authors to identify the pros and cons of the methods and come up with practical recommendations concerning their applicability.

In the third paper, entitled *On the predictive power of meta-features in OpenML*, B. Bilalli, A. Abelló, and T. Aluja-Banet consider the problems of choosing the best algorithm for learning classifiers from complex data. Their main idea is to exploit the methods of feature extraction and selection for a novel purpose of studying the predictive power of meta-features in a meta-learning scenario. First, they handcraft the latent features, which are abstract concepts that group together meta-features with common characteristics. Then, they analyze the relationship of the latent features with different performance measures for various classification algorithms on hundreds of benchmark datasets. This, in turn, allows them to select the latent features with higher predictive power for performing meta-learning. The experiments show that applying this method improves the meta-learning process. Furthermore, it is argued that the benefit of working at the latent feature level is easy generalization to new future datasets by omitting the absent meta-features or by including new ones, provided that they are related with the actual concepts present in the latent features. The meta-features and datasets considered come from OpenML, the leading collaborative effort in collecting machine learning experiments, datasets,

[1] `http://www.cs.put.poznan.pl/events/PAS-seminar-pl.html`.

and metadata. The authors also present their software tool that allows generating meta-datasets for meta-learning from OpenML.

The fourth paper, entitled *Efficient storage, retrieval and analysis of poker hands: An adaptive data framework*, by M. Gorawski and M. Lorek, proposes an architecture for storing and efficiently retrieving massive (but simple) data sets on poker games. Data items are divided into the so-called poker hands (of about 715 bytes each, when serialised and compressed). The total number of hands stored in a repository is about 40 billion. Data are stored in a NoSQL database distributed in a cluster, with over 4000 partitions, and can be accessed either via the Hive or Spark platforms. In order to efficiently retrieve poker hands of interest, the authors proposed three techniques for partition elimination, based on two alternative designs of global indexes, namely, the Player-Day Index (PDIX) and the To-Day Index (TDIX). The experimental evaluation of the developed architecture (in a 20-node cluster) with respect to data retrieval revealed that it offers shorter retrieval times than the competitive standard Hadoop architecture.

The fifth paper, entitled *CCR: A combined cleaning and resampling algorithm for imbalanced data classification*, by M. Koziarski and M. Woźniak, concerns the challenges of improving classifiers learned from class-imbalanced data. The authors focus on preprocessing methods that transform the original distribution of training examples in order to make it better for inducing classifiers. After summarizing the current research on related preprocessing techniques, the authors introduce combined cleaning and resampling (CCR). Its key concept is cleaning the neighborhoods of minority examples, which is realized by a novel energy-based technique for pushing away the majority examples located too closely to the minority ones. In the second phase of CCR, synthetic minority examples are additionally generated near the most difficult examples. The results of an experimental study demonstrate that the CCR algorithm outperforms several over-sampling methods with respect to recognition of the minority class.

The sixth paper, entitled *Interpretable decision-tree induction in a big data parallel framework*, by A.I. Weinberg and M. Last, is devoted to efficient training of classifiers in parallel, distributed big data frameworks, which allow handling large data sets efficiently. The focus of the paper is on decision tree algorithms, which typically perform well and offer comprehensible representation. The authors aim at inducing one representative decision tree in the big data distributed framework and introduce the syntactic similarity method (SySM) algorithm. Running as part of the combining (MapReduce) phase, it computes the similarity between the classifiers produced by parallel nodes and chooses the tree which is the most similar to other trees as the best representative of the entire dataset. Notably, this approach is different from ensemble methods that combine predictions of several decision trees that serve as base estimators. The empirical results demonstrate that SySM is not only more accurate than other classifiers, but also produces more compact and interpretable trees with lower computational effort.

We take this opportunity to thank all authors for submitting their papers to this special section. Our project came to the final success only because of their hard work and smooth cooperation in the revision phase. We also wish to express our gratitude to all colleagues who assisted us in the reviewing process, providing insightful feedback that helped the authors to significantly improve their manuscripts.

Continuous guidance and support of the *AMCS* Editorial Office Manager, Ms. Agnieszka Rożewska, and the entire editorial team is also appreciated. Finally, we owe a vote of thanks to Professor Józef Korbicz, the Editor-in-Chief, for his initiative of organizing this special section as well as his excellent co-operation and support for our efforts.

Jerzy Stefanowski
Krzysztof Krawiec
Robert Wrembel

December 2017

Poznań University of Technology, Poland

Jerzy Stefanowski is an associate professor in the Institute of Computing Science, Poznań University of Technology. His research interests include machine learning, data mining and intelligent decision support—in particular, rule induction, multiple classifiers, class imbalance, concept drift, classification of data streams and big data. For more information, refer to `http://www.cs.put.poznan.pl/jstefanowski`.

Krzysztof Krawiec is an associate professor in the Institute of Computing Science, Poznań University of Technology. His recent work includes program synthesis (in particular, by means of genetic programming) evolutionary computation for machine learning, learning game strategies and pattern recognition; deep learning for image analysis and game playing; coevolutionary algorithms and test-based problems. He is an associate editor of the *Genetic Programming and Evolvable Machines* journal, and has participated in research projects at the University of California and the Massachusetts Institute of Technology. For more information, refer to `http://www.cs.put.poznan.pl/kkrawiec`.

Robert Wrembel is an associate professor in the Faculty of Computing, Poznań University of Technology, Poland. His main research area is data warehouse systems. He has been involved in 7 research projects in the area of databases and data warehouses as well as in 7 industrial projects in the field of information technologies. He has prepared and delivered numerous courses in the area of programming languages, database administration, and information systems designing, for multiple companies and institutions in Poland, including Oracle, Microsoft, IBM, BAE Systems. He has been a visiting professor at Loyola University (New Orleans, USA), an invited lecturer at Universidad de Costa Rica (San Jose, Costa Rica), a graduate in the Stanford University post-graduate programme on entrepreneurship and innovation, an intern at the BI company Targit (Tampa, USA), a prizewinner of the IBM Faculty Award for highly competitive research. For more information, refer to `http://www.cs.put.poznan.pl/rwrembel`.