Editorial

Anna Kobusińska

Institute of Computing Science,
Poznań University of Technology,
ul. Piotrowo 3, 60-965 Poznań, Poland
Email: Anna.Kobusinska@cs.put.poznan.pl

Biographical notes: Anna Kobusińska received her MSc, PhD and Dr. Habil. degrees in Computer Science from the Poznań University of Technology. She currently works as an Associate Professor at the Laboratory of Computing Systems, Institute of Computing Science, Poznań University of Technology, Poland. Her research interests include big data and large-scale distributed systems, service-oriented systems and cloud computing. She focuses on distributed algorithms, big data analysis, replication and consistency models, as well as fault-tolerance, specifically checkpointing and rollback recovery techniques. She has served and is currently serving as a PC member of several international conferences and workshops. She is also the author and co-author of many publications in high quality peer-reviewed international conferences and journals. She participated to various research projects supported by national organisations and by EC in collaboration with academic institutions and industrial partners.

With the advent of big data and the need to store and analyse it, big data infrastructures and deep learning methods have become more important than ever. As a result, possessing the right tools for storing and processing rising data volumes is crucial. Also, analysing massive amounts of unsupervised data through deep learning algorithms is on demand. Therefore, this special issue presents the selected research papers that feature some of the most recent and emerging big data infrastructures and discusses tools and software systems for collecting and storing big data. It also proposes papers that exemplify real-life applications of deep learning algorithms that extract high-level, complex abstractions as data representations through a hierarchical learning process.

The papers chosen for the publication in this special issue are the extended versions of papers from the third International Conference on Big Data Intelligence and Computing (DataCom), held in 2017, in Orlando, Florida, which were selected after several rounds of meticulous review process.

The first paper 'A computational Bayesian approach for estimating density functions based on noise-multiplied data', written by Y-X. Lin provides a new data-mining technique and introduces the B-M L2014 approach for estimating the density function of the original data based on noise-multiplied microdata.

Y. Abduallah and J.T.L. Wang in the paper 'New algorithms for inferring gene regulatory networks from time-series expression data on Apache Spark' present two new algorithms for reverse engineering gene regulatory networks (GRNs) in a cloud environment. The algorithms, implemented in Spark, employ an information-theoretic approach to infer GRNs from time-series gene expression data.

Next, in the paper 'A scalable system for executing and scoring K-means clustering techniques and its impact on applications in agriculture' by N. Golubovic et al., the Centaurus clustering service for K-means clustering of correlated, multidimensional data and its application to a real-world, agricultural analytics is introduced.

In turn, S. Arifuzzaman and B. Pandey, in 'Scalable mining, analysis and visualisation of protein-protein interaction networks' propose and describe a lightweight framework on a distributed-memory parallel system, which includes scalable algorithmic and analytic techniques to study protein-protein interaction networks and visualise them.

L. Li et al. in the paper 'Optimising NBA player signing strategies based on practical constraints and statistics analytics' studied the key indicators used to measure player efficiency and team performance of NBA players. He used data analytics to formulate the prediction of the team winning rate in different schemes. Based on the proposed models that were trained and tested, two player selection strategies were proposed according to different objectives and constraints. The obtained results were proven experimentally.

The paper 'Text visualisation for feature selection in online review analysis' by K. Koka and S. Fang focuses on identifying the fake online reviews. The authors propose the feature selection through visualisation by applying radial chart visualisation technique to the online review classification into fake and genuine reviews. The presented visualisation technique helps the user get fast and straightforward insights into the high dimensional data, as well as to eliminate the worst features and pick best features without statistical aids.

D. Gligoroski et al. in their work 'Network traffic driven storage repair' proposed the explicit construction of the family of locally repairable and locally regenerating codes. In their solution they used in hashtag codes that can have different sub-packetisation levels. They claim that repair

strategy is network traffic driven and emphasise the importance of having two ways to repair a node: repair only with local parity nodes or repair with both local and global parity nodes. The repair duality is illustrated by a practical example implemented in Hadoop.

Next paper, called 'DeepSim: cluster level behavioural simulation model for deep learning' by Y. Shi et al. proposes a cluster-level behavioural simulation model for deep learning, called DeepSim. DeepSim is based on the Intel CoFluent framework. It enables scalable system design, deployment, and capacity planning through accurate performance insights. The paper presents the results from preliminary scaling studies

O.M. Kumar et al. propose in the paper 'MapReduce-based fuzzy very fast decision tree for constructing prediction intervals' the fuzzy version of very fast decision tree (VFDT) to predict the lower upper bound estimation (LUBE). The authors implemented VFDT, developed and implemented fuzzy VFDT using Apache Hadoop MapReduce framework, where multiple slave nodes build a VFDT and fuzzy VFDT model, and demonstrated that the proposed MapReduce-based fuzzy VFDT and VFDT can construct high-quality prediction intervals precisely and quickly.

The consecutive paper, entitled 'Real-time event search using social stream for inbound tourist corresponding to place and time' proposes a tourist information distribution system. R. Kudo et al. in their solution extract event information from social media streams in a per place and time manner and provide it to tourists. For this reason they used event classification using Twitter data.

H. Wang et al. in the paper 'Two-channel convolutional neural network for facial expression recognition using facial parts' proposed the design of a facial expression recognition system based on the deep convolutional neural network. In the proposed solution the authors used combination of algorithms for face detection, feature extraction and classification. The solution was evaluated by using the Japanese female facial expression dataset and the extended Cohn-Kanada dataset.

Next, in the paper 'Efficient clustering techniques on Hadoop and Spark', S. Al Ghamdi and G. Di Fatta present K-means optimisations in order to provide solution that copes with large-scale datasets. The proposed solution uses triangle inequality on two well-known distributed computing platforms: Hadoop and Spark. The performed evaluation tests proved that the efficiency of optimised K-means on Hadoop and Spark has been significantly improved.

Then, A. Azimzadeh et al. in the work 'A hybrid power management schema for multi-tier data centres' tackle the problem of energy efficiency in data centres. The authors use dynamic power management policy-based model and validate the results and energy efficiency by calculating the average power consumption of each server under specific sleep mode and setup time.

In the paper 'Predicting hospital length of stay using neural networks', T. Gentimis et al. utilise neural networks for predicting the length of hospital stay for patients with various diagnoses based on selected administrative and clinical attributes. Authors train an all-condition neural network and use MIMIC III database as a dataset. The prediction accuracy of the proposed solution outperforms any linear model.

The last paper of the special issue, called 'Towards an automation of the fact-checking in the journalistic web context' conducts a detailed study on the automation of the fact-checking. E.N. Sarr et al. discussed the state of the art based on the perspectives via sites, the projects, the algorithms and the obstacles. The paper presents numerous solutions and algorithms, gives the diagnosis of the obstacles, and tackles the perspectives on the question of the automation of the fact-checking.

As a guest editor of this special issue, I would like to thank all the authors for their contributions and the reviewers for their great effort in the SI review process. I would also like to thank the Editor-in-Chief, Prof. Ching-Hsien Hsu and his entire Inderscience team for their continuous support. I believe that the papers from this special issue will contribute to the further development of big data infrastructures and deep learning algorithms, and inspire future research in the field.