Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 258 (2015) 1–4

FUZZY sets and systems

www.elsevier.com/locate/fss

Editorial

Editorial: Uncertainty in learning from big data

Big data refers to datasets that are so large that conventional database management and data analysis tools are insufficient to work with them. Big data, which was called massive data [1], has become a bigger-than-ever problem with the quick developments of data collection and storage technologies. Nowadays, many complex processes can generate big data, for example, there are a greater number of Earth Observing Satellites than ever before, collecting many terabytes of data per day. Also, engineered systems have greatly increased capabilities for sensing their environment, that results in a considerable increase of sensing data. In addition, the Internet has greatly added to the volume and heterogeneity of data available, in other words, the World Wide Web contains an enormous volume of text, images, videos, and connections between these, etc.

The main feature of big data is its 4V-characteristics. Gartner in 2012 gave an updated definition of big data, i.e., big data is high Volume, high Velocity, high Variety and/or high Value information assets that require cost-effective and innovative forms of processing to enable enhanced decision making, insight discovery and process optimization [2]. In addition, big data has two more features, i.e., multimodality and changed-uncertainty. The former means that the types of data can be very complex while the latter indicates that the modeling and measure of uncertainty for big data is significantly different from that for normal sized data.

The mixed features of big data bring unprecedented challenges in processing and management of big data. ⟨⟨Nature⟩⟩ in 2010 published an article pointing out that combining text, image, speech, video and the other multimodal data will become the main format of future information [3]. Furthermore, ⟨⟨Science⟩⟩ in 2011 organized a special issue "*Dealing with Data*", which proposed that, multimodality and un-structure of information in medical treatment, finance, and network environment brought a new challenge to information processing, machine learning, and data mining [4].

Studies on big data are relevant to many current fields such as biological sciences [5–7] (e.g., a single human genome consists of about 3 billion base pairs of DNA) and social networks [8–10] (e.g., Facebook and Twitter respectively have about 2.2 billion and 0.5 billion users up to now). Traditional data analysis methods and computational models focusing on small/large data and running on single processor are incapable to deal with data repositories of such massive size. Thus, new methodologies and technologies are required for analyzing and handling big data. Developing efficient learning methodologies and techniques is crucial to analyze and understand big data so as to extract useful knowledge. Feasible and potential strategies for analyzing and processing big data may include (1) parallel and distributed computation, (2) instance selection and dimensionality reduction, and (3) incremental learning. Briefly and incompletely we give a summary on developments of schemes based on these strategies.

(1) The parallelization of learning algorithm with MapReduce programming model is a primary way to process big data. The following two tutorials [11,12] present some very useful instructions to conduct Hadoop based parallel big data processing. Dittrich and Quiané-Ruiz [11] highlighted the similarities and differences between Hadoop MapReduce and parallel DBMS, and further pointed out unresolved research problems and open issues in Hadoop based big data processing. Shim [12] discussed how to design the efficient MapReduce algorithms for data mining, machine learning and similarity joins. And, some researchers have implemented Hadoop based parallel decision trees [13], extreme learning machines [14], and Bayesian networks [15] for big data processing. The experimental results reported good performances for the corresponding parallel algorithms.

http://dx.doi.org/10.1016/j.fss.2014.10.010

(2) The central idea of instance selection and dimensionality reduction strategies is to turn big into small. Instance selection or sampling is an effective technique to reduce size of the original data sets by selecting the representative instances. In [16], Olvera-López et al. gave a good review on instance selection methods, where most of them focus on small and medium data sets and only two studies [17,18] for large data set. Dimensionality reduction including feature selection and feature extraction aims to obtain a representative subset which has much less features in comparison with original feature space. The review presented in [19] addresses all main techniques for dimensionality reduction. The latest study regarding dimensionality reduction for large data has been given in [20].

(3) Incremental learning gradually improves the parameters in learning algorithms by using new samples rather than training again the learning algorithms with all available samples. This makes it possible to deal with big data in the form of data stream. Recently, Yang and Fong [21] and Wang et al. [22] respectively provided different incremental learning strategies for noisy and spatial big data. Incremental leaning is becoming a potentially promising and effective way to handle big data and is attracting more and more scholars to the field of big data processing.

During the recent years, one can view a rapid growth in the hybrid study which connects together uncertainty and learning from data (e.g. [23–26]). How to express, model, and handle uncertainty has become a key challenge in learning from data. It has a very significant impact on the entire knowledge extraction process. It is worth noting that uncertainty is a common phenomenon in learning and mining, which can be embedded in the entire process of learning and reasoning including data acquisition (such as noisy, incomplete, heterogeneous and dynamic data), data representation (such as data structure, organization, topology, and transformation), and data learning (such as the choice of learning methods, extraction of core knowledge, determination of decision rules, and improvement of robustness and generalization capability), etc. The representation, measure, and handling of uncertainty have a significant influence on the performance of learning from big data. Without dealing properly with these uncertainties, the performance of learning strategies may be greatly degraded.

Many theories and methodologies have been developed to model different kinds of uncertainties. For example, fuzzy set theory for imprecise information, probability theory for randomness, classification entropy for the impurity of a set regarding the classes, rough set theory for the approximation of concepts, etc. In addition, quantified uncertainties can guide or assist when building a more accurate learning system. How to effectively model these uncertainties including their representation and processing has become the key to obtain a robust data-mining algorithm with good generalization capability in building a high-performance learning system. Here, focusing on learning from big data, one extremely important issue is how to adapt those models and methodologies for massive data? Thus, in order to share the latest progress, current challenges and potential applications of handling uncertainty in learning from big data, we edited this special issue of Fuzzy Sets and Systems.

The special issue includes six papers among which three belong to the scope of parallelization-based big data processing, two belong to the area of instance selection and dimensionality reduction, and one belongs to the filed of incremental learning. The following is a brief introduction to the 6 contributions.

The paper "Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data" authored by Victoria López, Sara del Río, José Manuel Benítez, Francisco Herrera, mainly considers the uncertainty for imbalanced big data classification. In this paper, a fuzzy rule based classification system named Chi-FRBCS-BigDataCS is proposed to deal with imbalanced big data. Chi-FRBCS-BigDataCS uses the MapReduce framework to distribute the computational operations of the fuzzy model while it includes cost-sensitive learning techniques in its design to address the imbalance that is present in the data. This paper belongs to the scope of parallelization-based big data processing.

The paper "Fuzzy-rough feature selection accelerator" authored by Yuhua Qian, Qi Wang, Honghong Cheng, Jiye Liang, Chuangyin Dang, focuses on uncertainty in fuzzy-rough feature selection algorithms for big data. This paper proposes an acceleration strategy for heuristic process of fuzzy-rough feature selection, which combines sample reduction and dimensionality reduction together. Through the use of this acceleration mechanism, three representative heuristic fuzzy-rough feature selection algorithms are enhanced when selecting the optimal feature subset from the given big data. This paper belongs to the scope of instance selection and dimensionality reduction based big data processing.

In the paper "Learning ELM-tree from big data based on uncertainty reduction" authored by Ran Wang, Yu-Lin He, Chi-Yin Chow, Fang-Fang Ou, Jian Zhang, an extreme learning machine tree (ELM-Tree) model based on the heuristics of uncertainty reduction is proposed. In ELM-Tree model, information entropy and ambiguity are used as uncertainty measures for splitting decision tree (DT) nodes. Besides, in order to solve the over-partitioning problem in

DT induction, ELMs are embedded as leaf nodes when the gain ratios of all the available splits are smaller than a given threshold. Then, a parallel ELM-Tree model is developed to reduce the computational time for big data classification. This paper belongs to the scope of parallelization based big data processing.

The paper "Mining of protein–protein interfacial residues from massive protein sequential and spatial data" authored by Debby D. Wang, Weiqiang Zhou, Hong Yan, studies uncertainty in the identification of protein–protein interfacial residues from big protein structural data. A series of popular learning procedures including neuro-fuzzy classifiers, CART, neighborhood classifiers, extreme learning machines and naive Bayesian classifiers are tested in order to predict the interfacial residues, aiming to investigate the sensitivity of these massive structural data to different learning mechanisms. This paper belongs to the scope of dimensionality reduction based big data processing.

In the paper "Parallel sampling from big data with uncertainty distribution" authored by Qing He, Haocheng Wang, Fuzhen Zhuang, Tianfeng Shang, Zhongzhi Shi, a parallel sampling method based on hypersurface for big data with uncertainty distribution, namely PSHS, is proposed, where PSHS adopts a universal concept of minimal consistent subset of hypersurface classification and is parallelized in the framework of MapReduce. The experimental results show that PSHS can shrink big data sets while maintaining identical distributions, which is useful for obtaining the inherent structure of the data sets. This paper belongs to the scope of parallelization based big data processing.

The paper "A fuzzy rough set approach for incremental feature selection on hybrid information systems" authored by Anping Zeng, Tianrui Li, Dun Liu, Junbo Zhang, Hongmei Chen, proposes a fuzzy-rough set based incremental feature selection method on hybrid information system (HIS) to handle different kinds of big data, e.g., Boolean, categorical, real-valued and set-valued big data. The fuzzy-rough set in HIS is constructed by using a new hybrid distance and Gaussian kernel. Experimental results indicate that the incremental approaches significantly outperform non-incremental approaches with feature selection in terms of computation time. This paper belongs to the scope of incremental learning based big data processing.

We would be happy if this special issue can provide some useful references for those who are investigating theoretically or practically the processing and analyzing of big data in different fields.

Acknowledgements

We would like to thank all the authors who submitted papers for consideration to the special issue. We totally received 21 submissions from which 6 papers are selected for publication in this special issue based on the review comments. Thanks go to our reviewers for their time, their valuable comments, and their suggestions to improve the manuscripts. Special thanks go to Eyke Hüllermeier, one of the editor-in-chiefs of this journal, for his hard work in processing submissions of this special issue. Finally, we would like to thank the editorial staff for their support in the production of this special issue. This work is partially supported by NSFC (71371063 and 61170040).

References

[1] National Research Council, Frontiers in Massive Data Analysis, The National Academies Press, Washington, DC, 2013.

[2] http://www.gartner.com/it-glossary/big-data/.

[3] P. Norvig, D.A. Relman, D.B. Goldstein, et al., 2020 Visions, Nature 463 (2010) 26–32.

[4] http://www.sciencemag.org/site/special/data/.

[5] V. Marx, Biology: the big challenges of big data, Nature 498 (2013) 255–260.

[6] J. Boyle, Biology must develop its own big-data systems, Nature 499 (2013) 7.

[7] N. Savage, Bioinformatics: big data versus the big C, Nature 509 (2014) S66–S67.

[8] D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google flu: traps in big data analysis, Science 343 (6176) (2014) 1203–1205.

[9] D.A. Broniatowski, M.J. Paul, M. Dredze, Twitter: big data opportunities, Science 345 (6193) (2014) 148.

[10] D. Lazer, R. Kennedy, G. King, A. Vespignani, Twitter: big data opportunities–response, Science 345 (6193) (2014) 148–149.

[11] J. Dittrich, J.A. Quiané-Ruiz, Efficient big data processing in Hadoop MapReduce, Proc. VLDB Endow. 5 (12) (2012) 2014–2015.

[12] K. Shim, MapReduce algorithms for big data analysis, Proc. VLDB Endow. 5 (12) (2012) 2016–2017.

[13] W. Dai, W. Ji, A MapReduce implementation of C4.5 decision tree algorithm, Int. J. Database Theory Appl. 7 (1) (2014) 49–60.

[14] Q. He, T. Shang, F. Zhuang, Z. Shi, Parallel extreme learning machine for regression based on MapReduce, Neurocomputing 102 (2013) 52–58.

[15] A. Basak, I. Brinster, X. Ma, O.J. Mengshoel, Accelerating Bayesian network parameter learning using Hadoop and MapReduce, in: Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, 2012, pp. 101–108.

[16] J.A. Olvera-López, J.A. Carrasco-Ochoa, J.F. Martínez-Trinidad, J. Kittler, A review of instance selection methods, Artif. Intell. Rev. 34 (2) (2010) 133–143.

[17] J.R. Cano, F. Herrera, M. Lozano, Stratification for scaling up evolutionary prototype selection, Pattern Recognit. Lett. 26 (7) (2005) 953–963.

[18] A. de Haro-García, N. García-Pedrajas, A divide-and-conquer recursive approach for scaling up instance selection algorithms, Data Min. Knowl. Discov. 18 (3) (2009) 392–418.

[19] L.J.P. van der Maaten, E.O. Postma, H.J. van den Herik, Dimensionality reduction: a comparative review, J. Mach. Learn. Res. 10 (1–41) (2009) 66–71.

[20] B. Hammer, M. Biehl, K. Bunte, B. Mokbel, A general framework for dimensionality reduction for large data sets, in: Lecture Notes in Computer Science, vol. 6731, 2011, pp. 277–287.

[21] H. Yang, S. Fong, Incrementally optimized decision tree for noisy big data, in: Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, 2012, pp. 36–44.

[22] L. Wang, L. Ke, P. Liu, R. Ranjan, L. Chen, IK-SVD: dictionary learning for spatial big data via incremental atom update, Comput. Sci. Eng. (2014), http://dx.doi.org/10.1109/MCSE.2014.52.

[23] X.Z. Wang, Y.L. He, D.D. Wang, Non-naive Bayesian classifiers for classification problems with continuous attributes, IEEE Trans. Cybern. 44 (1) (2014) 21–39.

[24] R. Wang, D.G. Chen, S. Kwong, Fuzzy rough set based active learning, IEEE Trans. Fuzzy Syst. (2014), http://dx.doi.org/10.1109/TFUZZ.2013.2291567.

[25] X.Z. Wang, L.C. Dong, J.H. Yan, Maximum ambiguity based sample selection in fuzzy decision tree induction, IEEE Trans. Knowl. Data Eng. 24 (8) (2012) 1491–1505.

[26] X.Z. Wang, C.R. Dong, Improving generalization of fuzzy if-then rules by maximizing fuzzy entropy, IEEE Trans. Fuzzy Syst. 17 (3) (2009) 556–567.

Xizhao Wang
Joshua Zhexue Huang
Big Data Institute,
College of Computer Science
and Software Engineering,
Shenzhen University,
China
E-mail address: xizhaowang@ieee.org (X. Wang)

6 September 2014

Available online 18 October 2014