Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 126740, 2 pages
http://dx.doi.org/10.1155/2015/126740

Editorial

Information Analysis of High-Dimensional Data and Applications

Xin-She Yang,[1] Sanghyuk Lee,[2,3] Sangmin Lee,[4,5] and Nipon Theera-Umpon[6,7]

[1] *School of Science and Technology, Middlesex University London, London NW4 4BT, UK*
[2] *Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, High Educational Town, SIP, Suzhou 215123, China*
[3] *Centre for Smart Grid and Information Convergence, Xi'an Jiaotong-Liverpool University, High Educational Town, SIP, Suzhou 215123, China*
[4] *Department of Electronic Engineering, Inha University, Inha-ro 100, Incheon, Republic of Korea*
[5] *Institute for Information and Electronics Research, Inha University, Inha-ro 100, Incheon, Republic of Korea*
[6] *Department of Electrical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand*
[7] *Biomedical Engineering Center, Chiang Mai University, Chiang Mai 50200, Thailand*

Correspondence should be addressed to Xin-She Yang; x.yang@mdx.ac.uk

Received 28 September 2015; Accepted 29 September 2015

1. Introduction

Big data is becoming one of the hottest topics in current research in computer science, data mining, engineering, and applicable mathematics. In fact, the diverse research activities surrounding the big data are so vast that they form a new discipline, namely, the data science. There are many challenging issues associated with big data [1, 2], and one very important issue is the high-dimensional data analysis. Even with some moderate size data, high-dimensionality can pose extra challenges. High-dimensionality in combination with large datasets can be extremely challenging. High-dimensional data are relevant to a wide range of fields such as biometric, medicine, e-commerce, network security, and industrial applications. In order to use data characteristics, proper techniques and methods are needed to handle such high-dimensional data [3]. Furthermore, data can have atypical characteristics and high-dimensional data structures, which means that conventional analysis techniques do not work well. To analyse extra useful information from high-dimensional data, novel approaches are required.

2. Main Challenges

Among the many challenging issues concerning big data and high-dimensional data, we highlight the following five major challenges:

(i) For high-dimensional datasets, there is the so-called curse of dimensionality: the searchable volume in the hyperspace becomes small, compared with the vast feasible search space. Thus, any solution procedure can only sample a subset of sparse points with essentially zero sampling volume in order to make sense of the vast datasets. Thus, it is a huge challenge with an almost impossible task for finding the global optimality. In addition, the distance measures required for problem formulations become less meaningful as any finite distance will result in an almost zero ratio between the distance measure and the vast distance needed to cover in the high-dimensional search space.

(ii) As the number of dimensions increases, the number of features also increases, often far more rapidly, which means that there is huge sparsity associated with such high-dimensional features. In addition, some correlation may exist between different dimensions, and thus features can be difficult to define.

(iii) For high-dimensional data, datasets tend to be unstructured, which may pose extra challenges to use. In addition, noise and uncertainties often exist in big datasets. Such noisy data can become more challenging to process and to apply any proper data mining techniques. For such problems, there is no analytical approach to provide insight even for a small subset of problems. Therefore, algorithms tend to be problem specific and even data specific. Thus, there is no generic approach in general.

(iv) As the number of dimensions increases, the possible combinations of clusters grow exponentially, and clustering becomes nondeterministic polynomial-time hard (NP-hard), and thus there are no efficient methods to deal with such challenging problems.

(v) Even with the steady increase in speed of modern computers and the availability of cheaper parallel and cloud computing facilities, this does not ease the challenges of high-dimensional information analysis. Efforts on developing new methods and tools are still highly needed. It may need a paradigm shift and a nonconventional way of thinking to problem-solving concerning high-dimensional data.

These challenges mean that new methods and alternative approaches are needed to solve such tough problems [4]. In fact, heuristic and metaheuristic algorithms have been proven to be a promising set of alternative methods, especially those metaheuristic approaches based on nature-inspired optimization algorithms [5].

3. Recent Developments

This special issue strives to provide a timely platform to discuss and summarize the latest developments in this area. The emphasis has been on the theoretical methodology and mathematical analysis, though applications concerning high-dimensional data are also the focus. The responses were well received with a high number of high-quality submissions. After the rigorous peer-review process, the accepted papers, with an acceptance rate of about 23%, represent an extensive snapshot of the recent developments.

From this special issue, we can see that topics include from information analysis of high-dimensional data and feature selection to biomedical applications and real-world applications in engineering. First, N. Eiamkanitchat et al. present a study on feature selection and rule extraction for high-dimensional data in the context of microarray classification and then J. G. Lim et al. study the feature selection related to motion segmentation data, while D. Peralta et al. present a MapReduce approach for feature selection and big data classification. In addition, M. Li et al. investigate the protein-named entity recognition and protein-protein interaction extraction and B.-J. Yi et al. study the effects of feature optimization concerning high-dimensional essay data, whereas I. Shin et al. study fall detection using three-axis acceleration data and K. Park et al. present a learner profiling system using multidimensional characteristics analysis.

On the other hand, S. U.-J. Lee uses automated product configuration for software product line development and P. Campadelli et al. estimate intrinsic dimensions and also propose a benchmark framework. Then, N. Mishra et al. present a framework for semantic knowledge analytics and B. Kim et al. apply genetic algorithms to generate high-dimensional item response data. In addition, S. Kim et al. study wireless sensor networks with high-dimensional data aggregates and Y. Piao et al. propose an ensemble method for feature space partitioning and high-dimensional data classification. Furthermore, Y. Jeong and S. Shin use data connection coefficients for presenting a multidata connection scheme in the context of big data. In the real-world engineering applications, C. Liu et al. use a modified firefly algorithm to plan three-dimensional paths for autonomous underwater vehicle navigation in the complex terrains and D. Xiao et al. present a study of process monitoring for shell rolling production and seamless tubes.

Obviously, there are many other interesting developments concerning the information analysis of high-dimensional data. This special issue can only cover a small fraction of the latest developments. It is hoped that this special issue can inspire more research in this area in the near future, especially about the five challenging issues outlined above. Any progress in these areas will no doubt provide more insight into the understanding of high-dimensional data and the development of more effective tools.

Xin-She Yang
Sanghyuk Lee
Sangmin Lee
Nipon Theera-Umpon

References

[1] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2nd edition, 2014.

[2] H. Samet, *Foundations of Multidimensional and Metric Data Structures*, Morgan Kaufmann Publishers, 2006.

[3] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2005.

[4] X. S. Yang, *Recent Advances in Swarm Intelligence and Evolutionary Computation*, Springer, 2015.

[5] X. S. Yang, *Nature-Inspired Optimization Algorithms*, Elsevier, 2014.