Guest Editorial for ACM TECS Special Issue on Effective Divide-and-Conquer, Incremental, or Distributed Mechanisms of Embedded Designs for Extremely Big Data in Large-Scale Devices

As the size of data grows at a rapid speed, the demand for big data analysis has significantly increased in recent years, ranging from decentralized data centers, cloud servers, and sensor networks to the Internet of Things (IoT). Big data usually come with two types. One has an excessively huge volume of samples, and the other exhibits extremely large dimensions. Both of them jeopardize the performance of computing systems. Data analysts usually have to deal with limited storage, processing speed, communication bandwidth, and power consumption, especially when resources are constrained. What is worse, in the era of big data, the scale that engineers are handling is beyond billions. Moreover, when plenty of devices interact with each other (e.g., mobile devices, sensors, and the IoT), they form a large-scale network. This has relatively deepened the difficulty of data analysis.

In view of this problem, many famous and off-the-shelf tools, such as Apache Hadoop and Apache Spark, are developed to handle large-scale analyses by using distributed processing. These tools usually employ divide-and-conquer architectures in their implementations. With such architectures, when a large-scale dataset is segmented into subsets, the original problem can accordingly be divided into subproblems, separately processed in each machine. Despite the convenient framework, there is still no effective solution to embedded devices. Besides, not every algorithm can be converted into a divide-and-conquer version and gives optimal solutions.

In contrast to divide-and-conquer mechanisms, incremental analysis is another approach for dealing with big data because it does not rely on distributed architectures. Incremental analysis allows the system to add new samples without reprocessing whole samples because earlier calculation results can be reserved for updating the system in the future. Furthermore, if the data size is far beyond the capability of one machine, the entire dataset is divided into several batches, subsequently processed by a single machine. The combination of incremental and divide-and-conquer mechanisms yields more flexibility than an individual one. Therefore, how to take advantage of such a combination to resolve big data problems is the prior concern.

The articles selected from submission presented a recent update that covers front-end data aggregation and back-end data postprocessing. The work "Distributed Multirepresentative Refusion Approach for Heterogeneous Sensing Data Collection" by A. Liu et al. examined data collection for sensor networks and the IoT. They developed an effective refusion method for handling data pooling in sink nodes. Regarding back-end techniques, the studies by X. Chen et al., B. Liu et al., and S.-Y. Kung et al. respectively presented new solutions to supervised analysis, unsupervised techniques, and dimensional reduction. These approaches are widely used in machine learning. The work "A Load-Balancing Divide-and-Conquer SVM Solver" by X. Chen et al. advanced the idea of divide-and-conquer Support Vector Machines (SVMs) by devising a load-balancing mechanism for distributed and embedded systems. B. Liu et al. presented "Decentralized Sparse Subspace Clustering" that partitioned a large data matrix into

1539-9087/2017/04-ART72
DOI: http://dx.doi.org/10.1145/3068457

column blocks, so that the original problem was resolved by the parallel multivariate LASSO (Least Absolute Shrinkage and Selection Operator) and the sample-wise operator. Clustering was carried out by the Alternating Direction Method of Multipliers (ADMM). The work "Collaborative PCA/DCA Learning Methods for Compressive Privacy" by S.-Y. Kung et al. investigated representative component problems in observed data. They furthered typical Principal Component Analysis (PCA) and Fisher Discriminant Analysis (FDA) by proposing Discriminant Component Analysis (DCA). Rather than performing subspace analysis in signal space, DCA searched discriminant dimensions in whitened space, where the effect of each dimension was normalized, and isotropic dimensions were created.

Through these papers, readers can obtain the state-of-the-art technologies of recent data analytics. Readers can also understand several practical issues. Finally, the guest editors would like to thank the anonymous reviewers, who gave valuable comments to the authors. Without their help, nothing can be smoothly achieved.

Bo-Wei Chen
School of Information Technology, Monash University, Australia

Wen Ji
Institute of Computing Technology, Chinese Academy of Sciences, China

Zhu Li
Department of Computer Science and Electrical Engineering,
University of Missouri-Kansas City, USA

Guest Editors