Big Data

Weike Pan, *Shenzhen University*

Qiang Yang, *Hong Kong University of Science and Technology*

Charu Aggarwal, *IBM T.J. Watson Research Center*

Christoph Koch, *Swiss Federal Institute of Technology*

Big data has been an enabler for innovation, reconstruction, and advancement of most sectors of our society, and it's receiving continuous and growing attention from researchers and practitioners in academia, industry, and government. There are, however, still lots of challenges spanning from theoretical foundations, systems, and technology to data policy and standards. This special issue focuses on how big data cuts across systems and applications arenas.

We received 30 submissions and accepted 5 articles after several rounds of reviewing. These five articles cover a wide spectrum of interesting topics, including feature selection for big data analytics, astronomical image analysis, large-scale network prediction, online URL filtering, and massive transaction clustering.

The first article, "Challenges of Feature Selection for Big Data Analytics," by Jundong Li and Huan Liu, gives a brief introduction of the feature selection problem and some typical algorithms that tackle it before exploring six challenges in the context of big

1541-1672/17/$33.00 © 2017 IEEE
Published by the IEEE Computer Society

data analytics. The authors paid particular attention to a feature selection repository called scikit-feature, an open source repository that's expected to be very useful for the machine learning and big data community.

The second article, "Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy," by Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, and Christian Igel, introduces the current status of big survey data in astronomy and the challenges of handling them, and then presents several successful stories and preliminary work of applying machine learning and image analysis techniques. The authors also discuss the differences and relationships between physical and machine learning models and provide some guidance on how to start exploration of astronomy data. We agree that there are many interesting research problems such as learning from biased and noisy astronomy data that provide great opportunities in new discoveries in astrophysics and cosmology.

The third article, "Structured Regression on Multiscale Networks," by Jesse Glass and Zoran Obradovic, focuses on speeding up a representative structure-based regression algorithm called Gaussian conditional random field (GCRF) via multiscaled networks. The authors first decompose a large network and integrate the resulting multiscale networks jointly via the Kronecker product of matrices. They then design a novel algorithm called GCRF-MSN. The authors conducted empirical studies on a large-scale real-life health-informatics dataset containing 35,844,800 inpatient discharge records from 500 hospitals collected over 9 years. They report that GCRF-MSN is 36 times faster than baselines with less memory requirement. The authors also report that the proposed solution can handle a network with millions of nodes and trillions of links efficiently, which should be very useful to real structured prediction problems that can be formulated as a network.

The fourth article, "Online URL Classification for Large-Scale Streaming Environments," by Neetu Singh, Narendra Chaudhari, and Nidhi Singh, describes a novel online and ensemble semisupervised classification algorithm for filtering URLs in streaming environments. The proposed method doesn't require exploiting the webpage's content or other external information, which enables faster prediction and wider applications. Empirical studies on six large-scale datasets from a live production environment show an improvement of 3 to 8 percent over multinomial logistic regression, support vector machines, and boosting. Furthermore, the authors claim that the proposed method can adapt to situations with concept drift in streaming settings, which is crucial for real deployment.

Last but not least, the fifth article, "Local PurTree Spectral Clustering for Massive Customer Transaction Data," by Xiaojun Chen, Si Peng, Joshua Zhexue Huang, Feiping Nie, and Yong Ming investigates an important problem in retail and e-commerce companies (clustering of massive customer transaction data) and proposes a novel algorithm called local PurTree spectral (LPS) clustering. The authors demonstrate the effectiveness of the proposed algorithm using six real-life datasets and report a significant improvement over state-of-the-art clustering methods such as PurTreeClust, concept hierarchy clustering, hierarchical agglomerative clustering (HAC), and DBSCAN. The proposed algorithm is promising to be applied to various other clustering tasks, in which the objects can be represented in hierarchical taxonomies or trees.

We thank all the authors for their submitted work to the special issue and all the reviewers for their effort in reviewing the submitted manuscripts. We also thank editor-in-chief Daniel Zeng for his advice on preparing the call for papers and handling some submissions, and editorial manager Jennifer Stout for her guidance and reminders throughout the entire process. ■

THE AUTHORS

Weike Pan is an associate professor in the College of Computer Science and Software Engineering at Shenzhen University. His research interests include transfer learning, intelligent recommendation, and machine learning. Pan received a PhD in computer science and engineering from the Hong Kong University of Science and Technology. Contact him at panweike@szu.edu.cn.

Qiang Yang is a chair professor and department head in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. His research interests are data mining and artificial intelligence. Yang has a PhD in computer science from the University of Maryland, College Park. Contact him at qyang@cse.ust.hk.

Charu Aggarwal is a Distinguished Research Staff Member at the IBM T.J. Watson Research Center. His research interests include graph mining and social networks, data stream mining, and uncertain data mining. Aggarwal has a PhD in operations research from the Massachusetts Institute of Technology. Contact him at charu@us.ibm.com.

Christoph Koch is a professor of computer science at the Swiss Federal Institute of Technology. His research interests are in data management. Koch received a PhD in artificial intelligence from TU Vienna and CERN. Contact him at christoph.koch@epfl.ch.