2018

Advances in Processing, Mining, and Learning Complex Data: From Foundations to Real-World Applications

Jia Wu

Shirui Pan

Chuan Zhou

Gang Li

Wu He
Old Dominion University, whe@odu.edu

See next page for additional authors

Repository Citation

Original Publication Citation

Authors

Jia Wu, Shirui Pan, Chuan Zhou, Gang Li, Wu He, and Chengqi Zhang

WILEY | Hindawi

Editorial

Advances in Processing, Mining, and Learning Complex Data: From Foundations to Real-World Applications

Jia Wu (iD),[1] **Shirui Pan** (iD),[2] **Chuan Zhou,**[3] **Gang Li,**[4] **Wu He,**[5] **and Chengqi Zhang**[2]

[1]*Macquarie University, Sydney, NSW, Australia*
[2]*University of Technology Sydney, Sydney, NSW, Australia*
[3]*Chinese Academy of Sciences, Beijing, China*
[4]*Deakin University, Melbourne, VIC, Australia*
[5]*Old Dominion University, Norfolk, VA, USA*

Correspondence should be addressed to Jia Wu; jia.wu@mq.edu.au

Processing, mining, and learning complex data refer to an advanced study area of data mining and knowledge discovery concerning the development and analysis of approaches for discovering patterns and learning models from data with a complex structure (e.g., multirelational data, XML data, text data, image data, time series, sequences, graphs, streaming data, and trees) [1–5]. These kinds of data are commonly encountered in many social, economic, scientific, and engineering applications. Complex data pose new challenges for current research in data mining and knowledge discovery as they require new methods for processing, mining, and learning them. Traditional data analysis methods often require the data to be represented as vectors [6]. However, many data objects in real-world applications, such as chemical compounds in biopharmacy, brain regions in brain health data, users in business networks, and time-series information in medical data, contain rich structure information (e.g., relationships between data and temporal structures). Such a simple feature-vector representation inherently loses the structure information of the objects. In reality, objects may have complicated characteristics, depending on how the objects are assessed and characterized. Meanwhile, the data may come from heterogeneous domains [7], such as traditional tabular-based data, sequential patterns, graphs, time-series information, and semistructured data. Novel data analytics methods are desired to discover meaningful knowledge in advanced applications from data objects with complex characteristics. This special issue contributes to the fundamental research in processing, mining, and learning complex data, focusing on the analysis of complex data sources.

1. Spatial Data

With the development of mobile communication technology, location-based services are booming prosperously. Meanwhile, privacy protection has become the main obstacle for the further development of services. The paper titled "Efficient Privacy-Preserving Protocol for k-NN Search over Encrypted Data in Location-Based Service" proposes an efficient private circular query protocol with high accuracy rate and low computation and communication cost. The Moore curve is adopted to convert two-dimensional spatial data into one-dimensional sequence and encrypt the points of interest (POIs) information with the Brakerski-Gentry-Vaikuntanathan homomorphic encryption scheme for privacy preserving. The scheme performs the secret circular shifts of the encrypted POI information to hide the location of the user without a trusted third party. The proposed scheme provides high-accuracy query results while maintaining low computation and communication cost.

2. Web Server Data

The paper titled "Deep Recurrent Model for Server Load and Performance Prediction in Data Center" proposes to use deep learning to predict web server performance and workload. The model can extract features automatically during the learning process without any prior knowledge or hand-generated features for segmentation. Experiments conducted on real web server data sets show that the model can achieve a good performance and generalization on predicting the performance of different kinds of servers. And the result also shows that the load generated by our model is very similar to the real one, which can be applied to test data center and other kinds of servers. Most servers in data center have log system. As long as the log file recording of the operation of the users is provided, the method can be used to generate load for the server and predict server performance under different load conditions.

3. Image Data

The paper titled "Unsupervised Domain Adaptation Using Exemplar-SVMs with Adaptation Regularization" has proposed an effective method for domain adaptation problems with regularization item which reduces the data distribution mismatch between domains and preserves properties of the original data. Furthermore, utilizing the method of integrating classifiers can predict target domain data with high accuracy. The proposed method mainly aims to predict in the setting that exists distribution mismatch across domains or instances and achieves desired results. Experiments conducted on the transfer learning datasets transfer knowledge from image to image.

Hyperspectral imaging has been proved as an effective way to explore the useful information behind the land objects. And it can also be adopted for biologic information extraction, by which the origin information can be acquired from the image repeatedly without contamination. The paper titled "Background Information Self-Learning Based Hyperspectral Target Detection" proposes a target detection method based on background self-learning to extract the biologic information from the hyperspectral images. The conventional unstructured target detectors are very difficult to estimate the background statistics accurately neither in a global nor local way. Considering the spatial spectral information, its performance can be further improved by avoiding the above problem. It is especially designed to extract fingerprint and tumor region from hyperspectral biologic images. The validity and the superiority of the method have been demonstrated on detecting the biologic information from hyperspectral images.

Segmentation of the prostate from magnetic resonance imaging plays an important role in prostate cancer diagnosis. However, the lack of clear boundary and significant variation of prostate shapes and appearances make the automatic segmentation very challenging. In the past several years, approaches based on deep learning technology have made significant progress on prostate segmentation. However, those approaches mainly paid attention to features and

contexts within each single slice of a 3D volume. As a result, these kinds of approaches face many difficulties when segmenting the base and apex of the prostate due to the limited slice boundary information. To tackle this problem, the paper titled "Exploiting Inter-Slice Correlation for MRI Prostate Image Segmentation: From Recursive Neural Networks Aspect" proposes a deep neural network with bidirectional convolutional recurrent layers for magnetic resonance imaging of prostate image segmentation. In addition to utilizing the intraslice contexts and features, the proposed model also treats prostate slices as a data sequence and utilizes the interslice contexts to assist segmentation. The proposed approach achieved significant segmentation improvement compared to other reported methods.

Early detection of *Lobesia botrana* is a primary issue for a proper control of this insect considered as the major pest in grapevine. The paper titled "A Distributed K-Means Segmentation Algorithm Applied to *Lobesia botrana* Recognition" proposes a novel method for *L. botrana* recognition using image data mining based on clustering segmentation with descriptors which consider gray scale values and gradient in each segment. This system allows a 95 percent of *L. botrana* recognition in nonfully controlled lighting, zoom, and orientation environments. The image capture application is currently implemented in a mobile application, and subsequent segmentation processing is done in the cloud.

The paper titled "Deep Hierarchical Representation from Classifying Logo405" introduces a logo classification mechanism which combines a series of deep representations obtained by fine-tuning convolutional neural network architectures and traditional pattern recognition algorithms. The experiments are carried out on both the Logo-405 dataset and the publicly available FlickrLogos-32 image datasets. The experimental results demonstrate that the proposed mechanism outperforms two popular ways used for logo classification, including the strategies that integrate hand-crafted features and traditional pattern recognition algorithms and the models.

The paper titled "Kernel Negative ε Dragging Linear Regression for Pattern Classification" proposes a kernel negative ε dragging linear regression method for pattern classification, which simultaneously integrated the negative ε dragging technique and the kernel method into linear regression for robust pattern classification under the condition that the consistency and compatibility between the test samples and training samples are poor. The negative ε dragging technique learns a classifier with a proper margin from noised and deformable data. Meanwhile, the kernel approach can make linearly nonseparable samples become linearly separable. Based on the effect of the negative ε dragging technique and kernel collaborating, the method can better perform classification for noised and deformable data. Comprehensive 24 experiments on image data sets demonstrate algorithm performance.

Recently, infrared human action recognition has attracted increasing attention for it has many advantages over visible light, that is, robust to illumination change and shadows. However, the infrared action data is limited until

now, which degrades the performance of infrared action recognition. Motivated by the idea of transfer learning, an infrared human action recognition framework using auxiliary data from visible light is proposed to solve the problem of limited infrared action data in the paper titled "Transferable Feature Representation for Visible-to-Infrared Cross-Dataset Human Action Recognition." The proposed method is evaluated on InfAR, which is a publicly available infrared human action dataset. To build up auxiliary data, we set up a novel visible light action dataset XD145. Experimental results show that the proposed method can achieve state-of-the-art performance compared with several transfer learning and domain adaptation methods.

4. Social Network Data

Social influence analysis is important for many social network applications, including recommendation and cyber security analysis. The influence of community including multiple users outweighs the individual influence. Existing models focus on the individual influence analysis, but few studies estimate the community influence that is ubiquitous in online social network. A major challenge lies in that researchers need to take into account many factors, such as user influence, social trust, and user relationship, to model community-level influence. The paper titled "Mining Community-Level Influence in Microblogging Network: A Case Study on Sina Weibo" aims to assess the community-level influence effectively and accurately; the problem is formulated as modeling community influence and constructs a community-level influence analysis model. Empirical studies on a real-world dataset from Sina Weibo demonstrate the superiority of the proposed model.

In the distributed cloud environment, a cloud platform is often not willing to share its recorded user-service invocation data with other cloud platforms due to privacy concerns, which decreases the feasibility of cross-cloud collaborative service recommendation severely. Besides, the user-service invocation data recorded by each cloud platform may update over time, which reduces the recommendation scalability significantly. In view of these two challenges, a novel privacy-preserving and scalable service recommendation approach based on SimHash, that is, SerRecSimHash, is put forward in the paper titled "Privacy-Preserving and Scalable Service Recommendation Based on SimHash in A Distributed Cloud Environment." A set of experiments are conducted based on a real distributed service quality dataset WS-DREAM. Experiment results show that SerRecSimHash outperforms the other up-to-date approaches in terms of recommendation accuracy and efficiency while guaranteeing privacy preservation.

5. Time Series/Signal Data

Complex systems is a broad concept that comprises many disciplines, including engineering systems. Regardless of their particular behavior, complex systems share similar behaviors, such as synchronization. The paper titled "Determining the Coupling Source on a Set of Oscillators from Experimental Data" presents different techniques for determining the source of coupling when a set of oscillators synchronize. It is possible to identify the location and time variations of the coupling by applying a combination of analysis techniques, namely, the source of synchronization. For this purpose, the analysis of experimental data from a complex mechanical system is presented. The experiment consisted of placing a 24-bladed rotor under an airflow. The vibratory motion of the blades was recorded with accelerometers, and the resulting information was analyzed with four techniques: correlation coefficients, Kuramoto parameter, cross-correlation functions, and the recurrence plot. The measurements clearly show the existence of frequencies due to the foreground components and the internal interaction between them due to the background components (coupling).
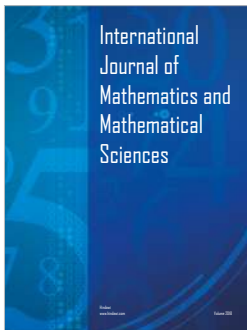
Acknowledgments

Jia Wu
Shirui Pan
Chuan Zhou
Gang Li
Wu He
Chengqi Zhang

References

[1] M. Shi, J. Liu, D. Zhou, and Y. Tang, "A topic-sensitive method for mashup tag recommendation utilizing multi-relational service data," *IEEE Transactions on Services Computing*, p. 1, 2018.

[2] H. Fan, Z. Ma, D. Wang, and J. Liu, "Handling distributed XML queries over large XML data based on MapReduce framework," *Information Sciences*, vol. 453, pp. 1–20, 2018.

[3] C. X. Zhai, "Probabilistic topic models for text data retrieval and analysis," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17*, pp. 1399–1401, Tokyo, Japan, 2017.

[4] L. B. Godfrey and M. S. Gashler, "Neural decomposition of time-series data for effective generalization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 7, pp. 1–13, 2017.

[5] H. Meyerhenke, P. Sanders, and C. Schulz, "Parallel graph partitioning for complex networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 9, pp. 2625–2638, 2017.

[6] K. Thearling, *An Introduction to Data Mining*, 2017.

[7] Y. Luo, Y. Wen, T. Liu, and D. Tao, "Transferring knowledge fragments for learning distance metric from a heterogeneous domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2018.

Advances in
Operations Research

Advances in
Decision Sciences

Journal of
Applied Mathematics

The Scientific World Journal

Journal of
Probability and Statistics

International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Optimization

International Journal of
Engineering
Mathematics

International Journal of
Analysis

Hindawi

Submit your manuscripts at
www.hindawi.com

Journal of
Complex Analysis

Advances in
Numerical Analysis

Mathematical Problems
in Engineering

International Journal of
Differential Equations

Discrete Dynamics in
Nature and Society

International Journal of
Stochastic Analysis

Journal of
Mathematics

Journal of
Function Spaces

Abstract and
Applied Analysis

Advances in
Mathematical Physics