International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-019-01010-8

EDITORIAL

Special issue on Machine learning approaches and challenges of missing data in the era of big data

Gwanggil Jeon[1] · Arun Kumar Sangaiah[2] · You-Shyang Chen[3] · Anand Paul[4]

With the proliferation of mobile computing technology in the rapidly growing IoT community we are bombarded with wide variety of data. As information and technology ear is gone and now it's for Big Data era were the questions arise about the veracity of the data that are generated. Thus these data are said to be 'missing at random' if the fact that they are missing is unrelated to actual values of the missing data Missing at Random: there is a pattern in the missing data but not on your primary dependent variables such as likelihood to recommend, once the data is missed it is vital to recover it by means of various machine learning methods and techniques as we have the historic data and its pattern. Missing Completely at Random: there is no pattern in the missing data on any variables. Many new techniques have offered very robust and hi-tech solutions for missing data and information analysis, collection, storage. Things get complicated with enormous amounts of valuable data in various formats. Since data are missed completely at random various data mining scheme and technique can be used to perform the task of data recovery. But still there is a challenge of fidelity of the data, how accurate are the data and how to verify its truthfulness and conformity of the facts. So data mining and AI based systems shall be used to evaluate the system. Today's scientists are trying to solve this issue with variety of new techniques and to analyze this data to help them understand their operations and management of data.

This special issue is intended to provide a highly recognized international forum to present recent advances in Machine Learning Approaches and Challenges of Missing Data in the Era of Big Data. The ultimate objective is to bring together well-focused, top quality research contributions, providing to the general machine learning community an opportunity to get an overall view of research results, projects, surveying works and industrial experiences that are dealing with theory and applications within the theme of Machine Learning Approaches and Challenges of Missing Data in the Era of Big Data. We invited authors to submit original research articles that would enhance our understanding of emerging and innovative technologies and the strategies and methods that contribute to the machine learning applied to sustainability of big data, and data mining in the assessment and management of missing data. We welcomed both theoretical contributions as well as papers describing interesting applications. Papers were invited for this special issue considering aspects of this problem, including:

- Methods to Evaluate and Understand the Missing Data
- Machine Learning Methods to Duplicate Data
- Interpolation, Extrapolation, Approximation and other approaches to analyze Missing data
- Innovative Learning Techniques to handle Missing Data
- Data Imputation and Pairwise deletion and other Techniques of missing data
- Historic Learning of Environment and Data Mining for Weather Data
- Mobile and Remote Sensing Big Data Evaluation and assessment using Machine Learning
- Infrastructure, organizational issues of Machine Learning for Big Data Analytics

✉ Anand Paul
paul.editor@gmail.com

Gwanggil Jeon
gjeon@inu.ac.kr

Arun Kumar Sangaiah
arunkumarsangaiah@gmail.com

You-Shyang Chen
ys_chen@cc.hwh.edu.tw

[1] Incheon National University, 119 Academy-ro, Incheon 22012, Korea

[2] VIT University, Vellore, Tamil Nadu 632014, India

[3] Hwa Hsia University of Technology, No. 111, Gongzhuan Road, New Taipei City, Taiwan

[4] Kyungpook National University, 80 Daehakro, Daegu 41566, Korea

Published online: 10 September 2019

- Reinforcement Learning for Sustainable and reliable Big Data analytics
- AI based Cloud for Big data architectures and real world applications;
- Case study, Models, methods, and tools for testing for Missing data

After review, a total of 6 papers out of 26 submissions have been accepted for publication in this issue.

Due to the limitation of the imaging system, it is hard to get Hyperspectral Image (HSI) with very high spatial resolution. Super-Resolution (SR) is a handling missing data technology to restore high-frequency information from the low-resolution image, can be used to solve this problem. Recently, Deep Learning (DL) has achieved great performance in computer vision, including SR. However, most DL-based HSI SR methods neglect the spectral disorder caused by normal 2D convolution. The contribution by Zheng et al. "Separable-spectral convolution and inception network for hyperspectral image super-resolution" proposes a novel end-end deep learning-based network named Separable-Spectral and Inception Network (SSIN) for HSI SR [1]. In SSIN, the feature extraction module independently extracts features of each band image, and then these features are fused together to further exploit residual image by using feature fusion module. In reconstruction module, a multi-path connection is built to obtain features of different levels to restore high spatial resolution image in a coarse-to-fine manner.

The Internet of Things (IoT) is an internet amongst things through advanced correspondence without human's operation. The effective use of data mining in exceedingly visible fields like, e-health, e-business and retail has led to its application in other industries and sectors. Due to the Health field, the Big Data knowledge could contribute with the plan to help with the end goal of investigation and management of the huge measures of health data. The contribution by Chilamkurti et al. "Random forest for big data classification in the internet of things using optimal features" implemented the e-health, big data in IoT utilizing innovative classifier and map reduce process [2]. A supervised, Random Forest Classifier(RFC) for grouping of e health data, this data are collected from by sensor devices and actuators, which will wear patients who suffer from different ailments, As medical data is with multiple attributes, medical data mining differs from other one. In order to verify the effectiveness of proposed model, different performance measures analyzed to compared with existing techniques, and additionally analyzed packet received rate and time taken for analysis.

With the rapid development of hyperspectral remote sensing technology, the spatial resolution and spectral resolution of hyperspectral images are continuously increasing, resulting the continuous increase of hyperspectral data's scale. At present, hyperspectral lossless compression technology has reached the bottleneck, at the same time, the rise of Deep Learning has provided us with new ideas. The contribution by Wu et al. "Lossless compression for hyperspectral image using deep recurrent neural networks" studies how to use deep learning for lossless compression of hyperspectral images [3]. In view of the shortcomings of the Differential Pulse Code Modulation (DPCM) method being insufficiently used for the prediction of the spectral band information, authors use the deep recurrent neural network to improve the traditional method DPCM, and improve the model's generalization ability and prediction accuracy.

Demosaickig is aiming at to approximate the missing color pixels by analysis the geometric structure between the given color pixels around and the missing color pixels. The contribution by Wang et al. "Reconstruction of missing color-channel data using a three-step back propagation neural network" introduces an efficient adaptive demosaicking method based on back propagation (BP) neural network (BP-NN) [4]. In the interpolation issue, different image features have totally different properties, such as smooth region, edges, and textures. To achieve high efficiency, authors provide the adaptive BP-NN based demosaicking algorithm which can reduce the blurring through recovering of the missing pixels by learning process, and also using pre-trained fixed network to reduce the computational complexity.

In big data applications, an important factor that may affect the value of the acquired data is the missing data, which arises when data is lost either during acquisition or during storage. The former can be a result of faulty acquisition devices or non-responsive sensors whereas the latter can occur as a result of hardware failures at the storage units. In the contribution by Kantarci et al. "Big data aggregation in the case of heterogeneity: A feasibility study for Digital Health," authors considered human activity recognition (HAR) as a case study of a typical machine learning application on big datasets [5]. Authors conduct a comprehensive feasibility study on the fusion of sensory data that is acquired from heterogeneous sources. Then, the authors present insights on the aggregation of heterogeneous datasets with minimal missing data values for future use.

An overhead view is often preferred in the cluttered environments because looking down from a top view can afford better coverage and much visibility of a scene. However, human detection in such or any other such type of extreme view can be challenging. The reason is that depending on the positions of people in the picture or image, there can be significant variations in the poses and appearances of a person. The contribution by Ahamd et al. "Person detector for different overhead views using machine learning" proposes a novel technique which transforms the region of interest containing a human to standardized the shape [6]. Authors

show the potential of our proposed RHOG algorithm across different scenes. When a classier trained on SCOVIS dataset and applied to our new recorded overhead datasets named SOTON and IMS, respectively.

We hope that this special issue would shed light on major developments in the area of Machine Learning and Cybernetics and attract attention by the scientific community to pursue further investigations leading to the rapid implementation of these technologies.

Acknowledgements We would like to express our appreciation to all the authors for their informative contributions and the reviewers for their support and constructive critiques in making this special issue possible. Finally, we would like to express our sincere gratitude to Professor Xizhao Wang, the Editor in Chief, for providing us with this unique opportunity to present our works in the *International Journal of Machine Learning and Cybernetics*.

References

1. Zheng K, Gao L, Ran Q, Cui X, Zhang B, Liao W, Jia S (2019) Separable-spectral convolution and inception network for hyperspectral image super-resolution. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-00911-4
2. Lakshmanaprabu S, Shankar K, Ilayaraja M, Nasir A, Vijayakumar V, Chilamkurti N (2019) Random forest for big data classification in the internet of things using optimal features. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-00916-z
3. Luo J, Wu J, Zhao S, Wang L, Xu T (2019) Lossless compression for hyperspectral image using deep recurrent neural networks. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-019-00937-2
4. Wang J, Anisetti M, Jeon G (2018) Reconstruction of missing color-channel data using a three-step back propagation neural network. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-0850-5
5. Obinikpo A, Kantarci B (2019) Big data aggregation in the case of heterogeneity: a feasibility study for digital health. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-00904-3
6. Ahmed I, Ahmad M, Adnan A, Ahmad A, Khan M (2019) Person detector for different overhead views using machine learning. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-019-00950-5