EDITORIAL

Guest Editorial: Non-IID Outlier Detection in Complex Contexts

Guansong Pang, *University of Adelaide, Adelaide, SA, 5005, Australia*
Fabrizio Angiulli, *University of Calabria, Rende 87036, Italy*
Mihai Cucuringu, *University of Oxford & The Alan Turing Institute, Oxford OX1 2JD, U.K.*
Huan Liu, *Arizona State University, Phoenix, AZ, 85004, USA*

Outlier detection, also known as *anomaly detection*, aims at identifying data instances that are rare or significantly different from the majority of instances. Due to its significance in many critical domains like cybersecurity, fintech, healthcare, public security, and AI safety, outlier detection has been one of the most active research areas in various communities, such as machine learning, data mining, computer vision, and statistics. Traditional outlier-detection techniques generally assume that data are independent and identically distributed (IID), which are significantly challenged in complex contexts where data are actually non-IID. These contexts are ubiquitous in not only graph data, sequence data, spatial data, temporal data, and streaming data,[1–3] but also traditional multidimensional, textual, and image data.[4–6] This demands for advanced outlier-detection approaches to address those explicit or implicit non-IID data characteristics. Motivated by this demand, we organized a Special Issue in IEEE Intelligent Systems to solicit the latest advancements in this topic in October 2019. In total, we received eight valid submissions from diverse countries, including China (2), Belgium (2), United States (1), Australia (1), India (1), and New Zealand (1). Each submission was reviewed by at least three reviewers. In the following, we provide a brief introduction to the five accepted articles.

In "Anomalous event sequence detection,"[7] Dong *et al.* study the problem of detecting abnormal event sequences in complex systems such as cyberphysical systems, aiming at identifying a sequential combination of event entities that are abnormal as a whole, but not for specific individual events. A graph diffusion-based method is introduced to tackle the problem and extensively evaluated on a large real-world system monitoring dataset containing 10 different types of intrusion attacks. This work showcases an interesting non-IID outlier-detection problem in sequential data.

In "Anomaly detection aided budget online classification for imbalanced data streams,"[8] Liang *et al.* study the problem of classification on imbalanced streaming data in the presence of outliers. The key challenge is to accurately classify the known majority and minority classes, while differentiating the outliers from the minority classes. A feedforward network with incrementally updated weights and an embedded distance-based outlier detector are jointly optimized to address this challenge. The method also incorporates a memory budget-aware module to efficiently maintain a set of active instances for classification and outlier detection in the current time windows. We believe this work provides an example of joint classification and outlier detection on imbalanced streaming data with potential concept drift.

In "How to optimize an academic team when the outlier member is leaving?,"[9] Yu *et al.* explore the problem of identifying "outlier" members in an academic team, in which the "outlier" members are defined as those whose collaboration network and skill sets are not or only weakly tied to those of the team. Both pairwise and high-order relations between the academic members are considered to identify the outliers. Extensive experiments are performed using large datasets from CiteSeerX and the Microsoft Academic Graph. This work provides a good real-world application of leveraging complex data dependence for outlier detection.

In "Isolation forest based anomaly detection framework on non-IID data,"[10] Xiang *et al.* study the problem of extending a widely used outlier detector, called isolation forest (iForest), to handle data drawn from a nonmetric space. The authors combine the underlying mechanism of isolation forest and extend distance-based hashing techniques to tackle this problem. The proposed method is evaluated on five non-IID datasets with trajectory anomalies, text anomalies, and discrete sequence

1541-1672 © 2021 IEEE
Digital Object Identifier 10.1109/MIS.2021.3072704
Date of current version 2 July 2021.

anomalies, in addition to a number of traditional commonly used IID datasets. This article demonstrates the issues of applying traditional outlier-detection methods to non-IID data, and presents a hashing-based solution for addressing these issues.

In "Outlier detection for foot complaint diagnosis: Modeling confounding factors using metric learning,"[11] Booth *et al.* explore an interesting application of outlier detection in computer-aided diagnosis of foot complaints using plantar pressure data. The work is motivated by the fact that the abnormality of the plantar pressure of a patient is dependent differently on individual demographic features, when comparing to healthy control samples. The authors further introduce a contextual metric-learning method to take into account this type of dependence in outlier detection, providing an example of re-inventing distance-based outlier detection for non-IID contexts.

Lastly, we would like to extend sincere appreciation to the Editor-in-Chief, Professor Venkatramanan Subrahmanian, the journal administrative team, authors, and reviewers for their support of this Special Issue.

REFERENCES

1. L. Akoglu, H. Tong, and D. Koutra, " Graph based anomaly detection and description: A survey," *Data Mining Knowl. Discov.*, vol. 29, no. 3, pp. 626–688, 2015.
2. F. Angiulli and F. Fassetti, " Distance-based outlier queries in data streams: The novel task and algorithms," *Data Mining Knowl. Discov.*, vol. 20, no. 2, pp. 290–324, 2010.
3. M. Gupta, J. Gao, C. C. Aggarwal, and J.Han, "Outlier detection for temporal data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2250–2267, Sep. 2013.
4. G. Pang, L. Cao, L. Chen, and H. Liu, "Learning homophily couplings from non-IID data for joint feature selection and noise-resilient outlier detection," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, pp. 2585–2591.
5. D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. Int. Conf. Learn. Representations*, 2018.
6. G. Pang, L. Cao, and L. Chen, "Homophily outlier detection in non-IID categorical data," *Data Mining Knowl. Discov.*, pp. 1–62, 2021.
7. B. Dong *et al.*, "Anomalous event sequence detection," *IEEE Intell. Syst.*, to be published, doi: 10.1109/MIS.2020.3041174.
8. X. Liang, X. Song, K. Qi, J. Li, J. Liu, and L. Jian, "Anomaly detection aided budget online classification for imbalanced data streams," *IEEE Intell. Syst.*, to be published, doi: 10.1109/MIS.2021.3049817.
9. S. Yu, J. Liu, F. Xia, H. Wei, and H. Tong, "How to optimize an academic team when the outlier member is leaving?," *IEEE Intell. Syst.*, to be published, doi: 10.1109/MIS.2020.3042871.
10. H. Xiang, J. Wang, K. Ramamohanarao, Z. Salcic, W. Dou, and X. Zhang, "Isolation forest based anomaly detection framework on non-IID data," *IEEE Intell. Syst.*, to be published, doi: 10.1109/MIS.2021.3057914.
11. B. G. Booth, N. L.W. Keijsers, and J.Sijbers, " Outlier detection for foot complaint diagnosis: Modeling confounding factors using metric learning," *IEEE Intell. Syst.*, to be published, doi: 10.1109/MIS.2020.3046431.

GUANSONG PANG is currently a research fellow with the Australian Institute for Machine Learning, University of Adelaide, Adelaide, Australia. His research interests include data mining and machine learning and their applications, with a research theme focused on abnormality and rarity learning for guarding the health, safety, and security in both digital and physical worlds. Contact him at guansong.pang@adelaide.edu.au.

FABRIZIO ANGIULLI is currently a professor of computer science at the University of Calabria, Rende, Italy. His research interests include data mining, machine learning, and artificial intelligence, with a focus on the design of anomaly detection approaches for various scenarios, efficient and effective large and high-dimensional data analysis, and explainable learning. Contact him at fabrizio.angiulli@unical.it.

MIHAI CUCURINGU is currently an associate professor with the Department of Statistics, University of Oxford, Oxford, and a Turing Fellow at The Alan Turing Institute, London, U.K. His research interests include the development and mathematical and statistical analysis of algorithms for data science, network analysis, and certain computationally hard inverse problems on large graphs, with applications to various problems in machine learning, statistics, and finance. Contact him at mihai.cucuringu@stats.ox.ac.uk.

HUAN LIU is currently a professor of computer science and engineering at Arizona State University, Phoenix, AZ, USA. His research interests include data mining, machine learning, social computing, and artificial intelligence, investigating interdisciplinary problems that arise in many real-world, data-intensive applications with high-dimensional data of disparate forms such as social media. He is a Fellow of ACM, AAAI, AAAS, and IEEE. Contact him at huanliu@asu.edu.