Guest Editorial for Selected Papers From BIOKDD 2020

Da Yan, Hong Qin, Hsiang-Yun Wu, and Jake Y. Chen

The 19th International Workshop on Data Mining in Bioinformatics (BIOKDD 2020) was held virtually on August 24, 2020 due to the COVID-19 pandemic. BIOKDD 2020 featured the special theme of "Battling COVID-19" which particularly welcomed paper submissions and invited talks related to COVID-19 research. As a whole-day workshop, altogether 15 submissions were accepted among a total of 35 submissions, and they were divided into 4 sessions: (1) Bioinformatics, (2) Data Curation, (3) Deep Learning with Biomedical Data, and (4) Data Mining & Statistical Methods. There are also 7 invited talks by domain experts.

This special section of the *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* features the extended versions of 6 quality papers presented in BIOKDD 2020. Each of the 6 invited papers was reviewed by 3 reviewers invited by the *TCBB* guest editors, and the BIOKDD workshop reviews were shared with the TCBB reviewers. The papers also went through 1 to 2 rounds of revisions.

The first invited paper, "A Gene Selection Method Based on Outliers for Breast Cancer Subtype Classification," by Rayol Mendonca-Neto, Zhi Li, David Fenyö, Claudio T. Silva, Fabíola G. Nakamura, and Eduardo F. Nakamura explores the use of gene expression data for cancer subtype classification by identifying relevant outlier genes (uncommonly over-expressed or under-expressed) that assist in gene signature identification. Their method, called Outlier-based Gene Selection (OGS), combines outlier detection techniques and feature elimination methods to find a small gene set capable of achieving high classification results. Experiments show that OGS presents an F1 score of 1.0 for basal and 0.86 for her 2, the two subtypes with the worst prognoses, respectively. Compared to other methods, OGS outperforms in the F1 score using 80% less genes. In general, OGS selects only a few highly relevant genes, speeding up the classification, and significantly improving the classifier's performance.

The second invited paper, "A KG-Enhanced Multi-Graph Neural Network for Attentive Herb Recommendation," by Yuanyuan Jin, Wendi Ji, Wei Zhang, Xiangnan He, Xinyu Wang, and Xiaoling Wang studies the problem of herb recommendation in Traditional Chinese Medicine (TCM) based on a set of symptoms. Since different symptoms have different importance, this work designs an attention network to discriminate the symptom importance and adaptively fuse the symptom embeddings to provide fine-grained syndrome representation. A TCM knowledge graph (KG) is also incorporated to enrich the input corpus and improve the quality of representation learning. The proposed KG-enhanced Multi-Graph Neural Network architecture performs attentive propagation to combine node features and graph structural information. Extensive experimental results on two TCM data sets show that their proposed model outperforms the other state-of-the-arts.

The third invited paper, "Learning Prognostic Models Using Disease Progression Patterns: Predicting the Need for Non-Invasive Ventilation in Amyotrophic Lateral Sclerosis," by Andreia S. Martins, Marta Gromicho, Susana Pinto, Mamede de Carvalho, and Sara C. Madeira predicts the need of Non-invasive Ventilation (NIV) as a treatment to Amyotrophic Lateral Sclerosis, which is a devastating neurodegenerative disease causing rapid degeneration of motor neurons and usually leading to death by respiratory failure. The work proposes to use itemset mining together with sequential pattern mining to unravel disease presentation patterns together with disease progression patterns by analyzing, respectively, static data collected at diagnosis and longitudinal data from patient follow-up. The learned prognostic models are promising. Pattern evaluation through growth rate suggests bulbar function and phrenic nerve response amplitude, additionally to respiratory function, are significant features towards determining patient evolution.

The fourth invited paper, "LitMC-BERT: Transformer-Based Multi-Label Classification of Biomedical Literature With An Application on COVID-19 Literature Curation," by Qingyu Chen, Jingcheng Du, Alexis Allot, and Zhiyong Lu proposes LITMC-BERT, a transformer-based multi-label classification method in biomedical literature. It uses a shared transformer backbone for all the labels while also captures label-specific features and the correlations between label pairs. The goal is to assist topic annotation in the production system of LitCovid which is a literature database of COVID-19 related papers in PubMed, where an article is assigned with up to eight topics, e.g., Treatment and Diagnosis. Experiments show that LITMC-BERT achieved the highest overall performance on two datasets than three baselines. Also, LITMC-BERT only takes 18% of the inference time taken by the previous best model for COVID-19 literature.

The fifth invited paper, "MGATRx: Discovering Drug Repositioning Candidates Using Multi-View Graph Attention," by Jaswanth K. Yella and Anil G. Jegga builds MGATRx, a novel approach to predict and identify drug repositioning candidates. The work systematically curated annotations from various sources and constructed a multi-view heterogeneous network. Leveraging this relatively current drug and disease annotations, MGATRx selectively aggregates relevant information from neighbors to learn node representations. Using the derived representation, links between drug nodes and disease nodes are predicted through multi-label classification.

Digital Object Identifier no. 10.1109/TCBB.2022.3176912

Comparative analysis with four state-of-the-art methods shows a substantial improvement in prediction performance, and several predicted drug-disease indication pairs overlap with drug indications that are either currently in clinical trials or are supported by literature references, demonstrating the overall translational utility of MGATRx.

The sixth invited paper, "SODA: Detecting COVID-19 in Chest X-rays With Semi-Supervised Open Set Domain Adaptation," by Jieli Zhou, Baoyu Jing, Zeya Wang, Hongyi Xin, and Hanghang Tong studies the problem of detecting COVID-19 disease in chest x-ray images. Prior works first train a Convolutional Neural Network (CNN) on an existing large-scale chest x-ray image dataset and then fine-tune the model on a smaller-scale newly collected COVID-19 chest x-ray dataset. However, simple fine-tuning may lead to poor performance due to the large domain shift present in chest x-ray datasets and the relatively small scale of the COVID-19 chest x-ray dataset. This work formulates the problem of COVID-19 chest x-ray image classification in a semi-supervised open set domain adaptation setting and proposes a novel domain adaptation method, called Semi-supervised Open set Domain Adversarial network (SODA). SODA is designed to align the data distributions across different domains in the general domain space and also in the common subspace of source and target data. Experiments show that SODA achieves a leading classification performance compared with other state-of-the-art models in separating COVID-19 with common pneumonia, and can produce better pathology localizations in the chest x-rays.

DA YAN,
HONG QIN,
HSIANG-YUNWU, and
JAKE Y. CHEN
Guest Editors

ACKNOWLEDGMENTS

As guest editors of this special section, we would like to thank the contributing authors, BIOKDD 2020 program committee, the *TCBB* reviewers who reviewed papers in this special section, and the *TCBB* staff for the support to make this special section possible.

Da Yan is an assistant professor of computer science with the University of Alabama at Birmingham (UAB). He was the sole winner of Hong Kong 2015 Young Scientist Award in physical/mathematical science. His research expertise lies in developing scalable systems and algorithms for Big Data analytics, with experience in Data Science projects on bioinformatics. He frequently publishes in conferences such as SIGMOD, VLDB, SIGKDD, ICDE, ICML, EMNLP, and AAAI, and he also regularly serves in the program committee of conferences such as SIGKDD 2020-2022, SIGMOD 2019-2021, VLDB 2018 and 2021, IJCAI 2017, 2021 (SPC) and 2022, AAAI 2021, ICDE 2020-2023 and serves as reviewers of journals such as the *ACM Transactions on Database Systems*, *VLDB Journal*, *IEEE Transactions on Parallel and Distributed Systems*, and *IEEE Transactions on Knowledge and Data Engineering*. He developed a series of systems for Big Data analytics, which are of high impact and are now being used by many researchers and companies. As an expert in Big Data and Data Science, he has been invited to publish surveys and books as the first author in prestigious venues such as Foundations and Trends in Databases and Springer Briefs in Computer Science. He has organized workshops including BIOKDD 2018-2022 with SIGKDD, and DMBIH 2019 with ICDM.

Hong Qin is an associate professor with the Department of Computer Science and Engineering, University of Tennessee at Chattanooga. He uses computational and mathematical approaches to investigate biomedical and biological questions. One focus is to develop probabilistic gene network models to infer network changes during cellular aging, where he builds gene network models from heterogeneous genomics data sets, including protein interactions, gene expression data sets, RNAseq data sets, protein mass-spec data sets, high-throughput phenotypic screens, and gene annotations. He is developing machine-learning methods to automatically estimate cellular lifespan from time-lapsed images. He is also applying engineering principles to study molecular, biological, and ecological networks. He is developing deep learning methods for better classification and prediction using heterogeneous biomedical and biological large data sets. He is a recipient of an NSF CAREER award and a lead-PI of an NSF Big Data Spoke project.

Hsiang-Yun Wu received the PhD degree from the University of Tokyo, Japan, in 2013, where she investigated several computational approaches for network analytics and visualization. She is a senior scientist with the St. Pölten University of Applied Sciences, Austria and TU Wien, Austria. Her principal research interests cover information visualization and human-centered techniques, focusing on exploring biological and geospatial data. She was awarded the European Union Marie Skłodowska-Curie Actions (MSCA) Individual Fellowship in 2017, with which she has investigated visualizations to effectively untangle complex relationships in biological context. She has also organized Dagstuhl and Shonan seminars as well as PhD school along this research direction and has volunteered many scientific activities (e.g., international PC members, editors, conference organizers, and etc.). More information about her can be found at http://yun-vis.net/.

Jake Y. Chen is the chief bioinformatics officer of the Informatics Institute, University of Alabama at Birmingham and, a tenured professor of genetics, computer science, and biomedical engineering, the past president of the Midsouth Computational Biology and Bioinformatics Society. He has more than 25 years of R&D experience in biological data mining and systems biology with more than 190 peer-reviewed publications and more than 200 invited talks worldwide on bioinformatics methodologies and biomedical applications. At UAB, he leads the AI.MED Laboratory (http://aimed-lab.org/) to advance multi-omics modeling and artificial intelligence application in medicine. He is an ACM distinguished scientist and an elected fellow of the American College of Medical Informatics (ACMI) and of the American Institute of Medical and Biological Engineering (AIMBE). He also serves on the editorial boards of the *BMC Bioinformatics*, *Journal of American Medical Informatics Association* (JAMIA), and *Frontiers in Artificial Intelligence and Big Data*. He was recognized as one of the "17 Informatics Experts Worth Listening To" by HealthTechTopia (2011), as a finalist for the "Indiana's Technology Educator of the Year" award (2012-14), and as one of the "Top 100 AI Leaders in Drug Discovery and Healthcare" by Deep Knowledge Analytics (2019).