Journal of Healthcare Informatics Research (2018) 2:367–369
https://doi.org/10.1007/s41666-018-0039-4

EDITORIAL

Special Issue on Data Mining in Health Informatics

Xia Hu[1] · Gregor Štiglic[2] · Fei Wang[3]

Published online: 7 November 2018

1 Data Mining and Health Informatics

Vast amounts of health-related data are captured in the form of Electronic Health Records (EHR), health insurance claims, medical imaging databases, disease registries, spontaneous reporting sites, clinical trials, as well as user-generated contents from social media and wearable devices. As a result, data mining has become critical to the healthcare world. On the one hand, the introduction of data mining techniques can bring many benefits, while on the other hand, we are facing many new challenges that did not exist before. Those challenges can range from collecting, preprocessing or transforming the data to building, making sense of the data, evaluating or interpreting the data mining models as well as applying the solutions in practice.

Today, many different methods and techniques have been studied in the context of health informatics. Based on the characteristics of the health and medical informatics, data mining techniques which were designed to tackle healthcare problems, are faced with new challenges. First, how to process large volume of data collected in datasets or data warehouses. In real-world applications, large amounts of clinical data including electronic patient records, medical images, or clinical trials data require new data mining techniques to process the data sources computationally effectively and efficiently. Second, how to properly handle noisy and incomplete data. Noise and missing values in the medical records are harmful to the performance of data mining algorithms and approaches. Consequently, there exists a very diverse set of research questions, approaches, and data sources in the literature, which cover different aspects of the applications of data mining techniques in health informatics. Based on the papers in this

✉ Xia Hu
hu@cse.tamu.edu

[1] Texas A&M University, College Station, TX, USA

[2] University of Maribor, Maribor, Slovenia

[3] Weill Cornell Medicine, Cornell University, New York, NY, USA

special issue, one can identify a few possible directions for future research, such as the following:

- Mining heterogeneous or multi-modality health data: With the rapid growth in different sources of medical information, the variety and diversity of data have become a new challenge for data mining and exploitation. Mining useful information from different sources of data effectively will be an important research problem.
- Interpretability: Human usually cannot understand the behaviors of complex machine learning models and the reasons behind the outcomes from the black-box algorithms. Prediction level explanation will help uncover the causal relations between a specific input and its corresponding model prediction, which is beneficial for clinical diagnosis, disease prevention, and health promotion.
- Information privacy: It is crucial to guarantee a high level of confidence in mining health information from clinical data without leaking or tampering sensitive personal information from patients.

2 The Special Issue

The goal of this special issue is to provide a forum for researchers to share their latest achievements, covering a diversity of articles and work that reflect the state-of-the-art in developing data mining techniques for medicine and healthcare. Following the open call for papers, the seven papers that comprise this special issue were selected from a total of 26 submissions. The selected papers underwent a rigorous extra refereeing and revision process.

The paper by Alyousef, Nihtyanova, Denton, Bosoni, Bellazzi, and Tucker presents a new method for clustering patients to make disease subclass diagnosis. The proposed model combines K-means with consensus clustering to provide additional features and build cohort-specific decision trees to improve the classification performance and introduce the interpretability of the underlying differences of the discovered groups. Experiments on real-world datasets indicate the effectiveness of the proposed nearest consensus clustering classification approach.

Abidi, Roy, Shah, Yu, and Yan propose a data analytics framework to improve diagnosis of glaucomatous optic discs. First, to distinguish between healthy and glaucomatous optic discs, the framework derives shape information from CSLO images using image processing, selects salient features, and then trains image classifiers. Second, to monitor glaucoma progression over time, the framework extracts morphological features from the original images and applies the clustering method to visualize subtypes of glaucomatous optic disc damage. Experiments on Zernike moments and morphological features of CSLO images indicate the framework can discriminate healthy and glaucomatous CSLO images and can discover glaucoma damage subtypes.

The paper by Manchanda, Meyer, Li, Liang, Li, and Kong presents a new mass spectrometry data analysis platform based on pattern recognition techniques. The main framework centers on a three-step process including feature extraction, binary classification, and feature ranking. First, the data preprocessing module generates inputs that can be fed into classifiers. Second, the classifiers are built. Third, feature selection

module is used to find the most important features for classification results. Experiments on human blood samples demonstrate the effectiveness of the proposed approach.

Feldman, Kotoulas, and Chawla propose an intervention question selection framework to identify a personalized subset of survey questions, to maximize the estimation of the need for user-defined targeted intervention. The framework is composed of three novel components to identify data, generate candidate questions, and evaluate question sets. The authors analyze the performance of the proposed framework using several real-world interventions. Empirical results validate the effectiveness of the proposed framework in frequent pattern mining, optimal set ranking, and the final intervention questions generation.

The paper by Ardywibowo, Huang, Gui, Xiao, Cheng, Liu, and Qian implements and carries out a comprehensive evaluation of population switching-state auto-regressive models with missing value imputation and outlier detection on real-world daily behavioral data. In particular, the paper handles missing data and outliers by simultaneously considering them while conducting model identification. Extensive experimental evaluations show the effectiveness and soundness of the proposed method in achieving interpretable health dynamic models with better prediction of health status changes.

The paper by Haddawy, Yin, Wisanrakkit, Limsupavanich, Promart, Lawpoolsri, and Sa-angchai proposes a new approach to predict the cases of malaria, which improves the performance in predicting malaria cases significantly. The approach is based on spatial hierarchical clustering could find compact geographic regions with good time series predictability. Experiments on malaria data from northern Thailand show the improvement in predictability.

The last paper by Cakin, Gorgulu, Baydogan, Zou, and Li proposes a data representation approach to extract features from DNA sequence, which could be utilized to monitor and understand gene expression. This plays a vital role in understanding the underlying mechanisms of gene regulation and thus the function of an organism.

The guest editors would like to thank everybody who contributed to the special issue, the authors for their high-quality work and contribution to this field, the reviewers for delivering detailed and thoughtful comments, and the Springer team for their support. We hope you find this special issue both inspiring and interesting.

All seven papers for this special issue can be found online here: https://link.springer.com/journal/41666/topicalCollection/AC_e6fcb1dde6c7e008e95fe17c5356b39c/page/1