Data Min Knowl Disc (2015) 29:867–870
DOI 10.1007/s10618-015-0414-1

EDITORIAL

Guest editorial: Special issue on data mining for medicine and healthcare

Fei Wang[1] · **Gregor Stiglic**[2] · **Zoran Obradovic**[3] · **Ian Davidson**[4]

Received: 7 March 2015 / Accepted: 4 April 2015 / Published online: 23 April 2015

1 Data mining and healthcare

Healthcare is a field that is closely related to everyone's daily life. Because of the high complexity in healthcare industry, every year a huge amount of money is wasted. In recent years researchers from different areas went into the healthcare world with the hope of helping to reduce the cost and improve the quality of care delivery. Among all those emerging trends, data driven technologies have received a lot of attentions due to the availability of more and more healthcare data. Data-driven healthcare is at the center of the vision of learning health systems and holds great promise for transforming the current healthcare status.

The goal of this special issue is to present novel approaches in the field of data mining that can be applied in different fields of healthcare. The vast majority of the

Responsible editor: Johannes Fürnkranz.

✉ Fei Wang
fei_wang@uconn.edu

Gregor Stiglic
gregor.stiglic@um.si

Zoran Obradovic
zoran@obradovic.temple.edu

Ian Davidson
davidson@cs.ucdavis.edu

[1] Department of Computer Science and Engineering, University of Connecticut, Storrs, USA

[2] Research Institute at Faculty of Health Sciences, University of Maribor, Maribor, Slovenia

[3] Computer and Information Sciences Department, Temple University, Philadelphia, USA

[4] Department of Computer Science, University of California, Davis, USA

papers presented in this special issue present novel methods that were empirically evaluated on medical datasets that are much larger than datasets we were used to meet in the results sections of similar papers a few years ago. We can observe similar trends in different conferences or workshops where novel methods in healthcare or medicine are usually presented. We are also happy to notice that most of the papers in this special issue include a medical doctor or other healthcare expert as a coauthor. This only emphasizes the fact that one should be careful when applying data mining in healthcare and stresses the importance of including the medical experts open to new approaches in the interdisciplinary teams working in this field.

2 The special issue

Recently, we started some activities to establish a platform for exchanging the ideas on data mining for medicine and healthcare. We have successfully organized the workshop on *Data Mining for Medicine and Healthcare* (DMMH) held in conjunction with KDD 2010, SDM 2013, SDM 2014. *Connected Health at Big Data Era* in conjunction with KDD 2014, and *Data Mining for Medical Informatics* (DMMI) in conjunction with AMIA 2014. This special issue provides a leading focused forum for timely, in-depth presentation of recent advances in algorithms, theory and applications on data mining technologies for medicine and healthcare. The selected papers underwent a rigorous extra refereeing and revision process.

Temporal mining is always an important problem in data mining applications, especially in medicine and healthcare. The paper by Moskovitch and Shahar presents a framework for classification of multivariate time series analysis, which implements three phases: (1) application of a temporal-abstraction process that transforms a series of raw time-stamped data points into a series of symbolic time intervals (based on either unsupervised or supervised temporal abstraction); (2) mining these time intervals to discover frequent temporal-interval relation patterns (TIRPs), using versions of Allen's 13 temporal relations; (3) using the patterns as features to induce a classifier. They evaluated the framework, focusing on the comparison of three versions of the new, supervised, temporal discretization for classification (TD4C) method, each relying on a different symbolic-state distribution-distance measure among outcome classes, to several commonly used unsupervised methods, on real datasets in the domains of diabetes, intensive care, and infectious hepatitis. The results clearly demonstrated their proposed TD4C framework.

The paper by Huang, Dong, Bath, Ji and Duan studied the problem of utilizing the heterogeneous EMRs to assist clinical pathway (CP) analysis and improvement, which plays an important role in health-care management in ensuring specialized, standardized, normalized and sophisticated therapy procedures for individual patients. More specifically, they developed a probabilistic topic model to link patient features and treatment behaviors together to mine treatment patterns hidden in EMRs. Discovered treatment patterns, as actionable knowledge representing the best practice for most patients in most time of their treatment processes, form the backbone of CPs, and can be exploited to help physicians better understand their specialty and learn from previous experiences for CP analysis and improvement. Experimental results on a real collection of 985 EMRs collected from a Chinese hospital show that the proposed approach can effectively identify meaningful treatment patterns from EMRs.

Sádz, Rodrigues, Gama, Robles and García-Gómen studied the temporal stability problem in biomedical data. They established the temporal stability as a data quality dimension and proposed new methods for its assessment based on a probabilistic framework. Concretely, they proposed methods for (1) monitoring changes, and (2) characterizing changes, trends and detecting temporal subgroups. First, a probabilistic change detection algorithm is proposed based on the Statistical Process Control of the posterior Beta distribution of the Jensen–Shannon distance, with a memoryless forgetting mechanism. This algorithm (PDF-SPC) classifies the degree of current change in three states: In-Control, Warning, and Out-of-Control. Second, a novel method is proposed to visualize and characterize the temporal changes of data based on the projection of a non-parametric information-geometric statistical manifold of time windows. This projection facilitates the exploration of temporal trends using the proposed IGT-plot and, by means of unsupervised learning methods, discovering conceptually-related temporal sub- groups. Methods are evaluated using real and simulated data based on the National Hospital Discharge Survey (NHDS) dataset.

The paper by Ji, He, Han and Spangler studied the problem of discovering strong relevance between heterogeneous typed biomedical entities from massive biomedical text data. They first build an entity correlation graph from data, in which the collection of paths linking two heterogeneous entities offer rich semantic contexts for their relationships, especially those paths following the patterns of top-k selected meta paths inferred from data. Equipped with meta path constrained relationship contexts, they designed a novel relevance measure to compute the strong relevance between two heterogeneous entities, named EntityRel. They evaluated their method on 20 millions of MEDLINE abstracts and five types of biological entities (Drug, Disease, Compound, Target, MeSH). A prototype of drug search engine for dis- ease queries is implemented. Extensive comparisons are made against multiple state-of-the-arts in the examples of Drug-Disease relevance discovery.

Henriques, Madeira and Antunes propose a novel predictive model that exploits cross-attribute and temporal dependencies in large hospital datasets. The proposed model is based on Markov models and aims to improve predictive performance and decoding of patterns of interest in data. Authors of this paper were granted a permission to use the well-known Heritage health prize dataset that is freely available which makes this paper even more interesting from the evaluation point of view. Prediction of the number of surgeries in the upcoming quarter, average number of drug prescriptions for the upcoming month and binary prediction of hospitalization needs were the tasks used in the evaluation on the real data.

The paper by Bandyopadhyay, Wolfson, Vock, Vazquez-Benitez, Adomavicius, Elidrisi, Johnson and O’Connor addresses a problem of building predictive models from large heterogeneous and contemporaneous patient populations in contrast to most predictive models that are in use today and were built using carefully selected groups of patients. Authors propose a novel risk model that can outperform the Cox proportional hazards model by taking missing risk factor, non-linear relationships between risk factors and cardiovascular event outcomes, and differing effects from different patient subgroups into account. The proposed Bayesian networks based approach was evaluated using calibration, concordance index and net reclassification improvement metrics on a large dataset of electronic medical records and claims data.

Another patient risk estimation approach is presented by Qian, Wang, Cao, Li and Jiang where active learning supports interactive queries to help domain experts compare similar patients by answering their questions. Performance and effectiveness of the proposed method is observed on benchmark and clinical datasets. In addition to traditional benchmarking datasets from UCI machine learning repository, authors explore the capabilities of the proposed method using structural MRI data as well as real world EHR based data with 319,650 patients observed over 4 years period. Using image and larger HER datasets, authors demonstrate the scalability of their method to allow solving of large healthcare learning problems.

The last paper by Li, Vinzamuri and Reddy studies transfer learning methods in the diagnosis of diabetes and proposes a novel framework that selectively transfers the knowledge across different sites. A sparse feature selection model based on constrained elastic net penalty is used to determine whether a transfer would positively impact the target task by transferring knowledge from the source site. The proposed model is evaluated on a public dataset consisting of EHR data for 9948 patients from all fifty states in the United States. Authors demonstrate a successful application of the proposed transfer learning framework to estimate the risk for diabetes on the state level.

Acknowledgments Guest editors of this special issue would like to thank all authors of 49 submitted papers, as well as all reviewers for their hard work and detailed reviews that led to eight accepted papers. Additionally, we would like to thank Geoff Webb, Editor-in-Chief and editorial staff of the Data Mining and Knowledge Discovery journal for their support and guidance in the production of this special issue.