Journal of Biomedical Informatics 90 (2019) 103092

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Guest Editorial

Temporal biomedical data analytics

1. Introduction

The large amounts of recorded longitudinal observational medical data hold a wealth of information, and have been recognized as potential sources for gaining valuable new knowledge. However, the application of data mining and machine learning methods to temporal data is still a developing research area, and extracting useful information from time series is a challenging task. Despite meaningful progress in research regarding the analysis of biomedical data [1,2], and especially analysis of longitudinal multivariate datasets [3–7], significant challenges remain. However, the work holds great promise for enhancing our understanding of clinical processes over time.

Typically the analysis of biomedical data, which are often stored in *electronic health records* [EHRs], is significantly constrained [8] because such data are sampled and stored in various forms over time: (a) some are sampled in a fixed frequency, such as data originating in electrical sensors; (b) some might be intended to be measured typically every month, but at a potentially varying frequency; (c) others might be sampled irregularly; and (d) some are usually represented as events (such as conditions or procedures) that may have a duration that needs to be recorded and captured, such as medication administrations, clinical conditions, and the like [9]. In addition, EHR data are often characterized by a large variety of concept types, such as conditions, procedures, medications, or laboratory tests, and are often sparsely distributed, which makes analyzing such data even more challenging.

Thus, in recent years, several approaches to temporal data analysis have been proposed, many of which are represented in this special issue. Approaches vary from those using raw time-series data, such as for forecasting, or classification and prediction, through approaches transforming the data using Fourier transformation, wavelets, or extracted Markovian models, sequential patterns, or time-interval related patterns (after performing temporal abstraction), and up to the use of the data within *Recurrent Neural Networks* (RNNs) [10], as part of the recent explorations of deep learning methods.

In this special issue, which includes nine accepted papers, there are three major research topics within temporal data analytics:

(1) Symbolic time intervals and sequential analysis
(2) Medication effects with irregular sampling
(3) Time series analysis and classification

The first topic refers to a recent trend in which raw, time-point-based variables are transformed into symbolic time intervals using temporal abstraction, enabling the researchers to incorporate uniformly all the various types of data in the analysis. Then recurring patterns of relations among time intervals can be discovered and used for clustering, classification, or prediction. The second topic refers to the use of time series analysis methods to analyze the effects of drugs on laboratory tests values, a task in which a typical challenge is the irregular sampling of laboratory measurements. The third topic deals with time series data from routine clinical settings, which can be represented as sequences of events and analyzed accordingly.

2. Symbolic time intervals and sequential data analysis

As mentioned, a major challenge in biomedical data analysis is the diversity of variables and their often unpredictable sampling over time, presenting issues such as sampling-frequency variance, determination of temporal duration, and dealing with sparsity. This reality makes it challenging to incorporate all variables within a single type of analysis because most methods, for time-point series analysis, typically expect all variables to be sampled in the same way or be of similar temporal types.

One idea, which was initially proposed years ago for interpretation and aggregation of time series, but which was more recently exploited for temporal data mining, is the idea of pre-processing time-series data before analyzing them by transforming all of the time-point based variables into meaningful symbolic time intervals through a process often called *temporal abstraction* [11]. The abstraction can be based on domain knowledge [12] or might be purely data driven [13]. In either case the process results in a uniform data representation of all data types, whether these were originally based on single, time-point measurements or on interval-based events. Then, *time intervals related patterns* (TIRPs), which represent frequently recurring temporal relations among the symbolic time intervals, can be discovered and can be used for clustering, classification, or prediction [9,14,15]. In this special issue four papers offer contributions to these topics.

Shkevinsky et al. [16] examined whether it is possible to discover frequent TIRPs consistently within different subsets of patients' data of the same type. Several measures for defining consistency were used, based on three criteria: (1) whether the patterns are frequent in each subset, (2) whether the patterns preserve their "local" metrics - the absolute frequency of each pattern in each subset, measured by a Proportion Test, and (3) whether the patterns preserve their "global" characteristics - their overall distribution, measured by a Kolmogorov-Smirnov test. The methodology was applied to three medical domains (oncology, infectious hepatitis, and diabetes).

The authors found that, when using the low frequency threshold ranges, 70–95% of the discovered TIRPs were consistently discoverable; 40–48% of them maintained their local frequency. TIRP global distribution similarity varied widely, from 0% to 65%. Increasing the threshold usually increased the percentage of TIRPs that were repeatedly discovered across different patient subsets within the same domain and increased the probability of a similar TIRP distribution.

https://doi.org/10.1016/j.jbi.2018.12.006
Received 22 December 2018; Accepted 24 December 2018
Available online 14 January 2019
1532-0464/ © 2019 Published by Elsevier Inc.

Limiting the discovered TIRPs to only those satisfying certain plausible semantic constraints further enhanced consistency.

Orphanou et al [17] applied the idea of generating temporal abstractions using Symbolic Aggregate approXimation (SAX) and discovering a subset of frequent TIRPs, which they refer to as *Temporal Association Rules* (TARs), as features to train a Naïve Bayes classifier. The authors examined also the idea of Horizontal Support for a TIRP within each patient record and the Mean Duration of each TIRP [12,18], to represent the features in the Coronary Heart Disease domain, and they found that horizontal support is more effective for prediction.

Moskovitch et al [19] present Maitreya - a framework for outcomes prediction in EHR data based on frequent TIRPs, which was evaluated on a large database of inpatients data, for procedures prediction. The framework learns frequent patterns from the Cohort set of patients (having the outcome), but not from the Controls that are used only for the evaluation. Then, the patients' TIRPs are detected at both classes and a classifier is induced. The data that are used in this paper were only symbolic, that is they included procedures, conditions and drug-exposures, but not lab-test results. Three novel TIRP metrics that are normalized versions of the horizontal-support, representing the number of TIRP instances per patient, were introduced. The evaluation of Maitreya was performed on 28 frequent and clinically important procedures, using the three novel TIRP representation metrics and comparing them to both a lack of temporal representation and to previous TIRP metrics. For 22 of these procedures, the use of TIRPs as predictors was superior to non-temporal features, and the use of the vertically normalized horizontal support metric to represent TIRPs as features was most effective.

Sequential events analysis offers a simpler way to analyze biomedical data, such as data found within the EHR and, in particular, data regarding events such as medical conditions, procedures, or medication administrations that might have a duration (possibly requiring a time intervals analysis). Bonomi and Jiang [20] focused on the problem of detecting sequential patterns of medical data. Thus, given a set of temporal patterns modeling a specific event of interest, one would like to return as output matching-data instances. In this work, the authors develop a new pattern-matching technique that aims to detect clinically useful knowledge, given a set of patterns, in patient's records. The matched instances are ranked according to a significance score based on a p-value threshold shown to be effective in a set of evaluation experiments on real datasets.

3. Medication effects with irregular sampling

When analyzing longitudinal data, one may wish to capture the relations of events, or trends of series over time, ideally in order to infer causal relations. Discovering causal relations is still a highly challenging problem, although investigators have undertaken multiple approaches, such as regression and lagging. In this special issue, two papers dealt with this topic.

Levine et al. [21] studied how lagged linear regression can be used to detect the physiologic effects of various drugs from data in EHRs. The authors systematically examined methodological variations, such as time-series construction, temporal parameterization, intra-subject normalization, differencing, use of explanatory variables, and regression models. They then assessed the effects of these variations on performance of lagged linear methods. After generating two somewhat similar gold standards (one based on general domain knowledge and the other based on modifications of the knowledge performed by a domain expert), the authors examined the pairwise relationships among seven medications and four laboratory measurements within their institutional EHR. They concluded that time-series analysis of EHR data will likely rely on some of the beneficial pre-processing and modeling methodologies identified, and will certainly benefit from continued careful analysis of methodological perturbations. They also found that methodological variations, such as pre-processing and representations, have a large effect on the results, emphasizing the importance of thoroughly evaluating these components when comparing machine-learning methods.

The motivation of Poh et al. [22] was to assist clinicians in managing disease progression, determination of drug dose, performance of follow up, and other clinical tasks. For that purpose, they wanted to be able to identify both short and long term trends in repeated measurements, which is challenging for current regression models due to the irregular sampling. Thus, the authors introduce an extension to broken-stick regression models. Since the model is parametric and is completely generative, its first derivative provides a long-term non-linear estimate of the annual rate of change in the measurements more reliably than does linear regression, a result that the authors demonstrate in a case study of managing patients with chronic kidney disease.

4. Time series analysis and classification

There are several approaches for analyzing time series, with the optimal approach often depending on the task. In this subsection, we refer to three tasks and approaches, from creating a score for a patient's risk, through classical time-series classification using factorization, and finally, the use of the trendy RNNs.

Huang et al. [23] develop a novel formulation for contemporaneous patient risk monitoring by exploiting the emerging data-rich environment in many healthcare applications. Unlike current typical risk scores that mainly calculate the likelihood of an outcome, their formulation translates multivariate longitudinal measurements into a Contemporaneous Health Index which does not necessarily require labeling, unlike typical risk scores. The authors provide several algorithms to mitigate the challenges associated with the non-smooth convex optimization problem, by first identifying its dual reformulation as a constrained smooth optimization problem, and then using the block coordinate descent algorithm to solve the optimization iteratively with a derived efficient projection at each iteration.

Hendryx et al. [24] use the CUR matrix factorization for dimension reduction to identify important sub-sequences in electrocardiogram (ECG) time series. Being able to represent these abstracted sequences makes it possible to share the results with physicians and use them in clinical settings as a summarization technique. Additionally, the authors found that, using case-based reasoning, it is possible to label the remaining unselected beats based on using CUR-selected beats.

Wu et al. [25] refer in their work to the synchronicity, evenness and cardinality of event sequences in the context of relative time. They show how event sequences and their properties—asynchronous, uneven, and multi-cardinal problem settings—can support explicit accountings of relative time. In the evaluation of their approach applied to pediatric asthma patients, the authors show several ways to incorporate relative time into an RNN model that improves the overall classification of patients who do or do not have asthma, or who have a persistent disease versus those who are in long-term remission.

In sum, the special issue on temporal data analytics deals with one of the most challenging and promising research topics in the field of biomedical informatics, and the papers in the special issue reflect the challenges and novel contributions that these investigators have made.

Declaration of interests

The authors declared that there is no conflict of interest.

Acknowledgements

We would like to thank the Israeli Ministry of Science and Technology that funded Robert Moskovitch during the editing of this special issue through grants 3-14360 and 3-14428.

References

[1] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (6) (2012) 395–405.

[2] J.L. Warner, P. Zhang, J. Liu, G. Alterovitz, Classification of hospital acquired complications using temporal clinical information from a large electronic health record, J. Biomed. Inform. 59 (2016) 209–217.

[3] G. Hripcsak, D. Albers, A. Perotte, Parameterizing time in electronic health record studies, J. Am. Med. Informatics Assoc. 22 (2015) 794–804.

[4] L. Sacchi, C. Larizza, P. Magni, R. Bellazzi, Precedence temporal networks to represent temporal relationships in gene expression data, J. Biomed. Inform. 40 (2007) 6.

[5] Y. Lin, H. Chen, R.A. Brown, MedTime: a temporal information extraction system for clinical narratives, J. Biomed. Inform. 46 (2013).

[6] A. Singh, G. Nadkarni, O. Gottesman, S.B. Ellis, E.P. Bottinger, J.V. Guttag, Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration, J. Biomed. Inform. 53 (2015) 220–228.

[7] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, A.F. Laine, A framework for mining signatures from event sequences and its applications in healthcare data, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 272–285.

[8] D. Albers, N. Elhadad, E. Tabak, A. Perotte, G. Hripcsak, Dynamical phenotyping: using temporal analysis of clinically collected physiologic data to stratify populations, PLoS ONE 9 (6) (2014).

[9] R. Moskovitch, C. Walsh, F. Wang, G. Hripsack, N. Tatonetti, Outcomes prediction via time intervals related patterns, IEEE International Conference on Data Mining (ICDM), Atlantic City, USA, (2015).

[10] L. Jain, L. Medsker, Recurrent Neural Networks: Design and Applications, CRC Press, FL, USA, 1999.

[11] Y. Shahar, A framework for knowledge-based temporal abstraction, Artif. Intell. 90 (1997) 79–133.

[12] R. Moskovitch, Y. Shahar, Classification of multivariate time series via temporal abstraction and time intervals mining, Knowl. Inf. Syst. 45 (1) (2015) 35–74.

[13] R. Moskovitch, Y. Shahar, Classification driven temporal discretization of multivariate time series, Data Min. Knowl. Disc. 29 (4) (2015) 871–913.

[14] D. Patel, W. Hsu, M. Lee, Mining relationships among interval-based events for classification, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008.

[15] Batal, D. Fradkin, J. Harrison, F. Moerchen, M. Hauskrecht, Mining recent temporal patterns for event detection in multivariate time series data, in: Proceedings of Knowledge Discovery and Data Mining (KDD), Beijing, China, 2012.

[16] A. Shkevinsky, Y. Shahar, R. Moskovitch, Consistent discovery of frequent interval-based temporal patterns in chronic patients' data, J. Biomed. Inform. 75 (2017) 83–95.

[17] K. Orphanou, A. Dagliati, L. Sacchi, A. Stassopoulou, E. Keravnou – Papailiou, R. Bellazzi, Incorporating repeating temporal association rules in Naive Bayes classifiers for coronary heart disease Diagnosis, J. Biomed. Inform. 81 (2018) 74–82.

[18] R. Moskovitch, Y. Shahar, Fast time intervals mining, Knowl. Inf. Syst. 42 (2015) 21–48.

[19] R. Moskovitch, F. Polubriaginof, A. Weiss, P. Ryan, N. Tatonetti, Procedure prediction from symbolic electronic health records via time intervals analytics, J. Biomed. Inform. 75 (C) (2017) 70–82.

[20] L. Bonomi, X. Jiang, Patient ranking with temporally annotated data, J. Biomed. Inform. 78 (2018) 43–53.

[21] M.E. Levine, D.J. Albers, G. Hripcsak, Methodological variations in lagged regression for detecting physiologic drug effects in EHR data, J. Biomed. Inform. 86 (2018) 149–159.

[22] N. Poh, S. Tirunagari, N. Cole, S. de Lusignan, Probabilistic broken-stick model: a regression algorithm for irregularly sampled data with application to eGFR, J. Biomed. Inform. 76 (2017) 69–77.

[23] Y. Huang, Q. Meng, H. Evans, B. Lober, Y. Cheng, X. Qian, J. Liu, S. Huang, CHI: a contemporaneous health index for degenerative disease monitoring using longitudinal measurements, J. Biomed. Inform. 73 (2017) 115–124.

[24] E. Hendryx, B. Riviere, D. Sorensen, C. Rusin, Finding representative electrocardiogram beat morphologies with CUR, J. Biomed. Inform. 77 (2017) 97–110.

[25] S.T. Wu, S. Liu, S. Sohn, S. Moon, C. Wi, Y.J. Juhn, H. Liu, Modeling asynchronous event sequences with RNNs, J. Biomed. Inform. 83 (2018) 167–177.

Guest Editors
Robert Moskovitch, Yuval Shahar
Department of Information Systems Engineering, Ben Gurion University of the Negev, Beersheba, Israel
E-mail addresses: robertmo@bgu.ac.il (R. Moskovitch), yshahar@bgu.ac.il (Y. Shahar).

Fei Wang
Department of Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, New York, NY, USA

George Hripcsak
Department of Biomedical Informatics, Columbia University, New York, NY, USA
E-mail address: hripcsak@columbia.edu.