Prediction of Clinical Procedures via Time Intervals Mining

Robert Moskovitch, Colin Walsh, George Hripcsak, Nicholas Tatonetti
Department of Biomedical Informatics, Columbia University
New York, USA
rm3198,cgw2106,gh13,npt2105@columbia.edu

ABSTRACT

The increasing availability of temporal data from Electronic Health Records (EHR) provides exceptional opportunities for the prediction of clinical outcome events. However, the nature of EHR data is sparse, non-randomly missing and heterogeneous, which is very challenging to analyze. We propose the use of temporal abstraction to transform the data into symbolic time intervals series. Then we use KarmaLego, a fast time intervals mining algorithm, to discover frequent Time Intervals Related Patterns (TIRPs) that are used as features to predict the outcome event. In this study we rigorously evaluate various aspects in the KarmaLego Outcome Events Prediction framework on an extraction of 32,168 patients from Columbia University Medical Center focusing on 6 procedures as outcome events. Our results show that the use of TIRPs for prediction significantly outperforms using only static concepts, and the use of our two TIRPs metrics for features representation outperform the use of the default Binary TIRP representation.

INTRODUCTION

The increasing availability of time-stamped electronic health records (EHR) enables researchers to perform classification and prediction tasks that leverage the EHR's temporal nature -- one of the most challenging research topics in biomedicine. This challenge arises since EHR data are sparse, in both the variables collected and across time, non-randomly missing, and stored in heterogeneous formats[1,2,3]. In addition, although medicine is practiced over temporal *intervals*, (e.g., "Rx of amoxicillin for 10 days") these time intervals are not well captured by the EHR and temporal abstraction is required to derive meaning from the coded data. Common methods for classification and prediction of multivariate temporal data, such as Hidden Markov Models or recurrent neural networks, time series similarity measures (e.g., Euclidean distance or Dynamic Time Warping and time series feature extraction methods (e.g., discrete Fourier transform, discrete wavelet transform or singular value decomposition), cannot be directly applied to such temporal data.

Therefore, in order to better analyze such data for purposes of prediction, we propose to transform the time point series into symbolic time interval series representation, through a process known as knowledge based temporal abstraction[1,2]. The use of knowledge based temporal abstraction enables the transformation of multivariate temporal data, having various forms, into a series of symbolic time intervals, based on existing domain knowledge, providing a uniform format of various temporal variables. Then, it is possible to discover frequent *Time Intervals Related Patterns* (*TIRPs*)[2,3,4]. The use of TIRPs as features for the classification of multivariate temporal data was proposed in. In our study, the discovered TIRPs from a given period of multivariate temporal data are used as features, to predict clinical procedures, which we introduce here and demonstrate on EHR data. Prediction and forecasting of procedure and visit utilization are important topics in the biomedical domain; these tasks span multiple clinical disciplines in medicine from general medicine to cardiology to neurology. Flexible methods to predict utilization of procedures might aid efficient resource management and – if coupled with tools to aid processing of the results of those procedures – clinical care more directly.

The increased attention to the subject of mining time intervals has led several research groups to quite simultaneously propose using the discovered temporal patterns for classifying multivariate time series[3,4,5]. Batal et al.[4] presented a study in which time intervals patterns classify multivariate time series, using an a priori approach, STF-Mine, for the discovery of temporal patterns, and the X^2 (chi-square) measure for pattern selections. In this paper we propose a more expressive representation metrics with the goal of better performance, inspired by their use in[3,5].

METHODS and RESULTS

Before introducing the proposed framework for Outcome Events Prediction, which we call Maitreya, we would like to describe its main components, which are the KarmaLego[2] and SingleKarmaLego[3] algorithms, used for frequent time intervals related patterns discovery and detection, respectively. KarmaLego is a fast time intervals mining algorithm that discovers frequent TIRPs, including all their instances, measured by *horizontal support* (the number of instances of a TIRP detected at a patient), *mean duration* (the average duration of these instances), which will be used to represent the features in the prediction talk. In this paper we propose a methodology for the prediction of outcome events (e.g., procedures, diagnoses, or adverse events) using time intervals mining. Our methodology consists of discovering TIRPs only in the set of patients having the outcome, which we call the Outcome Cohort. Thus, the prediction model is not learning or relying on any other type of patients. An alternative set of patients is defined that are similar to the Cohort in their concepts data, which is used to induce a binary classifier and evaluate the method. The process includes the following steps, as illustrated in figure 2: KarmaLego is applied *only* to the set of patients having the outcome procedure to discover frequent TIRPs, which are used later for the prediction task. Then the set of TIRPs is detected in each patient in the Cohort set and the control set of patients using SingleKarmaLego. Then, based on each TIRP representation (e.g., Binary, Horizontal Support, or Mean Duration) a matrix of values is generated that is later used to induce a classifier and evaluate it.

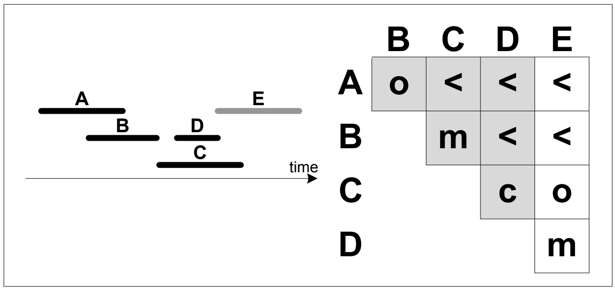

Figure 1. An example of a Time-Interval Related Pattern (TIRP), represented by a sequence of five lexicographically ordered symbolic time intervals and all of their pair-wise temporal relations. Time interval E is a candidate symbol that is being added to the current TIRP, and its relations with the other four symbolic intervals are shown in the last column of the half matrix.

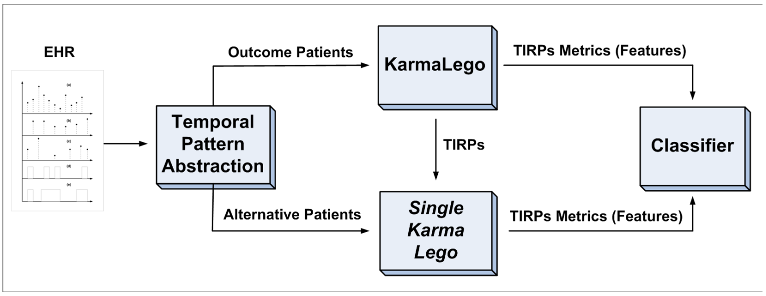

Figure 2. Outcome Events Prediction Framework - the overall outcome events prediction process. The EHR data is abstracted into symbolic time intervals. The patients having the outcome event are mined by KarmaLego and a set of TIRPs are discovered, which are later detected by SingleKarmaLego at the alternative set of patients. A matrix of all the patients represented by their TIRPs are used to induce a classifier and evaluate their performance.

In order to evaluate our Outcome Events Prediction method we used the data from the Columbia University Medical Center New York Presbyterian Hospital (CUMC-NYP). In total, the CUMC-NYP EHR contains medical record data for approximately 4.5 million patients going back to 1989, containing approximately 30 million diagnosis billing codes, 20 million prescription orders, 9 million procedures, and 500 million laboratory results. We used only coded data for this analysis, including drug exposures, conditions (billing codes), and procedures. Medical concepts are then transformed into symbolic time intervals (called "eras") in two ways: (i) if the medical concept contains coded time interval data (e.g. amoxicillin, 10 days) an era will be constructed for 10 days following the record date or (ii) if no coded time interval data are available (e.g. conditions and outpatient prescriptions) then they are assumed to last for 30 days. Finally, any overlapping eras are concatenated into a single era. The Cohort datasets were created based on a year of observation period ending a month prior to the predicted outcome event (procedure), which was the prediction period. A control set of patients was created of those who most similar to the Cohort patients. We selected 6 procedures as outcome events, which were the prediction task in each dataset. Procedures were selected for clinical relevance among the highest frequency procedures in the data. Table 1 presents the results for each of the outcome events prediction. The second column shows the dataset size, and then the optimal settings of number of temporal relations, epsilon and TIRP representation and finally the AUC achieved. These results are encouraging and we are now in the process of running the method on significantly larger datasets and many more procedures.

Outcome Event Procedure	**|Patients|**	**Relations**	**epsilon**	**TIRP Rep**	**AUC**
Colonoscopy	266	3	30	MeanD	0.732
Computerized Axial tomography of head	294	7	30	MeanD	0.726
Diagnostic ultrasound of heart	396	3	0	HS	0.764
Transfusion of packed cells	538	7	0	HS	0.857
Emergency department visit	600	7	60	MeanD	0.787
Injection of antibiotic	600	3	0	HS	0.730

Table 1. The results of each outcome event procedure given the optimal settings of number-of-relations/epsilon/TIRP-Representation and the corresponding AUC. HS and MeanD occur evenly, while HS favors epsilon 0 and MeanD favors epsioon 30.

REFERENCES

[1] G. Hripcsak, D. Albers, Next-Generation Phenotyping of Electronic Health Records, Journal of American Medical Informatics Association, 20: 117-121, 2013.

[2] R. Moskovitch, Y. Shahar, Fast Time Intervals Mining Using Transitivity of Temporal Relations, Knowledge and Information Systems, DOI 10.1007/s10115-013-0707-x, In Press, 2013.

[3] R. Moskovitch, Y. Shahar, Classification of Multivariate Time Series via Temporal Abstraction and Time Intervals Mining, Knowledge and Information Systems, In Press, 2014.

[4] I. Batal, D. Fradkin, J. Harrison, F. Moerchen, and M. Hauskrecht, Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data, Proceedings of *Knowledge Discovery and Data Mining (KDD)*, Beijing, China, 2012b.

[5] R. Moskovitch, Y. Shahar, Classification Driven Temporal Discretization of Multivariate Time Series, Data Mining and Knowledge Discovery, DOI: 10.1007/s10115-014-0784-5, 2014.