Journal of Biomedical Informatics 81 (2018) 74–82

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Incorporating repeating temporal association rules in Naïve Bayes classifiers for coronary heart disease diagnosis

Kalia Orphanou[a,*], Arianna Dagliati[c,f,g], Lucia Sacchi[b,c], Athena Stassopoulou[e], Elpida Keravnou[a], Riccardo Bellazzi[b,c,d]

[a] Department of Computer Science, University of Cyprus, Cyprus
[b] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy
[c] Centre for Health Technologies, University of Pavia, Italy
[d] IRCCS Istituti Clinici Scientifici Maugeri, Pavia, Italy
[e] Department of Computer Science, University of Nicosia, Cyprus
[f] Centre for Health Informatics, University of Manchester, Manchester, United Kingdom
[g] The Manchester Molecular Pathology Innovation Centre, University of Manchester, Manchester, United Kingdom

A R T I C L E I N F O

Keywords:
Bayesian models
Time series classification
Temporal abstraction
Temporal reasoning
Temporal association rules

A B S T R A C T

In this paper, we develop a Naïve Bayes classification model integrated with temporal association rules (TARs). A temporal pattern mining algorithm is used to detect TARs by identifying the most frequent temporal relationships among the derived basic temporal abstractions (TA). We develop and compare three classifiers that use as features the most frequent TARs as follows: (i) representing the most frequent TARs detected within the target class ('Disease = Present'), (ii) representing the most frequent TARs from both classes ('Disease = Present', 'Disease = Absent'), (iii) representing the most frequent TARs, after removing the ones that are low-risk predictors for the disease. These classifiers incorporate the horizontal support of TARs, which defines the number of times that a particular temporal pattern is found in some patient's record, as their features. All of the developed classifiers are applied for diagnosis of coronary heart disease (CHD) using a longitudinal dataset. We compare two ways of feature representation, using horizontal support or the mean duration of each TAR, on a single patient. The results obtained from this comparison show that the horizontal support representation outperforms the mean duration. The main effort of our research is to demonstrate that where long time periods are of significance in some medical domain, such as the CHD domain, the detection of the repeated occurrences of the most frequent TARs can yield better performances. We compared the classifier that uses the horizontal support representation and has the best performance with a Baseline Classifier which uses the binary representation of the most frequent TARs. The results obtained illustrate the comparatively high performance of the classifier representing the horizontal support, over the Baseline Classifier.

1. Introduction

Temporal abstraction (TA) is useful for abstracting time point data into interval-based sequences of events [1]. Temporal abstracted events were shown to be helpful in various clinical tasks and domains such as summarizing and managing patient data in oncology [2], monitoring of children's growth [3], management of insulin-dependent diabetes [4] and interpreting online patient data for monitoring purposes in intensive care units (ICUs) [5]. Bayesian networks (BNs) [6–8] belong to the family of probabilistic models and they were widely used in many clinical domains as they can handle well uncertainty in medical knowledge and data. Both Bayesian models and TAs demonstrated their effectiveness as standalone engines predominantly for medical problem solving and for medical data processing respectively, but not in conjunction. A detailed survey on TAs and BNs approaches applied to clinical domains and the benefits of their integration can be found in [9].

Temporal association rules (TARs) are special types of association rules extracted by applying a temporal operator between the antecedent and the consequent of the rule. TARs characterize the temporal relation between the time-interval events defined in the antecedent and consequent. Both the antecedent and consequent represent temporal abstraction events and the aim of the TARs is to extract complex abstractions that are mined from data in a knowledge-based fashion

* Corresponding author.
E-mail address: kalia.orphanou@gmail.com (K. Orphanou).

https://doi.org/10.1016/j.jbi.2018.03.002
Received 1 March 2017; Received in revised form 14 February 2018; Accepted 7 March 2018
Available online 16 March 2018
1532-0464/

[11,10]. Related works [12,13] introduced symbolic time-interval events that extract knowledge-based abstractions such as TIRPs which are discussed in Section 2.

In our previous work [14], we developed a Naïve Bayes classifier [15,16], using as features the most meaningful TARs from patient records and utilize them to improve the classification performance. We applied the developed classifier for predicting a future risk of Coronary Heart Disease (CHD). The notion of horizontal support that was introduced in [12,13] and defines the number of times that a particular temporal pattern is found in some patient's record, was incorporated in the values of the features. As an extension of that work, the focus of this paper is to demonstrate how the classification performance changes due to the feature selection and the feature representation process. In this paper, we introduce two more classifiers where the features represent the most frequent TARs that better discriminate the target class and where the TARs that are not good predictors for the disease, based on the medical knowledge for the specific domain, were excluded from the selected features. We also develop a classifier that uses the mean duration of the TARs as its features and finally we compare the classifier with the best performance, to a Baseline Classifier which uses a binary representation of the TARs.

The rest of the paper is organized as follows: In Section 2, we present an overview of temporal data mining in biomedicine and temporal patterns mining, while in Section 3 we describe the methods used for the development of the classifiers. A discussion of our experiments for selecting the most frequent and predictive TARs for the disease, to be represented as features to the network is given in Section 4. In the same section, we describe the dataset used as a testbed. In Section 5, the obtained results for the evaluation of all the developed classifiers are presented. Finally, a discussion about the results is given in Section 6, while conclusions and future work are presented in Section 7.

2. Background

2.1. Temporal data mining in biomedical data

The medical history of some individual is a repository of time-stamped data/information of diverse format/content and storage medium. Such data are invariably expressed at different levels of semantic detail and sampling frequency, they could have gaps or be excessively voluminous, and they are not amenable to direct processing and reasoning with. The recent advances in technology enabled the collection and storage on electronic platforms (e.g. electronic health records (EHR)) of large volumes of such data, while the development of temporal data mining techniques enabled the analysis, representation, interpretation and reasoning with the EHR longitudinal data [17–20].

A variety of temporal data mining techniques is proposed in the literature to deal with biomedical data, such as phenotyping [18], machine learning techniques such as Gaussian kernel smoothing and differential entropy [21], temporal abstraction over time intervals [1,9,22] and dynamic Bayesian networks [23,24]. The aim of temporal data mining techniques is to induce new temporal knowledge from the time-series in EHR data and to develop accurate classification and predictive models.

2.2. Temporal abstraction and time intervals mining

Temporal Abstraction (TA) refers to a set of techniques that allow describing a set of time series data and external events through sequences of context-specific temporal intervals [1,25]. TAs can be divided into two main categories: basic and complex. Basic TAs take as input time point events and return as output time intervals on the basis of some predefined rules, known as TA mechanisms. The derived symbolic time intervals can then be combined into complex temporal patterns representing their temporal relationships (complex TAs).

Although mining time-intervals is a relatively young research field, many automated tools have already been proposed in the literature to automatically discover frequent temporal patterns derived as the conjunction of temporal relations between pairs of intervals. Most of the methods use a subset or all of Allen's 13 interval relations to discover the temporal relationships among events. Kam and Fu [26] were the first who proposed the discovery of temporal patterns using all of Allen's interval relations, however their discovered patterns were ambiguous, since only the temporal relations among all the pairs of consecutive intervals were defined. Following, Hoppner [27] resolves this issue, by defining a non-ambiguous representation where all the possible pairwise temporal relations are represented in a *k*-intervals pattern.

The work proposed by Batal et al. [28] follows the approach by Hoppner by deriving temporal patterns using a subset of Allen's temporal operators. They detect temporal patterns based on the sliding window method. This algorithm is based on the assumption that events occurring far enough from each other, have no temporal relationship. Adam et al. [29,30] proposed 'Frequence', a web-based interface that integrates a data mining algorithm with a visualization tool. The interface aims at discovering frequent patterns from temporal event sequences and then to present the patterns mined in a user friendly way. The 'Frequence' system considers the sequence and the duration of temporal events, but it does not explicitly use any temporal relation to mine the frequent patterns.

In [31], a fast symbolic time intervals mining algorithm, KarmaLego, is presented to mine Time Intervals Related Patterns (TIRPs). In that work, KarmaLego is included in a process that implements a knowledge based temporal abstraction (KBTA) [13] framework for deriving basic TAs and then the algorithm is iteratively applied to the derived abstractions to detect TIRPs. In addition, other methods have been proposed that do not use Allen's temporal relations [29,30,32].

In the current work, we use the method proposed by Sacchi et al. [11] to extract temporal patterns as a set of TARs. As mentioned in the introduction, TARs are a special type of association rules extracted by applying a temporal operator between the antecedent and the consequent of the rule. In a TAR, the members of the antecedent are characterized by a co-occurrence of the temporal patterns that compose it. In the current work, we only use the *Precedes* operator which synthesizes the *Before*, *Meets*, *Overlaps*, *Equal*, *Starts* and *Finished-by* temporal relations, and we only extract TARs consisting of two symbolic time intervals. From the related work, TIRPs are the most similar technique to the one we are presenting. Differently from TARs, in TIRPs mining the temporal operator is applied among each interval that builds up the pattern. According to its definition, any TAR involves only two elements (an antecedent and a consequent), and as such only one temporal relation is considered between the two, albeit a disjunctive one. In addition, while TIRPs are mined by using as input an interval based representation, to mine the TARs herein proposed, raw time series go through a temporal abstraction step that results in the definition of arbitrarily complex abstractions that are extracted from data in a knowledge-based fashion. Thanks to this procedure, the extracted TARs provide more compact patterns, where the necessary complexity is included in the temporal abstractions that make up the antecedent and the consequent, rather than in the temporal operator that links them. The idea of defining a set of complex abstractions of interest that are then used to create the antecedents and the consequents of the TARs allows pursuing a knowledge-based approach, where the users can define the patterns on the basis of what they have in mind and interpret the final results in terms of such knowledge.

2.3. Time intervals related patterns based classification

The use of the discovered temporal patterns as features for classification is becoming increasingly popular in data mining literature. The classification task has been performed by using different methods such as decision trees [33], Naïve Bayes [12,28,34], recurrent neural