Knowledge-Based Systems 122 (2017) 148–158

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Extraction and optimization of classification rules for temporal sequences: Application to hospital data

M. Vandromme[a,c,*], J. Jacques[a], J. Taillard[a], A. Hansske[b], L. Jourdan[c], C. Dhaenens[c]

[a] Alicante, 50 Rue Philippe de Girard, 59113 Seclin, France
[b] Hospital Group of the Lille Catholic Institute (GHICL), Lomme, France
[c] CRIStAL, UMR 9189, University of Lille, CNRS, Centrale Lille, France

ARTICLE INFO

Article history:
Received 29 September 2016
Revised 30 January 2017
Accepted 1 February 2017
Available online 1 February 2017

Keywords:
Data mining
Classification
Temporal data
Optimization

ABSTRACT

This study focuses on the problem of supervised classification on heterogeneous temporal data featuring a mixture of attribute types (numeric, binary, symbolic, temporal). We present a model for classification rules designed to use both non-temporal attributes and sequences of temporal events as predicates. We also propose an efficient local search-based metaheuristic algorithm to mine such rules in large scale, real-life data sets extracted from a hospital's information system. The proposed algorithm, MOSC (Multi-Objective Sequence Classifier), is compared to standard classifiers and previous works on these real data sets and exhibits noticeably better classification performance. While designed with medical applications in mind, the proposed approach is generic and can be used for problems from other application domains.

1. Introduction

Many data mining problems involve temporal data; that is, data where each information is associated with the time at which it occurred. One of the best-known of such problems is the market basket analysis, where each element (or product, in this case) is described by its name and the date or time at which it was bought [1]. This study focuses on medical data as collected by a hospital's information system. Such data also include a strong temporal aspect. Indeed, all of the medical acts, diagnoses and prescriptions are associated with the time and date at which these were performed. Non-temporal information, such as patient information (age, gender, etc.) is also present in the databases. Medical data mining is an interesting topic, rapidly expanding as hospitals store more and more data, and find more and more practical questions for which they would like to scan their databases for answers. In this study, we focus on classification; more specifically, we build classifiers that exploit temporal sequences as primary elements, in addition to non-temporal attributes. By doing this, we expect to build a more faithful representation of reality and generate classifiers that are more useful and easier to interpret for a domain expert (in this case, medical staff). As an example, a basic, non-temporal classifier that says "if events A and B occur, then X is likely to occur" might be useful, but less than "if event A occurs before event B, then X is likely to occur". This second one carries more information and usually corresponds more precisely to the implicit domain knowledge of experts. An added benefit lies in an easier validation of obtained results, where we want to check whether the classifiers make sense and have a practical interest or not.

In addition to the heterogeneity and temporal aspect, hospital data tend to be very sparse, which makes their exploitation challenging. The only information recorded for a patient (in addition to personal data) is the set of events that occurred to him. This set is usually much smaller than the set of all possible medical events. Indeed, a patient is usually associated with a few dozens of events (at most), out of the tens of thousands possible events. This results in a very sparse data matrix (around 0.1% to 1% non-missing values). This is a problem for most data mining algorithms, but it allows for improvements on the optimization algorithm at several levels.

The main contributions of this article are:

- the proposition of a model for representing temporal sequences in classification rules
- the design of adequate neighborhood operators for *sequence* terms
- the definition of data structures for restricting the search space for these terms.

The remaining of this article is organized as follows. In Section 2, existing work in the field of temporal data mining is presented. Section 3 describes the proposed classification model using

* Corresponding author.
E-mail address: maxence.vandromme@inria.fr (M. Vandromme).

http://dx.doi.org/10.1016/j.knosys.2017.02.001
0950-7051/

sequences of events, and Section 4 details the optimization algorithm developed to build efficient classifiers through exploration of the search space. Section 5 compares the proposed approach with other algorithms on large size, real-life medical data sets. The main contributions of this work are summarized in the conclusion.

2. Related work

2.1. Overview

Temporal data mining has received significant attention since its emergence. Indeed, many practical applications deal with temporal data; designing tools to handle this type of data is therefore a necessity. Temporal data mining encompasses many different problems. We focus here on the classification of temporal sequences, which is relevant to the context exposed in the introduction of this article. Temporal sequences classification is one of the least-explored problems in the field of temporal data mining, as stated in [2]. Although this study was published in 2001, its conclusions are still valid nowadays. This problem has nonetheless been explored by researchers interested in computer security and intrusion detection. The earliest studies focused on predicting potential security breaches based on the sequence of previous events [3,4]. Most of the works on this topic use algorithms based on association rule mining to build the classification rules, but some use other techniques such as classification trees [5] or neural networks. A survey of the anomaly detection field showed that the main limitations of the current methods are the need for a label for each instance (in order to train the classifier), and performance issues on imbalanced data (which make the "minimal support" level hard to determine). Some studies proposed solutions to these limitations. A notable work develops a method focused on rare events prediction (i.e. for imbalanced data), but does not consider sequences of events but rather item sets occurring within a specific time window before the rare event [6].

2.2. Statistical methods

Statistical methods have also been applied to temporal sequences classification. In particular, Hidden Markov Models (HMM) [7] and Conditional Random Fields (CRF) [8] have proved useful, with notable applications in voice recognition [9] and motion detection [10] among others. Note that CRF are often applied to sequence labeling problems, and more rarely used for sequence classification. However, these methods can rarely use more than hundreds of attributes, which makes feature selection [11] a requirement for large scale applications. Unfortunately, feature selection for temporal data is a hard problem in itself, and has no definitive answer at the present time. A second limitation is that HMM and CRF are able to handle event sequentiality, but not event simultaneity, which is a strong requirement in the type of data we consider.

2.3. Sequence classification

An interesting work on temporal sequences classification proposed an approach where additional features are generated to represent the events occurring at various times in a sequence [12]. Feature selection is then used to extract useful features, which are then fed to a standard classifier. The authors considered two classifiers: Bayes and Winnow. The additional features significantly improve the classifiers' efficiency, but this process is only applicable when the starting set of features is small, since it grows exponentially when generating the additional features. For example, a starting set of dozens of features produces tens of thousands to millions of additional features. This approach is therefore not applicable to our case, where tens of thousands of features have to be considered. Another work focused on mining temporal medical data using Fuzzy Cognitive Maps (FCM) [13]. More precisely, the proposed system aims to describe and predict diseases using sequences of medical events (interventions, alterations in health status, etc.). The main limitation is that FCM require a constant time interval between a sequence's elements. Hence, it is not applicable to our case, for reasons similar to HMM and CRF, described above. This limitation also appears in a study of sequence classification using recurrent neural networks [14]. Here, the authors use neural networks to predict the next event, which serves as basis for classifying the sequence that led to this event. It also makes the explicit assumption that "temporal sequences are periodic and can be circularly extended", which does not hold in our case. Another more recent use of recurrent neural networks was done in [15], where the proposed system also aims to predict the next event to occur, although with more detail (mainly the estimated time before occurrence). However, this system is limited on the number of attributes; the authors recommend no more than around one hundred. One last related work tackles the open problem of Deterministic Finite Automata (DFA) induction, whose goal is to build automata describing natural languages [16]. The sequences of letters and words are in a way similar to the sequences of events in our data. However, like Gupta et al. [14], they do not allow for event simultaneity. DFA induction is also very specific and hard to solve, especially for larger instances of the problem. The experiments in [16] show that algorithms fail to find solutions for instances with hundreds of features. Sparse data also make the problem considerably harder, which means that DFA induction algorithms are not suited to deal with our problem.

2.4. Temporal rule mining

A closely related sub-field is that of *temporal rule mining*, where the goal is not to classify the instances, but to unveil interesting rules formed with temporal sequences. This sub-field has been thoroughly explored since the seminal work on the Apriori algorithm [17], and includes many applications in domains such as marketing, health, network security [18] or finance security [19]. The original Apriori algorithm dealt with sequential data, but other approaches and further developments extended its range to time series [20,21]. Another notable work uses SPEA2 [22], a well-known multi-objective optimization algorithm, to build a set of high-quality association rules [23]. A seminal work on this topic managed to bridge the gap between rule mining and classification, by using the rules found in the data to build the classification model [24]. Named CBA (Classification Based on Associations), this algorithm was not initially designed to handle temporal data. However, the extension seems quite straightforward given the work done on Apriori. The issue with this approach is that it is not well-suited for imbalanced data. Indeed, a recurrent problem with rule mining is choosing a suitable value for the support threshold. If one of the classes is much less frequent in the data, this threshold needs to be low enough to capture rules describing the "rare" class. On the other hand, doing so means that more rules are produced overall. For heavy imbalances between classes (for example, 1% of positives), the rule mining process typically produces thousands of low-value rules, and this in turn impairs the classifier's efficiency. Note that a more recent study on this topic proposed a modified Apriori algorithm to deal with imbalanced data [25]. However, the data used for validation was quite small (230 instances and 11 attributes), so the scalability to much larger data remains to be assessed.