Computational and Structural Biotechnology Journal 14 (2016) 185–199

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry

A. Anguera [a], J.M. Barreiro [a], J.A. Lara [b,*], D. Lizcano [b]

[a] Technical University of Madrid, School of Computer Science, Campus de Montegancedo, s/n - 28660, Boadilla del Monte, Madrid, Spain
[b] Open University of Madrid, UDIMA - Facultad de Enseñanzas Técnicas, Ctra. De la Coruña, km 38.500 – Vía de Servicio, 15 - 28400, Collado Villalba, Madrid, Spain

ARTICLE INFO

Article history:
Received 15 February 2016
Received in revised form 3 May 2016
Accepted 8 May 2016
Available online 18 May 2016

Keywords:
Medical Data Mining
Electronic Health Record
Time Series
Knowledge Discovery

ABSTRACT

One of the major challenges in the medical domain today is how to exploit the huge amount of data that this field generates. To do this, approaches are required that are capable of discovering knowledge that is useful for decision making in the medical field. Time series are data types that are common in the medical domain and require specialized analysis techniques and tools, especially if the information of interest to specialists is concentrated within particular time series regions, known as events.

This research followed the steps specified by the so-called knowledge discovery in databases (KDD) process to discover knowledge from medical time series derived from stabilometric (396 series) and electroencephalographic (200) patient electronic health records (EHR). The view offered in the paper is based on the experience gathered as part of the VIIP project.[1]

Knowledge discovery in medical time series has a number of difficulties and implications that are highlighted by illustrating the application of several techniques that cover the entire KDD process through two case studies.

This paper illustrates the application of different knowledge discovery techniques for the purposes of classification within the above domains. The accuracy of this application for the two classes considered in each case is 99.86% and 98.11% for epilepsy diagnosis in the electroencephalography (EEG) domain and 99.4% and 99.1% for early-age sports talent classification in the stabilometry domain. The KDD techniques achieve better results than other traditional neural network-based classification techniques.

1. Introduction

The quantity of information generated by the many different activities carried out in medicine is constantly on the increase. The efficient and responsible use of this information is one of the key challenges today.

In the healthcare field, information is generated at many different levels: management, planning, medical examinations, etc. In particular, the research described in this paper focuses on patient medical data, formally known as electronic health records (EHR).

EHRs may contain very wide-ranging data types: nominal (ICD9 codes, CPT codes), ordinal (pain scales, PEW scores), numerical (temperature, BP), unstructured clinical narratives (for which text mining techniques are required), etc. There is a lot of literature on clinical systems operating on these data types [1]. However, more and more EHRs contain a data type whose structure may, on occasions, be extremely complex and which has been found after investigation not to have been thoroughly researched: time series.

A time series can be defined as a sequence *TS* of time-ordered data $TS = \{TS_t, t = 1,\ldots,N\}$, where t represents time, N is the number of observations made during that time period and TS_t is the value measured at time instant t. The results of medical examinations (electroencephalogram, electrocardiogram, electromyogram, etc.) very often constitute a time series [2,3]. Such is the importance of time series in medicine today that important data types like medical images (radiodiagnosis) are also very often mapped as time series for later processing and analysis [4].

The analysis of time series for knowledge discovery is far from straightforward and requires the application of special-purpose tools, especially if the key information of interest to the expert is concentrated within particular time series regions, known as events. Data mining is an interesting option in this respect. As illustrated by the success stories described by Shadabi and Sharma [5], data mining techniques have a huge potential for analysing such large volumes of stored medical data in order to discover knowledge. Generally, the extraction of useful, tacit and previously unknown knowledge from large data volumes is

* Corresponding author at: Facultad de Enseñanzas Técnicas, Universidad a Distancia de Madrid, UDIMA Ctra. De la Coruña, km 38.500 – Vía de Servicio, 15, 28400, Collado Villalba, Madrid, Spain. Tel.: +34 630524530.
E-mail address: juanalfonso.lara@udima.es (J.A. Lara).

[1] This work was partially supported by the Spanish Ministry of Education and Science as part of the 2004–2007 National R&D&I Plan through the VIIP Project (DEP2005–00,232-C03).

http://dx.doi.org/10.1016/j.csbj.2016.05.002

what is known as knowledge discovery in databases (KDD). The KDD process ranges from the understanding and preparation of the data to the interpretation and use of the discovered knowledge (results of the KDD process). Data mining is the stage of the KDD process where the data are studied and useful information is extracted using a set of techniques and tools [6].

Traditional time series analysis techniques examine whole time series. However, the techniques applied in this case study were especially designed to address the analysis of time series events. As discussed in detail later, together, these techniques solve a classification problem, for example, by means of a strategy combining:

a) The identification of time series events
b) The generation of time series reference models for several subjects
c) The comparison of a subject (to be classified) with different reference models.

The aim of this paper is to report the results of two case studies applying the above techniques and also share with the scientific community the experience that we have gained in the field of medical time series analysis, highlighting the particularities of medical time series processing throughout the different stages of the KDD process. To do this the case study research methodology was used in order to propose and apply advanced knowledge discovery techniques on data from two branches of medicine: stabilometry and electroencephalography. In doing so, the above process was supervised in its entirety by medical specialists from the respective fields. A sample of their impressions is reported as lessons learned in Section 5. Other researchers may find the experience shared in this paper useful for more efficiently and successfully undertaking similar projects for extracting useful knowledge from other medical time series.

The remainder of the paper is organized as follows. Section 2 discusses some papers and concepts of interest related to our proposal. Section 3 describes the reference domains used in this research. Section 4 details the process enacted to extract knowledge from time series, as well as the results of its application. Section 5 briefly discusses different issues of interest related to the proposed techniques and the illustrated case study (applicability, relationship to other techniques, limitations, viewpoint of medical experts, etc.). Finally, Section 6 reports the conclusions of the research and states some challenges in this field.

2. Background

The literature covers different approaches based on the application of computer techniques applied to the domain of medicine. Some are based exclusively on expert knowledge [7–12]. Others, however, learn from previous problems (case-based approaches) [13,14] or are representations (e.g., decision trees) that support decision making (model-based approaches) [15]. There are also hybrid approaches, such as the one illustrated in this article, where expert knowledge is used to gain a better understanding of the domain and KDD techniques are then applied to build models for use in decision making (e.g., diagnosis) based on the medical data.

The KDD process includes the following stages (which may vary slightly from author to author) [6]:

1. **Domain and data understanding**. This first phase (which some authors consider to be outside the scope of the KDD process) studies the general characteristics of the data to be analysed and the source domain.
2. **Data selection**. This phase determines all the sources of data of interest, which are unified in a target dataset.
3. **Data preprocessing**. The goal of this stage is to assure the quality of the data. To do this, a series of tasks are performed on the dataset generated in the selection phase. These tasks include reducing noise, handling missing values, etc.
4. **Data transformation and reduction**. In this phase, the preprocessed data are subjected to a number of filters and operations in order to assure that the data format is suitable for running data mining algorithms.
5. **Data mining**. A series of techniques and machine learning algorithms can be applied to the correctly formatted data in order to discover knowledge. These techniques are applied in order to solve different problem types, known as tasks.
6. **Knowledge interpretation/evaluation**. The last step in the KDD process aims to evaluate the resulting models and, if the assessment is positive, interpret the knowledge inferred from the models.

Clearly, KDD is a well-established process divided into phases and tasks. It generally functions as a paradigmatic framework for discovering knowledge from the data of any domain. And medicine is not immune either to the beneficial effects of being able to access a highly standardized and widely documented framework such as the above. In fact, applying the KDD process to a branch of medicine by documenting and storing (whenever possible) the interim and final results could be a major step forward in medical research based on data analysis.

2.1. Time series analysis techniques

There are a great many techniques related to time series analysis in the literature.

There are techniques for comparing time series and extracting common subsequences. The most noteworthy are techniques based on Fourier [6] or wavelet [16,17,18] transforms. Others are based on comparing time series singularities, known as landmarks [19]. Unlike the above, another group of techniques address the time series directly, using concepts such as the time warping distance [20,21], minimum bounding rectangles (MBR) [22], Markovian models [23] or graph theory [24]. Of the above, the wavelet-based technique is most closely related to our proposal, as it is somehow capable of identifying events. The drawback of this technique, however, is that the events in question (wavelets) do not necessarily match up with the segments of interest to domain experts. The other techniques described in this section are useful for comparing two whole time series. These techniques apply different methods to extract information on the entire time series. In many domains, like EEG or stabilometry, the focus should be exclusively on regions of interest (events) in the time series.

There are techniques not only for comparison but also for generating transform-based reference models [25]. Again, however, they analyse the whole time series in order to output the transform coefficients (which are modelled). The same applies to other research aiming to find parts that a group of time series have in common but which are not necessarily of interest to the specialist [26,27,28,29]. Some techniques are based on previously transforming the series into a set of segments. Even so, their applicability is confined to specified domains [30].

On the other hand, there are some proposals in the literature related to event identification. They are linked to specified domains, which means that they are either not usually generally applicable [31,32] or are based on identifying the prominently shaped segments of the series [33,34,35] that do not necessarily match up with the events that are of interest to domain specialists.

Finally, this article illustrates an example of time series classification. Note, therefore, that most of the reviewed literature concerns traditional techniques like the simple nearest neighbour algorithm [36,37,38]. We have also found techniques that are more like the approach reported here and are based on distinctively identifying subsequences in time series (not necessarily events of interest for experts) [39].

2.2. Time series analysis techniques applied to medicine

Other authors have proposed different approaches to time series analysis techniques for the medical domain. Firstly, several authors