Artificial Intelligence in Medicine 70 (2016) 12–30

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Classification of auditory brainstem responses through symbolic pattern discovery

Marco E. Molina, Aurora Perez*, Juan P. Valente

Department of Languages, Information Systems and Software Engineering, School of Computer Engineering, Technical University of Madrid, Campus de Montegancedo, s/n, Boadilla del Monte, Madrid 28660, Spain

ARTICLE INFO

Article history:
Received 21 October 2015
Accepted 9 May 2016

Keywords:
Time series data mining
Decision support systems
Pattern-based classification
Symbolic pattern discovery
Auditory brainstem responses

ABSTRACT

Introduction: Numeric time series are present in a very wide range of domains, including many branches of medicine. Data mining techniques have proved to be useful for knowledge discovery in this type of data and for supporting decision-making processes.
Objectives: The overall objective is to classify time series based on the discovery of frequent patterns. These patterns will be discovered in symbolic sequences obtained from the time series data by means of a temporal abstraction process.
Methods: Firstly, we transform numeric time series into symbolic time sequences, where the symbols aim to represent the relevant domain concepts. These symbols can be defined using either public or expert domain knowledge. Then we apply a symbolic pattern discovery technique to the output symbolic sequences. This technique identifies the subsequences frequently found in a population group. These subsequences (patterns) are representative of population groups. Finally, we employ a classification technique based on the identified patterns in order to classify new individuals. Thanks to the inclusion of domain knowledge, the classification results can be explained using domain terminology. This makes the results easier to interpret for the domain specialist (physician).
Results: This method has been applied to brainstem auditory evoked potentials (BAEPs) time series. Preliminary experiments were carried out to analyse several aspects of the method including the best configuration of the pattern discovery technique parameters. We then applied the method to the BAEPs of 83 individuals belonging to four classes (healthy, conductive hearing loss, vestibular schwannoma—brainstem involvement and vestibular schwannoma—8th-nerve involvement). According to the results of the cross-validation, overall accuracy was 99.4%, sensitivity (recall) was 97.6% and specificity was 100% (no false positives).
Conclusion: The proposed method effectively reduces dimensionality. Additionally, if the symbolic transformation includes the right domain knowledge, the method arguably outputs a data representation that denotes the relevant domain concepts more clearly. The method is capable of finding patterns in BAEPs time series and is very accurate at correctly predicting whether or not new patients have an auditory-related disorder.

1. Introduction

Time series are a complex data type consisting of a potentially large time-ordered sequence of values. Time series appear in a wide range of domains, such as medicine, astronomy, seismography, economy, climatology, etc. Over the last few years, the community of data mining scientists has taken an interest in research on time series, and countless medical applications have been built [1–7]. Complex tests like electrocardiograms, electroencephalograms, evoked potentials, isokinetics tests, otoacoustic emissions, etc., are regularly performed in many fields of the medicine. These tests record data in the form of time series that provide specialist physicians with very useful information for patient diagnosis. Physicians use their expert knowledge to analyse these time series in search of the signs characteristic of a disease or of a healthy person. In many cases, time series analysis is very complex, and physicians may find data mining techniques, which automate the sign identification process, very helpful [8].

Many data mining techniques analyse time series numerically. In many medical domains, however, physicians are more interested in the morphology of the time series than in their numerical values.

* Corresponding author.
E-mail addresses: me.molina@alumnos.upm.es (M.E. Molina), aurora@fi.upm.es (A. Perez), jpvalente@fi.upm.es (J.P. Valente).

http://dx.doi.org/10.1016/j.artmed.2016.05.001
0933-3657/

Experts often look for peaks, troughs, sharp ascents, peak sequences or other types of shapes, that is, anything that represents the time series behaviour. We present a technique capable of analysing and transforming a numeric time series into a sequence of symbols that represent the time series behaviour. This symbolization process is capable of representing the time series in terms that are understandable for the domain expert. Additionally, it significantly reduces time series dimensionality.

Using this symbolic representation, we propose a method for classifying time series. This method searches for frequent subsequences (patterns) in the sequences of symbols that could, under certain conditions, be a key diagnostic aid for physicians. This method, called symbolic pattern-based classification (SPC), is divided into the following steps:

- Transform the numeric time series into symbolic time sequences that reflect key domain concepts.
- Mine sets of symbolic time sequences belonging to the same population group (class) to find their representative patterns.
- Classify new symbolic time sequences based on the class patterns.

For the purposes of evaluation, we applied our method to the domain of brainstem auditory evoked potentials (BAEPs) to detect conductive hearing loss and vestibular schwannoma abnormalities in a group of young people suspected of having hearing disorders. This study was performed in partnership with the Homero Castanier Crespo Hospital in Ecuador.

Auditory evoked potentials are the neuroelectric response of the auditory system to acoustic stimuli. They have been extensively used as a non-invasive electrophysiological tool for studying the human auditory system [9]. The measured signals are the result of many ionic currents associated with transduction processes that take place in the cochlear hair cells and with the generation of action potentials in the auditory nerve fibres. Depending on the time elapsed after stimulus onset, the evoked potentials can be classified as short-, medium- and long-latency potentials. The most commonly used potentials in clinical medicine are short-latency BAEPs. Short-latency BAEPs are measured during the first 10–15 ms after stimulus onset. A common use of BAEPs for clinical purposes is to diagnose hearing problems and brainstem tumours. The recorded data are time series with a duration of up to 15 ms that plot the evolution of the brainstem response to an acoustic stimulus against time.

The remainder of the paper is organized as follows. In Section 2, we briefly summarize key related work. In Section 3, we describe the proposed method and detail the symbolic transformation, pattern discovery and pattern classification techniques. In Section 4, we present the BAEPs domain on which the method has been applied and tested by means of several experiments. In Section 5, we report and briefly discuss the results. Finally, Section 6 outlines the conclusions.

2. Related work

Two major groups of techniques for processing time series are numeric approaches and techniques based on time series morphology.

The most commonly used numeric techniques for time series analysis are the discrete Fourier transform (DFT) [10] and discrete wavelet transform (DWT) [11]. They represent the time series using a small number of coefficients, thereby reducing dimensionality. The first k coefficients of the time series are used for comparison. Time series with more similar coefficients are more alike.

Another group of techniques focuses on the morphology (shape) of the time series [12,13] rather than on numerical values. The goal when comparing time series is to find series whose plots look similar even if they are on different scales. For example, two time series that peak at a similar timestamp might be similar even if they do not have the same values at that timestamp. According to this approach, time series whose behaviour is similar are regarded as being alike. Time series behaviours can be represented by symbols, thereby simplifying the time series. The resulting symbolic time sequence should represent the relevant information in the original time series.

The selection of the best technique largely depends on the domain and goals. On one hand, numeric techniques provide very accurate results, but domain experts find them hard to understand. On the other hand, symbolic sequences can be used to adopt relevant domain concepts in the data mining process. This is very important because the results can then be explained using the domain terminology [14]. An understandable explanation is a key issue on which computer system acceptance by a user hinges [15,16]. This is especially relevant in a medical domain, and it is the main reason that led us to opt for a symbolic approximation.

2.1. Symbolic representation of time series

A very important task in knowledge discovery processes from time series is to reduce data dimensionality while retaining their key characteristics. Symbolic transformation is one of the most promising alternatives since it can represent the data with domain terminology. It also has the advantage of implicitly removing noise through complexity reduction [17].

The idea of transforming a time series into a sequence of symbols (with a previously defined alphabet) that represents the original series with an acceptable loss of information is not new. Agrawal et al. [12] proposed a shape definition language (SDL) with symbols like up, down, stable, etc. André-Jönsson and Badal [18] applied a similar idea for time series indexation. Discrete representations of time series have been output through segmentation. Some such proposals are based on piecewise approximation, like piecewise linear approximation (PLA) [19,20], piecewise aggregate approximation (PAA) [21], adaptive piecewise constant approximation (APCA) [22] or multiresolution piecewise aggregate approximation (MPAA) [23]. Symbolic aggregate approximation (SAX) [24] performs a two-step time series transformation: first it discretizes the time series using PAA and then it converts this representation into a symbolic string. Cassissi et al. [25] propose a transformation of time series into symbolic sequences that retains the most important original characteristics, e.g., local max and min, peaks, valleys, etc.

2.2. Pattern discovery

Frequent pattern discovery was first proposed by Agrawal et al. [26] for discovering association rules in shopping cart transactions. A frequent pattern is a set of items that appear in a dataset with a frequency greater than or equal to a predefined threshold. Many papers on this topic have been published following this seminal work.

Agrawal and Srikant [27] stated an important property: a set of items with length k is frequent if and only if all of its subsets are frequent too. This property is called Apriori. Applying this property to frequent pattern discovery in time series, it can be said that frequent patterns of length 1 can be used to find patterns of length 2 and so on until no longer patterns can be found.

Han et al. [28] presented a proposal for discovering partial periodic patterns in time series. This proposal combines three elements: a data cube [29], a bitmap technique and the Apriori property. Rather than searching for patterns with perfect periodicity, it uses a minimum confidence value to provide some level of flexibility. Han et al. [30] propose the FP-growth (frequent pattern