Information Sciences 330 (2016) 125–144

Contents lists available at ScienceDirect

Information Sciences

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

A soft computing framework for classifying time series based on fuzzy sets of events

Juan Ares [b], Juan A. Lara [a,*], David Lizcano [a], Sonia Suárez [b]

[a] 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
[b] University of A Coruña, Campus de Elviña, s/n, 15071 A Coruña, Spain

ARTICLE INFO

Article history:
Received 7 May 2014
Revised 1 September 2015
Accepted 2 October 2015
Available online 19 October 2015

Keywords:
Data mining
Time series
Event
Soft computing
Fuzzy set

ABSTRACT

Time series are sequences of data gathered over a period of time that emerge in different domains and whose analysis requires the application of specialized techniques, like, for example, data mining. Many existing time series data mining techniques, like the discrete Fourier transform (DFT), offer solutions for analysing whole time series. Often, however, it is more important to analyse certain regions of interest, known as events, rather than whole time series. Event identification is a highly complex task, as it is not always possible to determine with absolute certainty whether or not a segment of a time series is an event. In such cases, the best practice is to establish the certainty of this segment being a time series event, thus outputting a fuzzy set of events.

In this paper we propose a framework that is capable of identifying events and establishing the degree of certainty that a domain expert would assign to the identified events based on a previous training process assisted by a panel of experts. Having identified the events, the proposed framework can be used to classify time series. This is done by means of a process that combines time series comparison and time series reference model generation by analysing the events contained in the respective time series and the certainties of the identified events. The proposed framework is an evolution of an earlier framework that we developed which did not apply soft computing techniques to identify and manage the time series events.

We have used our framework to classify times series generated in the electroencephalography (EEG) area. EEG is a neurological exploration used to diagnose nervous system disorders. The performance of the framework was evaluated in terms of classification accuracy. The results confirmed that, thanks to the use of soft computing techniques, the new framework substantially improves the time series classification results of its predecessor.

1. Introduction

A times series $X = \{x_t, t = 1, \ldots, N\}$ is defined as a sequence of data gathered over a period of time, usually at regular intervals, where N is the number of observations and x_t is the value measured at time t. The analysis of time series, which occur in countless domains, calls for the use of specialized techniques that use data mining algorithms, for example [19].

* Corresponding author. Tel.: +34 630524530.
E-mail address: juanalfonso.lara@udima.es (J.A. Lara).

http://dx.doi.org/10.1016/j.ins.2015.10.014
0020-0255/

In some domains, time series record information, of which only some regions are of interest. These regions of interest represent the occurrence of an event of interest to domain experts. Each of these regions of interest in a time series is known as an **event**.

From the viewpoint of information theory, the concept of time series event is closely related to the concept of entropy. The entropy of a system means the amount of information that a set of system symbols contains. If, in our case, the systems are time series, the events are the regions of the time series that contain more information, that is, have greater entropy.

In order to be able to extract useful knowledge from time series that contain events, it is necessary to first identify those events. Events are the only regions of the time series that provide information of interest. Say, for example, that we want to examine the seismicity of a particular geographical region. In this case, we would have to analyse the seismograph tracings corresponding to the occurrence of seismic phenomena because no information of interest to the expert is recorded outside the periods of seismic activity. The identification of such events is an open problem. Most existing techniques are either only applicable to particular domains for which they propose ad hoc mechanisms or identify time series breakpoints that are meaningless because they are devoid of expert knowledge.

We have addressed the problem of identifying events elsewhere [6,33], proposing a language for defining events in multidimensional times series. This language models the process of identifying events enacted by experts based on domain expert knowledge. The events definition language was designed to be generally applicable to multiple domains.

One of the future lines outlined in this proposal was to consider soft computing techniques in order to identify events. This was motivated by the fact that it is often not possible to determine beyond all doubt whether or not a segment of a time series is an event. This depends on the experience and expertise of the consulted domain expert. In view of the subjectivity of the event identification process and the impact of domain expert training, experience and opinion, it is better to determine the certainty of this segment of a time series being an event.

The use of soft computing techniques is a common practice in data mining, a discipline which tends to handle imprecise and uncertain data, like events [41].

1.1. Contribution

In this article we present a modification of an earlier event identification mechanism that we proposed. The modification applies soft computing techniques in order to establish the certainty of a segment of a time series being an event. To do this, our proposal is divided into two phases: (i) a training phase in which we have relied on a panel of experts to determine the certainty of a given segment being an event, generating fuzzy sets of events from the consensus opinion of a panel of experts; and (ii) a testing phase in which we evaluated how well our system automatically establishes event certainty values and its potential for classifying time series. As part of this second phase, we redesigned an earlier framework for classifying time series that we developed [3,6,31,32,34–36]. The framework is enhanced by adopting the concept of event certainty into the framework data mining algorithms.

Our framework for classifying time series containing events has been tested on electroencephalography data, a neurological examination that measures brain activity. The results confirm that the framework is valid for learning to establish and handling certainties during the process of classifying times series containing events. The new framework outperforms the original framework. As they handle the concept of event certainty, the classification models are more accurate than those that use Boolean logic and crisp sets.

Additionally, not only does the proposal offer a more precise and finer-grained mechanism for managing time series events, it also features a general-purpose mechanism for assigning event certainties. This mechanism is applicable in any field where event identification is an expert-mediated subjective and fuzzy process with which binary logic is incompatible.

1.2. Outline

Section 2 presents work related to this article. The proposed framework, described in Section 3, is applied to electroencephalography time series. The results of applying the framework are discussed in Section 4. Section 5 addresses the applicability of the proposal, whereas the findings and future work are detailed in Section 6.

2. Related work

In this article we propose a mechanism for discovering useful knowledge in time series containing events. This mechanism is an evolution of an earlier framework that we developed [3,6,31,32,34–36], which included a language for defining events in time series. This language was proposed to overcome the limitations of other existing event identification proposals, which are summarized in Section 2.1.

Event identification is not a trivial process, as it is not always possible to clearly define the conditions determining whether a segment of a time series is really an event. Under such circumstances, the use of soft computing techniques has proved to be very useful, as shown in Section 2.2.

Our proposal is based on establishing and handling event certainty. In order to calculate certainties, we have proposed a training mechanism based on the use of expert knowledge. As detailed in Section 2.3, expert participation in the data mining field is a common practice which outputs richer and more representative models. Expert participation is confined to the training