Applied Soft Computing 36 (2015) 589–598

Contents lists available at ScienceDirect

Applied Soft Computing

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

Representation model and learning algorithm for uncertain and imprecise multivariate behaviors, based on correlated trends

Miguel Delgado, Waldo Fajardo, Miguel Molina-Solana*

Department of Computer Science and AI, Universidad de Granada, Spain

ARTICLE INFO

Article history:
Received 24 February 2014
Received in revised form 26 May 2015
Accepted 27 July 2015
Available online 10 August 2015

Keywords:
Behavior modeling
Data series
Frequent Correlated Trends

ABSTRACT

The computational representation and classification of behaviors is a task of growing interest in the field of *Behavior Informatics*, being series of data a common way of describing those behaviors. However, as these data are often imperfect, new representation models are required in order to effectively handle imperfection in this context. This work presents a new approach, *Frequent Correlated Trends*, for representing uncertain and imprecise multivariate data series. Such a model can be applied to any domain where behaviors recur in similar—but not identical—shape. In particular, we have already used them to the task of identifying the performers of violin recordings with good results. The present paper describes the abstract model representation and a general learning algorithm, and discusses several potential applications.

1. Introduction

In recent years, behavioral sciences have received a lot of attention from the informatics perspective. This fact is mainly due to current demands for behavior analysis and understanding going faster than the capability of traditional methods and techniques in behavioral sciences. New computational tools for representing and working with behaviors are very welcomed and a growing field of research, namely *Behavior Informatics* [1], is receiving increasing recognition. A recent book [2] precisely points out this interest by gathering together a representative number of different works from several researchers in this area.

Intuitively, we can define a behavior as a set of actions that are characteristic of one individual or phenomenon. These actions are ordered (or partially ordered), and therefore indexed by a variable, which is generally *time*. By representing behaviors, two goals are aimed: (1) identification and tagging of the behaviors and (2) forecasting future actions within them.

Behaviors, as we have just defined them, can be found in many different domains. The following are some illustrative examples of such phenomena that can be represented and individually identified (some of them will be later described in more detail in Section 4.3):

- The weather in a given area, represented as a series of observations at different time instants, including information such as temperature, precipitations or wind speed [3].
- The movements of a (injured) knee when doing some rehabilitation exercises, by monitoring the position of several reference points at different time instants [4].
- The way of playing an instrument. A particular performance of a piece of music can be represented as a series of notes with its respective duration and volume, among others attributes [5].
- User modeling for personalization of applications or prediction. This task is specially complex due to the inherent imperfect and elusive nature of human behaviors [6].
- The way a human being behaves within Ambient Assisted Living, with the aim of identifying strange actions and situations of potential danger [7].
- The personalization of mobile services. As mobile devices increase their capacity, new services and applications are developed which need modeling the user behavior and context [8].
- The interactions between currency exchanges. Several works have studied how some currencies behave against each other at different financial situations [9].

What all these real-world phenomena have in common is that they can be naturally represented by data series. As databases from industrial and biological areas often contain timestamped

* Corresponding author at: Department of Computer Science and AI. ETSIIT, c/ Daniel Saucedo Aranda, s/n 18071 Granada, Spain.
E-mail addresses: mdelgado@ugr.es (M. Delgado), aragorn@ugr.es (W. Fajardo), miguelmolina@ugr.es (M. Molina-Solana).

http://dx.doi.org/10.1016/j.asoc.2015.07.033
1568-4946/© 2015 Elsevier B.V. All rights reserved.

or ordered records, data series are gaining weight as a suitable source of information, and working with them has become a relevant machine learning task. It is common to obtain those records in an automatic manner from different sensors.

Two main goals of data series analysis are found in literature [10]: forecasting and modeling. The aim of forecasting is to accurately predict the next values of the series, whereas modeling aims to describe the whole series in a compact form. Even though they can be sometimes related, they usually differ as finding a proper model for the long-term evolution might not be the best approach to predicting the short-term evolution and vice versa.

Interestingly enough, forecasting and modeling are also the main tasks concerning behaviors, as we previously said. Therefore, data series are a suitable representation for behaviors, being also the most common one. In either case, and whatever the goal of a particular data series analysis is, data representation is a crucial task anyway. It is hence required a formal representation capable of modeling the complexity of the particular data. This representation must be more reduced than representing all the observations of the phenomenon, but still describe it accurately enough.

An additional problem is that real-world information is hardly certain, complete and precise; more on the contrary, it is usually incomplete, imprecise, vague, fragmentary, not fully reliable, contradictory, or imperfect in some other way. Two ways of addressing imperfection have been historically employed for representing information in a computer [11]:

- The first solution consists in restricting the model to only that part of the available information of the real world that is accurate and reliable. Such a constrained approach avoids further complications of representation, but it lacks the capacity of capturing the whole rich notion of information in human cognition, being generally very limited.
- The second solution implies developing models capable of representing imperfect information. As this approach allows a greater number of applications, it is the one that developers usually implement in their systems. However, those models cannot successfully cope with the whole range of imperfections that generally appear in real life, and in many occasions data are simplified to a point that makes them easily treatable with traditional computational tools, but losing part of their meaning.

Due to this lack of general systems capable of dealing with any kind of imperfect data, developers have been forced to handle this information in an ad hoc manner; that is, by devising specific algorithms and systems for each new application, domain and representation. Therefore, in order to model the real world as accurately as possible, several approaches for dealing with imperfect information have been introduced and studied, and there are plenty of ongoing research efforts on the matter.

Although some schemes have been proposed for directly handling imperfect information coming as data series [12,13], most of the research has focused on similarity measures to deal with imperfection. Hence there is still a need for further research and new practical systems capable of accurately modeling imperfect data series, being the field of *Behavior Informatics*, in particular, greatly benefited by such advances.

This paper addresses this necessity by proposing a novel approach for representing imperfect behaviors—concretely uncertain and imprecise—that come in the form of multivariate data series. Our work shows how we can represent underlying local trends in the data in an easy and effective way, without a complicated formalism. Some other works have identified the necessity of focusing in frequent local cues for behavior modeling [14,15]. The further aim of the work is addressing the issue of soft data series[1] recognition and comparison.

Note here, that the main objective of our paper is representing behaviors (defined by data series) and to identify future instances of those behaviors. While certainly possible to use the model to predict future values within the series, that is not the focus of the current work.

Specifically, our proposal identifies given behaviors through capturing their general footprint by means of discovering repetitive patterns in one dimension and their interdependence with patterns in other dimensions. That is, by computing the frequency distribution of the co-occurrence of patterns in different dimensions. The process can be divided in the following three stages:

1. high-level abstraction of the observations within each dimension;
2. tagging according to the patterns identified in one of the dimensions;
3. characterization of behaviors as sets of frequency distributions.

The paper is organized as follows. Section 2 offers both an introduction to data series, and to imperfect data, describing the main forms of imperfection and the problems to address. In Section 3, we describe our proposed model, *Frequent Correlated Trends* for generically representing imperfect behaviors, and the developed system, including data gathering, representation and distance measurement. This section also introduces the formal notation and an illustrative example. Section 4 discusses the main advantages of our proposal and its computational complexity, and presents some potential domains of application, illustrating the use of the model on them. The paper concludes (Section 5) with a summary, some final considerations, and pointing out future work.

2. Background

This section aims to present the necessary background that frames the present work. Two main topics are described: data series, and imperfect information.

2.1. Data series

As data series[2] have an increasing popularity and they are the formalism we will use in this paper, we devote this section to briefly introduce them. The interested reader should refer to any basic reference on time series analysis (for instance [10]) for further information about ordered series. Databases in areas such as Engineering, Medicine or Finances often contain timestamped or ordered records. The analysis of data series (and time series in particular) is then of great interest in these areas, and searching for similarities between data series is fundamental for several data mining tasks (e.g. classification, rule gathering, clustering or finding patterns) within these domains.

A data series A can be intuitively defined as an ordered sequence (finite or not) of values obtained at successive intervals of an indexing variable, often *time*. Each one of these observations A_i takes values from a domain $\mathcal{U}^n$. According to the value of n, we have several kinds of series. If $n = 1$, each element is a single (scalar) value and we have an *univariable* series. On the other hand, if $n > 1$,

[1] The term *soft data series* is used for series whose values are not accurate or verifiable.

[2] The reader will notice that in the following we mainly focus our descriptions on time series. That is due to the fact that *time* is by far the most common indexing variable in data series. However, we will keep using the term *data series* to make explicit that our model is not limited to any particular kind of series.