Expert Systems with Applications 42 (2015) 596–611

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Characterization of time series for analyzing of the evolution of time series clusters

Ana P. Serra, Luis E. Zárate *

Department of Computer Science, Applied Computational Intelligence Laboratory-LICAP, Pontifical Catholic University of Minas Gerais, Av. Dom José Gaspar 500, Coração Eucarístico, Belo Horizonte 30535-610, MG, Brazil

ARTICLE INFO

Article history:
Available online 21 August 2014

Keywords:
Time series
Temporal database
Cluster analysis
Multivariate database
Data mining

ABSTRACT

This work proposes a new approach for the characterization of time series in databases (temporal databases – TDB) for temporal analysis of clusters. For the characterization of time-series it were used the level and trend components calculated through the Holt-Winters smoothing model. For the temporal analysis of those clusters, it was used in a combined manner the AGNES (Agglomerative Hierarchical Cluster) and PAM (Partition Clustering) techniques. For the application of this methodology an R-based script for generating synthetic TDBs was developed. Our proposal allows the evaluation of the clusters, both in the object movement such as in the appearance or disappearance of clusters. The model chosen to characterize the time-series is adequate because it can be applied for short periods of time in situations where changes should be promptly detected for quick decision making.

1. Introduction

It is currently observed an accelerated growth in data storage capacity. This increase has allowed the registration of data for long periods of time introducing a temporal character in it. It is then passed to handle huge volumes of historical data that should be exploited in the temporal point of view for a better understanding of a problem. This is also a characteristic scenario in Big Data.

Within this context there was an extension of conventional data mining for Temporal Data Mining (TDM), which in essence is based on mining of sequential data. According to Roddick and Spiliopoulou (2002), TDM brought the ability to mine activities and trajectories more than simply defined states or moments of time. TDM makes the extracted knowledge more complete.

Among the initial studies of manipulation of historical time series, the work of Last, Klein, Kandel, and Abraham (2001) stands as an important contribution to present a methodology for the application of data mining in time series. This methodology suggests that the conventional steps of a data mining process, applied to a database containing data collected over a particular time or time interval, called by many authors as a static database, were restructured for data mining on a temporal database (TDB).

As exposed previously, the need arises for algorithms, techniques, procedures and methods for data mining capable of dealing with temporal information. The conventional mining algorithms need to be adapted to treat temporal databases, or temporal data needs to be pre-processed and converted into "frozen" or summarized point values for an instant of time '*t*' before the application of conventional data mining techniques.

With the growing interest of the scientific community in recognition of the time value of data, several studies from TDM have emerged addressing not only the sequential ordering of data, but the time value itself contained in the historical data.

The work of Lin, Orgun, and Williams (2002) provides a simplified view of the TDM process and foundations to manipulate temporal data. The authors discuss the two fundamental problems of TDM which are to calculate the similarity between time series and the identification of periodicities in historical data. At the end, a challenge was launched for the search for a general theory for TDM which would represent a milestone in this area, since the work has not previously had an established theoretical foundation. In Last, Kandel, and Bunke (2004), the authors presented an important contribution to structuring the area containing an overview of relevant articles that present proposals for the biggest challenges of TDM. From this work, new categorizations arise where researchers eventually end up inducing the appearance of subareas within the TDM.

* Corresponding author. Tel.: +55 31 3319 4117; fax: +55 31 3319 4001.
E-mail addresses: serra.anapaula@yahoo.com.br (A.P. Serra), zarate@pucminas.br (L.E. Zárate).

http://dx.doi.org/10.1016/j.eswa.2014.08.012
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

On the other hand, in mining works of time series using clustering techniques is possible to observe the emphasis in the discussion of the best way to calculate similarity/dissimilarity measures. Approaches that generalize this issue are still under discussion, and the conclusion that most of the studies reached is that this choice depends on the domain and the database structure (Liao, 2005).

As contextualized previously, when clustering techniques are applied on the TDB, the main aspect to be considered is the characterization of the series and therefore it is necessary to extract features that represent the essence of a series. The best measures for the characterization of a series are extracted from structures present in a generalized model of a time series, which is essentially made up of the following components: level, trend, seasonality and periodicity. This study contributes in this direction presenting a methodology for clustering time series by temporal windows through the characterization of the series, based on the structural components of a generalized model, in this paper the Holt-Winters model. This model is easy to understand, has low computational cost, it is applicable in series with few observations and it allows the adjustment of the values of the smoothing constants for each component in the model, making it a flexible model. It is important to note that the ability of this model to deal with a few observations allows the selection of temporal windows with lower amplitudes, restricted to a minimum of three observations necessary to adjust the model. This is a useful feature when one wishes to apply this model in a historical data that is still relatively small and when there is a need to obtain information for decision making.

In this work the similarity shall be based on the characteristics (level and trend) extracted from each series so far from Holt-Winters model. With this, it is possible to understand the dynamics of moving objects and clusters over time, providing subsidies for explanation and prediction of behaviors and phenomena.

Due to the lack of real databases that are complete and comprehensive that allow analysis of proposal, a generator for a synthetic multivariate TDB was built. The generator allows the setting of the number of attributes in the database, how many observation points in time and how many records will be considered. For the generation of the synthetic database, the R[1] environment was used.

This paper is organized as follows: in Section 2 a brief review of the literature about TDM is presented; in Section 3, the formalization of our proposal and the problem of characterization of series are presented; Section 4, shows how the time series generator used to evaluate the proposal was developed; in Section 5, simulation results are presented and discussed; in Section 6, the final comments and conclusions are presented.

2. A brief review of the literature

As it can be observed through the extensive literature, the discussion of methodologies for handling temporal data is a latent area of research and with a strong tendency of growth especially motivated by new paradigms and trends in Big Data. Among contributions more recent one can cite: Xiong and Yeung (2004), Liao (2005), Gardner and Diaz-Saiz (2008), Zhang, Chen, Brijs, and Zhang (2008), Böttcher, Spott, Nauck, and Kruse (2009), Gullo, Ponti, Tagarelli, and Greco (2009).

In Roddick and Spiliopoulou (2002), the authors proposed taxonomy for TDM considering three dimensions: (i) the *type of data*; (ii) the *ordering of the data*; and (iii) the *mining paradigm* used. As for the *type data* dimension, three divisions were considered: (a) approaches that deal with scalar values; (b) approaches that deal with events; (c) approaches that deal with mining results. The latter one also called mining high-order (Roddick, Spiliopoulou, Lister, & Ceglar, 2008) is considered by the authors as a challenge in the TDM. To the dimension corresponding to the *ordering of the data* (ii), two divisions were defined: (a) works dealing with temporally ordered data and (b) works that deal with data where there is no sort order. For the *mining paradigm* dimension (iii), two divisions were created: (a) work for discovering temporal association rules (Zhang et al., 2008); and (b) work classification which may be supervised or unsupervised, the latter one corresponding to the clustering technique (Böttcher et al., 2009).

Considering the taxonomy presented by Roddick and Spiliopoulou (2002) this work is inserted in the categories *type data* (i) 'mining result' (c), and in the *mining paradigm* (iii) – 'classification work' (b). In this paper, a methodology to monitor the evolution of cluster models and their objects, after the application of an unsupervised classification on a TDB from a characterization of the time series, via Holt-Winters smoothing model (Gardner & Diaz-Saiz, 2008; Winters, 1960), that make up the database is proposed.

In Liao (2005) the author has reviewed the methodologies applied in the cluster analysis over time series. Most methodologies are restricted to a univariable time series, and the difference between them is how to calculate the similarity/dissimilarity measure between series, which depends on the type (regular and irregular) and the specific characteristics of TDB.

The approaches raised by Liao (2005) were organized into three categories: (1) cluster analysis applied directly to the TDB, with some modifications in conventional data mining algorithms; (2) cluster analysis on features extracted from the time series (Gullo et al., 2009); and (3) cluster analysis based on models built from the TDB (Xiong & Yeung, 2004). In the first category, it proposes a direct manipulation of the original data, resulting in a very high computational cost in implementing TDM techniques. In the third category, the proposal was to use information models (coefficients, residue, etc.), which is not very consistent since these coefficients have no direct relation with the problem domain. Many of these factors are merely adjustments of the models for the historical data. Among the three categories the second, which covers the applications of cluster analysis based on features extracted from the time series, are the most interesting, because with the data synthesis, it is possible to gain in computational cost and if well chosen, these extracted features can correctly represent the information contained in the TDB. For example, in Gullo et al. (2009) the authors propose a representation of time series based on dynamic time warping (DTW). The new approach permits to capture the main trends of time series and data compress, which can be used for similarity detection of time series. DTW is a method that searches an optimal match between two given sequences which may vary in time or speed. Other alternative ways of dealing with the problem of computational cost in processing a TDB consist in the use of techniques to reduce the dimensionality. Amongst the most relevant works Wang, Wirth, and Wang (2007), Wang and Megalooikonoumou (2008) and Al-Naymat and Taheri (2008) can be mentioned.

3. Characterization of time-series for temporal evolution of clusters

Fig. 1 illustrates the proposal of this work. Consider a multivariate TDB containing records that contain variables (attributes) expressed through temporal series. Then it is possible to characterize these series extracting the level and trend components and applying a clustering technique chosen for two time intervals, Window 1 and Window 2. Each point on the graph represents a record of the database shown for example through the first two

[1] www.r-project.org.br.