Contents lists available at ScienceDirect

Information Sciences

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

Piecewise two-dimensional normal cloud representation for time-series data mining

Weihui Deng, Guoyin Wang*, Ji Xu

Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China

ARTICLE INFO

Article history:
Received 5 December 2015
Revised 6 September 2016
Accepted 10 September 2016
Available online 14 September 2016

Keywords:
Two-dimensional normal cloud
Representation
Dimensionality reduction
Similarity measure
Cloud model
Time-series data mining

ABSTRACT

Many high-level dimensionality reduction approaches for mining time series have been proposed, e.g., SAX, PWCA , and Feature-based. Due to the rapid performance degradation of time-series data mining in much lower dimensionality and the continuously increasing amount of time series data with uncertainty, there remains a burning need to develop new time-series representations that can retain good performance in much lower reduced space and address uncertainty efficiently. In this work, we propose a novel time series representation, namely Two-dimensional Normal Cloud Representation (2D-NCR), based on cloud model theory. The representation achieves dimensionality reduction by transforming the raw time series into a sequence of two-dimensional normal cloud models. Moreover, a new similarity measure between the transformed time series is presented. The proposed method can reflect the characteristic data distribution of the time series and capture the variation with time. We validate the performance of our representation on the various data mining tasks of classification, clustering, and query by content. The experimental results demonstrate that 2D-NCR is an effective and competitive representation for time-series data mining.

1. Introduction

A time series is a collection of numerical values obtained from sequential measurements over time. It is generally used to describe the current status and future variations of objects over time lags. Enormous amounts of time-series data are being continually generated and collected from various domains, including the financial industry, the ecological environment, aviation, telecommunication, the medical field, and so on.

Time-series data mining (TSDM), as a research issue currently receiving significant amounts of attention, has been extensively studied to discover the knowledge and information hiding in time-series data. Various tasks are applied to mine time series, which can be summarized by 7 aspects [13]: classification [1,6,18], clustering [15,28], query of content [23,27,37], prediction [10,21,42], segmentation [16,43], anomaly detection [38] and motif discovery [3]. These tasks have been successfully applied to various fields, including power systems, smart grid, stock market time-series prediction and biomedicine. With the rapid growth of digital sources of information, all of the tasks of TSDM are unveiling various facets of complexity. The most prominent problems stem from the high dimensionality of time-series data and the difficulty of defining a form

* Corresponding author.
E-mail addresses: dengweihui@cigit.ac.cn (W. Deng), wanggy@ieee.org (G. Wang).

http://dx.doi.org/10.1016/j.ins.2016.09.027

of similarity measure based on human cognition. Therefore, the following two major issues must be addressed to perform TSDM tasks efficiently [13]:

- Time-series representation. The motivation of time-series representation is to gain additional benefits such as acceleration of processing, accuracy improvement and noise removal, with a simultaneous emphasis on the essential and concise characteristics of the raw time series. Thus, a representation technique should have the properties of significant dimensionality reduction, emphasis on fundamental data distribution and variation, good reconstruction quality from the reduced space, and insensitivity to noise.
- Similarity measure. How to distinguish or match any pair of time series and how to formalize an intuitive similarity or distance between two time series based on human cognition are the two core issues of similarity measure. Therefore, a similarity measure should have the following properties: recognition of perceptually similar time series, consistency with human intuition, emphasis on the most salient features on both local and global scales, and capability to identify or distinguish arbitrary objects without any assumptions.

In this work, we primarily focus on the problem of time-series representation and the corresponding similarity measure. The formalized problem of time-series representation can be described as follows:

Given a time series $T = (t_1, t_2, \ldots, t_n)$ of length n, a time-series representation of T is a model $\overline{T}$ of reduced dimensionality $w(w \ll n)$ such that $\overline{T}$ closely approximates T or extracts the essential features of T.

Over the past several decades, numerous time-series representations have been developed and successfully applied to various time-series data mining tasks. The well-known representations include the transformation domain methods (like discrete Fourier transform (DFT) [14], discrete wavelet transform (DWT) [4]), piecewise aggregate approximation (PAA) [22], piecewise linear approximation (PLA) [32], singular value decomposition (SVD) [30], symbolic aggregate approximation (SAX) [26], symbolic aggregate approximation based on trend distance (SAX-TD) [33], derivative segment approximation (DSA) [19], piecewise cloud approximation (PWCA) [25], learned pattern similarity (LPS) [2], and non-parametric symbolic approximate representation (NSAR) [20]. All of the methods mentioned above have in common the capability of representing time series. Additionally, particular time-series representations have been proposed for the specific TSDM tasks [8,17,31]. For instance, feature-based representations have been developed for time-series classification [17].

However, most of the existing dimension-reduction representations mentioned above have a major limitation; their performance degrades rapidly in much lower dimensionality, although they are the better approximate representations for mining time-series data in the appropriate reduced space. That is, a larger compression ratio means worse performance. PWCA was proposed to overcome this limitation by representing raw time series as a series of cloud models. However, because it only emphasizes the distribution of raw time series without considering variation with time, its performance still must improve. In addition, the prominent state-of-the-art methods for time-series dimension-reduction representation cannot address the vagueness and uncertainty inherent in certain time-series data because of inaccuracies in measurements, incomplete sets of observations, or difficulties in obtaining the measurements.

The motivation of this paper is to develop a novel time-series representation and the corresponding similarity measure in a reduced space to tackle the three problems mentioned above. First, the new representation must not only be able to approximate time series well for the TSDM tasks but also guarantee its performance in much lower dimensionality. In other words, the new method must guarantee its efficiency and accuracy in both high and much lower reduced space. Second, the proposed approach should consider the distribution and variation of the time series. Finally, because increasing numbers of time-series datasets with uncertainty are being employed, the new approach must be capable of handling uncertainty so that it can be successfully applied to various TSDM tasks.

In this paper, we propose a novel time-series representation to reduce the dimensionality of time series based on cloud model theory. The proposed method first calculates the first-order difference of the raw time series to obtain the variation series over time. Second, we simultaneously partition the raw series and variation series into equal-length “subsequence pairs”. Each “subsequence pair” is then transformed into one two-dimensional normal cloud model. Thus, a two-dimensional normal cloud model sequence for a time series is obtained. Finally, we propose a new similarity measure between two two-dimensional normal cloud models and the corresponding similarity measure for the transformed time series.

To compare with prominent state-of-the-art methods for time-series representation and dimensionality reduction, our approach has the following advantages:

- Using six numerical characteristics of a two-dimensional normal cloud model to represent a subsequence in the reduced space can retain more characteristics of the raw time series than state-of-the-art methods such as PAA, SAX, or transformation domain methods (DFT, DWT), which determines that our approach will have better performance in much lower reduced space.
- The proposed approach considers the distribution of raw time series and the variation with time rather than only the distribution of the time series (PWCA); thus, the proposed approach outperform PWCA. In addition, the similarity measure between the two cloud models in this paper is more simple and efficient than that of is PWCA.
- The normal cloud model can effectively integrate the randomness and fuzziness of concepts via its numerical characteristics, which will be further discussed in Section 2.2. This effectiveness guarantees that the proposed method can efficiently address the uncertainty inherent in time-series data. In addition, 2D-NCR is insensitive to abnormal data or noise because the cloud model emphasizes the whole distribution rather than a single point.