Physica A 427 (2015) 10–25

Contents lists available at ScienceDirect

Physica A

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

Piecewise aggregate representations and lower-bound distance functions for multivariate time series

Hailin Li

College of Business Administration, Huaqiao University, Quanzhou 362021, China

HIGHLIGHTS

- Novel methods reduce the dimensionality of MTS from two dimensions.
- MTS with different lengths is represented by extended sequences with same length.
- Some functions lower bounding on the Euclidean distance and DTW are proposed.
- Fast similarity measure can be achieved by the lower-bound functions on DTW.

ARTICLE INFO

Article history:
Received 18 November 2014
Received in revised form 28 January 2015
Available online 4 February 2015

Keywords:
Piecewise aggregate representation
Similarity measure
Lower bound function
Multivariate time series
Dynamic time warping

ABSTRACT

Dimensionality reduction is one of the most important methods to improve the efficiency of the techniques that are applied to the field of multivariate time series data mining. Due to multivariate time series with the variable-based and time-based dimensions, the reduction techniques must take both of them into consideration. To achieve this goal, we use a center sequence to represent a multivariate time series so that the new sequence can be seen as a univariate time series. Thus two sophisticated piecewise aggregate representations, including piecewise aggregate approximation and symbolization applied to univariate time series, are used to further represent the extended sequence that is derived from the center one. Furthermore, some distance functions are designed to measure the similarity between two representations. Through being proven by some related mathematical analysis, the proposed functions are lower bound on Euclidean distance and dynamic time warping. In this way, false dismissals can be avoided when they are used to index the time series. In addition, multivariate time series with different lengths can be transformed into the extended sequences with equal length, and their corresponding distance functions can measure the similarity between two unequal-length multivariate time series. The experimental results demonstrate that the proposed methods can reduce the dimensionality, and their corresponding distance functions satisfy the lower-bound condition, which can speed up the calculation of similarity search and indexing in the multivariate time series datasets.

1. Introduction

Time series is a ubiquitous data existed in different domains including finance, medicine, business and other industrial fields. Recently, time series data mining attracts much attention [1,2]. The type of the data includes univariate time series (UTS) and multivariate time series (MTS). They have a time-based dimension that makes the dimensionality very large as time goes on. Compared to UTS, MTS has more than one variables (or attributes) to describe the system that produces the data. The variable-based dimension places much burden on the process of data mining. In views of the two kinds of

E-mail address: hailin@mail.dlut.edu.cn.

http://dx.doi.org/10.1016/j.physa.2015.01.063
0378-4371/

dimensions, it is important to simultaneously reduce their dimensionality of MTS. In addition, some distance functions [3–6] are also designed to improve the quality of time series data mining. Some techniques of dimensionality reduction often combine with their corresponding distance functions [7–9] or correlations [10] for measuring the similarity between two different representations.

The tasks of time series data mining often combine some valid processes to improve the results [1,5,11]. Two kinds of the processes are the dimensionality reduction and distance function. So far, there are some existing methods used to reduce the dimensionality of MTS before data mining, such as singular value deposition (SVD) [12], principal component analysis (PCA) [13], independent component analysis (ICA) [14] and wavelet-based method [15,16]. They transform the original MTS into some important components or features. However, they often take much time to obtain the representations. In addition, some related distance functions are designed to clustering and classification. But most of the functions dissatisfy the lower-bound condition. It means that false dismissals for time series similarity search and indexing may be happened.

Fortunately, there are some ways to reduce the dimensionality for UTS, and their distance functions satisfying the lower-bound condition are used to fast measure the similarity between any two different representations [17]. It means that the dissimilar objects can be removed in advance through executing the algorithm of lower-bound function when applied to similarity search. The typical methods include parameter representation [18], piecewise aggregate approximation (PAA) [19], symbolic aggregate approximation (SAX) [7], piecewise linear approximation (PLA) [20], segmentation representation [21], piecewise vector quantized approximation [22] and so on. Especially, PAA is one of the most important methods to reduce the dimensionality of time series. Moreover, SAX based on PAA is one of the most popular methods to symbolize UTS. In addition, the corresponding distance functions are lower-bound on Euclidean distance and their extensions are also lower-bound on dynamic time warping (DTW) [9,23].

There exist many applicable functions to measure the similarity between two different representations. They must be lower-bound on dynamic time warping so that the false dismissals are avoided when time series are indexed by DTW. *LB_Yi* defined that the sum of distances between the maximum (minimum) of a time series and points in the other time series that are larger (smaller) than the maximum (minimum) was a lower-bound function [24,25]. *LB_Kim* is based on four features extracted from each time series, the maximum distance of the corresponding feature was reported as another lower-bound distance [26]. *LB_Keogh* [3] has good tightness and is the Euclidean distance between any parts of the candidate matching sequence not falling within an envelope and the nearest corresponding section of the envelope. It had already been successfully used to resolve the problems including music retrieval, handwriting retrieval, images indexing, web mining, and so on. LB_Improved [27] based on two-pass pruning technique was compared to LB_Keogh and was 2–3 times faster over random-walk and shape time series. It also showed that pruning candidates left from the Zhu–Shasha R*-tree with the full LB_Keogh alone were enough to significantly boost the speed and pruning power. *LB_ECorner* [28] is based on corner-like boundaries in warping matrix. Its performance also depends on a warping window, which means that the corner-like boundaries are chosen from the warping window. The experimental results [28] demonstrated that in most cases *LB_ECorner* is better than *LB_Keogh*. Lately, we proposed the extensions of *LB_Kim* and *LB_Keogh* and discussed the relationships of the existing lower-bound functions [29].

Beside the time-based dimension, MTS has the variable-based one, which makes some troubles to design a method of dimensionality reduction and a lower-bound function. The motivation of this research is to address the above mentioned problems and propose a novel method based on PAA to reduce the dimensionality of MTS from the two dimensions. At the same time, we use the algorithm of the traditional symbolization to obtain symbol string. Meanwhile, some distance functions that satisfy the lower-bound condition for dynamic time warping are proposed. In this way, we can use the lower-bound functions to index MTS as the traditional ones [3,24,26–29] index univariate time series. Especially, the traditional methods only focus on the comparison between two time series of which the length is equal. But in our methods, MTS with different lengths can be transformed into the equal-length ones by a representation method. Moreover, the distance functions measuring the two representations are also lower-bound on dynamic time warping. Our contribution can be summarized as follows:

(1) Two novel methods based on PAA and SAX are proposed to reduce the dimensionality of MTS from two kinds of the dimensions, the time-based and the variable-based.
(2) MTS with different lengths is represented by extended sequences with same length, which makes the two MTS items with different length can be conveniently compared to each other.
(3) Some fast lower-bound functions on Euclidean distance and DTW are developed to measure the similarity between two MTS, which contributes to multivariate time series indexing.

The remainder of the paper is organized as follows. In Section 2, Background and related work are introduced. The proposed methods are presented in Section 3. Some experiments are arranged to evaluate the proposed methods in Section 4. In the last section we conclude our work and discuss the future work.

2. Background and related work

Piecewise aggregate representations are often used to reduce the dimensionality of univariate time series. The typical two methods are piecewise aggregate approximation and symbolic aggregate approximation. They are two of the most popular methods to represent time series. In addition, their distance functions are lower-bound on dynamic time warping which is a robust similarity measurement.