Physica A 468 (2017) 622–637

Contents lists available at ScienceDirect

Physica A

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

Distance measure with improved lower bound for multivariate time series

Hailin Li *

College of Business Administration, Huaqiao University, Quanzhou 362021, China
Research Center for Applied Statistics and Big Data, Huaqiao University, Xiamen 361021, China

HIGHLIGHTS

- The proposed function LB_CMK is lower bound on DTW based on square local distance.
- The tightness is better than the existing methods for multivariate time series.
- The efficiency and effectiveness based on LB_CMK is improved for similarity search.

ARTICLE INFO

Article history:
Received 14 April 2016
Received in revised form 13 October 2016
Available online 21 October 2016

Keywords:
Multivariate time series
Lower bound function
Piecewise aggregate approximation
Center sequence
Data mining

ABSTRACT

Lower bound function is one of the important techniques used to fast search and index time series data. Multivariate time series has two aspects of high dimensionality including the time-based dimension and the variable-based dimension. Due to the influence of variable-based dimension, a novel method is proposed to deal with the lower bound distance computation for multivariate time series. The proposed method like the traditional ones also reduces the dimensionality of time series in its first step and thus does not directly apply the lower bound function on the multivariate time series. The dimensionality reduction is that multivariate time series is reduced to univariate time series denoted as center sequences according to the principle of piecewise aggregate approximation. In addition, an extended lower bound function is designed to obtain good tightness and fast measure the distance between any two center sequences. The experimental results demonstrate that the proposed lower bound function has better tightness and improves the performance of similarity search in multivariate time series datasets.

1. Introduction

Knowledge discovery from time series datasets is one of the most challenging research directions in the field of data mining [1]. Time series often includes two aspects of dimensionality, the time-based and the variable-based. The former refers to the length of time series and the latter means that there exist several variables (or features) to describe the feature of time series. When time series has only one variable, it is often regarded as univariate time series. In most cases, sequences with more than one variables to describe their features are regarded as multivariate time series. It means that multivariate time series generally has two aspects of dimensionality, which leads to much trouble to discover knowledge. Thus dimensionality reduction should be taken into consideration in the process of data mining. In addition, many tasks [2–6] such as classification, clustering, similarity search, patterns and motifs discovery often require distance (or similarity)

* Correspondence to: College of Business Administration, Huaqiao University, Quanzhou 362021, China.
E-mail address: hailin@mail.dlut.edu.cn.

http://dx.doi.org/10.1016/j.physa.2016.10.062
0378-4371/

measurement. The quality of distance function directly impacts the performance of the algorithms and techniques in the field of data mining [7,8].

There are many techniques to reduce the dimensionality of multivariate time series, such as singular value decomposition (SVD) [9], principal component analysis(PCA) [10,11], independent component analysis (ICA) [12], wavelet-based method [13] and key points extraction [14]. PCA is one of most popular methods applied to dimensionality reduction, which can transform the data to another feature space so that the first k principal components in the new space can retain most of the information about the original time series and represent the major features, where k is often much less than the number of the variables of multivariate time series. In this way, the variable-based dimensionality of the original multivariate time series is reduced. Furthermore, some methods [9,15] can be combined with time-based dimensionality reduction that are often applied to the field of univariate time series. So far, there are many refined methods used to reduce the time-based dimensionality, such as piecewise aggregate approximation (PAA) [16], symbolic aggregate approximation (SAX) [17], piecewise linear approximation (PLA) [18], discrete Fourier transform (DWT) [19], discrete fourier transform (DFT) [20] and so on. In addition, there are some approaches to achieve the reduction of the variable-based and the time-based dimensionality simultaneously [9,15].

Dimensionality reduction is often regarded as one of the important processes for time series data mining. Similarity measurement (or distance function) design is also an important and necessary work for some tasks of time series data mining, such as classification, clustering, pattern discovery and similarity search. Euclidean distance (Euc) and dynamic time warping (DTW) are two of the most popular and important distance functions [21,22]. They are often regarded as the true distance between any two time series [23]. With comparison between them, DTW is more robust than Euc [24,21]. The reason is that DTW not only deals with time series with unequal-length but also is insensitive to the abnormal points that exist in time series. In addition, DTW can match the similar trends at the different time points and obtain the minimal distance between two time series. However, the time and space complexity of DTW is the square of the lengths of the two compared series, that is $O(mn)$, where m and n are the lengths of the two compared series. The high cost causes much trouble in the field time series data mining. Specially, it is not suitable to directly apply to similarity search and indexing of long time series.

To search and index the similar objects in the time series dataset, the well-known method is the lower bound function to remove the dissimilar series before measuring time series using DTW. It means that the function should be lower bound on the true distance measured by DTW, which can avoid the false dismissals when similarity search and indexing. Now there are many lower bound functions used for univariate time series, such as *LB_Yi* [25], *LB_Kim* [26], *LB_Keogh* [27], *LB_Improved* [28] and *LB_ECorner* [29]. In addition, we discussed the relationships among the existed lower bound functions and proposed another two extended versions of *LB_Kim* and *LB_Keogh* respectively, that is *LB_NKim* and *LB_NKeogh* [30]. Theoretical proofs and experimental results show that both of the extended lower bound functions have better tightness than the old ones. However, due to the influence of the variable-based dimension, it is difficult to design an efficient and effective lower bound function for multivariate time series. Rath and Manmatha [31] presented lower bound function (*LB_MV*) for multivariate time series based on *LB_Keogh* [27]. We had designed some functions to fast search a similar object in the dataset through removing a large number of dissimilar objects. The functions in the previous work [32] are lower-bound on DTW based on absolute distance between two points that respectively derive from the two original multivariate time series.

The motivation of this work includes two aspects. One is that the variable-based dimensionality is reduced so that the traditional lower bound functions are valid for similarity search and indexing of multivariate time series. The other is that we theoretically prove and experimentally demonstrate that the proposed function is lower bound on DTW, and comparing to the previous methods, the tightness and the performance of similarity search are improved. To reduce the variable-based dimensionality, the elements of each point of multivariate time series are regarded as the members of each piecewise segment of a univariate time series. According to the principle of piecewise aggregate approximation, a center sequence is calculated to represent the original multivariate time series. Meanwhile, an extended lower bound function is regarded as the estimated distance between any two center sequences. In this work, some contributions can be obtained. Firstly, the reason why we use the center sequence to represent multivariate time series is explained. Secondly, the way dealing with unequal-length time series is taken into consideration so that the proposed lower bound function can measure the estimated distance between two different-length based sequences. Thirdly, we design some experimental approaches to discuss the performance of the existing methods and the proposed one.

The remainder of the paper is organized as follows. In Section 2, we discuss background and related work. In Section 3, the proposed methods are introduced. In Section 4, we arrange some experiments to evaluate the performance of the lower bound functions. In the last section we conclude our work and discuss the future work.

2. Background and related work

Piecewise aggregate approximation (PAA) [16] is a well known and popular method used to reduce the time based dimensionality of univariate time series. It improves the performance of time series similarity search and indexing [17,27]. Dynamic time warping (DTW) [21,22] is one of the most robust distance measurements in the field of time series data mining. DTW in its original form is too slow for most time series applications so that a technique is needed to speed it up, e.g. by indexing the time series data to find the best match without examining every candidate time series. Some lower