Neurocomputing 138 (2014) 189–198

Contents lists available at ScienceDirect

Neurocomputing

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

An improvement of symbolic aggregate approximation distance measure for time series

Youqiang Sun [a,b,*], Jiuyong Li [c], Jixue Liu [c], Bingyu Sun [a,b], Christopher Chow [d,e]

[a] School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui, China
[b] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, China
[c] School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, SA, Australia
[d] Australian Water Quality Centre, SA Water, Adelaide, SA, Australia
[e] SA Water Centre for Water Management and Reuse, University of South Australia, Adelaide, SA, Australia

ARTICLE INFO

Article history:
Received 25 September 2013
Received in revised form
23 January 2014
Accepted 24 January 2014
Communicated by D. Wang
Available online 18 February 2014

Keywords:
Time series
Trend distance
Symbolic Aggregate approXimation
Lower bound
Classification

ABSTRACT

Symbolic Aggregate approXimation (SAX) as a major symbolic representation has been widely used in many time series data mining applications. However, because a symbol is mapped from the average value of a segment, the SAX ignores important information in a segment, namely the trend of the value change in the segment. Such a miss may cause a wrong classification in some cases, since the SAX representation cannot distinguish different time series with similar average values but different trends. In this paper, we firstly design a measure to compute the distance of trends using the starting and the ending points of segments. Then we propose a modified distance measure by integrating the SAX distance with a weighted trend distance. We show that our distance measure has a tighter lower bound to the Euclidean distance than that of the original SAX. The experimental results on diverse time series data sets demonstrate that our proposed representation significantly outperforms the original SAX representation and an improved SAX representation for classification.

1. Introduction

Mining time series has attracted an increasing interest due to its wide applications in finance, industry, medicine, biology, and so on. There are a number of challenges in time series data mining, such as high dimensionality, high volumes, high feature correlation and large amount of noises. In order to reduce execution time and storage space, many high level representations or abstractions of the raw time series data have been proposed. The well-known representations include Discrete Fourier Transform (DFT) [1], Discrete Wavelet Transform (DWT) [2], Discrete Cosine Transform (DCT) [3], Singular Value Decomposition (SVD) [4], Piecewise Aggregate Approximation (PAA) [5] and Symbolic Aggregate approXimation (SAX) [6].

The SAX has become a major tool in time series data mining. The SAX discretizes time series and reduces dimensionality/numerosity of data. The distance in the SAX representation has a lower bound to the Euclidean distance. In other words, the error between the distance in the SAX representation and the Euclidean distance in the original data is bounded [7]. Therefore, the SAX representation speeds up the data mining process of time series data while maintaining the quality of the mining results. The SAX has been widely used for applications in various domains such as mobile data management [8], financial investment [9] and shape discovery [10].

The SAX representation has a major limitation. In the SAX representation, symbols are mapped from the average values of segments. The SAX representation does not consider the trends (or directions) in the segments. Different segments with similar average values may be mapped to the same symbols, and the SAX distance between them is 0. For example, in Fig. 1, time series (a) and (b) are different but their SAX representations are the same as 'feacdb'. This drawback causes misclassifications when using distance-based classifiers.

The ESAX representation overcomes the above limitation by tripling the dimensions of the original SAX [11]. To distinguish the two time series in Fig. 1, the ESAX representation adds additional symbols for the maximum and minimum points of a segment. The ESAX representations of time series (a) and (b) are 'efffe-caaaacffdbcbb' and 'effcefbaafcaadfbbc' respectively.

We propose to store one value along with a symbol in the SAX to improve the distance calculation of the SAX. Time series (a) and (b) in our representation are represented as '$_{0.2}f_{1.2}e_{-0.1}a_{-1.2}c_1d_{-0.2}b_{-0.3}$' and '$_{0.3}f_{-0.8}e_0a_{1.3}c_{-1.4}d_{0.4}b_{0.3}$' respectively. Note that we store one additional value for the last segment. For the same number of

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.01.045
0925-2312

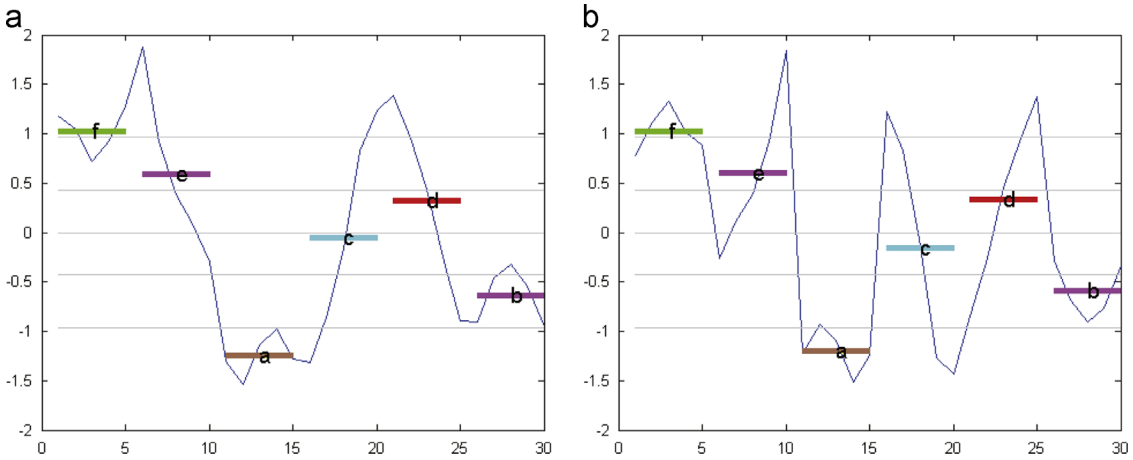

Fig. 1. (a) and (b) have the same SAX symbolic representation 'feacdb' in the same condition where the length of time series is 30, the number of segments is 6 and the size of symbols is 6. However, they have different time series. (a) Time series 1, (b) time series 2.

Table 1
A lookup table for breakpoints with the alphabet size from 3 to 10.

β_i	3	4	5	6	7	8	9	10
β_1	−0.43	−0.67	−0.84	−0.97	−1.07	−1.15	−1.22	−1.28
β_2	0.43	0	−0.25	−0.43	−0.57	−0.67	−0.76	−0.84
β_3	–	0.67	0.25	0	−0.18	−0.32	−0.43	−0.52
β_4	–	–	0.84	0.43	0.18	0	−0.14	−0.25
β_5	–	–	–	0.97	0.57	0.32	0.14	0
β_6	–	–	–	–	1.07	0.67	0.43	0.25
β_7	–	–	–	–	–	1.15	0.76	0.52
β_8	–	–	–	–	–	–	1.22	0.84
β_9	–	–	–	–	–	–	–	1.28

segments, our representation doubles the dimensions of the SAX representation. In contrast, the ESAX triples the dimensions of the SAX representation. Our presentation improves the precision of calculating the distances greatly over the SAX and the ESAX representations.

In this work, we have made three main contributions. Firstly, we present an intuitive trend distance measure on time series segments. Because of the approximately linear trend in a short segment, the average value of the segment and its starting and ending points help measure different trends. Our presentation captures the trends in time series better than the SAX and the ESAX representations. Secondly, we propose a distance measure of two time series by integrating the SAX distance with our weighted trend distance. Our improved distance measure not only keeps a lower-bound to the Euclidean distance, but also achieves a tighter lower bound than that of the original SAX distance. Thirdly, comprehensive experiments have been conducted to show that, in comparison with the SAX and the ESAX representations, our representation has improved the classification accuracy. In addition, for achieving the best classification accuracy, our representation has attained a similar dimensionality reduction to the SAX.

The remainder of this paper is organized as follows: Section 2 provides the background knowledge of the SAX. Section 3 reviews the related work. Section 4 introduces our improved distance measure and its lower bounding property. Section 5 presents experimental evaluation on several time series data sets. Finally, Section 6 concludes the paper and points out the future work.

2. Background

The SAX is the first symbol representation of time series with a dimensionality reduction and a lower bound of the Euclidean

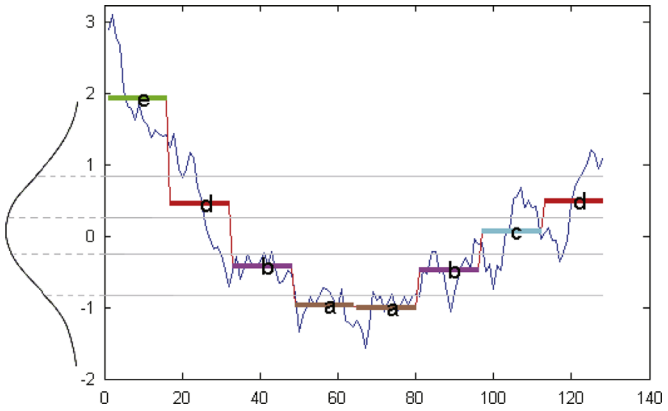

Fig. 2. A time series of length 128 is mapped to the word 'edbaabcd', where the number of segments is 8 and the size of alphabetic symbols is 5.

distance. For instance, to convert a time series sequence C of length n into w symbols, the SAX works as follows. Firstly, the time series is normalized. Secondly, the time series is divided into w equal-sized segments by Piecewise Aggregate Approximation (PAA) [5]. That is, $\overline{C}=\overline{c_1},\ldots,\overline{c_w}$, the ith element of $\overline{C}$ is the average of the ith segment and is calculated by the following equation:

$$\overline{c_i}=\frac{w}{n}\sum_{j=(n/w)(i+1)+1}^{(n/w)i} c_j, \tag{1}$$

where c_j is one point of time series C, j is from the starting point to the ending point for each segment. Next, the 'breakpoints' that divide the distribution space into α equiprobable regions are determined. Breakpoints are a sorted list of numbers $B=\beta_1,\ldots,\beta_{\alpha-1}$ such that the area under a $N(0,1)$ Gaussian curve from β_i to $\beta_{i+1}=1/\alpha$. A lookup table that contains the breakpoints is shown in Table 1.

Finally, each region is assigned a symbol using the determined breakpoints. The PAA coefficients are mapped to the symbols corresponding to the regions in which they reside. The symbols are assigned in a bottom-up fashion, i.e. the PAA coefficient that falls in the lowest region is converted to 'a', in the one above to 'b', and so forth. These symbols for approximately representing a time series are called a 'word'. Fig. 2 illustrates a sample time series converted into the SAX word representation.

For the utilization of the SAX in classic data mining tasks, the distance measure was proposed. Given two original time series Q and C with the same length n, $\widehat{Q}$ and $\widehat{C}$ are their SAX