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 'efffecaaaacffdbcbb' 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 © 2014 Elsevier B.V. All rights reserved.

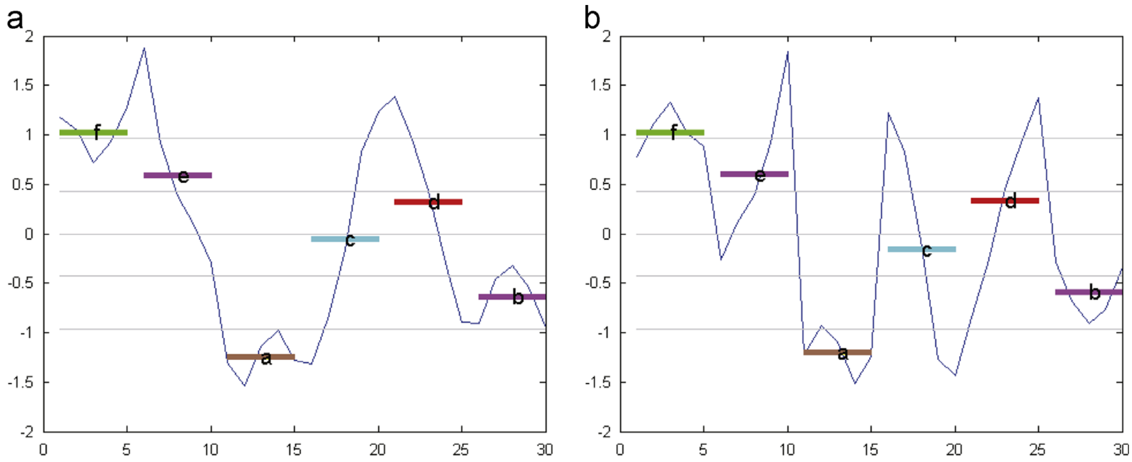


Fig. 1. (a) and (b) have the same SAX symbolic representation 'feacdb' in the same condition where the length of time series is 30, the number of segments is 6 and the size of symbols is 6. However, they have different time series. (a) Time series 1, (b) time series 2.

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 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, $\overline{C} = \overline{c_1}, \dots, \overline{c_w}$, the ith element of $\overline{C}$ is the average of the ith segment and is calculated by the following equation:

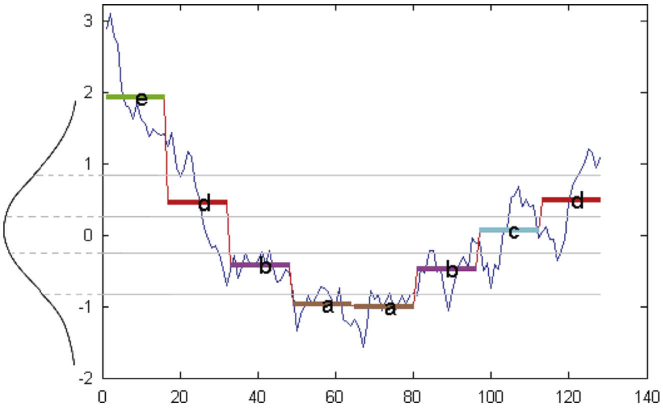


Fig. 2. A time series of length 128 is mapped to the word 'edbaabcd', where the number of segments is 8 and the size of alphabetic symbols is 5.

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

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 $B = \beta_1, \dots, \beta_{\alpha-1}$ 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 representations respectively with the word size w, their SAX distance function MINDIST is defined as follows:

$$MINDIST(\widehat{Q},\widehat{C}) = \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(dist(\widehat{q},\widehat{c}))^2}, \tag{2}$$

where the $dist()$ function can be implemented using the lookup table as illustrated in Table 1, and is calculated by the following equation:

$$dist(\widehat{q},\widehat{c}) = \begin{cases} 0 & \text{if } |\widehat{q}-\widehat{c}| \le 1 \\ \beta_{max(\widehat{q},\widehat{c})-1} - \beta_{min(\widehat{q},\widehat{c})} & \text{otherwise.} \end{cases} \tag{3}$$

3. Related work

The SAX has a generality of the original presentation and works well in many problems. There have been some improvements of the SAX representation recently.

Some methods improve the SAX by adaptivity choosing the segments. The method in [12] uses the discretization of Adaptive Piecewise Constant Approximation (APCA) [13] to replace the PAA [5] in the SAX. The method in [14] makes use of an adaptive symbolic representation with the adaptive vector of 'breakpoints'. While the two methods above reduce the reconstruction error on some data sets, they still use the average values as the basis for approximation (the latter method uses the same distance measure as the SAX) and do not consider the differences of value changes between segments.

Some methods improve the SAX by enriching the representation of each segment. The method in [11] uses three symbols, instead of a single symbol, to represent a segment in time series. This method triples the dimensions of the SAX and the high dimensionality increases the computational complexity. The method in [15] utilizes a symbolic representation based on the summary of statistics of segments. The method considers the symbols as a vector, including the discretized mean and variance values as two elements. However, it is may be inappropriate to transform the variances to symbols using the same breakpoints for the transformation of the mean values to symbols.

Trend estimation of time series is an important research direction. Many methods have been proposed to represent and measure trends. It is a common way to fit a line and then characterize the trend of the line. The least square fitting and Piecewise Linear Approximation (PLA) [16] are two normally used fitting methods [17,18], as well as their extensions [19,20]. The least square method chooses the best fitting line with the minimal sum of the squared errors from a given segment of data. The PLA splits the series into many representative segments, and then fits a polynomial line model for time series. The trends are categorized as two (up and down) or three (up, down and level) directions using the slopes of the lines. The distance is measured by some artificial instantiations, such as -1, 0 and 1. These approaches are simple and easily understandable. However, they have several drawbacks. Firstly, it is not possible to capture the difference between the segments with the same direction. For example, distance between two up trends is zero. Secondly, there is not an appropriate measure to characterize the difference of segments with different directions, such as the distance between 'up' and 'down' trends. Some other statistic approaches have been proposed in [21–23], but their models are complex.

4. SAX-TD: improved SAX based on trend distance

As we reviewed above, the time series segments are mapped to symbols by their average values when using the SAX. This representation is imprecise when the trends of the segments are different but with similar average values. Fig. 3 lists several typical segments with the same average: (a) level and slight up, (b) obvious up, (c) down, up and then down, (d) level and slight down, (e) obvious down, (f) up, down and then up.

Trends are important characteristics of time series, and they are crucial for the analysis of similarity and classification of time series [24]. Although there are no common definition of trend and a measurement of trend distance in time series, the starting and

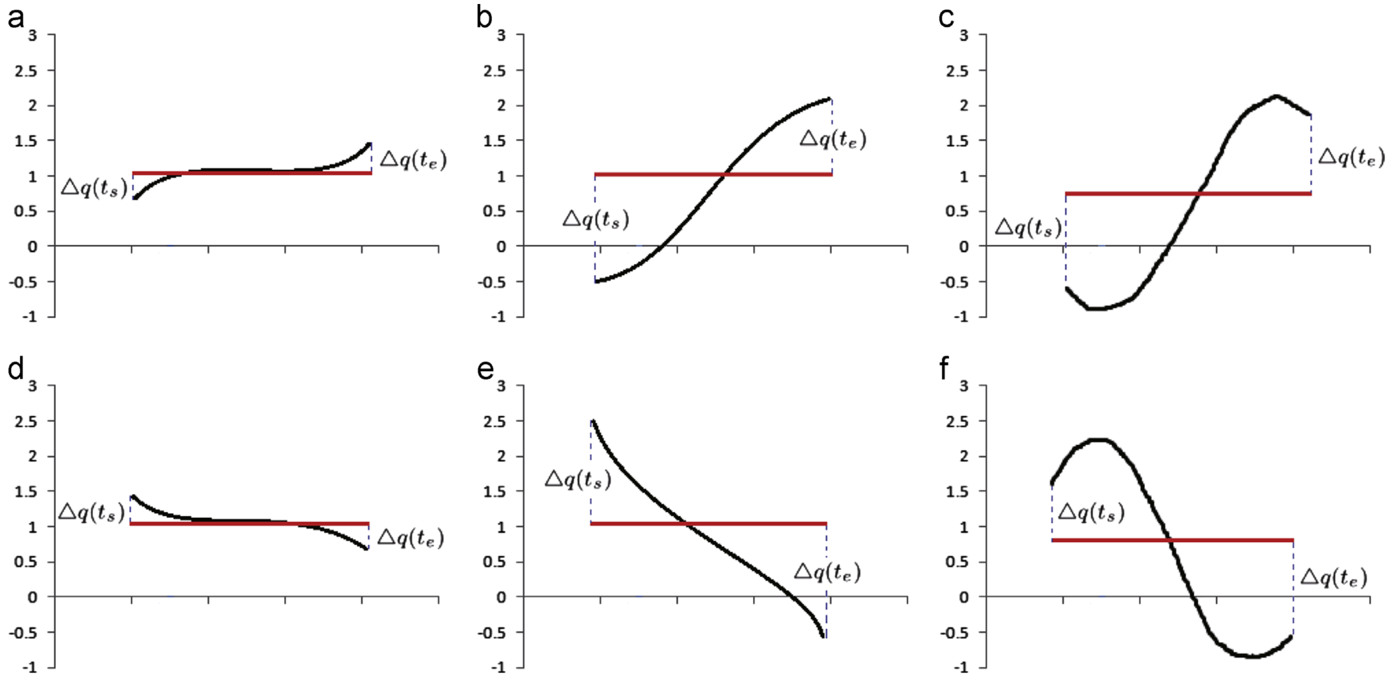


Fig. 3. Several typical segments with the same average value but different trends.

the>ending points are important in segment trend estimation. For example, a trend is up when the value of the ending point is significantly larger than the value of the starting point, while the trend is down when the value of the ending point is significantly smaller than the value of the starting point. It is difficult to qualitatively define a trend, such as the definitions of ‘significant up’ and ‘significant down’, ‘significant down’ and ‘slight down’. However, if the trend information of a segment is not utilized, the representations of a time series containing many segments are rough.

In this paper, we do not use symbols to capture the trends of time series, but quantitatively measure the trends by calculating their distance, called trend distance. Because the divided segments are short, the trend in a segment approximates a linear relationship in most of cases. Therefore, the starting and the ending points approximatively determine a trend. When more data points are used, the trend will be represented more precisely. However, the dimensions of the representation will be increased significantly. When we use the starting and the ending points, because of the continuity of time series data, only one extra dimension is added to one segment.

4.1. Distance measures

We use the difference of changes between the average and the values of starting and ending points to quantify the distance of segments. For an illustration, given two time series segments q and c, let $\Delta q(t_s)$ and $\Delta c(t_s)$ represent the changes between the average and the starting point's value, and $\Delta q(t_e)$ and $\Delta c(t_e)$ represent the changes between the average and the ending point's value. $|\Delta q(t_s)-\Delta c(t_s)|$ and $|\Delta q(t_e)-\Delta c(t_e)|$ indicate the difference between the trends of the two segments, and is called the divergence of two trends. For example in Fig. 3, the divergence of (b) and (e) is larger than that of (b) and (c), that means the trends of (b) and (c) are more similar to the trends of (b) and (e); the trends of (a) and (d) are more similar to the trends of (a) and (b) although (a) has slight up trend.

Based on the discussion above, we define the trend distance between two segments.

Definition 1 (*trend distance*). Given two time series segments q and c with the same length, the trend distance $td(q,c)$ between them is defined as follows:

$$td(q,c)=\sqrt{(\Delta q(t_s)-\Delta c(t_s))^2+(\Delta q(t_e)-\Delta c(t_e))^2}, \tag{4}$$

where t_s and t_e are the starting and ending time points of q and c, respectively. $\Delta q(t)$ ($\Delta c(t)$) means the difference between $q(t)$ ($c(t)$) and its average value, and can be calculated by the following:

$$\Delta q(t)=q(t)-\overline{q}. \tag{5}$$

The $\Delta c(t)$ is calculated in a similar way. Note that in Eq. (5), $\overline{q}$ is a known value obtained by the PAA discretization, we just need to calculate the $\Delta q(t_s)$ and $\Delta q(t_e)$. We call $\Delta q(t_s)$ and $\Delta q(t_e)$ as the trend variations of a segment.

We incorporate the trend variations into the SAX representation. Because the continuity of time series data, the ending point of a segment is the starting point of the following segment. One segment needs only one trend variation (except the last segment). For an illustration, given two time series Q and C with the length of n, the representations with w words of them are

$$Q:{}_{\Delta q(1)}\widehat{q}_1\,{}_{\Delta q(2)}\widehat{q}_2\,{}_{\Delta q(3)}\cdots{}_{\Delta q(w)}\widehat{q}_w\,{}_{\Delta q(w+1)},$$
$$C:{}_{\Delta c(1)}\widehat{c_1}\,{}_{\Delta c(2)}\widehat{c_2}\,{}_{\Delta c(3)}\cdots{}_{\Delta c(w)}\widehat{c_w}\,{}_{\Delta c(w+1)},$$

where $\widehat{q}_1$, $\widehat{q}_2 \ldots \widehat{q}_w$ are the symbolic representations by the SAX, $\Delta q(1)$, $\Delta q(2) \ldots \Delta q(w)$ are the trend variations, and $\Delta q(w+1)$ is the change of the last point. Compared to the original SAX, our representation adds $w+1$ dimensions for trend variations.

We define the distance between two time series based on the trend distance as follows:

$$TDIST(\widehat{Q},\widehat{C})=\sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}\left((dist(\widehat{q}_i,\widehat{c}_i))^2+\frac{w}{n}(td(q_i,c_i))^2\right)}, \tag{6}$$

where $\widehat{Q}$ and $\widehat{C}$ are the new representations of time series Q and C with the same length n. w is the number of segments (or words), $\widehat{q}_i$ and $\widehat{c}_i$ are the symbolic representations of segments q_i and c_i, respectively.

From Eq. (6), we see that the influence of the trend distance on the overall distance is weighted by the ratio of dimensionality reduction w/n. w/n is larger when there are more divided segments and each segment is shorter. w/n is smaller when there are fewer divided segments and each segment is longer. This is because in a short segment, the trend is likely to be linear and can be largely captured by two points and hence the weight for the trend distance is high. When the segment is long, the trend is complex, two points are unlikely to capture the trend and hence the weight of the trend distance is low.

We use an example to show the difference of the SAX distance and our SAX-TD distance. Two time series are given from data set CBF [25], the lengths of both are 128. The Euclidean distance between them is 11.88. We show the distances calculated by the SAX and the SAX-TD while w are assigned from 2 up to 64 (2 to $n/2$, and we double the value each time) in Table 2. The distances of the SAX-TD are closer to the true distance than that of the SAX.

4.2. Lower bound

One of the most important characteristics of the SAX is that it provides a lower bounding distance measure. Lower bound is very useful for controlling errors and speeding up the computation. Below, we will show that our proposed distance also lower bounds the Euclidean distance.

According to [5,7], the authors have proved that the PAA distance lower bounds the Euclidean distance as follows:

$$\sqrt{\sum_{i=1}^{n}(q_i-c_i)^2}\geq\sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(\overline{q_i},\overline{c_i})^2}. \tag{7}$$

For proving the TDIST also lower bounds the Euclidean distance, we repeat some of the proofs here. Let $\overline{Q}$ and $\overline{C}$ be the means of time series Q and C respectively. We first consider only the single-frame case (i.e. $w=1$), Ineq. (7) can be rewritten as follows:

$$\sqrt{\sum_{i=1}^{n}(q_i-c_i)^2}\geq\sqrt{n}\sqrt{(\overline{Q}-\overline{C})^2}. \tag{8}$$

Squaring both sides we get

$$\sum_{i=1}^{n}(q_i-c_i)^2\geq n(\overline{Q}-\overline{C})^2. \tag{9}$$

Recall that $\overline{Q}$ is the average of the time series, so q_i can be represented in terms of $q_i=\overline{Q}-\Delta q_i$. The same applies to each

Table 2
The distances of the SAX and the SAX-TD with different w. The Euclidean distance is 11.88.

w	2	4	8	16	32	64
SAX	0	4.72	4.54	6.21	6.60	7.24
SAX-TD	2.84	6.30	6.44	7.94	8.49	9.12

point c_i in C. Thus, Ineq. (9) can be rewritten as follows:

$$\sum_{i=1}^{n}((\overline{Q}-\Delta q_i)-(\overline{C}-\Delta c_i))^2 \ge n(\overline{Q}-\overline{C})^2. \tag{10}$$

Rearranging the left-hand side, we get

$$\sum_{i=1}^{n}((\overline{Q}-\overline{C})-(\Delta q_i-\Delta c_i))^2 \ge n(\overline{Q}-\overline{C})^2. \tag{11}$$

We can expand and then rewrite Ineq. (11) by the distributive law as follows:

$$\sum_{i=1}^{n}(\overline{Q}-\overline{C})^2+\sum_{i=1}^{n}(\Delta q_i-\Delta c_i)^2-\sum_{i=1}^{n}2(\overline{Q}-\overline{C})(\Delta q_i-\Delta c_i) \ge n(\overline{Q}-\overline{C})^2. \tag{12}$$

Note that $\overline{Q}$ and $\overline{C}$ are independent to n, Ineq. (12) can be transformed as follows:

$$n(\overline{Q}-\overline{C})^2+\sum_{i=1}^{n}(\Delta q_i-\Delta c_i)^2-2(\overline{Q}-\overline{C})\sum_{i=1}^{n}(\Delta q_i-\Delta c_i) \ge n(\overline{Q}-\overline{C})^2. \tag{13}$$

It was also proved that $\sum_{i=1}^{n}(\Delta q_i-\Delta c_i)=0$ in [5]. Therefore, after substituting 0 into the third term on the left-hand side, Ineq. (13) becomes

$$n(\overline{Q}-\overline{C})^2+\sum_{i=1}^{n}(\Delta q_i-\Delta c_i)^2 \ge n(\overline{Q}-\overline{C})^2. \tag{14}$$

Because $\sum_{i=1}^{n}(\Delta q_i-\Delta c_i)^2 \ge 0$, Ineq. (14) holds true. Recall the definition in (4), $(td(q,c))^2=(\Delta q(t_s)-\Delta c(t_s))^2+(\Delta q(t_e)-\Delta c(t_e))^2$, we can obtain an inequality as follows ($i=1$ is the starting point and $i=n$ is the ending point):

$$\sum_{i=1}^{n}(\Delta q_i-\Delta c_i)^2 \ge (\Delta q_1-\Delta c_1)^2+(\Delta q_n-\Delta c_n)^2. \tag{15}$$

Substituting Ineq. (15) into Ineq. (14), we get

$$n(\overline{Q}-\overline{C})^2+\sum_{i=1}^{n}(\Delta q_i-\Delta c_i)^2 \ge n(\overline{Q}-\overline{C})^2+(td(q_i,c_i))^2. \tag{16}$$

According to [7], the MINDIST lower bounds the PAA distance, that is

$$n(\overline{Q}-\overline{C})^2 \ge n(dist(\widehat{Q},\widehat{C})^2 \tag{17}$$

where $\widehat{Q}$ and $\widehat{C}$ are symbolic representations of Q and C in the original SAX, respectively. By transitivity, the following inequality is true

$$n(\overline{Q}-\overline{C})^2+\sum_{i=1}^{n}(\Delta q_i-\Delta c_i)^2 \ge n(dist(\widehat{Q},\widehat{C}))^2+(td(q_i,c_i))^2. \tag{18}$$

Recall Ineq. (9), this means

$$\sum_{i=1}^{n}(q_i-c_i)^2 \ge n\left((dist(\widehat{Q},\widehat{C}))^2+\frac{1}{n}(td(q_i,c_i))^2\right). \tag{19}$$

N frames can be obtained by applying the single-frame proof on every frame, that is

$$\sqrt{\sum_{i=1}^{n}(q_i-c_i)^2} \ge \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}\left((dist(\widehat{q}_i,\widehat{c}_i))^2+\frac{w}{n}(td(q_i,c_i))^2\right)}. \tag{20}$$

The right-hand side of the above inequality is $TDIST(Q,C)$ and the left-hand side is the Euclidean distance between Q and C. Therefore, the TDIST distance lower bounds the Euclidean distance.

The quality of a lower bounding distance is usually measured by the tightness of lower bounding (TLB).

$$TLB=\frac{Lower\ Bounding\ Distance\ (Q,C)}{Euclidean\ Distance\ (Q,C)}.$$

The value of TLB is in the range [0,1]. The larger the TLB value, the better the quality. Recall the distance measure in Eq. (6), we can obtain that $TLB(TDIST) \ge TLB(MINDIST)$, which means the SAX-TD distance has a tighter lower bound than the original SAX distance.

In conclusion, our improved SAX-TD not only holds the lower bounding property of the original SAX, but also achieves a tighter lower bound.

5. Experimental validation

In this section, we will present the results of our experimental validation. Firstly we introduce the data sets used, the comparison methods and parameter settings. Then we evaluate the performances of the proposed method in terms of classification error rate, dimensionality reduction and efficiency.

5.1. Data sets

We performed the experiments on 20 diverse time series data sets, which are provided by the UCR Time Series repository [25]. Some summary statistics of the data sets are given in Table 3. Each data set is divided into a training set and a testing set. The data sets contain classes ranging from 2 to 50, are of size from dozens to thousands, and have the lengths of time series varying from 60 to 637. In addition, the types of the data sets are also diverse, including synthetic, real (recorded from some processes) and shape (extracted by processing some shapes).

Table 3
The description of data sets used.

No.	Data set name	# Classes	Size of training set	Size of testing set	Length of time series	Type
1	Synthetic Control	6	300	300	60	Synthetic
2	Gun-Point	2	50	150	150	Real
3	CBF	3	30	900	128	Synthetic
4	Face (all)	14	560	1690	131	Shape
5	OSU Leaf	6	200	242	427	Shape
6	Swedish Leaf	15	500	625	128	Shape
7	50 Words	50	450	455	270	Real
8	Trace	4	100	100	275	Synthetic
9	Two Patterns	4	1000	4000	128	Synthetic
10	Wafer	2	1000	6174	152	Real
11	Face (four)	4	24	88	350	Shape
12	Lightning-2	2	60	61	637	Real
13	Lightning-7	7	70	73	319	Real
14	ECG	2	100	100	96	Real
15	Adiac	37	390	391	176	Shape
16	Yoga	2	300	3000	426	Shape
17	Fish	7	175	175	463	Shape
18	Beef	5	30	30	470	Real
19	Coffee	2	28	28	286	Real
20	Olive Oil	4	30	30	570	Real

5.2. Comparison methods and parameter settings

Since our method aims to improve the SAX by modifying the distance measure, we do the evaluation on the classification task, of which the accuracy is determined by the distance measure. We compare the accuracies with the classic Euclidean distance and the original SAX. To the best of our knowledge, there is no other research improving the SAX distance by measuring trends. We choose an extension of SAX called as Extended SAX (ESAX) [11] to compare with. The ESAX adds two additional symbols for the maximum and minimum values in a segment, but uses the same distance measure as the SAX after mapping. For example in Fig. 3, let us assume that the SAX words of sub-figure (b) and (e) are 'b', the ESAX representations of them are 'abc' and 'cba', respectively.

Table 4
1-NN classification error rates of EU (Euclidean distance); 1-NN best classification error rates, w lengths, α numbers and dimensionality reduction ratios of the SAX, ESAX and SAX-TD on 20 data sets. The lowest error rates are highlighted in bold.

Data set no.	EU error	SAX error	SAX w	SAX α	SAX ratio	ESAX error	ESAX w	ESAX α	ESAX ratio	SAX-TD error	SAX -TD w	SAX -TD α	SAX-TD ratio
1	0.12	**0.02**	16	10	0.27	0.157	16	10	0.80	0.077	8	10	0.28
2	0.087	0.18	64	10	0.43	0.193	64	10	1.28	**0.073**	4	3	0.06
3	0.148	0.104	32	10	0.25	0.138	64	10	1.50	**0.088**	4	10	0.07
4	0.286	0.33	64	10	0.49	0.333	64	9	1.47	**0.215**	16	8	0.25
5	0.483	0.467	128	10	0.30	**0.446**	16	9	0.11	**0.446**	32	7	0.15
6	**0.211**	0.483	32	10	0.25	0.4	64	10	1.50	0.213	16	7	0.26
7	0.369	0.341	128	10	0.47	**0.321**	32	10	0.36	0.338	128	9	0.95
8	0.24	0.46	128	10	0.47	**0.16**	4	10	0.04	0.21	128	3	0.93
9	0.093	0.081	32	10	0.25	0.129	64	10	1.50	**0.071**	16	10	0.26
10	0.0045	0.0034	64	10	0.42	**0.0031**	64	9	1.26	0.0042	32	8	0.43
11	0.216	**0.17**	128	10	0.37	0.182	128	7	1.10	0.181	32	9	0.18
12	0.246	0.213	256	10	0.40	**0.18**	32	7	0.15	0.229	8	9	0.02
13	0.425	0.397	128	10	0.40	0.37	128	8	1.20	**0.329**	16	10	0.10
14	0.12	0.12	32	10	0.33	**0.09**	32	10	1.00	**0.09**	16	5	0.34
15	0.389	0.89	64	10	0.36	0.89	64	10	1.09	**0.273**	32	9	0.37
16	**0.17**	0.195	128	10	0.30	0.2	128	10	0.90	0.179	128	10	0.60
17	0.217	0.474	128	10	0.28	0.469	128	10	0.83	**0.154**	64	9	0.28
18	0.467	0.567	128	10	0.27	0.533	32	9	0.20	**0.2**	64	9	0.27
19	0.25	0.464	128	10	0.45	0.179	4	5	0.04	**0**	8	3	0.06
20	0.133	0.833	256	10	0.45	0.833	2	3	0.01	**0.067**	64	3	0.22
Average	0.234	0.340	–	–	0.36	0.310	–	–	0.82	**0.172**	–	–	**0.31**

Table 5
The sign test results of the SAX-TD vs. other methods. A p-value less than or equal to 0.05 indicates a significant improvement.

Methods	n_+	n_-	n_0	p-value
SAX-TD vs. Euclidean	18	2	0	$p<0.01$
SAX-TD vs. SAX	16	4	0	$0.01<p<0.05$
SAX-TD vs. ESAX	14	4	2	$p=0.05$

Thus the distance calculated by the ESAX is more accurate than that calculated by the SAX.

To compare the classification accuracy, we conduct the experiments using the 1 Nearest Neighbor (1-NN) classifier. The main advantage is that the underlying distance metric is critical to the performance of 1-NN classifier, hence, the accuracy of the 1-NN classifier directly reflects the effectiveness of a distance measure. Furthermore, the 1-NN classifier is parameter free, allowing direct comparisons of different measures.

To obtain the best accuracy for each method, we use the testing data to search for the best parameters w and α. For a given time series of length n, w and α are picked using the following criteria (to make the comparison fair, the criteria are the same as those in [7]):

- For w, we search for the value from 2 up to $n/2$, and double the value of w each time.
- For α, we search for the value from 3 up to 10.
- If two sets of parameter settings produce the same classification error rate, we choose the smaller parameters.

After obtaining the values of parameters, we will measure the dimensionality reduction ratios as follows:

$$Dimensionality\ Reduction\ Ratio = \frac{Number\ of\ the\ reduced\ data\ points}{Number\ of\ the\ original\ data\ points}.$$

The dimensionality reduction ratio of the SAX is w/n, and the dimensionality reduction ratios of the ESAX and the SAX-TD are $3*w/n$ and $(2*w+1)/n$ respectively.

5.3. Results

The overall classification results are listed in Table 4, where entries with the lowest classification error rates are highlighted. SAX-TD has the lowest error in the most of the data sets (12/20), followed by the ESAX (6/20).[1] We use the sign test to test the significance of our method against other methods. The sign test results are displayed in Table 5, where n_+, n_- and n_0 denote the numbers of data sets where the error rates of the SAX-TD are lower, larger than and equal to those of another method respectively. The p-values (the smaller a p-value, the more significant the improvement) demonstrate that our distance measure achieves a significant improvement over the other three methods on classification accuracy. On average, SAX-TD reduces the classification error by almost a half from the original SAX, with a slight decrease of the dimensionality reduction ratio due to the smaller parameter w used in the SAX-TD than the others.

To show the performance of our method in comparison with other methods using different parameters, we run the experiments on data sets Gun-Point and Yoga. Specifically, on Gun-Point, we firstly compare the classification error rates with different w while α is fixed at 3, and then with different α while w is fixed at 4 (to illustrate the classification error rates using small parameters); on Yoga, w varies while α is fixed at 10, and then α varies while w is fixed at 128 (to illustrate the classification error rates using large parameters). The comparison lines are shown in Fig. 4. SAX-TD has lower error rates than the other two methods when the

[1] These numbers include two cases that the ESAX and the SAX-TD have the same error rate.

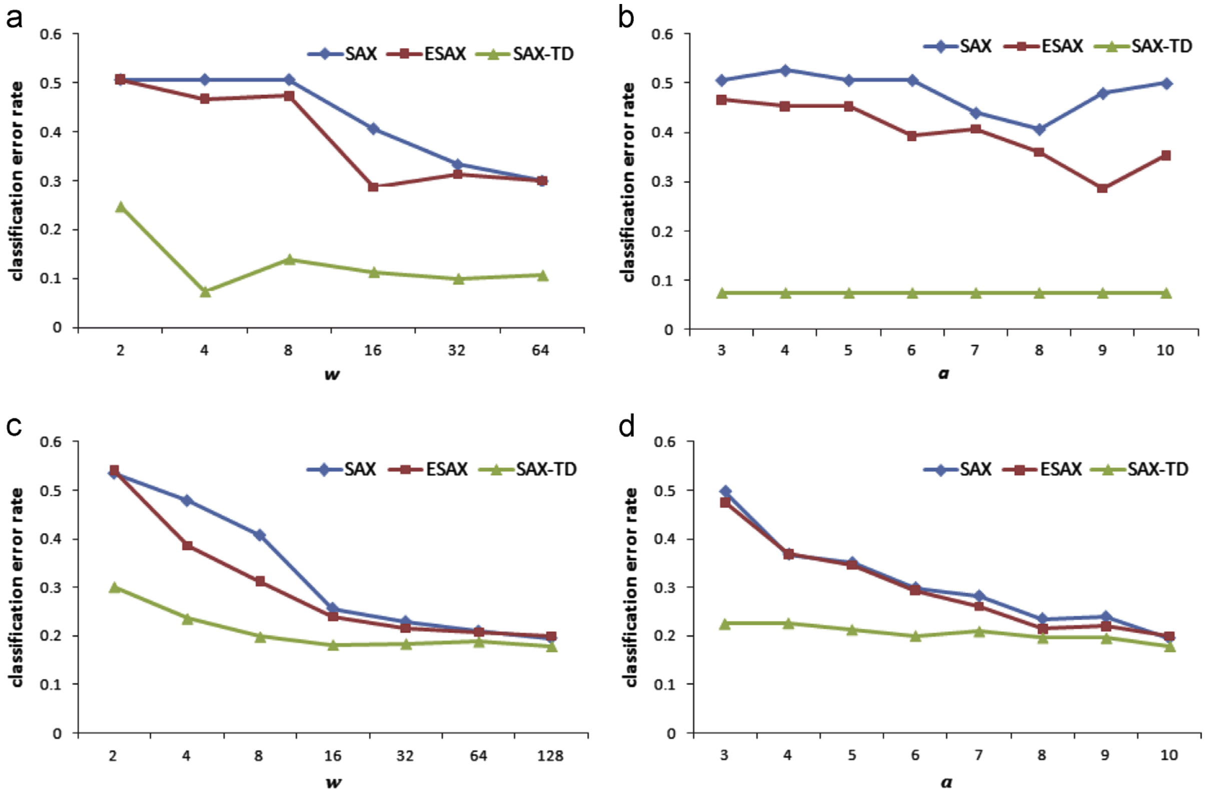


Fig. 4. The classification error rates of the SAX, ESAX and SAX-TD with different parameters w and α. (a) On Gun-Point, w varies while α is fixed at 3, (b) on Gun-Point, α varies while w is fixed at 4, (c) on Yoga, w varies while α is fixed at 10, (d) on Yoga, α varies while w is fixed at 128.

parameters are small and large. The superiority of the SAX-TD is more significant when the parameters are small. In addition, unlike the SAX and the ESAX, our method is not very sensitive to the size of α. These demonstrate that our method can achieve high accuracy with low parameter values.

To provide an illustration of the performance of the different measures compared in Table 4, we use scatter plots for pair-wise comparisons. In a scatter plot, the error rates of two measures under comparison are used as the x and y coordinates of a dot, where a dot represents a particular data set [26]. When a dot falls within a region, the corresponding method in the region performs better than the other method. In addition, the further a dot is from the diagonal line, the greater the margin of an accuracy improvement. The region with more dots indicates a better method than the other.

In the following, we explain the results in Fig. 5.

First, we review that the Euclidean distance versus the SAX distance in Fig. 5(a). Although the number of data points (data sets) in two regions is similar, the errors on some data sets by the Euclidean distance are much smaller than the errors by the SAX. The SAX has very high classification error rates on some data sets such as Adiac and Olive Oil (0.89 and 0.833). Therefore, the SAX distance is not superior over the Euclidean distance.

Secondly, we illustrate the performances of our distance measure against the Euclidean distance, the SAX distance and the ESAX distance in Fig. 5(b), (c) and (d) respectively. Our method outperforms the other three methods by a large margin, both in the number of points and the distance of these points from the diagonals. From these figures, we can see that most of the points are far away from the diagonals, which indicates our method has much lower error rates on the most of the data sets. For example, on Coffer and Beef data sets, the error rates of our method are 0 and 0.2 respectively, but the error rates of the Euclidean distance are 0.25 and 0.467 respectively. On Adiac and Olive Oil data sets, the error rates of our method are 0.273 and 0.067 respectively, but both the error rates of the SAX and ESAX are 0.89 and 0.833 respectively.

5.4. Dimensionality reduction and efficiency

Since one major advantage of the SAX representation is its dimensionality or numerosity reduction, we shall compare the dimensionality reduction of our method with the SAX and the ESAX. The dimensionality reduction ratios are calculated using the w when the three methods achieve their smallest classification error rates on each data set. The results are shown in Fig. 6. The SAX-TD is competitive with the SAX on dimensionality reduction. For each segment, the SAX-TD representation uses more values than the SAX but fewer symbols than the ESAX. However, to achieve the lowest classification error rate, the SAX-TD needs a smaller number of words than the SAX and the ESAX do. For example, using the SAX, the values of w on data sets Gun-Point, CBF, Lightning-2 and Coffee are 64, 32, 256 and 128 respectively. When using the SAX-TD, the values of w on them are 4, 4, 8 and 8 respectively. On average, the dimensionality reduction ratio by the SAX-TD (0.31) is similar to that by the SAX (0.36), but much lower than that by the ESAX (0.82).

Finally, we compare the computation time of the SAX, ESAX and SAX-TD. The experimental environment is a machine with 2×2.53 GHz processors and 4 GB RAM running 32-bit Windows Operating System. Four data sets are used to show the running time with different w: Synthetic Control, ECG, CBF and Yoga. The maximum w values of the data sets are 16, 32, 64 and 128 respectively. The α is fixed[2] at the maximum value, i.e. 10. The

[2] Since the efficiency is mainly determined by the value of parameter w, we just compare the computation time with different w while α is fixed.

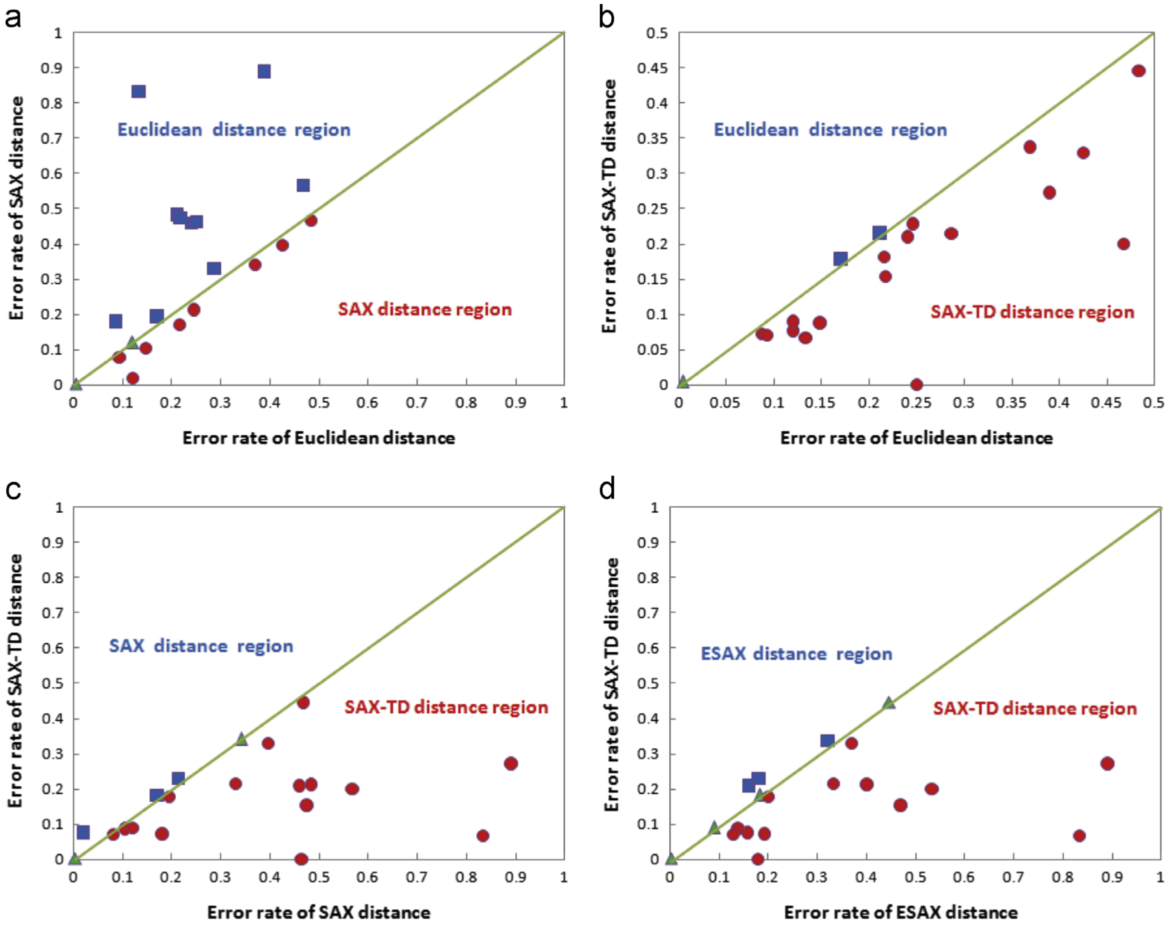


Fig. 5. A pairwise comparison of classification error rates for the Euclidean distance, the SAX distance, the ESAX distance and the SAX-TD distance on 20 data sets. The round dots indicate the values that are better in the lower triangle region, the square dots indicate the values that are better in the upper triangle region and the triangular ones indicate the values in both regions that are similar or equal. (a) The Euclidean distance vs. the SAX distance, (b) the Euclidean distance vs. the SAX-TD distance, (c) the SAX distance vs. the SAX-TD distance, (d) the ESAX distance vs. the SAX-TD distance.

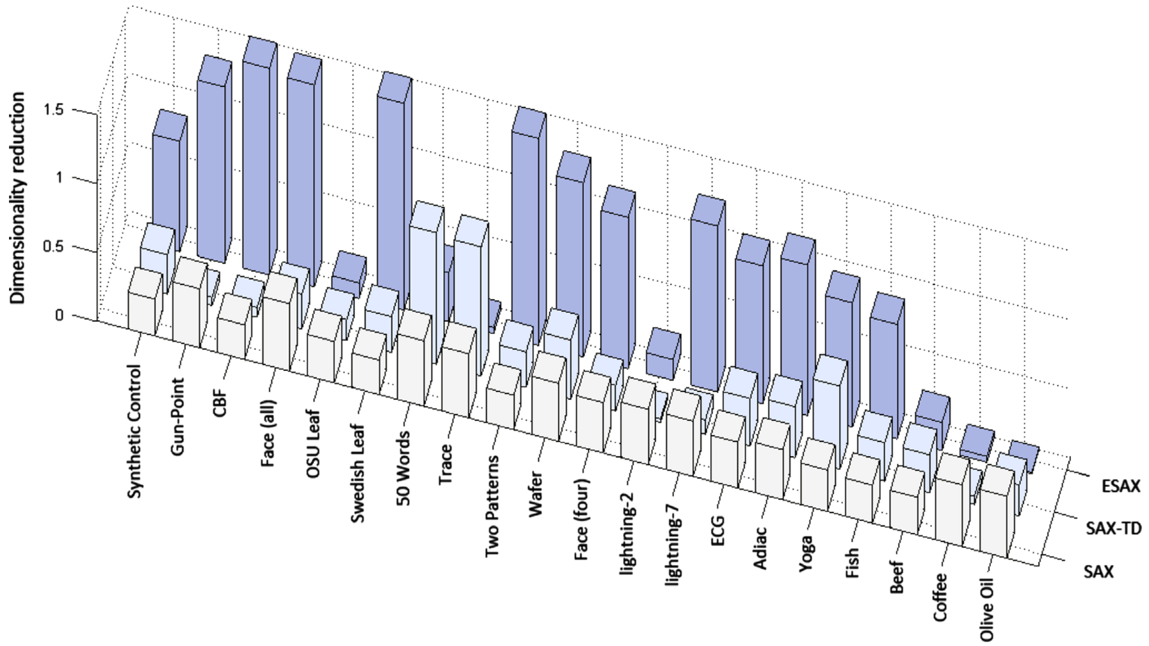


Fig. 6. Dimensionality reduction ratios of the SAX, the ESAX and the SAX-TD on 20 data sets with their smallest error rates.

results are shown in Fig. 7. Note that, the running time includes the transformation time (mapping values into words) and classification time (training and testing). We see that the running time increases with the increase of w. The SAX and the SAX-TD take similar amount of time while the ESAX takes more time than both especially when w becomes larger. Since the SAX-TD needs smaller

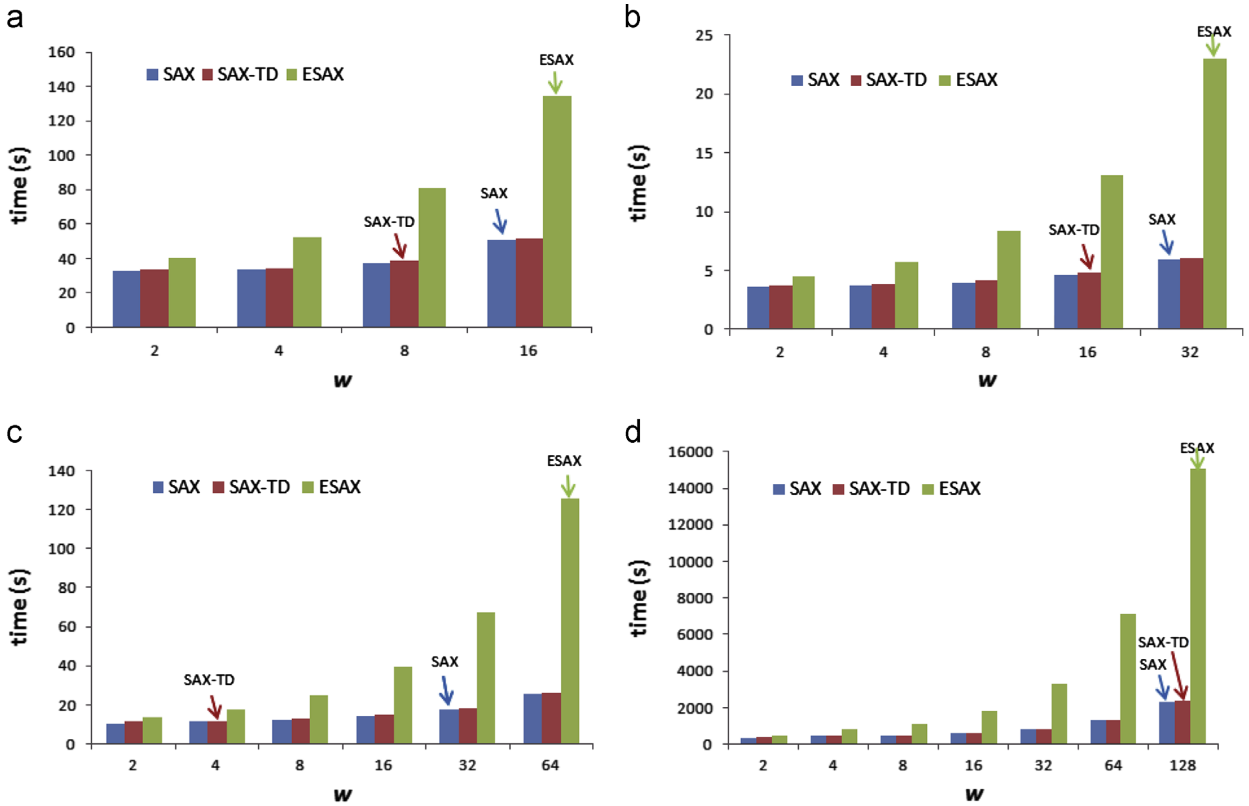


Fig. 7. The running time of different methods with different values of *w* on data sets Synthetic Control, ECG, CBF and Yoga. We marked the *w* used in different methods when they achieve the best classification accuracy. (a) Synthetic Control running time ($w \leq 16$), (b) ECG running time ($w \leq 32$), (c) CBF running time ($w \leq 64$), (d) Yoga running time ($w \leq 128$).

parameter *w* for achieving the best classification accuracy in most cases, the computation time of SAX-TD is shorter than that of the SAX and ESAX in many data sets.

6. Conclusions and future work

We have proposed an improved symbolic aggregate approximation distance measure for time series. We firstly define a trend distance using the divergences between the starting and ending points and the average. We then modify the original SAX distance measure by integrating a weighted trend distance. The new distance measure keeps the important property that lower bounds the Euclidean distance. Furthermore, the lower bound of our proposed measure is tighter than that of the original SAX. According to the experimental results on diverse data sets, our improved measure decreases the classification error rate significantly and needs a smaller number of words and alphabetic symbols for achieving the best classification accuracy than the SAX does. Our improved method has similar capability of dimensionality reduction and has similar efficiency as the SAX.

For the future work, we intend to extend the method to other data mining tasks such as clustering, anomaly detection and motif discovery. The proposed method may be utilized in improving the indexable Symbolic Aggregate approXimation (*i* SAX) [27] for terabyte sized time series.

Acknowledgments

This work has been partially supported by Australian Research Council Discovery grant DP130104090 and the National Natural Science Foundation of China (Nos. 41101516 and 31171456).

References

[1] C. Faloutsos, M. Ranganathan, Y. Manolopoulos, Fast subsequence matching in time-series databases, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1994, pp. 419–429.
[2] K. Chan, A.W. Fu, Efficient time series matching by wavelets, in: Proceedings of the IEEE International Conference on Data Engineering, 1999, pp. 126–133.
[3] F. Korn, H.V. Jagadish, C. Faloutsos, Efficiently supporting ad hoc queries in large datasets of time sequences, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1997, pp. 289–300.
[4] K.V. Ravi Kanth, D. Agrawal, A. Singh, Dimensionality reduction for similarity searching in dynamic databases, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1998, pp. 166–176.
[5] E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotra, Dimensionality reduction for fast similarity search in large time series databases, Knowl. Inf. Syst. 3 (3) (2001) 263–286.
[6] J. Lin, E. Keogh, S. Lonardi, B. Chiu, A symbolic representation of time series, with implications for streaming algorithms, in: Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003, pp. 2–11.
[7] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: a novel symbolic representation of time series, Data Min. Knowl. Discov. 15 (2) (2007) 107–144.
[8] H. Tayebi, S. Krishnaswamy, A.B. Waluyo, A. Sinha, M.M. Gaber, RA-SAX: resource-aware symbolic aggregate approximation for mobile ecg analysis, in: the IEEE International Conference on Mobile Data Management, 2011, pp. 289–290.
[9] A. Canelas, R. Neves, N. Horta, A new SAX-GA methodology applied to investment strategies optimization, in: Proceedings of the ACM International Conference on Genetic and Evolutionary Computation Conference, 2012, pp. 1055–1062.
[10] T. Rakthanmanon, E. Keogh, Fast shapelets: a scalable algorithm for discovering time series shapelets, in: Proceedings of the SIAM Conference on Data Mining, 2013.
[11] B. Lkhagva, Y. Suzuki, K. Kawagoe, New time series data representation ESAX for financial applications, in: the Workshops on the IEEE International Conference on Data Engineering, 2006, pp. 17–22.
[12] B. Hugueney, Adaptive segmentation-based symbolic representations of time series for better modeling and lower bounding distance measures, in: the European Conference on Principles and Practice of Knowledge Discovery in Databases, 2006, pp. 545–552.
[13] E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotra, Locally adaptive dimensionality reduction for indexing large time series databases, in: Proceedings of the

ACM SIGMOD International Conference on Management of Data, 2001, pp. 151–162.

[14] N.D. Pham, Q.L. Le, T.K. Dang, Two novel adaptive symbolic representations for similarity search in time series databases, in: the IEEE International Asia-Pacific Web Conference, 2010, pp. 181–187.

[15] Z.X. Cai, Q.L. Zhong, The symbolic algorithm for time series data based on statistic feature, Chin. J. Comput. 10 (2008) 1857–1864.

[16] H. Shatkay, S.B. Zdonik, Approximate queries and representations for large data sequences, in: Proceedings of the IEEE International Conference on Data Engineering, 1996, pp. 536–545.

[17] P. Ljubič, L. Todorovski, N. Lavrač, J.C. Bullas, Time-series analysis of UK traffic accident data, in: Proceedings of the International Multi-conference Information Society, 2002, pp. 131–134.

[18] G.Z. Yu, H. Peng, Q.L. Zheng, Pattern distance of time series based on segmentation by important points, in: Proceedings of the IEEE International Conference on Machine Learning and Cybernetics, vol. 3, 2005, pp. 1563–1567.

[19] M. Kontaki, A.N. Papadopoulos, Y. Manolopoulos, Continuous trend-based clustering in data streams, in: Proceedings of the International Conference on Data Warehousing and Knowledge Discovery, 2008, pp. 251–262.

[20] M. Kontaki, A.N. Papadopoulos, Y. Manolopoulos, Continuous trend-based classification of streaming time series, in: Advances in Databases and Information Systems, 2005, pp. 294–308.

[21] G.P.C. Fung, J.X. Yu, W. Lam, News sensitive stock trend prediction, in: Advances in Knowledge Discovery and Data Mining, 2002, pp. 481–493.

[22] H. Wu, B. Salzberg, D. Zhang, Online event-driven subsequence matching over financial data streams, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2004, pp. 23–34.

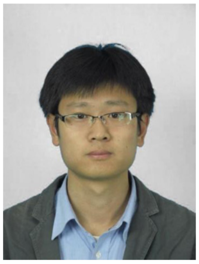
[23] W.B. Wu, Z. Zhao, Inference of trends in time series, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 69 (3) (2007) 391–410.

[24] T.-c. Fu, A review on time series data mining, Eng. Appl. Artif. Intell. 24 (1) (2011) 164–181.

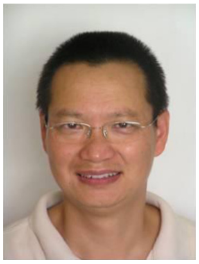
[25] E. Keogh, Q. Zhu, B. Hu, Y. Hao, X. Xi, L. Wei, C.A. Ratanamahatana, The UCR time series classification/clustering homepage. ⟨http://www.cs.ucr.edu/~eamonn/time_series_data/⟩, 2011.

[26] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, E. Keogh, Querying and mining of time series data: experimental comparison of representations and distance measures, Proc. VLDB Endow. 1 (2) (2008) 1542–1552.

[27] J. Shieh, E. Keogh, *i* SAX: indexing and mining terabyte sized time series, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 623–631.



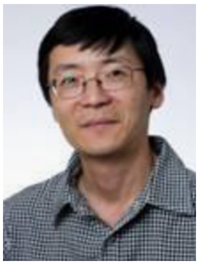
Youqiang Sun received the B.E. degree from Shandong University, China, in 2009. He is currently a Ph.D. candidate in pattern recognition and intelligent system at the University of Science and Technology of China. His research interests include time series data mining and pattern recognition.



Jiuyong Li is a Professor at the School of Information Technology and Mathematical Sciences of University of South Australia. He leads the Data Analytics Group at the University. His main research interests are in data mining, bioinformatics and data privacy. He has led multiple Australian Research Council Discovery projects. He has published more than 80 papers, mostly in leading journals and conferences in the areas. His software tools have been used in several real world projects. He has played a major role in many Australasian and International conferences and workshops.

Jixue Liu got his Ph.D. degree in computer science from the University of South Australia in 2001. He currently works in the University of South Australia. His research interests include the transformation and integration of data, constraints, and queries between XML and relational data, data privacy, online trust modeling, and integrity constraint discovery from data, patterns from sequential data. He has published in world's top journals in Databases (TODS, JCSS, TKDE, Acta Informatica, etc.).

Bingyu Sun is a professor at the Institute of Intelligent Machines, Chinese Academy of Sciences. He also holds an Adjunct Professor position at the department of automation, University of Science and Technology of China. His main research interests are in pattern recognition, data mining and bioinformatics. He has published more than 50 papers, mostly in leading journals and conferences in the areas.



Chris Chow is a Senior Research Specialist, Water Treatment and Distribution Research, at the Australian Water Quality Centre (AWQC), SA Water. He also holds an Adjunct Professor position at UniSA and the Leader of the Advanced Water Quality Sensing and Optimization group, SA Water Centre for Water Management and Reuse. He has co-authored over 250 technical papers in water science journals and conferences.