AN IMPROVED SYMBOLIC AGGREGATE APPROXIMATION DISTANCE MEASURE BASED ON ITS STATISTICAL FEATURES

Chaw Thet Zan
Waseda University
Tokyo
Japan
chawtzan@fuji.waseda.jp

Hayato Yamana
Waseda University
Tokyo
Japan
yamana@waseda.jp

ABSTRACT

The challenges in efficient data representation and similarity measures on massive amounts of time series have enormous impact on many applications. This paper addresses an improvement on Symbolic Aggregate approXimation (SAX), is one of the efficient representations for time series mining. Because SAX represents its symbols by the average (mean) value of a segment with the assumption of Gaussian distribution, it is insufficient to serve the entire deterministic information and causes sometimes incorrect results in time series classification. In this work, SAX representation and distance measure is improved with the addition of another moment of the prior distribution, standard deviation; SAX_SD is proposed. We provide comprehensive analysis for the proposed SAX_SD and confirm both the highest classification accuracy and the highest dimensionality reduction ratio on University of California, Riverside (UCR) datasets in comparison to state of the art methods such as SAX, Extended SAX (ESAX) and SAX Trend Distance (SAX_TD).

CCS Concepts

•**Mathematics of computing** → **Dimensionality reduction;** Time series analysis; •**Information systems** → *Information retrieval;*

Keywords

Symbolic representation; dimension reduction; time series; statistical features; classification

1. INTRODUCTION

Most of the real-world application domains such as financial assessment, weather monitoring, medical data examination, and multimedia systems, create huge amounts of time-series data daily. One of the main features of time series data, curse of dimensionality, is highly required to develop efficient data representation techniques that not only can reduce the high dimensionality of time series data, but also can preserve its significant characteristics. In addition, the significant influence on the desirable distance measures for reduced time series data have to be defined carefully for indexing and searching, classification, clustering, motif discovery, anomaly detection, rule discovery and other mining tasks of time series.

Some of the well-known time series data representations with dimension reduction techniques are Discrete Fourier Transform (DFT) [1][7], Discrete Wavelet Transform (DWT) [5], Discrete Cosine Transform (DCT) [11], Singular Value Decomposition (SVD) [8], Piecewise Aggregate Approximation (PAA) [10], Adaptive Piecewise Constant Approximation (APCA) [9] and Symbolic Aggregate Approximation (SAX) [13].

Most of the above mentioned techniques except for SAX adopt a data reducing technique in real-valued representation that are more expensive in term of storage and computational complexity than symbolic representation for high dimensional time series data. In SAX representation, SAX transforms real-valued time series data into symbolic string based representation. It has two main steps: (1) transforming the original time series to Piecewise Aggregate Approximation (PAA), and (2) converting the PAA represented values to alphabetic symbols based on the assumption that the given normalized time series data follows a highly "Gaussian distribution". Symbolic representation allows us to use the available rich set of existing string-based algorithms and data structures to work with time series mining tasks. In addition, the SAX distance measure allows a lower bound to the popular distance measures defined on the original data. Because of its criteria such as storage efficiency, time efficiency and correctness of answer sets (no false dismissals), SAX has been widely used in various application domain such as semantic sensor network [2], mobile data management [20], and data visualization tool[12].

Even though symbolic aggregate approximation has worthy characteristics concerning dimensionality reduction and minimum distance measure, it still has some weaknesses in identifying the dis/similar or frequent/rare patterns for recognition and classification. SAX adopts a lookup table that contains breakpoints for mapping the alphabetic symbols with their average (mean) values. These breakpoints can be generated from a Gaussian curve in an arbitrary num-

iiWAS '16, November 28-30, 2016, Singapore, Singapore

DOI: http://dx.doi.org/10.1145/3011141.3011146

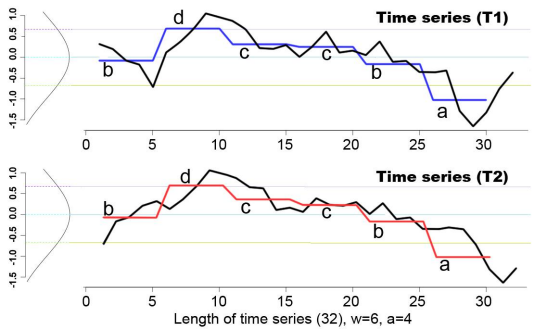


Figure 1: SAX Representation for sample time series of length (32), segment size (w=6) and symbol size(s=4)

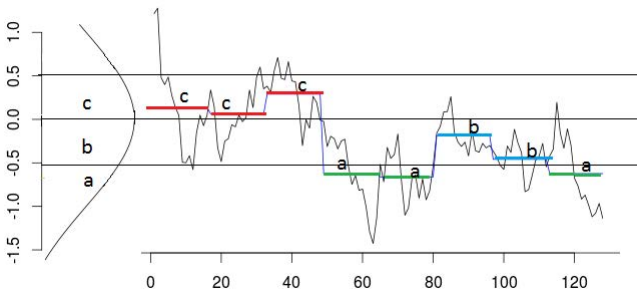


Figure 2: An Example SAX Representation (A time series of length 128 is mapped to the word 'cccaabba' where the segment size is 8 and symbol size is 3.)

ber of alphabets with equally probable regions. In fact, there is a question as to how to recognize the different segments of time series data (Example: T1 and T2) with the same mean values as shown in Figure 1. These same mean values are mapped with the same symbols in the SAX representation and the similarity between them is produced as "0" for same time series data. In reality, the result of similarity between them is incorrect and it may lead to wrong decision making in time series mining.

As only the mean value is utilized, SAX is insufficient to deliver the entire deterministic information of time series data. To solve the problem, SAX_TD [19] proposed a modified distance measure by integrating the SAX distance with the weighted trend distance to classify the time series data more precisely. Alternatively, Extended SAX (ESAX) [15] overcomes such difficulties by using two new discrete points to divide the time series data finely in addition with the method used in original SAX.

In our work, we propose to adopt one more value, its statistical feature, Standard Deviation (SD), along with the SAX symbol that can identify the amount of variability between the actual and average value such that a lower SD corresponds to smaller variability and a higher SD with larger variability. We have made three contributions as follows:

1. Standard deviation (SD) is adopted with the original SAX representation as another feature to classify (dis)similar patterns more precisely.

2. The SAX distance measure is improved via its statistical feature (SD) that can provide tightness of the lower bound distance; it guarantees no false negatives in searching.

3. Comprehensive experiments have been conducted to show that, in comparison with previous works (SAX, ESAX, and SAX_TD); and our proposed method has achieved better classification accuracy, highest dimensionality reduction ratio and, tightness of the lower bound distance.

The rest of the paper is organized as follows. Section 2 briefly discusses background knowledge and related work for the SAX representation technique. Section 3 explains our proposed method, how it works, and its theoretical analysis. Section 4 presents the experimental evaluation by using 20 UCR datasets [6] comparing our proposed method with existing state of the art methods. Finally, we conclude our work by discussing further plans in section 5.

2. BACKGROUND AND RELATED WORK

We first review the foundation of symbolic aggregate representation and some extended works on it in the following sub-sections.

2.1 SAX Representation

SAX is the first powerful dimensionality/numerosity reduction and lower bounding approach in the time series domain. SAX transforms a time series of length (n) into the string of arbitrary length by using a represented symbol size s (s > 2). The operation of SAX involves two main steps. In step one, the given time series is normalized followed by being transformed into Piecewise Aggregate Approximation (PAA) [10]. In PAA representation, time series T of length n (n-dimensions), $T = < t_1, t_2, \ldots, t_n >$,is divided into w-dimensional equal-sized segments as $\overline{T} = < \overline{t_1}, \overline{t_2}, \ldots, \overline{t_w} >$. Each segment $i^{th}$of $\overline{T}$ is represented as its average (mean) value and is calculated by the following equation:

$$\overline{t_i} = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} t_j, \quad (1)$$

where n is the length of time series and w is the segment size. In brief, PAA reduces the data from n-dimensional space to w-dimensional space represented by segment-wise mean value.

In step two, PAA coefficients are mapped into alphabetic symbols by using a lookup table that contains "breakpoints" for separating the symbols as shown in Table 1 where the symbol size is from 3 to 7. Breakpoints are a sorted list of numbers $B =< \beta_1, \ldots, \beta_{a-1} >$such that the area under a N(0,1) Gaussian curve from β_i to $\beta_{i+1} = \frac{1}{a}$($\beta_0$ and β_a are defined as $-\alpha$ and α, respectively)[13].

By using the defined breakpoints, for example, if we define to use the breakpoints with 3 alphabetic symbols, PAA coefficients that are below the smallest breakpoints are mapped to the symbol "a", all coefficients greater than or equal to the smallest breakpoints and less than the second smallest breakpoint are mapped to the symbol "b" and all the coefficients greater than or equal to the second smallest breakpoint are mapped to the symbol "c". This mapping is illus-

Table 1: Lookup table for breakpoints for separating the symbols (symbol size is from 3 to 7)

$\beta \backslash s$	3	4	5	6	7
β_1	-0.43	-0.67	-0.84	-0.97	-1.07
β_2	0.43	0	-0.25	-0.43	-0.57
β_3		0.67	0.25	0	-0.18
β_4			0.84	0.43	0.18
β_5				0.97	0.57
β_6					1.07

Table 2: A sample lookup table for MINDIST function

	a	b	c	d
a	0	0	0.67	1.34
b	0	0	0	0.67
c	0.67	0	0	0
d	1.34	0.67	0	0

trated in Figure 2. Here, symbol size (s) is defines as the number of different symbols.

2.2 Distance Measure

Given two time series of the same length n, $S =< s_1, \ldots, s_n >$ and $Q =< q_1, \ldots, q_n >$, the most commonly used similarity measure for time series, the Euclidean Distance, is calculated as

$$D(S,Q) = \sqrt{\sum_{i=1}^{n}(s_i - q_i)^2}. \quad (2)$$

For PAA representation of time series of length w, the distance between $\overline{S}$ and $\overline{Q}$ is defined as

$$Dpaa(\overline{S},\overline{Q}) = \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(\overline{s_i} - \overline{q_i})^2}. \quad (3)$$

PAA [10] proved that the distance function, Dpaa($\overline{S}$,$\overline{Q}$), guarantees the lower bound to the true Euclidean distance D(S,Q).

For the SAX representation, the data is transformed into the symbolic representation. Given the two original time series S and Q of the same length n, and $\hat{S}$ and $\hat{Q}$ are the symbolic representation with the segment size w. The MINDIST function of symbolic representation that returns the minimum distance between the original time series of two words is defined as[13]:

$$MINDIST(\hat{S},\hat{Q}) = \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(dist(\hat{s_i},\hat{q_i}))^2}. \quad (4)$$

The dist() function is implemented using a distance-lookup table as illustrated in Table 2 and it can be calculated by the following expression.

$$dist(\hat{s},\hat{q}) = \begin{cases} 0 & \text{if } |\hat{s} - \hat{q}| \leq 1 \\ \beta_{max(\hat{s},\hat{q})-1} - \beta_{min(\hat{s},\hat{q})} & ,otherwise \end{cases}. \quad (5)$$

2.3 Related Works

SAX representation and distance measure have been improved from various aspects recently, such as modifying the lookup table, trend or slope based symbolic representation, regression and least mean square fitting, integrating with original SAX representation. Some improvements related with SAX representation and distance measure are explored in this section.

In [17], Marwan et al. proposed GASAX (Genetic Algorithm SAX) to determine breakpoints that are based on the genetic algorithm (GA). The authors argued that the assumption of Gaussianity oversimplifies the problem of SAX representation and may result in high values of error when performing time-series mining tasks. The objective of GA is to find the configuration that gives the best value of fitness function [17]. The configuration means the optimal or nearly optimal solution for the problem. It has the following elements: a population of individuals, selection according to fitness, crossover to produce new offspring, and random mutation of new offspring [18]. Although GASAX works well on both normalized and non-normalized time series data, GA needs to define the suitable control parameters for its elements and it is important to keep a balance between elements to obtain an optimal solution. Extended SAX (ESAX) [15] enhanced SAX by adding two new points, maximum and minimum, along with the original SAX representation. ESAX proved with financial time-series data that its representation is more precise than SAX without losing the symbolic nature of the original SAX. On the one hand, the storage cost of ESAX is triple that of the original SAX and it is necessary for locating the maximum, minimum and mean values for each segment. In [19], Sun et al. argued that SAX representation is imprecise in distinguishing different time series with similar average values, but with different trends. Because of some difficulties in defining a trend qualitatively, such as slight up/down and significant up/down, SAX_TD [19] defined trend distance quantitatively by using the starting and ending point of the segment and improved the original SAX distance with the weighted trend distance.

In [21], Yin et al. proposed a method called Trend Feature Symbolic Approximation (TFSA). It uses a two-step segmentation technique to segment long time series data rapidly. TFSA proved that it guarantees the lower bounding distance, better segmentation, and classification accuracy. S. Malinowski et al. [16] also represent a time series as a sequence of symbols that consists of the average and trend of the series on each segment. It is based on the quantization of the linear regression on time series sub-segments and symbols take into account information about the average value and slope values. That is why the method called 1d-SAX [16] improves retrieval performance and the compression ratio is the same as original SAX.

In [4], Butler et al. points out that SAX discretization does not guarantee equally probable symbols owing to the intermediate step of Piecewise Aggregate Approximation (PAA). As PAA is applied before SAX representation, the distribution of the data is altered and results in a shrinking standard deviation. In addition, the shrinking distribution negatively affects the symbolic representation of the time series with respect to the targeted distribution.

SAX has been improved in different aspects for different application areas. The original SAX representation assumes that the given normalized time series follows a highly Gaussian distribution and it determines breakpoints that produce equal probability for each symbol under the Gaussian curve. In fact, there is uncertainty in allocating symbols,

such that SAX produces the same symbols for the values that are within the same range of break points even if the similarity of original values is very small. In order to fix this kind of uncertainty, we propose to adopt one additional statistical feature, standard deviation, to assist in identifying wrong decision-making caused by this uncertainty.

3. IMPROVED SYMBOLIC AGGREGATE APPROXIMATION

3.1 SAX_SD:SAX with Standard Deviation

In this section, our improved symbolic aggregate approximation distance measure by using its statistical features is introduced. When comparing the spread of different time series that have the approximately the same mean value, standard deviation (SD) provides the most valuable information for a Gaussian distribution. The dataset with the smaller SD has a narrow spread of measurements around the mean and the one with higher SD has wide spread values. We considered standard deviation (SD) as one additional feature along with the original SAX symbols, both to improve the distance measure and to become more precise in time series classification. For instance, for the time series shown in Figure 2, each of the SAX representation and the new SAX_SD representations can be described as follows:

SAX representation: Q = c c c a a b b a,

SAX_SD representation:$\overline{Q}$ = $c_{1.11}$ $c_{0.61}$ $c_{0.57}$ $a_{0.87}$ $a_{0.52}$ $b_{0.39}$ $b_{0.39}$ $a_{0.84}$.

In the SAX_SD representation, $c_{0.61}$ represents the SAX symbol (c) with its standard deviation value. The standard deviation of each PAA segment (s) of length (n) can be calculated as the following equations:

$$sd(n) = \sqrt{var(n)}, \tag{6}$$

$$var(n) = \frac{1}{n}\sum_{i=1}^{n}(s_i - mean(n))^2, \tag{7}$$

$$mean(n) = \frac{1}{n}\sum_{i=1}^{n} s_i, \tag{8}$$

where sd(n) means standard deviation, var(n) means variance and mean(n) is the average value of each PAA segment of length (n).

3.2 Distance Function

As the original SAX representation can be improved by adding one more feature (SD), we need to modify the distance function of SAX as well. First, SD distance between two segments of the same length is defined as:

Definition 1(SD distance) Given two time series $S =< s_1, \ldots, s_n >$ and $Q =< q_1, \ldots, q_n >$ of the same length n, with equal width segments w, SD distance $SD(\tilde{S}, \tilde{Q})$ can be defined as follows:

$$SD(\tilde{S}, \tilde{Q}) = \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(\tilde{s_i}, \tilde{q_i})^2}. \tag{9}$$

where the standard deviation of $\tilde{s}$ and $\tilde{q}$ can be calculated using equation (6).

After defining the SD distance, our improved SAX_SD distance can be defined as the summation of MINDIST and $SD(\tilde{S}, \tilde{Q})$ by the following equation:

$$SAX\_SD(S, Q) = MINDIST(\hat{S}, \hat{Q}) + SD(\tilde{S}, \tilde{Q}). \tag{10}$$

By substituting equation (4) and (9) into equation (10), the SAX_SD can be obtained as follows:

$$SAX\_SD(S, Q) = \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}((dist(\hat{s_i}, \hat{q_i}))^2 + (\tilde{s_i}, \tilde{q_i})^2)}, \tag{11}$$

where $dist(\tilde{s_i}, \tilde{q_i})$ is the minimum distance of the SAX representation, w is the segment size and n is the length of the time series. As described in Section 2, the dimension (n) of the time series is reduced by the segment size (w). In equation (11), we have noticed that the size of w influences not only the dimensionality reduction ratio of the time series but also the appearance of SD. If the size of w is larger, w/n for each segment becomes larger. The larger number of equal-sized segments means that the length of each segment is shorter and the appearance of SD can be defined with finer grain and different time series can be classified precisely. In terms of the dimension reduction ratio, the original dimension cannot be reduced as much as expected. If the size of (w) becomes smaller, the length of each segment becomes longer and it reflects the appearance of SD. For instance, we figured out the influences of segment size (w) on our proposed SAX_SD and SAX distance with the two different time series of ECG (electrocardiogram) dataset [6] in Table 3. The results shown in Table 3 illustrates that our SAX_SD distance is closer to the true (Euclidean) distance than the original SAX. The true (Euclidean) distance for the ECG dataset is 9.95. Our proposed SAX_SD distance is closer to the true distance than the SAX distance and the larger w size influences the appearance of SD.

Table 3: Distance for SAX and SAX_SD with different (w) size and symbol size (s=10) of ECG dataset with length 96.

w size	2	4	8	16	32
SAX	0.00	2.53	5.80	5.51	6.58
SAX_SD	1.52	5.38	6.07	5.90	7.29

3.3 Lower Bounding Guarantee

Most of the data representation methods like symbolic representation and others techniques need to determine on how closely the approximated symbol can represent the features of the original data. An important conclusion is drawn to ensure there are no false dismissals in the distance measure between the symbolic string and the true distance [7]. In SAX representation, MINDIST($\hat{S}, \hat{Q}$)guarantees the lower bound to the PAA distance function, $Dpaa(\overline{S}, \overline{Q})$. PAA distance function also guarantees the lower bound to the true Euclidean distance D(S,Q). The lower bounding property of MINDIST, PAA distance function and true distance can be illustrated by the following inequality which is proven in [10][14]:

$$MINDIST(\hat{S}, \hat{Q}) \leq Dpaa(\overline{S}, \overline{Q}) \leq D(S, Q). \tag{12}$$

Now, we will confirm that our proposed distance measure guarantees the lower bound distance to the true Euclidean distance and tighter bound to the original SAX.

Based on the proof in [10], we first prove as follows that

our proposed distance function $SD(\hat{S},\hat{Q})$ guarantees lower bound distance to the D(S,Q). Let S and Q be two-time series, with |S|=|Q|=n. For the inequality (13),

$$D(S,Q) \geq SD(\hat{S},\hat{Q}). \tag{13}$$

Here, we have $s_i = \overline{s} - \Delta s_i$ and $q_i = \overline{q} - \Delta q_i$, where $\overline{s}$ and $\overline{q}$ are the mean values of S and Q respectively, and ΔS and Δq are the difference between mean value and its original values s and q respectively, where $\sum \Delta s_i = 0$ and $\sum \Delta q_i = 0$ have been proved in [10]. Then, D(S,Q), which is the left-hand side of inequality (13), is rewritten as follows by squaring:

$$\begin{aligned}
D(S,Q) &= \sqrt{\sum_{i=1}^{n}(s_i - q_i)^2} \\
D^2(S,Q) &= \sum_{i=1}^{n}(s_i - q_i)^2 \\
&= \sum_{i=1}^{n}((\overline{s}-\overline{q}) - (\Delta s_i - \Delta q_i))^2 \\
&= n(\overline{s}-\overline{q})^2 - 2(\overline{s}-\overline{q})\sum_{i=1}^{n}(\Delta s_i - \Delta q_i) \\
&\qquad + \sum_{i=1}^{n}(\Delta s_i - \Delta q_i)^2 \\
&= n(\overline{s}-\overline{q})^2 - 2(\overline{s}-\overline{q})(0-0) + \sum_{i=1}^{n}(\Delta s_i - \Delta q_i)^2 \\
&= n(\overline{s}-\overline{q})^2 + \sum_{i=1}^{n}(\Delta s_i - \Delta q_i)^2.
\end{aligned} \tag{14}$$

where we noted that,

$$\sum_{i=1}^{n}(\Delta s_i - \Delta q_i)^2 = \sum_{i=1}^{n}({\Delta s_i}^2 + {\Delta q_i}^2 - 2\Delta s_i \Delta q_i). \tag{15}$$

Next,$SD(\hat{S},\hat{Q})$, which is the right-hand side of inequality (13), is rewritten as follows by squaring. Here, we consider only one segment (w=1) case in this proof,

$$\begin{aligned}
SD(\widetilde{S},\widetilde{Q}) &= \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{n}(\widetilde{s_i} - \widetilde{q_i})^2} \\
SD^2(\widetilde{S},\widetilde{Q}) &= n(\widetilde{s} - \widetilde{q})^2 \\
&= n\left(\sqrt{\frac{1}{n}(\sum_{i=1}^{n}(\overline{s} - s_i)^2)} - \sqrt{\frac{1}{n}(\sum_{i=1}^{n}(\overline{q} - q_i)^2)}\right)^2 \\
&= n\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}{\Delta s_i}^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n}{\Delta q_i}^2}\right)^2 \\
&= \sum_{i=1}^{n}{\Delta s_i}^2 + \sum_{i=1}^{n}{\Delta q_i}^2 - 2\sqrt{\sum_{i=1}^{n}{\Delta s_i}^2 \sum_{i=1}^{n}{\Delta q_i}^2}.
\end{aligned} \tag{16}$$

Using Cauchy-Schwarz inequality, we have:

$$\sqrt{\sum_{i=1}^{n}{\Delta s_i}^2 \sum_{i=1}^{n}{\Delta q_i}^2} \geq \sum_{i=1}^{n}\Delta s_i \Delta q_i \tag{17}$$

Then, by using the equation (15), (16), and (17), we have the following inequality:

$$\sum_{i=1}^{n}(\Delta s_i - \Delta q_i)^2 \geq SD(\widetilde{S},\widetilde{Q}). \tag{18}$$

Therefore, by using the equation (14) and the inequality (18), the proof of inequality (13) is accomplished.

From the equation (3), (4), and (12), the relationship is $Dpaa(\overline{S},\overline{Q}) \geq MINDIST(\hat{S},\hat{Q})$ is transformed as follows,

$$n(\overline{s} - \overline{q})^2 \geq n(dist(\hat{s},\hat{q}))^2. \tag{19}$$

By summarizing the inequality (18) and (19), we have:

$$n(\overline{s} - \overline{q})^2 + \sum_{i=1}^{n}({\Delta s_i}^2 - {\Delta q_i}^2) \geq n(dist(\hat{s},\hat{q}))^2 + SD(\widetilde{S},\widetilde{Q}). \tag{20}$$

Recall equation (14), it comes

$$\sum_{i=1}^{n}(s_i - q_i)^2 \geq n\left((dist(\hat{s},\hat{q}))^2 + \frac{1}{n}SD(\widetilde{S},\widetilde{Q})\right) \tag{21}$$

Segment size (w) can be extended by applying the single segment proof to multiple segments, that is

$$\sqrt{\sum_{i=1}^{n}(s_i - q_i)^2} \geq \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}\left((dist(\hat{s_i},\hat{q_i}))^2 + (\widetilde{s_i},\widetilde{q_i})^2\right)}. \tag{22}$$

According to the inequality (21) and the equation (11), we are able to ensure that our proposed distance measure, i.e., SAX_SD, right-hand side of (21), guarantees lower bound to true Euclidean distance. Besides, because SAX_SD is the sum of MINDIST plus SD distance, it is assured that MINDIST (the original SAX distance) guarantees the lower bounds to SAX_SD distance as follows:

$$MINDIST(\hat{S},\hat{Q}) \leq SAX_SD(\widetilde{S},\widetilde{Q}) \leq D(S,Q). \tag{23}$$

In summary, our proposed distance SAX_SD achieves not only lower bound to true Euclidean distance but also tighter bounding property to the original SAX distance.

4. EXPERIMENTAL EVALUATION

In the evaluation work, we first introduce the diversity of time series data used, and then compare the results with our proposed method and previously proposed state of the art methods in classification error rate, dimensionality reduction ratio, and efficiency.

4.1 Dataset

We performed a comprehensive experiment on 20 diverse UCR time series datasets [6] in order to compare with the previous methods that have been done on the same datasets. Each dataset is split into training and testing parts. Table 4 describes each dataset characteristics such as name, number of classes, training data size, testing data size, length of time series, and its category.

4.2 Comparison of Classification Results

As classification is one of the major tasks in time series data mining, we compare classification accuracy of our improved SAX_SD distance measure with the classic Euclidean distance, state of art SAX [14], ESAX [15] and SAX_TD

Table 4: 20 UCR datasets [6]

No.	Name	#of classes	Training Set Size	Testing Set Size	Time Series Length	Type
1	Synthesis Control	6	300	300	60	Simulated
2	GunPoint	2	50	150	150	Motion
3	CBF	3	30	900	128	Simulated
4	FaceAll	14	560	1690	131	Image
5	OSULeaf	6	200	242	427	Image
6	SwedishLeaf	15	500	625	128	Image
7	50Words	50	450	455	270	Image
8	Trace	4	100	100	275	Sensor
9	TwoPatterns	4	1000	4000	128	Simulated
10	Wafer	2	1000	6164	152	Sensor
11	FaceFour	4	24	88	350	Image
12	Lighting2	2	60	61	637	Sensor
13	Lighting7	7	70	73	319	Sensor
14	ECG	2	100	100	96	ECG
15	Adiac	37	390	391	176	Image
16	Yoga	2	300	3000	426	Image
17	Fish	7	175	175	463	Image
18	Beef	5	30	30	470	Spectro
19	Coffee	2	28	28	286	Spectro
20	OliveOil	4	30	30	570	Spectro

[19]. As we discussed in our related work, ESAX [15] improves its representation preciseness by tripling the original SAX representation and SAX_TD [19] improves the classification accuracy by integrating trend distance with original SAX. For the experimental method and parameter setting, the 1-nearest neighbor classification method is applied for comparing different distance measures. We used a training dataset to find out the best parameters of w (segment size) and s (symbol size), and the same paradigm as applied in [14] for fairness of comparison:

1. For segment size (w), we search from 2 up to n/2 by doubling the value of w each time. Here, n is the length of the time series.
2. For symbol size (s), we search each value from 3 to 10.
3. If different parameter settings resulted in the same result, we will use the smallest parameter values primary based on segment size (w).

As the dimensionality reduction ratio, we will reference the same ratio used in [19] as follows:

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

Based on the above formula, we can work out each reduction ratio as shown in Table 5 where w represents the segment size. Here, each symbol requires at least $log_2(s)$ bits per word to be represented, where s is symbol size and r is a number of bits to represent a real number. Then, the space complexity of each method is shown in Table 5. SAX represents each segment as one symbol. ESAX requires three symbols for each segment. SAX_TD needs one symbol and one real trend value for each segment plus another real value for the last segment. Our proposed SAX_SD uses one symbol and one standard deviation value for each segment.

We classify 20 UCR time series' test datasets with different distance measures based on the training dataset and report the results (classification error rate) in Table 6. We highlighted in the lowest classification error rate for each distance measure in that table. As the result, SAX_SD has the lowest classification error rate (12 out of 20), SAX_TD (4 out of 20), Euclidean distance (2 out of 20), SAX and ESAX is (1 out of 20). In addition, we summarize the results by using a scatter plot with respect to comparison of the two distance measures (SAX, SAX_SD) as two-dimensional points as shown in Figure 3. If the points (the rounded dots) fall within the lower triangle, it indicates that SAX_SD is more accurate than the SAX distance and vice versa for the upper triangle (square block). If the points (triangle block) fall on the diagonals, both of the distance measures have the same precision rate. From these comparison results, our improved distance measure SAX_SD was achieved in more than half of the datasets over its competitive methods. According to Figure 3 and Table 6, we reviewed our distance measure SAX_SD with Euclidean distance, SAX distance, ESAX, and SAX_TD as follows.

Table 5: Dimension reduction ratios of SAX, ESAX, SAX_TD and SAX_SD

Name	Reduction ratio	Space Complexity
SAX	w/n	$w(log_2(s))/n$
ESAX	$3w/n$	$3w(log_2(s))/n$
SAX_TD	$(2w+1)/n$	$w(log_2(s)+r)/n+r$
SAX_SD	$2w/n$	$w(log_2(s)+r)/$n

First, we can see that three out of five distance measure including our proposed method get the same classification error rate on the "coffee" and "oliveoil" datasets. Especially, the classification error rate is zero in the coffee dataset. Second, the lowest classification error rate mostly occurs in our proposed distance measure, meaning that it works well on diversity of time series datasets.

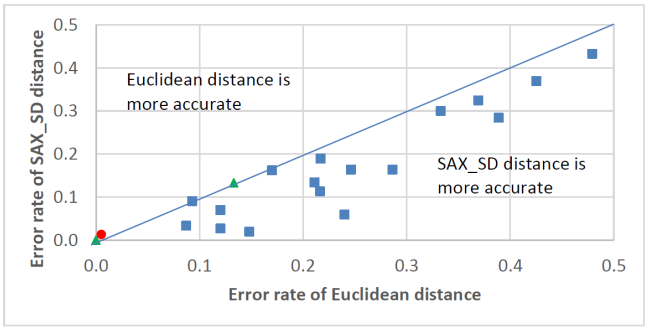
4.3 Comparison of Dimension Reduction and Efficiency

SAX is famous for its dimensionality/numerosity reduction in diverse domains. Its dimensionality reduction depends on the segment size (w) divided by the time series length (n) besides that the segment size (w) has to achieve the lowest classification error rate on the dataset. The result of the dimensionality reduction ratio on each of the representation methods (SAX, ESAX, SAX_TD, and SAX_SD) are illustrated in Table 6. Based on Table 6, we can see the dimensionality reduction ratio in the right most column of each distance measure. On the average of all of the datasets, ESAX has the lowest dimensionality reduction ratio (0.82), SAX and SAX_TD are roughly similar at (0.29) and (0.36) receptively, and our proposed SAX_SD has the highest reduction ratio at (0.21). From the dimensionality reduction ratio point of view, SAX_SD also achieved high dimensionality reduction with its competitive methods. The segment size w highly affects the classification accuracy and its dimensionality reduction. If the segment size w becomes large for a given time series data, the dimensionality reduction ratio becomes lower and the segment size significantly impacts on its prior distribution function (Gaussian).

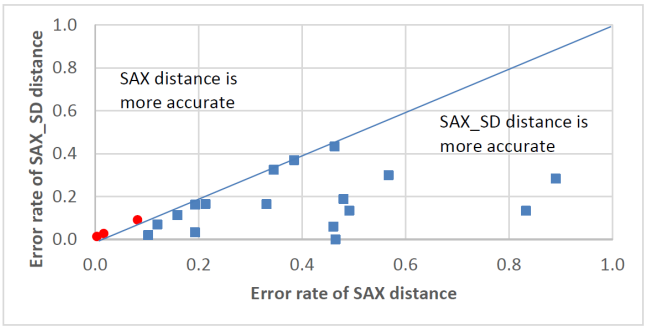
For the comparison of computational efficiency, we compared the execution time of our improved distance measure SAX_SD with the other distance measures ESAX and SAX_TD with the following environment and settings.

Table 6: 1-NN comparison between Euclidean Distance, SAX[14], ESAX[15], SAX_TD[19] and SAX_SD (proposed)

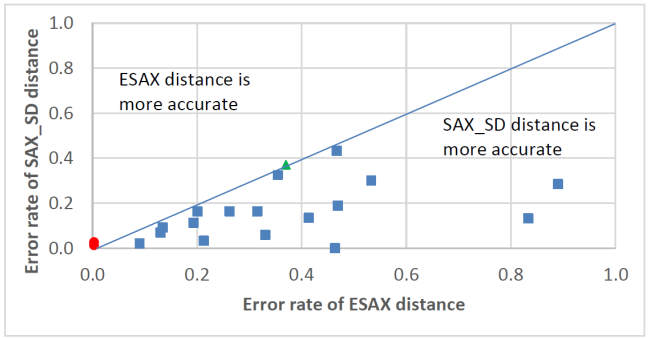
Data Set No	1NN EU Error	SAX				ESAX				SAX_TD				SAX_SD			
		1-NN SAX Error	SAX w	SAX a	SAX ratio	1NN ESAX Error	ESAX w	ESAX a	ESAX ratio	1NN SAX_TD Error	SAX_TD w	SAX_TD a	SAX_TD ratio	1NN SAX_SD Error	SAX_SD w	SAX_SD a	SAX_SD ratio
1	0.120	0.017	16	10	0.27	**0.003**	16	10	0.80	0.050	16	10	0.55	0.027	16	10	0.53
2	0.087	0.193	64	10	0.43	0.213	64	10	1.28	0.047	4	3	0.06	**0.033**	32	3	0.43
3	0.148	0.102	32	10	0.25	0.090	64	10	1.50	0.089	4	10	0.07	**0.020**	4	10	0.06
4	0.286	0.331	64	10	0.49	0.315	64	9	1.47	0.201	16	8	0.25	**0.164**	64	3	0.98
5	0.479	0.463	128	10	0.30	0.467	16	9	0.11	0.438	32	7	0.15	**0.433**	32	3	0.15
6	0.211	0.491	32	10	0.25	0.413	64	10	1.50	0.211	16	7	0.26	**0.134**	32	3	0.50
7	0.369	0.345	128	10	0.47	0.354	32	10	0.36	0.358	64	10	0.48	**0.325**	8	9	0.06
8	0.240	0.460	128	10	0.47	0.330	4	10	0.04	0.230	32	7	0.24	**0.060**	8	7	0.06
9	0.093	0.082	32	10	0.25	0.134	64	10	1.50	**0.071**	16	10	0.26	0.091	8	10	0.13
10	0.005	**0.003**	64	10	0.42	**0.003**	64	9	1.26	0.004	64	7	0.85	0.013	4	9	0.05
11	0.216	0.159	128	10	0.37	0.193	128	7	1.10	0.159	32	9	0.19	**0.114**	16	7	0.09
12	0.246	0.213	256	10	0.40	0.262	32	7	0.15	0.197	32	7	0.10	**0.164**	4	10	0.01
13	0.425	0.384	128	10	0.40	0.370	128	8	1.20	**0.301**	8	10	0.05	0.370	4	10	0.03
14	0.120	0.120	32	10	0.33	0.130	32	10	1.00	0.090	32	7	0.68	**0.070**	4	9	0.08
15	0.389	0.890	64	10	0.36	0.890	64	10	1.09	**0.284**	16	8	0.19	**0.284**	16	5	0.18
16	0.170	0.193	128	10	0.30	0.201	128	10	0.90	0.169	128	3	0.60	**0.162**	16	10	0.08
17	0.217	0.480	128	10	0.28	0.469	128	10	0.83	**0.189**	64	9	0.28	**0.189**	64	9	0.28
18	0.333	0.567	128	10	0.27	0.533	32	9	0.20	**0.300**	64	9	0.27	**0.300**	16	9	0.07
19	**0.000**	0.464	128	10	0.45	0.464	4	5	0.04	**0.000**	16	3	0.12	**0.000**	8	3	0.06
20	**0.133**	0.833	256	10	0.45	0.833	2	3	0.01	**0.133**	64	3	0.23	**0.133**	128	3	0.45
Avg	**0.214**	**0.340**			**0.360**	**0.333**			**0.818**	**0.176**			**0.293**	**0.154**			**0.213**



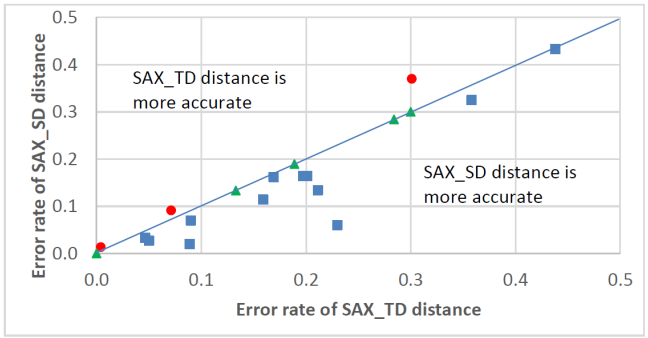
(a) SAX_SD and Euclidean distance



(b) SAX_SD and SAX distance

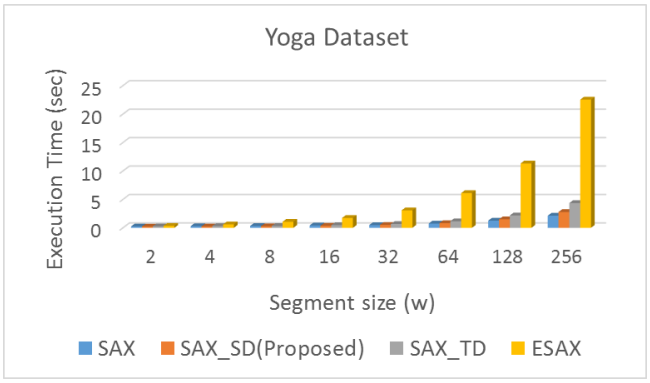


(c) SAX_SD and ESAX distance

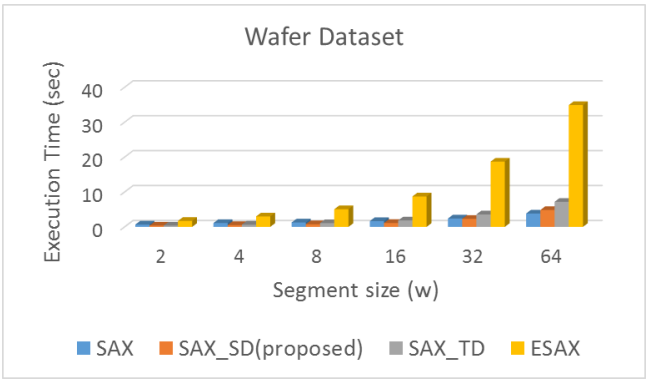


(d) SAX_SD and SAX_TD distance

Figure 3: Comparisons of error rate between SAX_SD (proposed) and other existing methods (EU, SAX, ESAX, SAX_SD) on 20 UCR datasets. Lower triangle (square blocks) is the region where SAX_SD is more accurate than other distance measure, upper triangle (circle dots) is the region where other distance measure is more accurate than SAX_SD and diagonal line (triangle blocks) is the region where both distance measure is same accuracy rate.



(a)



(b)

Figure 4: The execution time of four distance measures with different w size and fixed symbol size (s=10) on two different time series dataset (a) Yoga dataset with length of 426 (b) Wafer dataset with length of 152

1. A middle and large sized dataset from the 20 UCR datasets is selected for the experiment: Wafer for a large dataset and Yoga for a middle-sized dataset.
2. The segment size (w) is varied on its time series length and the symbol size is constant (s=10).
3. The experiment is performed on a machine with a 2.20GHz core i7 processor and 8 GB RAM running the 64 bit Windows operating system. The source codes for all methods are implemented in Java.
4. The execution time is measured from receiving the test dataset to outputting the search result followed by averaging of 25 times measurements. Here, the execution time consists of both the transformation time of test dataset and the time for classifying based on the distance measure.

We compare the execution time of the two different datasets with different (w) segment sizes and a fixed symbol size (s=10) in Figure 4. As we noticed in Figure 4, when the segment size (w) becomes larger, the execution time takes longer on each distance measure. Besides, among the four distance measures, SAX has the lowest computation time because only one value is represented. Execution time of SAX_SD is lower than that of SAX_TD on the same closely representation method and ESAX time is the highest execution time for its representation preciseness.

5. CONCLUSIONS

In this work, we presented the improved symbolic aggregate approximation distance measure (SAX_SD) on its statistical features (mean and standard deviation). SAX_SD proved that the appearance of standard deviation (SD) is well suited to identifying the diversity of time series datasets. Additionally, SAX_SD has achieved both the highest classification accuracy and the highest dimensionality reduction ratio on UCR datasets in comparison with state of the art methods such as SAX, ESAX, and SAX_TD. It also has better computational efficiency than other two methods (ESAX, SAX_TD) and is comparable with the original SAX representation. It guarantees not only the lower bound but also tighter bound distance over the state of the art SAX distance. At present, SAX_SD method does not support a lower bounding approximation of Dynamic Time Warping [3].

As a future direction, we intend to exploit our method to apply it to motif discovery and anomaly detection for multidimensional time series like climate data.

6. REFERENCES

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. In *Proc. of the 4th Int'l Conf. on Foundations of Data Organization and Algorithms*, pages 69–84, October 1993.

[2] P. Barnaghi, F. Ganz, and C. Henson. Computing perception from sensor data. In *Proc. of IEEE Sensors*, pages 1–4, 2014.

[3] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Proc. of the 3rd Int'l Conf. on Knowledge Discovery and Data Mining*, pages 359–370, 1994.

[4] M. Butler and D. Kazakov. Sax discretization does not guarantee equiprobable symbols. *IEEE Trans. on Knowledge and Data Engineering*, 27(4):1162–1166, April 2015.

[5] K. P. Chan and A. W. C. Fu. Efficient time series matching by wavelets. In *Proc. of the IEEE Int'l Conf. on Data Engineering*, pages 126–133, March 1999.

[6] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.

[7] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time series databases. In *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pages 419–429, June 1994.

[8] K. V. R. Kanth, D. Agrawal, and A. Singh. Dimensionality reduction for similarity searching in dynamic databases. In *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pages 116–176, June 1998.

[9] E. Keogh, K. Chakrabarti, S. Mehrotra, and M. Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. In *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pages 151–162, June 2001.

[10] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3):263–286, August 2001.

[11] F. Korn, H. V. Jagadish, and C. Faloutsos. Efficiently supporting ad hoc queries in large datasets of time sequences. In *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pages 289–300, June 1997.

[12] H. Li and L. Yang. Time series visualization based on shape features. *Knowledge-Based Systems*, 41:43–53, March 2013.

[13] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proc. Of the ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, 2003.

[14] J. Lin, E. Keogh, L. Wei, and S. Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144, October 2007.

[15] B. Lkhagva, Y. Suzuki, and K. Kawagoe. New time series data representation esax for financial applications. In *Proc. of the 22nd Int'l Conf. on on Data Engineering Workshops*, pages 17–22, 2006.

[16] S. Malinowski, T. Guyet, R. Quiniou, and R. Tavenard. 1d-sax: a novel symbolic representation for time series. In *Proc. of the 12th Int'l Symp. on Intelligent Data Analysis*, pages 273–284, 2013.

[17] M. Marwan and M. Fuad. Genetic algorithms-based symbolic aggregate approximation. In *Proc. of the 14th Int'l Conf. on Data Warehousing and Knowledge Discovery*, pages 105–116, 2012.

[18] M. Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge MA, 1998.

[19] Y. Sun, J. Li, J. Liu, B. Sun, and C. Chow. An improvement of symbolic aggregate approximation distance measure for time series. *Neurocomputing*, 138:189–198, August 2014.

[20] H. Tayebi, S. Krishnaswamy, A. B. Waluyo, A. Sinha, and M. M. Gaber. Ra-sax: Resource-aware symbolic aggregate approximation for mobile ecg analysis. In *Proc. of the IEEE Int'l Conf. on Mobile Data Management*, pages 289–290, 2011.

[21] H. Yin, S. qiang Yang, X. qian Zhu, S. dong Ma, and L. min Zhang. Symbolic representation based on trend features for knowledge discovery in long time series. *Frontiers of Information Technology and Electronic Engineering*, 16(9):744–758, September 2015.