IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 21, 2019, accepted August 6, 2019, date of publication August 9, 2019, date of current version August 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2934109

A New Time Series Similarity Measurement Method Based on the Morphological Pattern and Symbolic Aggregate Approximation

JIANCHENG YIN, RIXIN WANG, HUAILIANG ZHENG, YUANTAO YANG, YUQING LI, AND MINQIANG XU

Deep Space Exploration Research Center, Harbin Institute of Technology, Harbin 150001, China

Corresponding author: Minqiang Xu (xumqhit15@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 11172078.

ABSTRACT Aiming at the problem that the traditional similarity measurement methods cannot effectively measure the similarity of the time series with the difference both in the trend and detail, this paper proposes a new time series similarity measurement method (MP-SAX) based on the morphological pattern (MP) and symbolic aggregate approximation (SAX). According to the empirical mode decomposition (EMD), the time series are decomposed and reconstructed into the trend component and the detail component. Then, the similarity of the trend component under morphological pattern coding and that of the detail component under symbolic aggregate approximation coding are respectively calculated by the longest common subsequence (LCS). Finally, the similarity of the time series is obtained by weighted aggregation of the similarity of trend component and detail component. The MP-SAX is verified by the simulation time series and the time series from UCR Time Series Classification / Clustering Homepage. The results show that the MP-SAX can effectively measure the similarity of the time series with the changes both in trend and detail.

INDEX TERMS Similarity measure, morphological pattern, symbolic aggregate approximation, longest common subsequence, empirical mode decomposition.

I. INTRODUCTION

With the development of modern industry and information technology, massive data through various sensing devices are generated. These monitoring data which are discrete time series, in essence, have been widely analyzed to mine useful potential information in many application fields, such as finance, medicine, aerospace, and meteorology, etc. The similarity measurement between two time series is a core requirement for the data mining of time series and knowledge discovery tasks, such as clustering and classification [1].

The similarity measurement of time series was first proposed by Agrawal *et al.* [2], in which the similarity of time series was measured by Euclidean distance. Euclidean distance is the most widely used similarity measurement method which is easy to calculate and has clear meaning [3], [4]. It has been widely used in data mining tasks of time series [5]. However, Euclidean distance requires that the time series to be measured should have the same length. For calculating the similarity of time series with different length, some elastic measurement methods [6]–[9] based on the manner of 'one-to-many' or 'one-to-zero' are proposed. For example, Hsu *et al* [10] proposed a Flexible Dynamic Time Wrapping (FDTW). Folgado *et al.* [11] proposed Time Alignment Measurement to measure the similarity of time series in the temporal domain after aligning two time series by Dynamic Time Wrapping (DTW). Silva *et al.* [12] adapted the DTW with pruned warping paths to improve the internal efficiency of the DTW calculation. Tao *et al.* [13] proposed dynamic spatial time warping which can maintain the invariance of curve similarity to the rotations and translations of curves for predicting the capacity degradation of the battery. Putpuek *et al.* [14] utilized a simple signature and the longest common subsequence (LCS) algorithm to improve the efficiency of the automatic retake detection. Rivault *et al.* [15] proposed a generalization of the LCS to measure the events' semantic similarity. Ayad *et al.* [16] extended the cyclic edit distance based on q-gram to improve the computational speed and

The associate editor coordinating the review of this article and approving it for publication was Bora Onat.

accuracy. Zhang *et al.* [17] embedded the edit distance with a real penalty into difference-weighted KNN classifiers to realize the classification of the pulse waveform.

In modern industry, the dimensionality of time series is becoming higher gradually, and the rapid and accurate processing of time series is a new requirement in the data mining of time series [18]. Therefore, in order to improve the computational efficiency of similarity measurement, some methods which express the time series in a simple and feature-rich manner have been proposed, such as, describe time series from the following aspects of time series: symbolization [19], change trend [20] and shape [21]. For example, Tamura and Ichimura [22] proposed a hybrid symbolic aggregate approximation by combining the symbolic aggregate approximation strings of time series and moving average convergence divergence histogram. For the time series of IoT, Gonzalez-Vidal *et al.* [23] proposed an undeclared mutation segmentation algorithm for data drift. Baldini *et al.* [24] proposed a novel approach to Radio Frequency fingerprinting based on the symbolic aggregate approximation. Wang and Tang [25] proposed the fluctuating pattern based on the trend change information of the original time series.

However, the similarity measurement methods mentioned above can only evaluate the similarity of the time series with either change in trend component or changes in detail component. In the scenarios which both trend and detail component have differences, these methods could not give the most comprehensive measurement. For example, in the life cycle of a product, the degradation of the product is reflected in the change of trend component, and the real-time operating status and environment of the product are reflected in the change of detail component. Therefore, it is helpful to improve the results of time series similarity measurement by considering both the trend component and detail component of the time series. Where the trend component of time series is the general trend of change which is formed by some fundamental factors in a long-term period and the detail component is constituted by seasonal variations, calendar variations, and irregular component [26].

In this paper, a new similarity measurement method which aggregates the similarity of the trend and detail component of time series by weighted manner is proposed. According to the Empirical Mode Decomposition (EMD), the original time series are decomposed into Intrinsic Mode Functions (IMFs). Then the IMFs are reconstructed into trend component and detail component according to the multi-scale permutation entropy (MPE) of IMFs. Then the trend component and the detail component are coded by morphological pattern (MP) and symbolic aggregation approximation (SAX) respectively. The similarity of the trend component and the detail component are respectively calculated by LCS. Finally, the similarity of the time series is obtained by weighted aggregation of the similarity of trend component and detail component.

The main contributions of our work are summarized as follows:

(a) The similarity measurement of the time series with the difference both in trend and detail can be realized by using Empirical Mode Decomposition to obtain the trend and detail component of a signal.
(b) The proposed method, MP-SAX, can achieve superior performance under more extensive application scopes, especially under the scenario that useful information exists both in the trend and detail component of time series. To some extent, the combination of MP and SAX overcomes the issues that MP cannot measure trend differences when the time series contain detail component and SAX cannot effectively measure the detail differences when the time series contain obvious trend component.

The rest of the paper is organized as follows. Section II introduces the necessary background knowledge, In Section III, the new similarity measure method is proposed. The results and discussion are given in Section IV. In Section V, the conclusion of this paper is drawn.

II. BACKGROUND KNOWLEDGE

A. THE MORPHOLOGICAL PATTERN AND SIMILARITY MEASUREMENT

The morphological pattern (MP) is an encoding method of time series, which can encode the change rate of uptrend and downtrend into a series of discrete values and reflect the overall trend of time series very carefully. Meanwhile, the morphological characteristics of the original series are considered by MP, and the numerical size of the time series is ignored. For a time series $X = (x_1, x_2, \cdots, x_n)$, the morphological pattern series $F = (f_1, f_2, \cdots, f_n)$ can be obtained as follows [27]:

$$f_i = \begin{cases} 3, & (x_i - x_{i-1})/t > 1 \\ 2, & (x_i - x_{i-1})/t = 1 \\ 1, & (x_i - x_{i-1})/t < 1 \\ 0, & x_i = x_{i-1} \\ -1, & (x_i - x_{i-1})/t > -1 \\ -2, & (x_i - x_{i-1})/t = -1 \\ -3, & (x_i - x_{i-1})/t < -1 \end{cases} \tag{1}$$

where t is the interval between the two sampling points.

As the morphological pattern series is constituted by $\{-3, -2, -1, 0, 1, 2, 3\}$ alphabetic strings, the LCS can be used to measure the similarity between the two morphological pattern series. For the time series $X = (x_1, x_2, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots, y_m)$, the corresponding morphological pattern series F_x and F_y can be calculated by (1). Then the longest common subsequence of the morphological pattern series can be obtained by dynamic programming. Specifically, the LCS can be obtained as follows [28]:

$$L(i,j) = \begin{cases} \max(L(i-1,j), L(i,j-1)), & F_x(i) \neq F_y(j) \\ L(i-1,j-1) + 1, & F_x(i) = F_y(j) \end{cases} \tag{2}$$

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

β \ α	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

where $F_x(i)$ is the i^{th} element of F_x, and $F_y(j)$ is the j^{th} element of F_y, $L(i,j)$ is the length of the longest common subsequence of F_x and F_y.

Then the similarity based on the longest common subsequence can be calculated as follows

$$sim\left(F_x, F_y\right) = \frac{L}{\min(l_{F_x}, l_{F_y})} \times 100\% \tag{3}$$

where $L$is the length of the longest common subsequence, $\min\left(l_{F_x}, l_{F_y}\right)$ is the length of the shortest series in F_x and F_y.

B. THE SYMBOLIC AGGREGATE APPROXIMATION AND SIMILARITY MEASUREMENT

Lin *et al.* [29], based on the piecewise and central limit theorems of Piecewise Aggregate Approximation (PAA) and the normal distribution characteristics of time series, proposed a symbolic representation of time series – Symbolic Aggregate approXimation (SAX). SAX has been verified as a fast and effective tool for solving time series mining problems. The SAX can convert a time series X of length n into a symbol sequence of length $w(w \ll n)$, in which the compression and noise reduction of raw time series can be realized. The SAX works as follows [30], [31].

Step 1: The time series is normalized in order to make it obey the standard normal distribution by the equation as follows

$$NX = \frac{X - \mu}{\sigma} \tag{4}$$

where NX is the normalized series of X, μ is the mean of all the points in X and σ is its standard deviation.

Step 2: The normalized series is divided into w equal-sized segments by PAA. That is, the time series $X = \{x_1, x_2, \cdots, x_n\}$ can be represented by the average of each segment and the average is calculated by the following equation

$$\overline{x_i} = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} x_j \tag{5}$$

where $\overline{x_i}$ is the average of the i^{th} segment, x_j is one point of time series X, j is the sequence number of the time series from the starting point to the ending point for each segment.

Step 3: The breakpoints β that divide the distribution space into α equiprobable regions are determined by a lookup table shown in Table 1.

Step 4: After the aforementioned steps, each region is assigned a symbol using the determined breakpoints. All PAA coefficients below the smallest breakpoints are mapped to the symbol ''a'', all coefficients greater than or equal to the smallest breakpoint and less than the second smallest breakpoint are mapped to the symbol ''b'', etc.

As the SAX converts the original time series into symbol sequence, the similarity of the symbol sequence can also be calculated by the LCS described as above.

C. THE THEORY OF EMPIRICAL MODE DECOMPOSITION

EMD is an adaptive approach to decompose non-linear and non-stationary time series into a set of intrinsic mode functions (IMFs) and a residual, which the IMFs should satisfy two conditions: (a). the number of zero-crossing and extreme points differs at most by one, (b). at any point in time, the mean value of the upper envelope determined by the local maximum and the lower envelope determined by the local minimum of the series must be zero.

The residual and some higher-order IMFs can describe the trend component of the original time series, meanwhile other lower-order IMFs are the representations of the detail component of original time series. The procedures of EMD decomposition are shown as follows [32], [33]:

Step 1: The upper and lower envelopes of the original time series $x(t)$ are obtained by all local extreme points of the original time series (including the maximum and the minimum points). Then the mean value series of the envelopes $m(t)$ are obtained.

Step 2: The mean value series of the envelopes $m(t)$ are subtracted from the original series $x(t)$ until the result $h_1(t)$ satisfies the two conditions of the IMF as follows:

$$h_1 = x(t) - m(t) \tag{6}$$

the result $h_1(t)$ satisfying the two conditions of the IMF is the first IMF $c_1(t)$.

Step 3: Subtracting the first IMF $c_1(t)$ from the original series $x(t)$.

$$r_1 = x(t) - c_1(t) \tag{7}$$

where the $r_1(t)$ is the residual series after decomposing the first IMF.

Step 4: Then $r_1(t)$ is used as the 'original series' to repeat the above steps until $r_1(t)$ becomes a monotone function from which the IMF can no longer be extracted. Then the original series $x(t)$ can be express as:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{8}$$

III. THE PROPOSED METHOD

A. THE PROCESS OF THE PROPOSED METHOD

For accurately measuring the similarity of time series under the scenarios that there are changes both in the trend component and detail component, a new similarity measurement method of time series based on the MP and SAX is proposed in this paper. The similarity of time series is obtained by the weighted aggregation of the similarity of the trend and detail component. The essential discrepancy between the trend and detail components is their complexity. The change of the trend component is the long-term tendency with lower complexity, while the change of the detail component is rapid variation with larger complexity. Therefore, the trend and detail component can be distinguished by the complexity of time series. Meanwhile, the multi-scale permutation entropy (MPE) is an effective method to measure the complexity of time series. It is widely used to measure the complexity of the time series [34]. The detailed calculation process refers to [34]. And the MPE gradually increases from 0 to 1 with the increase of the complexity of time series. Therefore, the trend and detail component can be determined by the MPE.

As shown in Fig. 1, given the two time series $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2, \cdots, y_m\}$, the process of the MP-SAX is given as follows:

Step 1: Decompose the two time series to IMFs $\{IMF_i^X\}_{i=1}^{N_1}$ and $\{IMF_j^Y\}_{j=1}^{N_2}$ according to the description in Section II.C. Then, calculate the MPEs $\{MPE_i^X\}_{i=1}^{N_1}$ and $\{MPE_j^Y\}_{j=1}^{N_2}$ of each IMFs.

Step 2: According to the MPE, the IMFs of the two time series are reconstructed into trend component and detail component respectively as follows:

$$\begin{aligned} TR_X &= \sum_{i \in \{i | MPE_i^X < 0.4\}} IMF_i^X \\ TR_Y &= \sum_{j \in \{j | MPE_j^Y < 0.4\}} IMF_j^Y \\ DE_X &= \sum_{i \in \{i | MPE_i^X \geq 0.4\}} IMF_i^X \end{aligned} \tag{9}$$

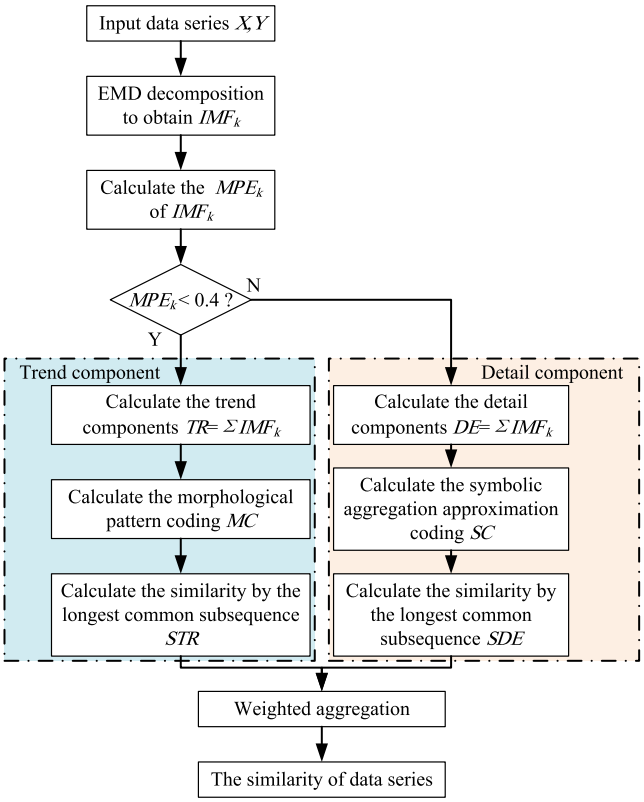


FIGURE 1. The calculation process of the MP-SAX.

$$DE_Y = \sum_{j \in \{j | MPE_j^Y \geq 0.4\}} IMF_j^Y \tag{10}$$

where TR_X and TR_Y is the reconstructed trend component of time series X and Y respectively, DE_X and DE_Y is the reconstructed detail component of time series X and Y respectively.

Step 3: The symbol sequences of trend component MC_X and MC_Y are calculated by (1). The symbolic sequences of detail component SC_X and SC_Y are obtained by the SAX.

Step 4: According to the LCS, the similarity of trend component $STR$and that of detail component SDE are calculated by (2) and (3). Finally, the similarity between the two time series is obtained by weighted aggregating the trend similarity STR and the detail similarity SDEas follows:

$$Sim_{tol} = W_{TR} \cdot STR + W_{DE} \cdot SDE \tag{11}$$

where W_{TR} is the weight of the trend similarity, W_{DE} is the weight of the detail similarity, Sim_{tol} is the similarity of the two time series.

And the weight can be obtained by one of the following two methods.

(a). Subjective method: Set the corresponding weight according to the focus of concern. Such as the weight of trend similarity is greater than that of detail similarity if the impact of trend on time series similarity is more concerned, and vice versa.

(b). Objective method: Set the corresponding weight according to the proportion of IMF in reconstruction. The

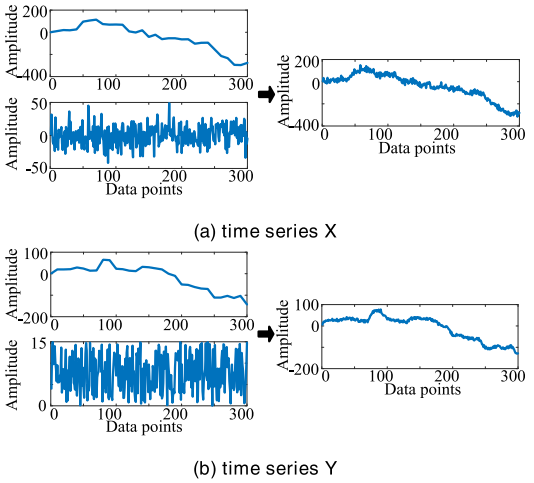


FIGURE 2. The simulation data with difference in trend and detail.

TABLE 2. The summary of datasets.

Dataset	Length of time series	Size of training set	Size of testing set	Number of classes
CBF	129	30	900	3
ECG200	97	100	100	2
Symbols	399	25	995	6
Trace	276	100	100	4

weight can be obtained as follows.

$$W_{TR} = \frac{N_{TR}}{N_{tol}} \tag{12}$$

$$W_{DE} = \frac{N_{DE}}{N_{tol}} \tag{13}$$

where N_{tol} is the number of IMFs decomposed by EMD, N_{TR} is the number of IMFs reconstructed to trend component, N_{DE} is the number of IMFs reconstructed to detail component.

B. THE EFFECT OF EMD RECONSTRUCTION IN THE MP-SAX

Next, two simulation time series are used to illustrate the effect of EMD decomposition and reconstruction in the MP-SAX. The simulation time series X, Y (as shown in the right side of Fig. 2 (a) and (b)) are constituted by two different groups of trend time series and detail time series (as shown in the left side of Fig. 2 (a) and (b)). Then the process and result of EMD are as follows.

The IMFs of time series X, Y (as shown in the left side of Fig. 3 (a) and (b)) can be calculated by the method in section II.C.

As shown in the left side of Fig. 3 (a) and (b), in order to reconstruct the trend component and detail component, it is necessary to determine which IMFs describe the trend component and which describe the detail component. Because the complexity of trend and detail component is different and the complexity of time series can be effectively distinguished by

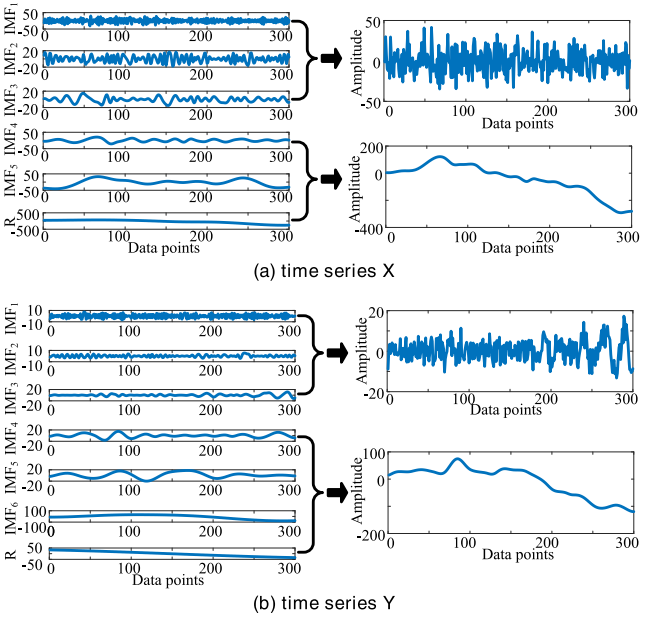


FIGURE 3. The decomposition and reconstruction of the simulation data.

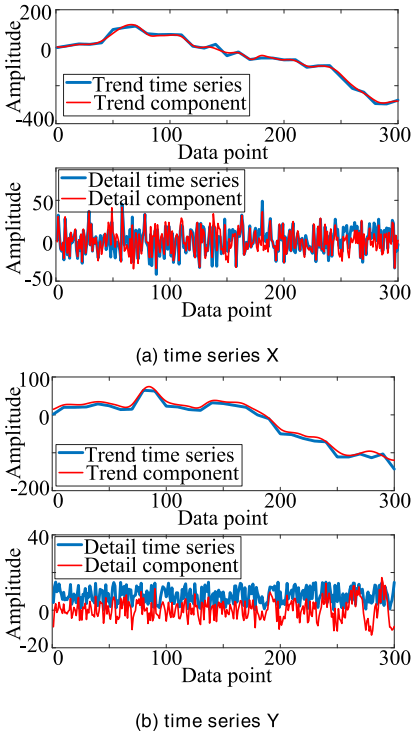
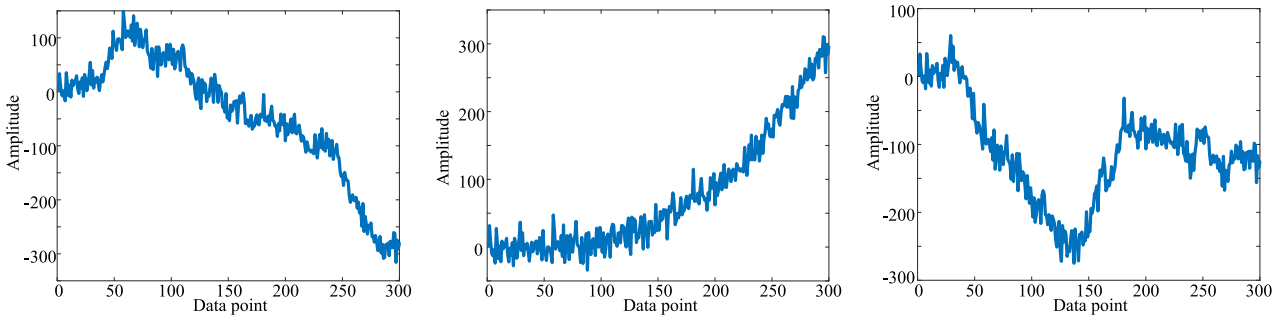


FIGURE 4. The comparison between the reconstructed time series and original time series.

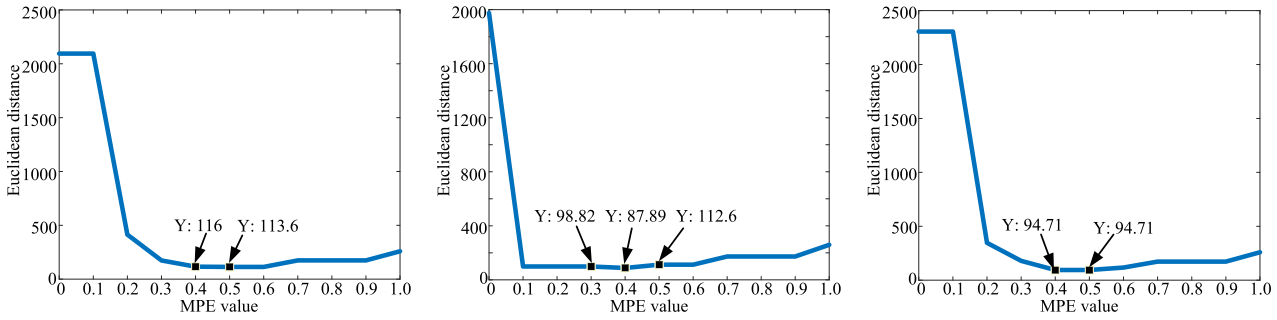
MPE, the trend component and detail component of original time series can be reconstructed according to their MPE values.

After EMD decomposition, the MPE of each IMF is calculated respectively. Then, as shown in Fig. 3, the IMFs with the MPE greater than 0.4 are reconstructed into the detail component of the original time series, and the IMFs with the



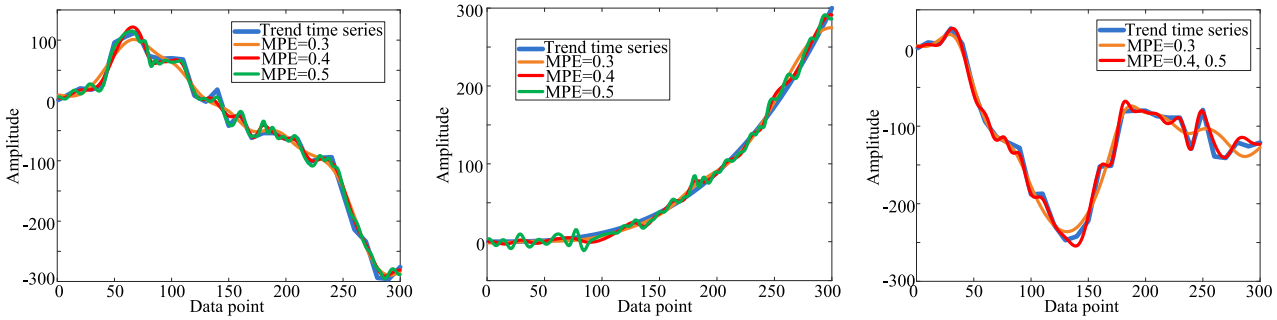
(a) Constructed Time series I (b) Constructed Time series II (c) Constructed Time series III

FIGURE 5. The constructed time series.



(a) Constructed Time series I (b) Constructed Time series II (c) Constructed Time series III

FIGURE 6. The calculation results under different MPEs.



(a) Constructed Time series I (b) Constructed Time series II (c) Constructed Time series III

FIGURE 7. The reconstructed result under different MPEs.

MPE less than 0.4 are reconstructed into the trend component of the original time series. The comparison between the reconstructed time series and the original time series is shown in Fig. 4.

As shown in Fig. 4, it is found that the reconstructed trend component of time series X and $Y$are similar to the original trend time series. However, due to the criterion of stopping iterative and the end effect of EMD, there are some differences between the reconstructed trend component and the original trend time series, the reconstructed detail component and the original detail time series. And the difference is mainly reflected in the reconstructed detail component. Although the original trend and detail time series cannot be completely restored by EMD decomposition and reconstruction, the reconstructed trend and detail component can also effectively reflect the relevant information of the original trend and detail time series. Therefore, using EMD and MPE criterion, the trend and detail component are reconstructed to measure the similarity in these two aspects respectively.

C. THE ESTIMATION OF PARAMETERS IN THE MP-SAX

The main parameters of the MP-SAX are the threshold of the MPE and the weights of the trend and detail component. As the weights of the trend and detail component can be determined according to the description in Section III.A, in this section, the estimation of the MPE threshold will

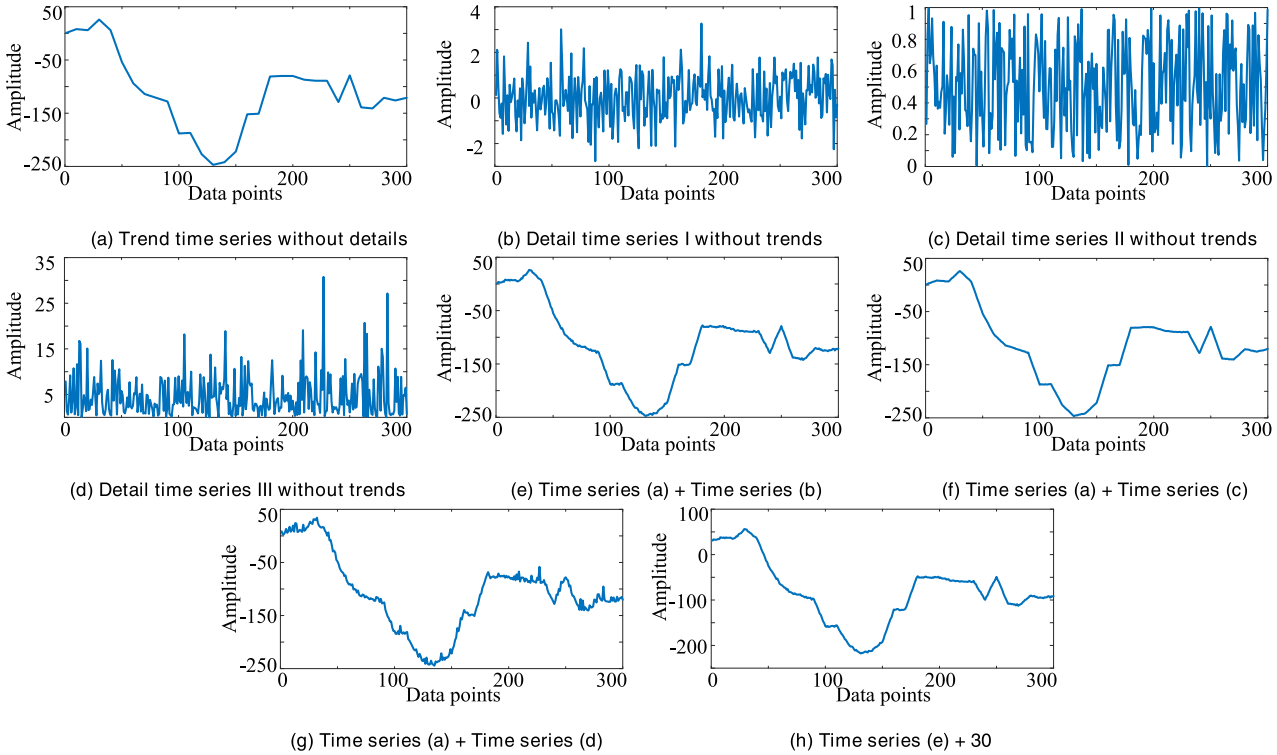


FIGURE 8. The 8 sets of simulation time series.

be discussed. The threshold of the MPE is determined by parameter optimization. Firstly, construct the time series with different trend components by using multiple trend time series and noise time series. As shown in Fig. 5, three of them are selected as examples to explain the estimation.

Then, reconstruct the trend components based on different MPE which is selected by step 0.1 in [0, 1]. The objective function is the Euclidean distance between the reconstructed trend components and the original trend time series. Then, select the MPE corresponding to the minimum objective function. As shown in Fig. 6, the objective function is at the low position when the MPE is 0.3 to 0.5. Therefore, the range 0.3 to 0.5 is selected as the initial optimization results of MPE.

Finally, based on the above optimization results, the trend components are reconstructed and compared with the original trend time series. Select the optimal MPE which the reconstructed trend components contain fewer fluctuations. As shown in Fig. 7, the reconstructed trend components are closer to the original trend time series and contain less fluctuation when the MPE is 0.4. Therefore, 0.4 is selected as the threshold of MPE.

IV. RESULT AND DISCUSSION

In this section, both simulation datasets and real datasets are conducted to evaluate the effect of the MP-SAX by comparing with Euclidean distance, morphological pattern, and symbolic aggregate approximation. During the verification, the parameters of the MP-SAX are set as follows:

(a) The threshold of MPE is 0.4.

(b) The weight is obtained by the subjective method as the time series used in Section IV are mainly different in trend component. Where the weight of the trend component is 0.75 and that of detail component is 0.25.

A. THE SIMULATION DATASETS

To verify the effectiveness of the MP-SAX, 8 sets of simulation time series are employed firstly. The 8 sets of simulation time series are described in Fig. 8.

Then, the similarities of the 8 sets simulation time series are calculated by cross-computing. Meanwhile, in order to show a unified result, the Euclidean distance is transformed into similarity as:

$$sim_{ED} = \left(1 - \frac{D_i}{D_{\max}}\right) \times 100\% \tag{14}$$

where sim_{ED} is the similarity of Euclidean distance, D_i is the i^{th} Euclidean distance, $D_{\max}$ is the maximum of Euclidean distance.

The similarity of Euclidean distance (ED), morphological pattern (MP), symbolic aggregate approximation (SAX), and the MP-SAX are shown in Fig. 9.

As illustrated in Fig. 9, the similarity of MP, SAX, and MP-SAX are different in the symmetrical positions. The reason is that the LCS is obtained by dynamic programming and the LCS of two time series might be a difference because of the calculation order of the two time series. However, the difference of the similarity in the symmetrical positions

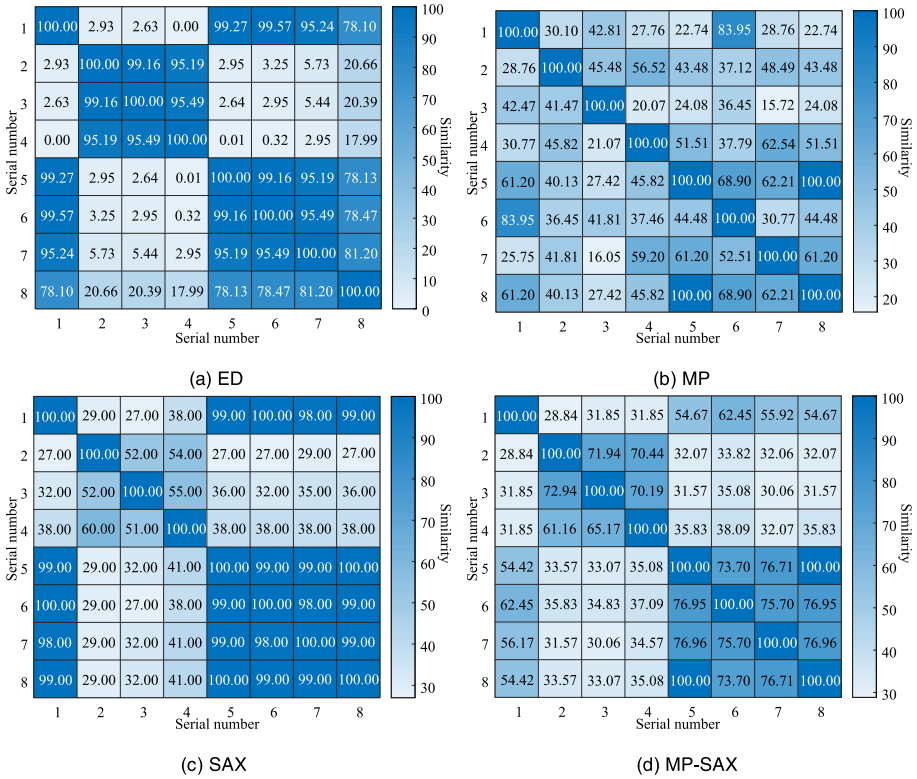


FIGURE 9. The similarity of simulation data.

is small. Besides, comparing the similarities of time series 2 and 5, 3 and 6, 4 and 7, the color of ED, SAX and MP-SAX are the lightest. And there is no obvious difference in the color of MP. In addition, comparing the similarity between the time series 2, 3, 4, 5, 6 and 7 which are obtained by ED, the similarities of time series 2 and 3, 2 and 4, 3 and 4 are the same as that of time series 5 and 6, 5 and 7, 6 and 7. These show that the time series cannot be effectively distinguished by ED and MP, and can be effectively distinguished by SAX and MP-SAX when the trend components of time series are obviously different. Although the MP can effectively identify the time series with different trend component, the MP can only reflect the change in detail component of time series rather than the change in trend component when there are different kinds of detail component in the time series. Therefore, only SAX and MP-SAX can effectively be used to measure the similarity of the time series with different trend component and same detail component.

Meanwhile, in the same trend, comparing the similarities between the time series 1, 5, 6 and 7, the similarities of ED, MP and the MP-SAX are difference under different detail component. Although the disparity of the similarity between time series 5 and 6 and that between time series 5 and 7 by MP-SAX are small, the time series with different detail component can also be distinguished according to the similarity of MP-SAX. However, the SAX will not be able to distinguish the time series with different detail component when the amplitude of the trend component of time series is much larger than that of detail component. In addition, the similarities of the time series 2 and 5, 3 and 6, 4 and 7 by ED are same. Therefore, MP and MP-SAX can effectively be used to measure the similarity of the time series with different detail component and same trend component.

By comparing the similarity of time series 5 and 8, ED can be affected by the amplitude translation of time series. However, MP, SAX, and MP-SAX can be used to identify the same time series with different amplitude translation.

Consequently, the comparisons on the simulation time series illustrate that the MP-SAX not only can measure the similarity of the time series with the differences in trend and detail component effectively, but also can obtain robust results when the amplitude translation exists.

B. THE REAL DATASETS

In the previous section, the simulation datasets are used to validate the effectiveness of the MP-SAX in measuring the similarity of the time series with the difference in trend and detail component. Then, the validity of the MP-SAX is validated by the classification accuracy of the real datasets.

Therefore, the four datasets with obvious trends are selected from the UCR Time Series Classification / Clustering Homepage, called CBF, ECG200, Symbols, Trace respectively. The summary of the four datasets is shown in Table 2.

The accuracy of classification results will be directly affected by the similarity algorithm of time series. The more accurate the similarity measurement is, the higher the classification accuracy is. Therefore, the accuracy of classification can be used to verify the effectiveness of the similarity

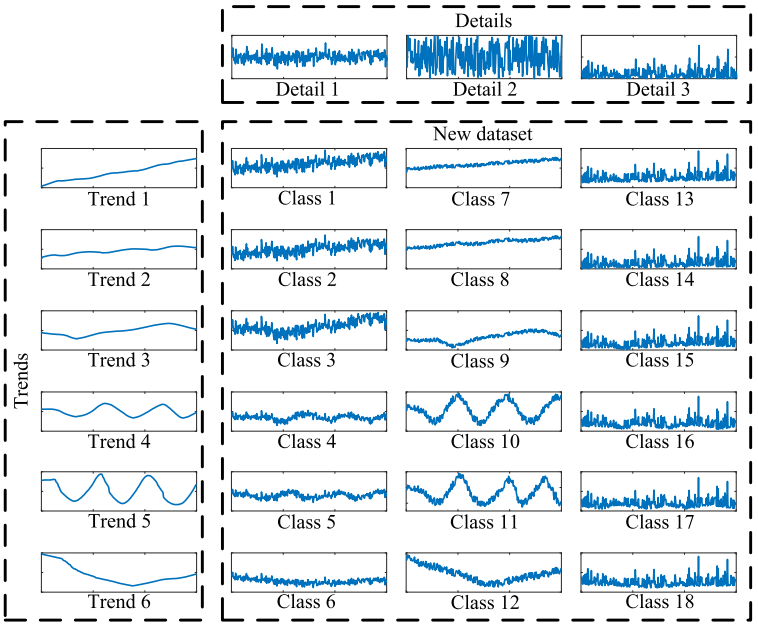


FIGURE 10. The new dataset under scale factor 1.

TABLE 3. The classification accuracy of the four methods.

Dataset	ED	MP	SAX	MP-SAX
CBF	86.78%	39.67%	98.78%	88.00%
ECG200	88.00%	88.00%	82.00%	84.00%
Symbols	89.65%	97.79%	92.16%	93.97%
Trace	76.00%	50.00%	94.00%	81.00%

algorithm. Then, ED, MP, SAX, and MP-SAX are respectively used as similarity measurement algorithms to classify the above mentioned time series. The accuracy of classification is shown in Table 3.

As illustrated in Table 3, ED, SAX, and MP-SAX all have high and stable classification accuracy for the four datasets. However, the classification accuracy of MP is unstable due to the existence of some detail component in the above four datasets. Therefore, MP can only effectively measure the similarity of the time series with only trend component. Besides, with respect to the four time series datasets, although the classification accuracy of the MP-SAX is not the highest, it can provide compared results for all the four datasets which contain different characteristics. That is to say, the MP-SAX is able to achieve acceptable similarity measurement without a priori considering where the discrepancies between different time series may occur (in trend component or detail component). Therefore, the application scope of MP-SAX is larger than that of other methods. Most importantly, it is unnecessary to know the characteristics of the time series before the similarity measurement.

C. THE SIMULATION DATASET BASED ON THE REAL DATASET

For the above-mentioned datasets, as the differences between the time series are only contained in the trend component, the classification accuracy of MP-SAX is not the highest when compared to other methods. Therefore, in order to further illustrate the advantages of the MP-SAX in measuring the similarity of the time series with differences both in trend and detail component, a new dataset is constructed based on the above real dataset. Besides, in order to illustrate the effectiveness of the MP-SAX in similarity measurement of the trend time series with the different degrees of detail component, the new datasets with the different scaling detail component are simulated as follows:

$$NewData = TR + SF \cdot DE + AT \tag{15}$$

where $NewData$ is the new datasets, TR is the trend component of the time series, SF is the scale factor, DE is the detail component of the time series, AT is the random amplitude translation quantity. Since the dataset Symbols only contain simple trend component without obvious detail component, the time series of Symbols are chosen as the trend component of the new time series. Then, the three detail time series in section IV.A are randomly added to the time series of Symbols. The new dataset under scale factor 1 is shown in Fig. 10.

Meanwhile, as there are few training sets in the dataset Symbols, the testing sets of Symbols are used as the training sets of the new dataset and vice versa. The summary of the new dataset is shown in Table 4.

Then, ED, MP, SAX, and MP-SAX are respectively used as similarity measurement algorithms to classify the new dataset. The accuracy of the classification is shown in Fig. 11.

As illustrated in Fig. 11, the ED cannot achieve high and stable classification accuracy due to the influence of the random amplitude translation. The MP can also not achieve the highest classification accuracy mentioned above as different kinds of detail component are added to the dataset

TABLE 4. The summary of the new dataset.

Length of time series	Size of training set	Size of testing set	Number of classes
300	995	25	18

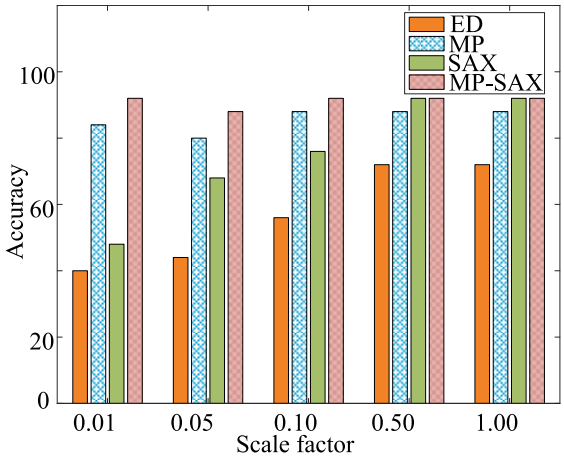


FIGURE 11. The classification accuracy of the new dataset.

Symbols. Meanwhile, the distribution of time series is mainly affected by the distribution of trend component when the scale of detail component is small. Therefore, the SAX can only distinguish the kinds of different trend component, but cannot distinguish the kinds of different detail component. However, as the scale of detail component increases, the distribution of time series is affected by both trend component and detail component. Thus the classification accuracy of SAX increases with the increase of scale factor. However, the MP-SAX maintains the highest classification accuracy under the different scaling detail component and different amplitude translation. Besides, the classification accuracy of the MP-SAX is relatively stable. Therefore, the MP-SAX cannot only be effectively used to measure the similarity of the time series with differences in trend and detail component, but also the results cannot be affected by different quantities of amplitude translation.

D. THE DISCUSSION OF PARAMETERS

The main parameters of the MP-SAX are the threshold of the MPE and the weights of the trend and detail components. Therefore, in this section, the influence of these parameters on the results of similarity measurement is discussed, respectively.

The dataset Symbols in the Section IV.B (represented by time series in Section IV.B) and the dataset with scale factor 0.01 in the Section IV.C (represented by time series in Section IV.C) is applied to illustrate the influence of the parameters on the results respectively. Where the time series in Section IV.B is the time series with difference in trend component and the time series in Section IV.C is the time series with difference both in trend and detail component. Then, the classification accuracies of the two datasets are used to illustrate the effect of the similarity measurement. So, the classification accuracy of different thresholds of MPE and the methods of determining weight (Subjective method and Objective method) is calculated respectively. Where the weight of the trend component is 0.75 and that of detail component is 0.25 when the weight is obtained by subjective method as mentioned above. The classification accuracy under different parameters is shown in Table 5.

In theory, the similarity measurement of the proposed method is that of symbolic aggregate approximation when the threshold of MPE is 0. And the similarity measurement of the proposed method is that of morphological pattern when the threshold of MPE is 1. However, in the proposed method, the trend component is reconstructed by the IMFs which the MPE of them is less than the threshold of MPE. Therefore, the similarity measurement of the proposed method may be different from that of morphological pattern when the threshold of MPE is 1. So, the classification accuracy of the time series in Section IV.B is different between the two methods of determining weight in Table 5.

As shown in Table 5, the classification accuracy is mainly affected by the threshold of the MPE. (a). For the time series with the difference only in trend or detail component, the best results may be obtained when the threshold of MPE tends to 0 or 1. Take the time series in Section IV.B as an example. The more IMFs will be reconstructed to the trend component when the threshold of MPE tends to 1. Therefore, the reconstructed trend component will contain more information of the original time series. So, the classification accuracy increases generally when the threshold of MPE tends to 1. (b). For the time series with the difference both in the trend and detail component, the best results may be obtained when the threshold of MPE tends to middle. This is because the similarity of trend and detail component is both taken into account. Altogether, an ideal result can be obtained for the time series with any characteristics when the threshold of MPE is 0.4. This further proves the effectiveness of the threshold 0.4.

Besides, the classification accuracy can be improved by the weight of the trend and detail component. (a). For the time series with the difference only in trend or detail component, the results can be further improved by the subjective method which considers the characteristics of time series. Take the time series in Section IV.B as an example. The classification accuracy of the subjective method which highlights the proportion of trend component is higher than that of objective method generally. (b). For the time series with the difference both in the trend and detail component, the objective method is better than the subjective method. This is because the objective method can determine the weights based on the proportion of the information which is contained in the reconstructed trend and detail component. So, the important component can be highlighted by the objective method. Altogether, the subjective method is applicable to the time series with the difference only in trend or detail component, while

TABLE 5. The classification accuracy under different parameters.

The threshold of MPE	Time series in Section IV.B		Time series in Section IV.C	
	Subjective method	Objective method	Subjective method	Objective method
0.0	92.16%	92.16%	48.00%	48.00%
0.1	92.46%	91.76%	56.00%	60.00%
0.2	81.31%	81.61%	60.00%	68.00%
0.3	92.06%	89.45%	88.00%	96.00%
0.4	93.97%	93.17%	92.00%	100.00%
0.5	94.67%	93.67%	88.00%	92.00%
0.6	95.08%	94.87%	96.00%	92.00%
0.7	95.08%	95.98%	88.00%	88.00%
0.8	88.14%	95.98%	80.00%	84.00%
0.9	92.56%	92.86%	76.00%	80.00%
1.0	96.68%	97.09%	84.00%	84.00%

the objective method is applicable to the time series with the difference both in trend and detail component.

Therefore, comprehensive analysis of the above simulation and real datasets, the MP-SAX can measure the similarity of time series more effectively than Euclidean distance, morphological pattern, and symbolic aggregate approximation when the differences between time series exist in the trend component, detail component, and amplitude translation simultaneously.

V. CONCLUSIONS

This paper proposes a new method for similarity measurement of time series with the differences both in trend component and detail component. The MP-SAX can comprehensively consider the influence of the trend component and detail component of time series on similarity measurement. Using EMD, the original time series are decomposed and reconstructed into trend component and detail component. Then, the similarity of the trend component under morphological pattern coding and that of the detail component under symbolic aggregate approximation coding is respectively calculated by LCS. Finally, the similarity of the original time series can be obtained by weighted aggregation of the similarity of trend component and detail component. Through the verification of the simulation time series and the real time series from UCR Time Series Classification / Clustering Homepage, it is proved that the MP-SAX can effectively measure the similarity of the time series with the differences both in trend component and detail component. The results of similarity measurement are mainly affected by the threshold of MPE, while the results can be improved by the weight of trend and detail component. Besides, the application scope of MP-SAX is larger than that of the MP and SAX. And it is unnecessary to know the characteristics of time series before the similarity measurement.

Although some important problems associated with the proposed method have been investigated in this paper, there are still a few questions that are worthy of further consideration. Frist, due to the limitation of EMD, the original trend and detail of the time series cannot be perfectly represented by the reconstructed trend and detail component. Second, as the proposed method needs to compute the similarity of SAX and MP at the same time, for the high-dimensional time series, the proposed method may cost more time. Therefore, the two questions deserve further study.

REFERENCES

[1] V. Kurbalija, M. Radovanović, Z. Geler, and M. Ivanović, "The influence of global constraints on similarity measures for time-series databases," *Knowl.-Based Syst.*, vol. 56, no. 3, pp. 49–67, Jan. 2014.

[2] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in *Proc. Int. Conf. Found. Data Org. Algorithms (ICFDOA)*. Berlin, Germany: Springer-Verlag, 1993, pp. 69–84.

[3] S. Ruzic, A. Vuckovic, and N. Nikolic, "Weather sensitive method for short term load forecasting in electric power utility of Serbia," *IEEE Trans. Power Syst.*, vol. 18, no. 4, pp. 1581–1586, Nov. 2003.

[4] J. P. Caraça-Valente and I. López-Chavarrías, "Discovering similar patterns in time series," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2000, pp. 497–505.

[5] R. R. Isnanto, A. A. Zahra, and E. D. Widianto, "Palmprint recognition system based on principle-lines feature using euclidean distance and neural network," in *Proc. 2nd Int. Conf. Inf. Technol., Comput., Elect. Eng. (ICITACEE)*, Oct. 2016, pp. 153–158.

[6] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 358–386, 2005.

[7] T. Górecki, "Using derivatives in a longest common subsequence dissimilarity measure for time series classification," *Pattern Recognit. Lett.*, vol. 45, pp. 99–105, Aug. 2014.

[8] C. Lei, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proc. ACM SIGMOD*, Jun. 2005, pp. 491–502.

[9] L. Chen and R. Ng, "On the marriage of Lp-norms and edit distance," in *Proc. 30th ICVLDB*, Sep. 2004, pp. 792–803.

[10] C.- J. Hsu, K.-S. Huang, C.-B. Yang, and Y.-P. Guo, "Flexible dynamic time warping for time series classification," *Procedia Comput. Sci.*, vol. 51, pp. 2838–2842, Jan. 2015.

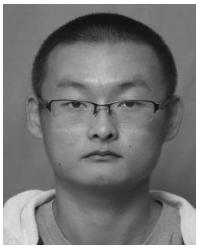
[11] D. Folgado, M. Barandas, R. Matias, R. Martins, M. Carvalho, and H. Gamboa, "Time alignment measurement for time series," *Pattern Recognit.*, vol. 81, pp. 268–279, Sep. 2018.

[12] D. F. Silva, R. Giusti, E. Keogh, and G. E. A. P. A. Batista, "Speeding up similarity search under dynamic time warping by pruning unpromising alignments," *Data Mining Knowl. Discovery*, vol. 32, no. 4, pp. 988–1016, Jul. 2018.

[13] L. Tao, C. Lu, and C. Yang, "Battery capacity degradation prediction using similarity recognition based on modified dynamic time warping," *Struct. Control Health Monit.*, vol. 25, no. 3, p. e2024, Jan. 2018.

[14] N. Putpuek, N. Cooharojananone, and S. Satoh, "A modification of retake detection using simple signature and LCS algorithm," in *Proc. 18th IEEE/ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput. (SNPD)*, Jun. 2017, pp. 257–261.

[15] Y. Rivault, N. Le Meur, and O. Dameron, "A similarity measure based on care trajectories as sequences of sets," in *Proc. 16th CAIM*. Cham, Switzerland: Springer, 2017, pp. 278–282.
[16] L. A. K. Ayad, C. Barton, and S. P. Pissis, "A faster and more accurate heuristic for cyclic edit distance computation," *Pattern Recognit. Lett.*, vol. 88, pp. 81–87, Mar. 2017.
[17] D. Zhang, W. Zuo, D. Zhang, H. Zhang, and N. Li, "Classification of pulse waveforms using edit distance with real penalty," *EURASIP J. Adv. Signal Process.*, vol. 2010, Dec. 2010, Art. no. 303140.
[18] M. Zhang and D. Pi, "A novel method for fast and accurate similarity measure in time series field," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 569–576.
[19] C. D. Stylios and V. Kreinovich, "Symbolic aggregate approXimation (SAX) under interval uncertainty," in *Proc. NAFIPS*, Aug. 2015, pp. 1–7.
[20] R. Xiao and G. H. Liu, "Research on trend-based time series similarity measure and cluster," *Appl. Res. Comput.*, vol. 31, no. 9, pp. 253–256, 2014.
[21] T. Nakamura, K. Taki, H. Nomiya, K. Seki, and K. Uehara, "A shape-based similarity measure for time series data with ensemble learning," *Pattern Anal. Appl.*, vol. 16, no. 4, pp. 535–548, Nov. 2013.
[22] K. Tamura and T. Ichimura, "Clustering of time series using hybrid symbolic aggregate approximation," in *Proc. IEEE SSCI*, Nov./Dec. 2018, pp. 1–8.
[23] A. González-Vidal, P. Barnaghi, and A. F. Skarmeta, "BEATS: Blocks of eigenvalues algorithm for time series segmentation," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 11, pp. 2051–2064, Nov. 2018.
[24] G. Baldini, R. Giuliani, G. Steri, I. Sanchez, and C. Gentile, "The application of the symbolic aggregate approximation algorithm (SAX) to radio frequency fingerprinting of IoT devices," in *Proc. IEEE SCVT*, Nov. 2017, pp. 1–6.
[25] Z. Wang and Z. Tang, "Study of time series similarity measure based on fluctuate pattern," *Appl. Res. Comput.*, vol. 34, no. 3, pp. 697–701, 2017.
[26] E. B. Dagum and S. Bianconcini, *Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation*. Cham, Switzerland: Springer, 2016, pp. 29–57.
[27] R. Wang and R. Jia, "Similarity measure algorithm of time series based on morphological pattern," *Comput. Appl. Softw.*, vol. 34, no. 9, pp. 253–256, 2017.
[28] W. J. Hsu and M. W. Du, "Computing a longest common subsequence for a set of strings," *BIT Numer. Math.*, vol. 24, no. 1, pp. 45–59, 1984.
[29] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proc. ACM SIGMOD*, Jun. 2003, pp. 2–11.
[30] X. He, C. Shao, and X. Yan, "A non-parametric symbolic approximate representation for long time series," *Pattern Anal. Appl.*, vol. 19, no. 1, pp. 111–127, Feb. 2016.
[31] Y. Sun, J. Li, J. Liu, B. Sun, and C. Chow, "An improvement of symbolic aggregate approximation distance measure for time series," *Neurocomputing*, vol. 138, no. 11, pp. 189–198, Aug. 2014.
[32] D. Yu, J. Cheng, and Y. Yang, "Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings," *Mech. Syst. Signal Process.*, vol. 19, no. 2, pp. 259–270, Mar. 2005.
[33] Y. Li, M. Xu, Y. Wei, and W. Huang, "An improvement EMD method based on the optimized rational Hermite interpolation approach and its application to gear fault diagnosis," *Measurement*, vol. 63, pp. 330–345, Mar. 2015.
[34] Y. Li, M. Xu, Y. Wei, and W. Huang, "A new rolling bearing fault diagnosis method based on multiscale permutation entropy and improved support vector machine based binary tree," *Measurement*, vol. 77, pp. 80–94, Jan. 2016.



JIANCHENG YIN received the B.S. degree in mechanics from the Harbin Institute of Technology, Harbin, China, in 2014, where he is currently pursuing the Ph.D. degree in mechanics with the Deep Space Exploration Research Center.

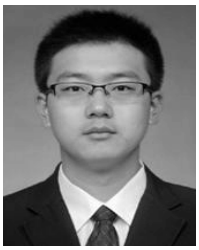
His research interests include fault diagnosis of machinery, residual life prediction method, and signal processing.

RIXIN WANG received the B.E. degree in computer science from the Harbin University of Science and Technology, Harbin, China, in 1985, and the M.E. degree in computer science and the Ph.D. degree in spacecraft design from the Harbin Institute of Technology, Harbin, in 1991 and 2003, respectively, where he is currently an Associate Professor with the Department of Engineering Mechanics.

His research interests include fault detection and diagnosis for machinery and spacecraft.

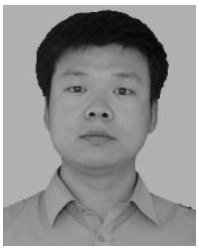
HUAILIANG ZHENG received the B.S. and M.S degrees in mechanics from the Harbin Institute of Technology, Harbin, China, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the Deep Space Exploration Research Center.

His research interests include fault diagnosis of machinery, intelligent fault diagnosis method, and transfer learning.



YUANTAO YANG received the B.S. and M.S degrees in mechanics from the Harbin Institute of Technology, Harbin, China, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the Deep Space Exploration Research Center.

His research interests include fault diagnosis of machinery, intelligent fault diagnosis method, and deep learning.



YUQING LI received the B.E. degree in mechanical design manufacturing and automation, and the M.E. and Ph.D. degrees in general mechanics from the Harbin Institute of Technology, Harbin, China, in 2002, 2004, and 2008, respectively, where he is currently an Associate Professor.

His main research interests include planning and scheduling of spacecraft and spacecraft fault detection and diagnosis.

MINQIANG XU received the B.E. degree in electronics from Peking University, Beijing, China, in 1983, the M.E. degree in nuclear physics from Northeast Normal University, Changchun, China, in 1989, and the Ph.D. degree in general and fundamental mechanics from the Harbin Institute of Technology, Harbin, China, in 1999.

Since 2000, he has been a Professor with the Harbin Institute of Technology. His research interests include machinery and spacecraft fault detection and diagnosis, signal processing, and space debris modeling.