ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20200107171121
Original scientific paper

# A New Time Series Similarity Measurement Method Based on Fluctuation Features

Hailan CHEN, Xuedong GAO

**Abstract:** Time series similarity measurement is one of the fundamental tasks in time series data mining, and there are many studies on time series similarity measurement methods. However, the majority of them only calculate the distance between equal-length time series, and also cannot adequately reflect the fluctuation features of time series. To solve this problem, a new time series similarity measurement method based on fluctuation features is proposed in this paper. Firstly, the fluctuation features extraction method of time series is introduced. By defining and identifying fluctuation points, the fluctuation points sequence is obtained to represent the original time series for subsequent analysis. Then, a new similarity measurement (D_SM) is put forward to calculate the distance between different fluctuation points sequences. This method can calculate the distance of unequal-length time series, and it includes two main steps: similarity matching and the distance calculation based on similarity matching. Finally, the experiments are performed on some public time series using agglomerative hierarchical clustering based on D_SM. Compared to some traditional time series similarity measurements, the clustering results show that the proposed method can effectively distinguish time series with similar shapes from different classes and get a visible improvement in clustering accuracy in terms of F-Measure.

**Keywords:** clustering; fluctuation features; similarity measurement; time series

## 1 INTRODUCTION

Time series data come up in a variety of domains [1], including financial data [2], environmental data [3, 4], telecommunication data [5], and medical data [6]. A time series is a set of observations arranged in sequence according to the occurrence of time [7]. Data mining is a mechanism of discovering hidden knowledge from a spurious amount of data, and aims to extract interesting rules or patterns from huge databases [8]. It includes some research directions: clustering [9], similarity search [10], classification [11] and prediction [12]. Among them, clustering plays an essential role in the field of data mining, and it can also be considered as a pre-processing stage for other data mining tasks, especially decision making [13, 14].

Time series similarity measurement is the basis for time series clustering [15], which is used to calculate the distance of two time series. The common time series similarity measurements are divided as follows [16, 17]: Euclidean distance (ED),Dynamic Time Warping (DTW) distance, segmented representation distance, symbolic distance, model distance, and compression distance.

In time series similarity measurement, Euclidean distance [18] is the most commonly used distance, but it cannot measure the time series of unequal-length. In 1994, Berndt and Clifford [19] firstly introduced DTW distance widely used in speech recognition to the study of time series similarity measurement. DTW allows a certain degree of offset on the time axis, and can measure time series of unequal-length. However, it has a disadvantage that the computational complexity of DTW is high, and it cannot reach the requirement of the distance triangle inequality [20].

The similarity measurement based on segmented representation distance, such as Piecewise Linear Approximation (PLA) [21], Piecewise Aggregate Approximation (PAA) [22], and Derivative Segment Approximation (DSA) [23], segments the long time series into several short sequences and uses the features of segmented sequences to represent the original time series. PLA uses many short segment sequences, which results in a rough representation and approximate degree, so the representation is not accurate. PAA needs to define the indicator of dimensionality reduction in advance. Moreover, the segmented sequences are of fixed length, and are represented as the mean value, ignoring important information such as the shape changes and key points of time series.

The similarity measurement based on symbolic distance converts the original time series into a string sequence, and then calculates the distance between string sequences. Symbolic aggregate approximation (SAX) [24] is the most typical symbolic representation method. Because SAX is a symbolic representation method based on PAA, it also inherits the shortcomings of PAA.

The similarity measurement based on model distance includes Auto-Regressive Model (AR) [25], Auto-Regressive and Moving Average Model (ARMA) [26], and Hidden Markov Model (HMM) [27]. This method describes the original time series by solving the appropriate parameter to fit model, and then expresses the distance between the parameters as a similarity index. The disadvantage of the model-based method is that time series needs to be defined in advance to satisfy certain assumptions.

Compression-based Dissimilarity Measure (CDM) [28] is one of the similarity measurements based on compression distance. It combines the results of bioinformatics and compression theory, and is suitable for the measurement and discovery of subsequence similarity. However, the calculation process of this method is very complicated and time-consuming. Many parameters need to be set correctly and reasonably, so its application is limited.

To avoid the above shortcomings in the above time series similarity measurement methods, in this paper, a new time series similarity measurement method is proposed based on fluctuation features. On the one hand, this method can extract the fluctuation features of time series to reduce the dimension of the original time series. On the other hand, it can also calculate the distance between unequal-length time series.

The chapter structure of this article is as follows: Section 2 introduces the method of fluctuation points identification to represent the fluctuation features of time series. Section 3 proposes a new similarity measurement to calculate the distance of fluctuation points sequences. Section 4 agglomeration hierarchical clustering based on the proposed distance is used to perform experiments on some public time series and analyze the clustering results. Section 5 provides conclusions.

## 2 FLUCTUATION FEATURES EXTRACTION METHOD OF TIME SERIES

In this section, the fluctuation features extraction method of time series is proposed, which mainly includes three steps: identification of extreme points, selection of extreme points, and determination of fluctuation points. After completing the above steps, the fluctuation points are got to represent the fluctuation features of time series.

### 2.1 Identification of Extreme Points

In this subsection, we firstly introduce the definition of an extreme point.

Definition 1 (Extreme Point) Given a time series $X = \{x_1, x_2, \ldots, x_n\}$, if there is a relationship $(x_i - x_{i-1})(x_{i+1} - x_i) < 0$ among three adjacent points $x_{i-1}, x_i, x_{i+1}$ in time series $X$ , the point $x_i$ is called an Extreme Point.

Where $i = 2, 3, \ldots, n-1$, the starting and ending points of a time series are also considered as extreme points.

A time series is used as an example for discussion. According to the above definition, extreme points can be identified (Fig. 1).

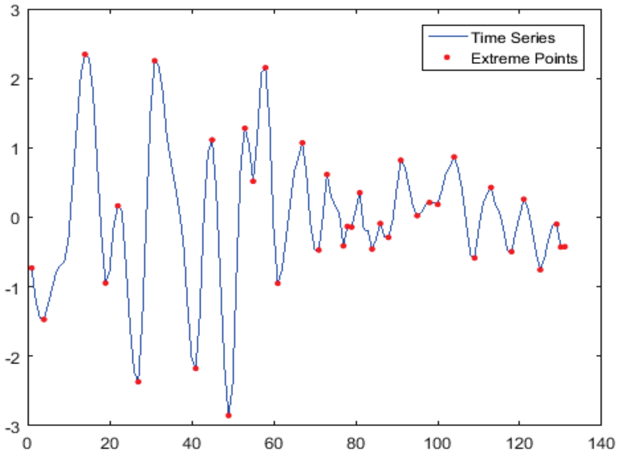


**Figure 1** Identification of Extreme Points in a time series

After obtaining the extreme points, the attributes of them are marked, where the attribute of the maximum point is 1 and the attribute of the minimum point is −1.

### 2.2 Selection of Extreme Points

In Fig. 1, we can see that some extreme point distributions are concentrated, and the fluctuations are relatively small in some subsequences.

Definition 2 (Candidate Fluctuation Point) For an extreme points sequence $E = \{e_1, e_2, \ldots, e_m\}$, given a threshold set $\varepsilon = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_q\}$, if there is a relationship $|e_j - e_{j-1}| > \varepsilon_k$ between two adjacent points $e_{j-1}$, $e_j$ in the sequence $E$, the point $e_j$ is called a candidate fluctuation point.

Where $j = 2, 3, \ldots, m$, $\varepsilon_k$ is a certain threshold in the threshold set $\varepsilon$, and the starting point of a time series is also considered as a candidate fluctuation point.

According to the above definition, the extreme points with small changes are filtered to obtain candidate fluctuation points (Fig. 2).

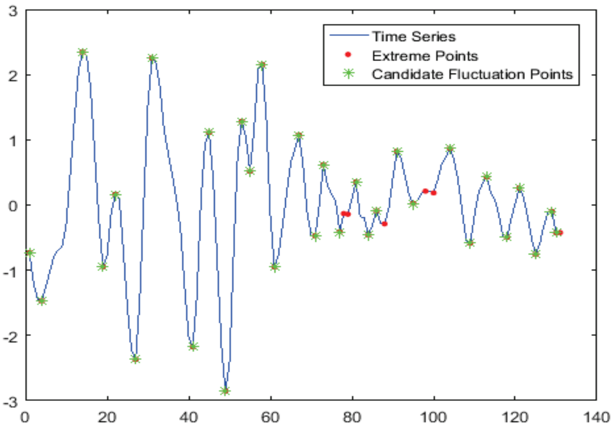


**Figure 2** Selection of extreme points in a time series

### 2.3 Determination of Fluctuation Points

Definition 3 (Fluctuation Point) For a candidate fluctuation points sequence $C = \{c_1, c_2, \ldots, c_p\}$, if there is a relationship $Attrc_{z-1}Attrc_z = -1$ between two adjacent points $c_{z-1}$, $c_z$ in the sequence $C$, the point $c_z$ is called a fluctuation point.

Where $z = 2, 3, \ldots, p$, $Attrc_{z-1}Attrc_z = -1$ represents the attributes of $c_{z-1}$ and $c_z$ are opposite, that is, one point is maximum and the other one is minimum. Meanwhile, the starting point of a time series is also considered as a fluctuation point.

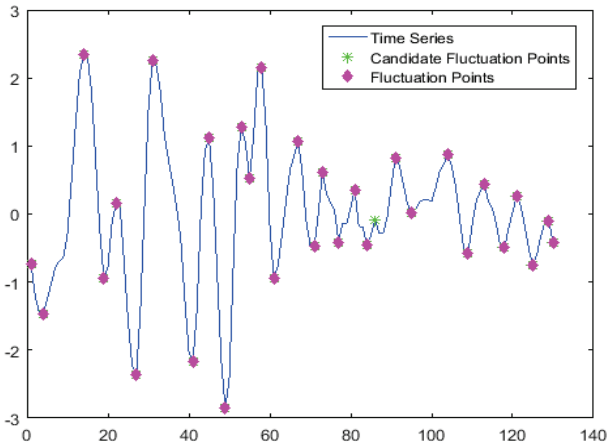


**Figure 3** Determination of Fluctuation Points in a time series

The candidate fluctuation points inherit the attributes of extreme points.The product of the attributes is −1 between the two adjacent points in the extreme point sequence. However, after deleting some extreme points with small changes, the outcome of the attributes may be 1between the two adjacent points in the candidate fluctuation point sequence. So further operation is needed to get the fluctuation points. For $Attrc_{z-1}Attrc_z = 1$, there are the following two cases, and the corresponding operations are as follows:

(1) If the attributes of two adjacent points are 1, meaning that they are both maximum points, delete the minimum point of them;

(2) If the attributes of two adjacent points are −1, meaning that they are both minimum points, remove the maximum point of them.

For candidate fluctuation points, fluctuation points can be obtained (Fig. 3) by operating according to the above corresponding cases.

In summary, the flowchart of the fluctuation points identification algorithm is given in Fig. 4.

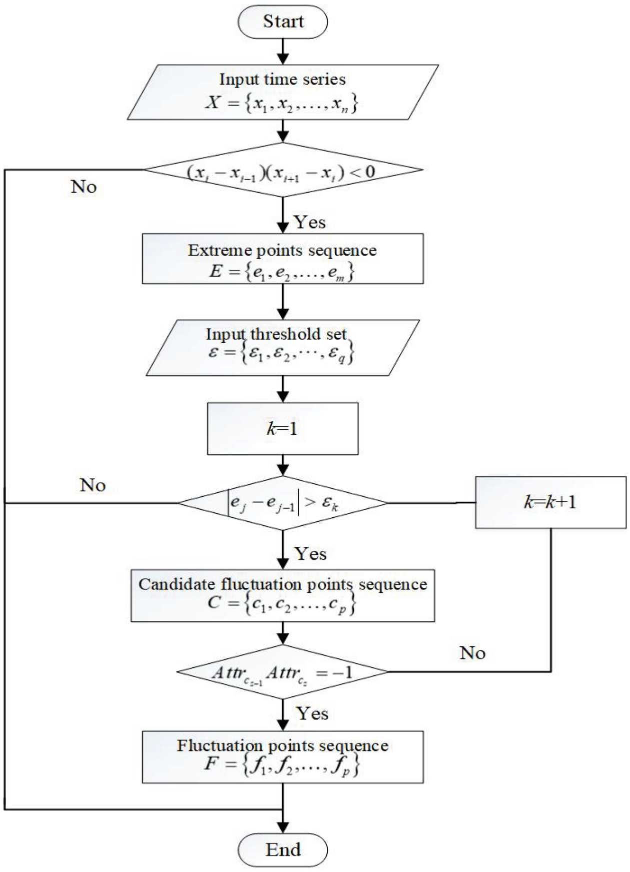

Start
Input time series
$X=\{x_1,x_2,\ldots,x_n\}$
$(x_i-x_{i-1})(x_{i+1}-x_i)<0$
No
Yes
Extreme points sequence
$E=\{e_1,e_2,\ldots,e_m\}$
Input threshold set
$\varepsilon=\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_q\}$
$k$=1
No
$|e_j-e_{j-1}|>\varepsilon_k$
$k$=$k$+1
Yes
Candidate fluctuation points sequence
$C=\{c_1,c_2,\ldots,c_p\}$
No
$Attr_{c_{z-1}}Attr_{c_z}=-1$
Yes
Fluctuation points sequence
$F=\{f_1,f_2,\ldots,f_p\}$
End

**Figure 4** Flowchart of fluctuation points identification algorithm

Next, we discuss the time complexity of this algorithm. The time complexity of identifying extreme points of a time series would be $O(n)$, and the computational burden for identifying fluctuation points based on extreme points would be $O(mpq)$. Therefore, the total time complexity is $O(n) + O(mpq)$.

Where $n$ is the number of data points in the time series, $m$ is the number of extreme points, $p$ is the number of candidate fluctuation points, and $q$ is the number of thresholds in the threshold set. Obviously, $m \leq n$, $p \geq 1$, $q \geq 1$.

The sequence which consists only of fluctuation points is called fluctuation points sequence and represents the fluctuation features of time series. For fluctuation points sequence $F = \{f_1, f_2, \ldots, f_p\}$, a point of $F$ is represented as $f_i = (t_i, v_i, Attr_i)$.

## 3 TIME SERIES SIMILARITY MEASUREMENT BASED ON FLUCTUATION FEATURES

The traditional time series similarity measurements are usually used to measure equal-length sequences, such as Euclidean distance, which is a point-to-point calculation method. DTW distance can be used to measure the time series of unequal-length, but its computational complexity is high and always leads to over warping. As the calculation method is not point-to-point, it cannot satisfy the triangular inequality of distance.

Since the length between any two fluctuation points sequences is usually unequal, and they also correspond to different time points, in this section, a new time series similarity measurement is proposed for unequal-length time series. It includes two main steps: similarity matching and the distance calculation based on similarity matching.

### 3.1 Similarity Matching

For time series, most of the traditional similarity measurements belong to point-to-point matching mode, which requires the matching points to have the same time stamp. This matching mode is defined as precise matching.

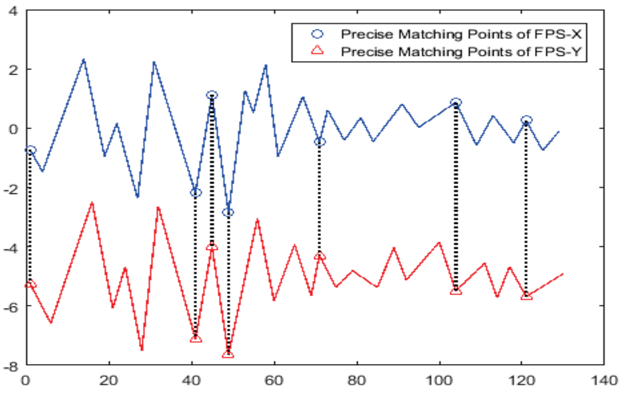


**Figure 5** Precise matching of two fluctuation points sequences

Using this method to match two fluctuation points sequences, as shown in Fig. 5, the points from two sequences are few matched. Although some matched points have the same timestamp, their attributes are different. The specific information of precise matching points is shown in Tab. 1. Fluctuation points sequence X (FPS-X) has 29 time points, and fluctuation points sequence Y (FPS-Y) has 27 time points. However, after precise matching, there are only seven pairs of matched points, and the last three pairs have the opposite attributes. In fact, the two fluctuation points sequences are very similar in shape.

**Table 1** Precise matching points of two fluctuation points sequences

| $x_i = (t_i, v_i, Attr_i)$ | | | $y_j = (s_j, u_j, Attr_j)$ | | |
|---|---|---|---|---|---|
| $t_i$ | $v_i$ | $Attr_i$ | $s_j$ | $u_j$ | $Attr_j$ |
| 1 | −0.73 | 1 | 1 | −0.25 | 1 |
| 41 | −2.18 | −1 | 41 | −2.11 | −1 |
| 45 | 1.11 | 1 | 45 | 1.00 | 1 |
| 49 | −2.85 | −1 | 49 | −2.65 | -1 |
| 71 | −0.47 | −1 | 71 | 0.70 | 1 |
| 104 | 0.87 | 1 | 104 | −0.49 | −1 |
| 121 | 0.26 | 1 | 121 | −0.69 | −1 |

Precise matching is too rigid, which will not effectively match the time series with a similar shape, while DTW allows a certain degree of offset on the time axis. Therefore, we take advantage of DTW and propose the following similarity matching method. On the one hand, this method allows the time series to be matched to have a certain degree of offset on the time axis. On the other hand, it meets each pair of matched points in a one-to-one relationship, that is, it satisfies the triangle inequality of distance.

Given two fluctuation points sequences $X = \{x_1, x_2, \ldots, x_m\}$; $(x_i = (t_i, v_i, Attr_i))$ and $Y = \{y_1, y_2, \ldots, y_n\}$; $(y_j = (s_j, u_j, Attr_j))$ of equal or unequal length, it needs to meet both of the following conditions to successfully match. This matching mode is defined as similarity matching.

Condition 1:

$$t_i \in \left[s_j - \varepsilon, s_j + \varepsilon\right]$$

Condition 2:

$$Attr_i = Attr_j$$

where $t_i$ and $s_j$ represent the time stamp of $x_i$ and $y_j$, $v_i$ and $u_j$ represent the value of $x_i$ and $y_j$, $Attr_i$ and $Attr_j$ represent the attribute of $x_i$ and $y_j$, $\varepsilon$ is the threshold that is used to control how much time the axis is allowed to shift.

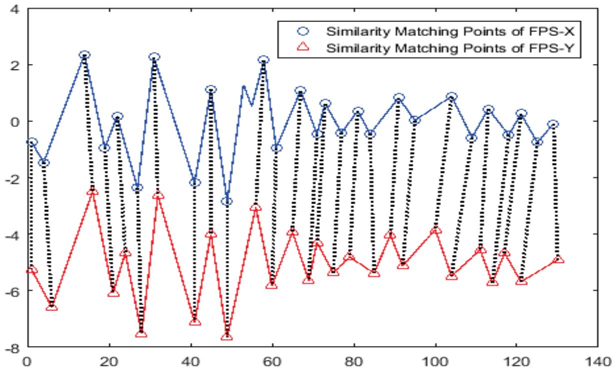


**Figure 6** Similarity matching of two fluctuation points sequences

In Fig. 6, the points marked by circle and triangle are the points after similarity matching. For fluctuation points sequence $X$, there are only two unmatched points. And all points in fluctuation points sequence $Y$ are matched. Compared with precise matching, this matching method improves efficiency.

### 3.2 The Distance Calculation Method based on Similarity Matching

After similarity matching, we need to calculate the distance of two fluctuation points sequences. Firstly, the distance between each pair matched points are calculated by Eq. (1).

$$d(x_i, y_j) = \sqrt{(v_i - u_j)^2} \tag{1}$$

Where $x_i$ and $y_j$ is a pair of matched points, $t_i$ and $s_j$ represent the timestamp of $x_i$ and $y_j$, $v_i$ and $u_j$ represent the value of $x_i$ and $y_j$.

Next, in this subsection, some concepts are proposed, such as fluctuation degree, information weight, and the distance calculation method based on similarity matching is given as well.

Definition 3 (Fluctuation Degree， FD) Given the fluctuation points sequences $X = \{x_1, x_2, \ldots, x_m\}$ $(x_i = (t_i, v_i, Attr_i))$, suppose there are three consecutive points $x_{i-1}$, $x_i$, $x_{i+1}$, the fluctuation degree of $x_i$ is

$$FD_{x_i} = \frac{\left|v_{i+1} - v_i\right| + \left|v_i - v_{i-1}\right|}{2} \tag{2}$$

where $i = 2, 3, \ldots, m$, especially, $FDx_1 = |v_2 - v_1|$ and $FDx_m = |v_m - v_{m-1}|$.

Definition 4 (Information Weight， IW) Given the fluctuation points sequences $X = \{x_1, x_2, \ldots, x_m\}$ $(x_i = (t_i, v_i, Attr_i))$, suppose there are three consecutive points $x_{i-1}$, $x_i$, $x_{i+1}$,, the information weight of $x_i$ is

$$IW_{x_i} = \frac{FD_{x_i}}{\sum FD_{x_i}} \tag{3}$$

Definition 5 (Similarity Matching Degree， SMD) Given two fluctuation points sequences $X = \{x_1, x_2, \ldots, x_m\}$ $(x_i = (t_i, v_i, Attr_i))$ and $Y = \{y_1, y_2, \ldots, y_n\}$ $(y_j = (s_j, u_j, Attr_j))$. $\sum\limits_X IW_{x_i}$ is the sum of information weights for matched points of $X$, and $\sum\limits_Y IW_{y_i}$ is the sum of information weights for matched points of $Y$. Then the similarity matching degree of $X$ and $Y$ is

$$SMD_{X,Y} = \min\left(\sum_X IW_{x_i}, \sum_Y IW_{y_j}\right) \tag{4}$$

Then the distance of $X$ and $Y$ based on similarity matching is:

$$d\left(X, Y\right) = \frac{1}{SMD_{X,Y}} \sum d\left(x_\mathrm{i}, y_j\right) \tag{5}$$

where $n$ is the number of matched points.

Based on the given concepts, we summarise the distance calculation method based on similarity matching (D_SM), and the algorithm steps are as follows.

```
Algorithm 1The Distance Calculation Method based on Similarity Matching (D_SM)
Input: Fluctuation points sequences $X$ and $Y$.
Output: The distance between $X$ and $Y$.
Step 1: Similarity matching. Set the neighbourhood to $\varepsilon$, if $t_i \in [s_j - \varepsilon, s_j + \varepsilon]$ and $Attr_i = Attr_j$, $x_i$ of $X$ and $y_j$ of $Y$ will be matched. As a result, the similarity matching sequence is achieved;
Step 2: After similarity matching, the distance between each pair matched points is calculated by Eq. (1).
```

**Step 3:** Calculate *FD* of *X* and *Y* according to Eq. (2), and then calculate *IW* according to Eq. (3). At last, calculate *SMD* of *X* and *Y* using Eq. (4).
**Step 4:** According to Eq. (5), calculate the distance of *X* and *Y*.

The above similarity measurement algorithm is not only suitable for the similarity measurement between fluctuation points sequences, but also for the similarity measurement between unequal-length time series.

## 4 EXPERIMENTAL STUDIES

In this section, the time series clustering is used to demonstrate the performance of the proposed similarity measurement, and the clustering method we used is agglomerative hierarchical clustering. Because it does not need to predefine number of classes in advance, we can draw a dendrogram to visualize the clustering results. The experiments are carried on UCR time series [29].

### 4.1 Experiment 1 on Face All Dataset

In experiment 1, FaceAll dataset isused for clustering. As shown in Fig. 7, we randomly select nine time series from three classes. And the comparison similarity measurements are Euclidean distance and DTW distance. Fig. 8 shows the clustering results of different methods.

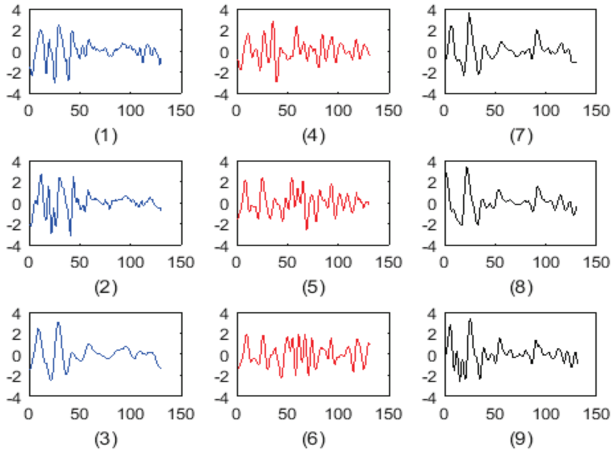
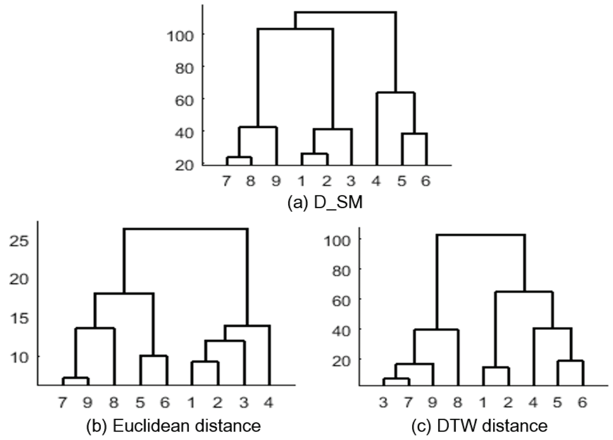


**Figure 7** Experimental time series of FaceAll dataset



(a) D_SM

(b) Euclidean distance

(c) DTW distance

**Figure 8** Clustering results of the dataset in Fig. 7

The dataset shown in Fig. 7 exhibits three clusters: $C_1 = \{1, 2, 3\}$, $C_2 = \{4, 5, 6\}$, $C_3 = \{7, 8, 9\}$. The first third of time series in cluster $C_1$ has obvious fluctuations, while the remaining length fluctuates slightly. For cluster $C_2$, the time series have significant fluctuations over its entire length. The shape of time series in cluster $C_3$ is similar to that in cluster $C_1$, but the last two-thirds length changes more obviously.

From Fig. 8, we can see that the clustering result based on Euclidean distance is {{1, 2, 3, 4}, {5, 6}, {7, 8, 9}}, and the clustering result based on DTW distance is {{1, 2}, {4, 5, 6}, {3, 7, 8, 9}}. However, the clustering result based on D_SM is {{1, 2, 3}, {4, 5, 6}, {7, 8, 9}}, and it is the same with the real classes division. This experimental result demonstrates that, for time series with similar shapes, the hierarchical clustering method based on D_SM can more accurately capture the fluctuation features of time series, and then effectively distinguish them.

### 4.2 Experiment 2 on Swedish Leaf Dataset

In experiment 2, we randomly select nine time series from three classes in SwedishLeaf dataset for clustering, see as in Fig. 9. And the comparison similarity measurements are Euclidean distance and DTW distance. Fig. 10 shows the clustering results of different methods.

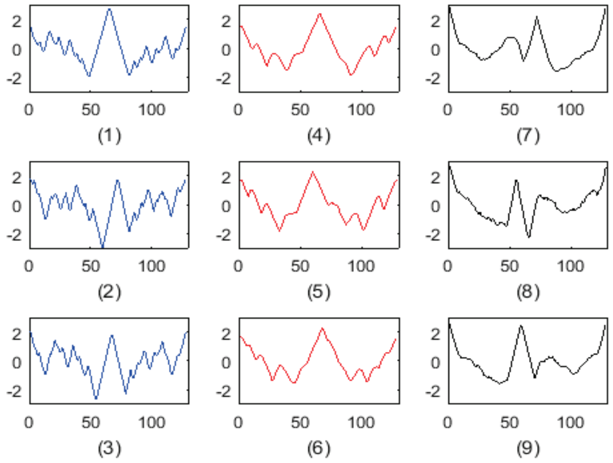
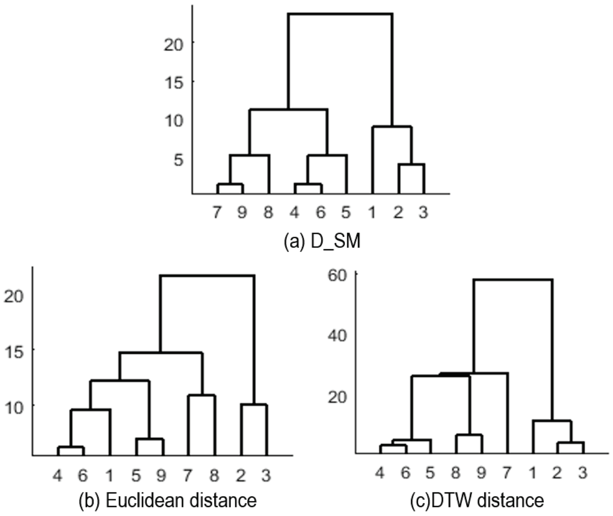


**Figure 9** Experimental time series of SwedishLeaf dataset



(a) D_SM

(b) Euclidean distance

(c)DTW distance

**Figure 10** Clustering results of the dataset in Fig. 9

In Fig. 9, there are three clusters: $C_1$ = {1, 2, 3}, $C_2$ = {4, 5, 6}, $C_3$ = {7, 8, 9}. The trend and shape of nine time series are very similar, and the difference between 3 classes is not very obvious.

From Fig. 10, it can be seen that the clustering result based on Euclidean distance is {{2, 3}, {1, 4, 5, 6, 9}, {7, 8}}, and the clustering result based on DTW distance is {{1, 2, 3}, {4, 5, 6}, {8, 9}, {7}}. However, D_SM based clustering method is used to divide nine time series into three classes: {{1, 2, 3}, {4, 5, 6}, {7, 8, 9}}. Obviously, only D_SM based clustering method can reveal correct clusters, while other clustering methods cannot produce correct clusters.

## 4.3 Experiment 3 on Synthetic Control Dataset

In experiment 3, we randomly select nine time series from three classes in Synthetic Control dataset for clustering, as shown in Fig. 11. And the comparison similarity measurements are Euclidean distance and DTW distance. Fig. 12 shows the clustering results of different methods.

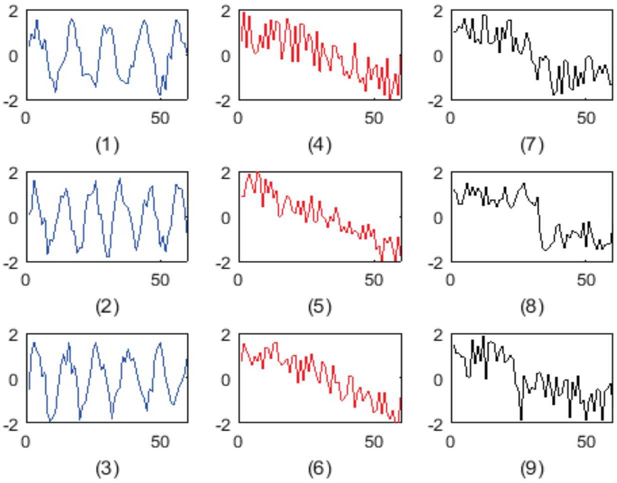
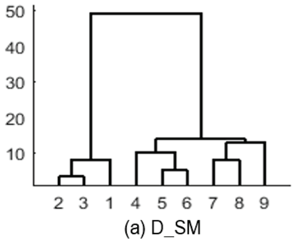
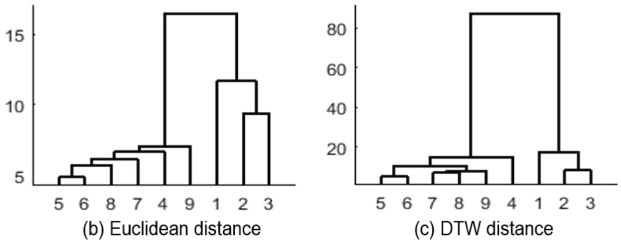


**Figure 11** Experimental time series of Synthetic Control dataset



(a) D_SM



(b) Euclidean distance (c) DTW distance

**Figure 12** Clustering results of the dataset in Fig. 11

As can be seen from Fig. 11, there are three clusters: $C_1$ = {1, 2, 3}, $C_2$ = {4, 5, 6}, $C_3$ = {7, 8, 9}. Cluster $C_1$ shows 'cyclic trend', Cluster $C_2$ exhibits 'decreasing trend', and Cluster $C_3$ reflects 'downward trend'.

As shown in Fig. 12, nine time series are clustered as {{1, 2, 3}, {4, 5, 6, 7, 8, 9}} using Euclidean distance, which is the same as the clustering result of DTW distance. These two methods failed to distinguish the last two classes. However, the clustering method based on D_SM is used to divide nine time series into {{1, 2, 3}, {4, 5, 6}, {7, 8, 9}}. The results are consistent with the real division.

## 4.4 Experiment 4 on Four UCR Datasets

In experiment 4, four datasets are conducted to experiment. Tab. 2 shows the basic information of these datasets.

The compared methods [30] are Dynamic Time Warping distance (DTW), Piecewise Linear Approximation based Dynamic Time Warping distance (PLA_DTW), Piecewise Aggregate Approximation based Dynamic Time Warping distance (PAA_DTW), Symbolic Aggregate Approximation based Dynamic Time Warping distance (SAX_DTW) and Derivative Segment Approximation based Dynamic Time Warping distance (DSA_DTW). And we also use the agglomerative hierarchical clustering method to do experiments based on these similarity measurements.

**Table 2** Basic information of UCR datasets in experiment 4

| Dataset | Samples | Length | Classes |
|---|---|---|---|
| Gun-Point | 50 | 150 | 2 |
| Trace | 100 | 275 | 4 |
| Synthetic Control | 300 | 60 | 6 |
| CBF | 30 | 128 | 3 |

In clustering analysis, *F-Measure* is usually used as an evaluation metric to validate the quality of the clustering results [31].

After clustering, we calculate the index of *F-Measure* and Tab. 3 presents the calculated results of 6 clustering methods.

As shown in Tab. 3, D_SM achieves the highest *F-Measure* on all datasets (except Trace). Especially in Synthetic Control dataset and CBF dataset, the *F-Measure* value of D-SM is much higher than that of other methods. However, in Trace dataset, the *F-Measure* value of D-SM is only 0.1 less than that of DSA. To make it easier to compare the clustering results of these methods, we rank the values of *F-Measure* for all of the approaches and then calculate their average ranking. It is obvious that the average ranking of D_SM is the highest.

**Table 3** F-Measure comparison of clustering base on six similarity measurements on four UCR datasets

| Dataset | DTW | PLA_ DTW | PAA_ DTW | SAX_ DTW | DSA_ DTW | D_SM |
|---|---|---|---|---|---|---|
| Gun-Point | 0.61 | 0.61 | 0.61 | 0.61 | 0.73 | 0.75 |
| Trace | 0.48 | 0.63 | 0.61 | 0.60 | 0.82 | 0.72 |
| Synthetic Control | 0.48 | 0.4 | 0.36 | 0.48 | 0.54 | 0.88 |
| CBF | 0.51 | 0.51 | 0.51 | 0.56 | 0.60 | 0.78 |
| Average Ranking | 4 | 3.75 | 4.25 | 3.5 | 1.75 | 1.25 |

Based on the above experimental results, the following conclusions can be summarised. In experiment 1, 2, and 3, our proposed similarity measurement (D_SM) can effectively distinguish time series with similar shapes from different classes compared to Euclidean distance and DTW distance. In experiment 4, compared to DTW, PLA_DTW, PAA_DTW, SAX_DTW, and DSA_DTW, D_SM achieves the highest average ranking on all datasets in terms of *F-Measure*, which means that D_SM has significantly improved the accuracy of the clustering results.

## 5 CONCLUSIONS

In this paper, we propose a new time series similarity measurement based on fluctuation features, which can calculate the distance between the unequal-length time series. Firstly, the fluctuation features extraction method of time series is proposed, which mainly includes three steps: identification of extreme points, selection of extreme points, and determination of fluctuation points. Completing the above steps, the fluctuation points are got to represent the fluctuation features of time series. After that, the fluctuation points sequence that consists of fluctuation points can replace the original time series. Since the length between any two fluctuation points sequences is usually unequal, a new similarity measurement (D_SM) is proposed to calculate the distance of fluctuation points sequences. This method allows the matched time series to have a certain degree of offset on the time axis, and it also satisfies the triangle inequality of distance. Finally, the experiments are performed on some public time series using agglomerative hierarchical clustering based on the proposed method. And the clustering results show that the proposed method can effectively distinguish time series with similar shapes from different classes and has significantly improved the clustering accuracy.

However, the fluctuation features extraction method of time series and the proposed time series similarity measurement both rely on setting reasonable thresholds. The solution to this problem is going to be solved and presented in future research.

## 6 REFERENCES

[1] Keogh, E. & Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, *7*(4), 349-371. https://doi.org/10.1023/A:1024988512476

[2] Chen, Y., Chen, Y., Tsao, S., & Hsieh, S. (2018). A novel technical analysis-based method for stock market forecasting. *Soft Computing*, *22*(4), 1295-1312. https://doi.org/10.1007/s00500-016-2417-2

[3] Burgan, H. I. & Aksoy, H. (2018). Annual flow duration curve model for ungauged basins. *Hydrology Research, 49*(5), 1684-1695. https://doi.org/10.2166/nh.2018.109

[4] Ong, B., Sugiura, K., & Zettsu, K. (2016). Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. *Neural Computing & Applications*, *27*(6), 1553-1566. https://doi.org/10.1007/s00521-015-1955-3

[5] Xu, F., Lin, Y., Huang, J., Wu, D., Shi, H., Song, J., & Li, Y. (2016). Big data driven mobile traffic understanding and forecasting: a time series approach. *IEEE Transactions on Services Computing*, *9*(5), 796-805. https://doi.org/10.1109/TSC.2016.2599878

[6] Li, S. (2017). Estimating time-dependent ROC curves using data under prevalent sampling. *Statistics in Medicine*, *36*(8), 1285-1301. https://doi.org/10.1002/sim.7184

[7] Fu, T. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), 164-181. https://doi.org/10.1016/j.engappai.2010.09.007

[8] Cao, X. & Li, Q. (2016). A RFID-based similarity cluster approach for detecting abnormal logistics paths and its performance evaluation. *International Journal of Information Technology & Management*, *15*(4), 387-400. https://doi.org/10.1504/IJITM.2016.079601

[9] Wang, X., Yu, F., Pedrycz, W., & Wang, J. (2019). Hierarchical clustering of unequal-length time series with area-based shape distance. *Soft Computing*, *23*(15), 6331-6343. https://doi.org/10.1007/s00500-018-3287-6

[10] Li, Z., Guo, J., Li, H., Wu, T., Mao, S., & Nie, F. (2019). Speed up similarity search of time series under dynamic time warping. *IEEE Access*, *2019*(7), 163644-163653. https://doi.org/10.1109/ACCESS.2019.2949838

[11] Abanda, A., Mori, U., & Lozano, J. A. (2019). A review on distance based time series classification. *Data Mining and Knowledge Discovery*, *33*(2), 378-412. https://doi.org/10.1007/s10618-018-0596-4

[12] Mao, S. & Xiao, F. (2019). Time Series Forecasting Based on Complex Network Analysis. *IEEE Access*, *2019*(7), 40220-40229. https://doi.org/10.1109/ACCESS.2019.2906268

[13] Xu, W., Liu, L., Zhang, Q., & Liu, P. (2018). Location decision-making of equipment manufacturing enterprise under dual channel purchase and sale mode. *Complexity, 2018*(2), 1-16. https://doi.org/10.1155/2018/3797131

[14] Xu, W. & Yin, Y. (2018). Functional objectives decision-making of discrete manufacturing system based on integrated ant colony optimization and particle swarm optimization approach. *Advances in Production Engineering & Management*, *13*(4), 389-404. https://doi.org/10.14743/apem2018.4.298

[15] Fakhrazari, A. & Vakilzadian, H. (2017). A survey on time series data mining. *IEEE International Conference on Electro Information Technology*, Lincoln, NE, 476-481.

[16] Li H. & Guo C. (2013). Survey of feature representations and similarity measurements in time series data mining. *Application Research of Computers*, *30*(5), 1285-1291. https://doi.org/10.3969/j.issn.1001-3695.2013.05.002

[17] Li, H. & Liang, Y. (2017). Similarity measure based on numerical symbolic and shape feature for time series. *Control & Decision*, *32*(3), 451-458. https://doi.org/10.13195/j.kzyjc.2016.0326

[18] Chen, H., Liu, C., & Sun, B. (2017). Survey on similarity measurement of time series data mining. *Control and Decision*, *32*(1), 1-11. https://doi.org/10.13195/j.kzyjc.2016.0462

[19] Berndt, D. J. & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. *International Conference on Knowledge Discovery and Data Mining*, WA, USA, 359-370.

[20] Mueen, A., Chavoshi, N., Abu-El-Rub, N., Hamooni, H., Minnich, A., & Maccarthy, J. (2018). Speeding up dynamic time warping distance for sparse time series data. *Knowledge & Information Systems*, *54*(1), 237-263. https://doi.org/10.1007/s10115-017-1119-0

[21] Lin, J., Keogh, E. J., Lonardi, S., & Chiu, Y. C. (2003). A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. *8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, San Diego, California, USA. https://doi.org/10.1145/882082.882086

[22] Keogh, E., Chakrabarti, K., & Pazzani, M. (2001). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, *3*(3), 263-286.https://doi.org/10.1007/PL00011669

[23] Francesco, G., Giovanni P., Andrea, T., & Sergio, G. (2009). A time series representation model for accurate and fast similarity detection. *Pattern Recognition, 42*(11), 2998-3014. https://doi.org/10.1016/j.patcog.2009.03.030

[24] Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery, 15*(2), 107-144. https://doi.org/10.1007/s10618-007-0064-z

[25] Kini, B. V. & Sekhar, C. C. (2013). Large margin mixture of AR models for time series classification. *Applied Soft Computing, 13*(1), 361-371. https://doi.org/10.1016/j.asoc.2012.08.027

[26] Xiong, Y. & Yeung, D. Y. (2004). Time series clustering with ARMA mixtures. *Pattern Recognition, 37*(8), 1675-1689.https://doi.org/10.1016/j.patcog.2003.12.018

[27] Holzmann, H., Munk, A., Suster, M., & Zucchini, W. (2006). Hidden Markov Models for circular and linear-circular time series. *Environmental & Ecological Statistics, 13*(3), 325-347.https://doi.org/10.1007/s10651-006-0015-7

[28] Keogh, E., Lonardi, S., Ratanamahatana, C. A., Li, W., Lee, S. H., & Handley, J. (2007). Compression-based data mining of sequential data. *Data Mining & Knowledge Discovery, 14*(1), 99-129.https://doi.org/10.1007/s10618-006-0049-3

[29] http://www.cs.ucr.edu/~eamonn/time_series_data/.

[30] Deng, W., Wang, G., & Xu, J. (2016). Piecewise two-dimensional normal cloud representation for time-series data mining. *Information Sciences, 2016*(374), 32-50. https://doi.org/10.1016/j.ins.2016.09.027

[31] Chen, H., Gao, X., & Guo, Y. (2019). Hierarchical Clustering of Time Series Based on Linear Information Granules. *Tehnički vjesnik, 26*(2), 478-485. https://doi.org/10.17559/TV-20190103125702

**Contact information:**

**Hailan CHEN,** PhD Candidate,
(Corresponding author)
Donlinks School of Economics and Management,
University of Science and Technology Beijing,
No. 30 Xueyuan Road, Haidian District, Beijing, China
E-mail: chl_hld@163.com

**Xuedong GAO,** Full Professor,
Donlinks School of Economics and Management,
University of Science and Technology Beijing,
No. 30 Xueyuan Road, Haidian District, Beijing, China
E-mail: gaoxuedong@manage.ustb.edu.cn