Cao and Liu *Journal of Cloud Computing: Advances, Systems and Applications* (2016) 5:11
DOI 10.1186/s13677-016-0062-z

Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH Open Access

# Research on dynamic time warping multivariate time series similarity matching based on shape feature and inclination angle

Danyang Cao* and Jie Liu

**Abstract**

Different sets of research mainly focus on one variable time series now, while researches involving multivariate time series have been insufficient. In this paper, combined linear segments and fitting error for multivariate time series, we present a new method to reduce the time complexity of DTW distance metric algorithm. Based on the shape feature and the tilt angle, we propose a new approach for similarity matching of DTW multivariate time series. Experimental results demonstrate that this method is helpful for ensuring accuracy and for reducing the time complexity of similarity matching.

**Keywords:** Similarity matching, Dynamic time warping, Multivariate time series, Shape feature

## Introduction

Currently, time series are widely used in economics, management, computers, mathematics, electronics and many other interdisciplinary researches [1]. The study about time series has been developed rapidly since 1990. Time series similarity search is used to research clustering, classification, pattern matching, rule discovery, content anomaly detection and many other aspects etc. Time series similarity is a fundamental problem of time series data mining [2], and its research mainly covers: how to judge the similarity of different time series, how to measure the degree of similarity, and how to find the most-like time sequences.

Time series is a series of recorded values in chronological order. This kind of data is quite common in our life, such as transactions data in stock industry, the vehicles' running state data produced during driving, statistics data of clicking times on webpages, the description data of human body posture matching the postural action, and measurement data of planetary motion trajectory in astronomy industry [1].

Time series analysis is mainly used to extract meaningful statistics and other characteristics of the data, or in other words, to extract the potentially useful information from time series [3]. For example, timing data of power system loading contains plentiful information about characteristics of the power load, and stock timing data contains laws of stock price fluctuation etc. Time series mining is of significant value. It can help people understand the information implied in time series and support people to make the right decision. However, current studies are mostly limited to one variable time series and the results sometimes may have deviation from the real situation. There are many studies on one variable time series, and mature theories and methods in this area have gradually formed. Multivariate time series are composed of several different data vectors [4], and its structure is more complex than one variable. Up to now, the theories and methods to study the multivariate time series are not well developed. However, the study of multivariate time series is more meaningful for many practical applications. For example, when we evaluate the weather conditions of a place, we need to considerate the temperature, pressure, humidity, and other factors to get more reliable results.

In recent years, with growth of massive information and data, more and more multivariate time series were produced. How to extract potential information from these multivariate time series has gradually attracted the

* Correspondence: danyangcaojcc@163.com
College of Computer Science and Technology, North China University of Technology, Shijingshan, Beijing 100144, China

attention of scholars. There are various variables in multivariate time series and they normally work together. If we only consider one variable at a time, we may loss valid information. In addition, for the same time period, multivariate time series data sets tend to occupy a larger space than one variable data set. Even after some dimension reduction process, they still take up very large storage space. Thus, data processing of multivariate time series needs an efficient data representation method. This method should focus on two aspects: pattern representation and similarity matching.

### Related work

Pattern representation indicates some kind of changing features, which can summarize and represent the time series. Pattern representation may be the slope after time series is segmented, the mean or variance of the sampling points during a period of time, the symbol representation after discretization, or some function representation [5]. Compared with the original time series, the time series through pattern extraction can be manifested in a more concise form, which can effectively avoid the phenomena of "Curse of Dimensionality". Currently, the typical pattern representation is Discrete Wavelet Transform (DWT) [6], Singular Value Decomposition (SVD) [7], Piecewise Aggregate Approximation (PAA) [8, 9], and Symbolic Aggregate Approximation (SAX) [10] etc. Among them, PAA can approximate the time series best, and it is simple, intuitive, and efficient [11]. Time series pattern representation is the basis of time series similarity matching research, while similarity matching is the core of their research. The similarity of two time series refers to: two time series are transformed, then calculated the values with similarity function, if the values satisfy stipulated error threshold, we think that the two time series under the condition of this kind of transformation are similar to each other. Although similarity matching has been studied as a key object, there still exist many unsolved difficult problems. For many current algorithms, the experimental results showed that they made a great improvement in operation efficiency. However, these improvements still have many limitations, if the parameters of the experiments are modified slightly, the results will be changed greatly. Currently, the common methods of multivariate time series similarity matching are Minkowski Distance [12], Dynamic Time Warping (DTW) distance [13, 14], Edit Distance [15–17], and Longest Common Subseries (LCS) etc. [17–19].

In similarity matching, DTW distance was first introduced by Berndt and Clifford to time series mining [20]. It could match the time series of equal or unequal length at the same time, support their stretching and bending on the time axis, and identify matching sequence effectively. In the traditional research, people usually changed multivariate time series into the one variable time series, and then calculated DTW distance of each variable directly and independently by making use of the methods achieved in the one variable time series studies. This had been used in many studies of multivariate time series. However, there are different relevant relationships among multivariate time series [21], which can provide more information for multivariate time series similarity matching and help us to improve the accuracy of similarity matching.

In [22], the authors use the Trend Distance (TD) approach based on DTW distance to measure time series. The method was proposed to fit time series by the first-order polynomial of Chebyshev and used TD method to measure the similarity simultaneously. A first order polynomial of Chebyshev is more suitable for one variable time

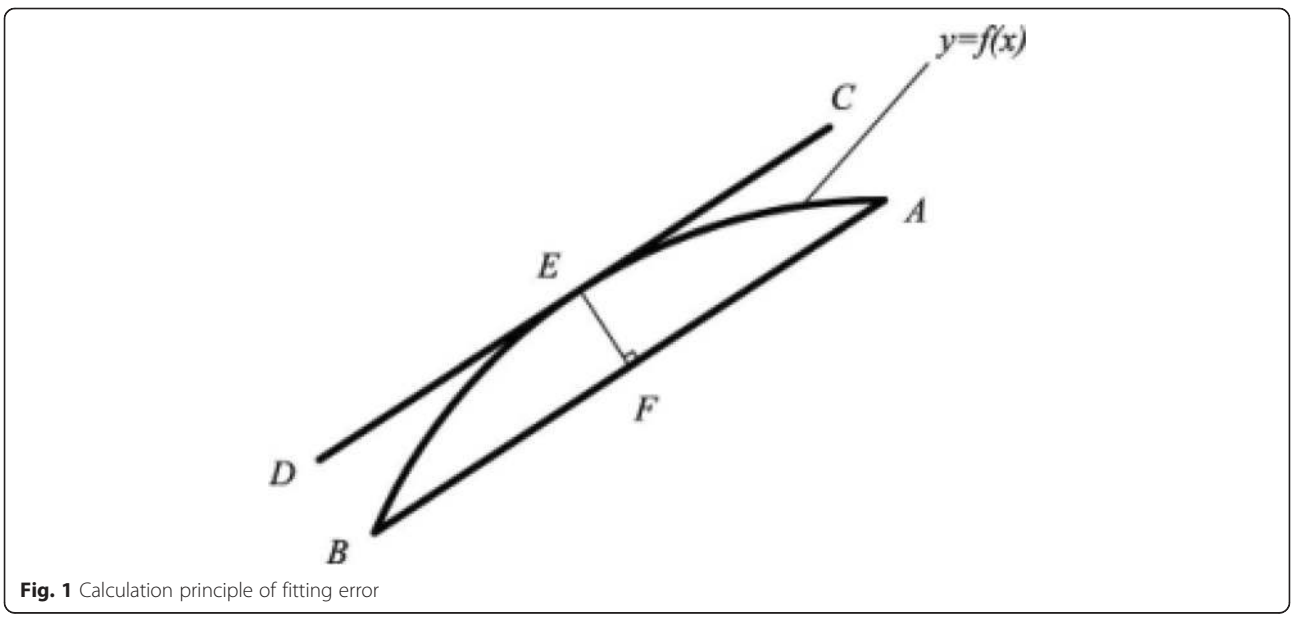


**Fig. 1** Calculation principle of fitting error

series, but for multivariate time series, it is much more complex to fit. Moreover, since TD method selected different parameters, the accuracy rate is also quite different in their similarity. In view of this situation, we use a more simple method which combines PAA piecewise linear representation and fitting errors methods to extract pattern features (Piecewise Aggregate Approximation and Error, PAA_ERR), and then select the "tilt angle" and "shape feature value" of each segment which is produced by PAA_ERR method in a variable as the feature representation of the segment, finally, we propose a new similar measure method based on shape feature and inclination angle (Shape and Angle Dynamic Time Warping, SA_DTW) to measure the DTW distance between two time series to match their similarity.

## Related concepts and theories

Definition 1 (time series) A time series, in a statistical perspective, refers to different samples values of certain index in the timeline. In a mathematics perspective, it is a series of record values by measuring a certain variable X (t) according to the time sequence. In a broad sense, it means the attribute values during a period of time. Similarly, multivariate time series means time series with multiple variables, based on the previous concept. For Multivariate time series $X_1$, $X_2$,..., $X_M$, Where $X_i = (\chi_{i,1}, \chi_{i,2}, \cdots, \chi_{i,N})$, i = 1,2,...,N, especially if M = 1, the given series is a univariate time series (UTS) and if M > 1, it's a multivariate time series(MTS).

We assume that there is a multivariate time series A which contains m variables, and the length of the time series is n, then it can be expressed as a matrix form of m × n:

$$A = \begin{bmatrix} a_{11}, a_{12}, a_{13}, \cdots, a_{1n} \\ a_{21}, a_{22}, a_{23}, \cdots, a_{2n} \\ \vdots \\ a_{m1}, a_{m2}, a_{m3}, \cdots, a_{mn} \end{bmatrix} \tag{1}$$

Definition 2 (MTS' piecewise line representation) Suppose that there is a multivariate time series A in formula 1, then its piecewise linear representation can be defined as follows:

$$X_i(t) = \begin{cases} f_i(t, w_{i1}) + e_i(t), t \in [1, t_1] \\ f_i(t, w_{i2}) + e_i(t), t \in [t_1, t_2] \\ \cdots \\ f_i(t, w_{in}) + e_i(t), t \in [t_{n-1}, t_n] \end{cases}, i \in [1, m] \tag{2}$$

Where $W_{ij}$ denotes the coordinate of two endpoints of the i-th variable from $t_{k-1}$ to $t_k$, the j-th segment. $f_i(t)$ denotes the $W_{ij}$ linear function linking the starting and ending point of the i-th variable from $t_{k-1}$ to $t_k$. $e_i(t)$ denotes the fitting error between the original time series and the new fitting segment from $t_{k-1}$ to $t_k$.

Definition 3 (the fitting error of MTS' piecewise line representation), We assume that there is an MTS A in the formula (1) and let $d_1$, $d_2$,..., $d_j$ denotes the vertical distance from j observations in the original time series to fitting line segment. Then the fitting error of the segments for variable i can be represented by the following formula:

$$\begin{aligned} e_i &= \max(d_k) \\ &= \max\left(\left|\frac{A\chi_k + By_k + C}{\sqrt{A^2 + B^2}}\right|\right), \quad k \in [1, j] \end{aligned} \tag{3}$$

Where i = 1, 2, ..., m, and it indicates a total of m variables. The calculation principle of fitting error is shown in Fig. 1.

Where | EF | is the length of the line fitting error of each segment. For a segment of all variables, the sum of the fitting error on all variables of multivariate time series is as follow: $e_{total} = \sum_{i=1}^{m} e_i$, where $e_{total}$ denotes a total fitting error of all the variables in a segment.

Definition 4 (Dynamic time warping distance [23]) It is also called DTW distance. Assuming two time series $A = (a_1, a_2, ..., a_m)$, $B = (b_1, b_2, ..., b_n)$, then the DTW distance formula of the two time series is as follows:

$$D_{dtw} = \begin{cases} 0 & m = n = 0 \\ \infty & m = 0 \text{ or } n = 0 \\ D_{base}(a_1, b_1) + \min \begin{cases} D_{dtw}(A, B[2, -]) \\ D_{dtw}(A[2, -], B) \\ D_{dtw}(A[2, -], B[2, -]) \end{cases} & others \end{cases} \tag{4}$$

DTW algorithm is to use the classic dynamic programming to find an optimal path with a minimum cost of bending, and the time complexity is O(|A| · |B|). In brief, it is to find the shortest path by constructing an adjacency matrix. The cumulative distance calculated by DTW distance method is shown in Table 1:

**Table 1** The calculation of the cumulative distance of time series

| | | | | | |
|---|---|---|---|---|---|
| 5 | 22 | 16 | 16 | 16 | *17* |
| 8 | 19 | 14 | 14 | *14* | 15 |
| 4 | 13 | 9 | *9* | 9 | 10 |
| 3 | 11 | *8* | 8 | 8 | 9 |
| 9 | 10 | *8* | 8 | 8 | 9 |
| 4 | 3 | *2* | 2 | 2 | 3 |
| 3 | *1* | 1 | 1 | 1 | 2 |
| A/B | 2 | 3 | 3 | 3 | 2 |

The first column represent the data of series A{3,4,9,3,4,8,5}. The last row represent the data of series B{2,3,3,3,2}. Other cells represent the DTW distances. The underlined data represent the mininum distance, they make up warping path

In the above table, the DTW distance between time series A and B [1:i] ($D_{dtw}$(A,B[1:i])) is stored in the cell of the top of the i-th column in Table 1. Similarly, $D_{dtw}$ (A [1: i], B) is stored in the cell of the right of the i-th row in Table 1. The same principle can be used to calculate multivariate time series, but in multivariate time series, $a_i$, $b_j$ indicate the corresponding column vectors.

## Multivariate time series model representation

From the TD method of Lee et al., we can see that when the weight values corresponding to the appropriate tilt angle and the time maintaining length were changed, the accuracy rate would be changed significantly. Particularly when the value of tilt angle was 0, the value of time maintaining length was 1, then the accuracy rate was only 0.27, which indicated the time maintaining length might not distinguish differences and similarities of the time series very well. Therefore, in order to reduce the time complexity, to make the measurement more accurate and to make the calculation more efficient, we combined PAA with fitting error to divide the multivariate time series into several segments. Then, on the basis of each tilt angle, we extracted the value of tilt angle and shape feature as pattern representations of the segments.

PAA method is a classic representation of time series, which is able to extract the relatively independent pattern based on morphological changes. The more segments, the higher segmentation accuracy will be. On the contrary, the less segments, the lower segmentation accuracy. In addition, PAA method is capable of data abstraction and noise filtering. Meanwhile, after the combination of PAA and fitting error, it can further screen segments of larger deviation, getting more detailed segmentation. And more importantly, fitting error can combine all the variables of time series, considering the overall situation, which can effectively prevent the loss of information, and further filter noise data.

### PAA_ERR: the segmentation method to multivariate time series

When we use the PAA_ERR method on the multivariate time series, we take all the variables as a whole into consideration, that is, we use PAA to divide all variables into segments, if the value of overall fitting error $e_{total}$ of a segment is greater than the threshold value on all the variables, then reprocess the segment, until the $e_{total}$ is less than threshold value.

The steps of this algorithm are as follows:

Assume that time series MTS processed by PAA_ERR can be divided into s segments, then the linear representation of these multivariate time series after processing can be denoted as $L(MTS) = \{L(x_{i1}, x_{i2}), L(x_{i2}, x_{i3}), \ldots, L(x_{ik-1}, x_{ik}), i \in [1,m], k \in [1,s]\}$, where i represents a variable of the segmentation, where $x_{ik}$ represents the record value of the time series, $L(x_{ik-1}, x_{ik})$ represents a straight line connecting two points.

### The shape feature representation of multivariate time series

In similarity query process, there are two situations needed to be carefully considered: the first is that if there exist "zooming in" or "zooming out" modes similar to the given mode when checking the time series, namely, similar to "zooming in" or "zooming out" mode, then it is an approximate proportional scaling of the length and magnitude to the given mode. For instance, the trend of the stock price change with different time durations could be similar, but this cannot be found with the general equal length pattern queries method. Another case is that in an enlarged similar model, there are some small "vibrating" intervals, and they do not affect the overall trends. In this case, we should focus on the main trend and ignore these small intervals.

An important feature of time series is the changes of growing or declining sequence rate. If the growth or decline rate becomes bigger, it indicates the sequence morphological changes tend to increase; if the growth rate becomes smaller, it indicates that the sequence curves tend to be flat and its morphological changes start to decrease. Hence, these turning points can be found through the changes of line segment slopes.

In the [24] literature, according to the changes of slope, the authors described the shape characteristics of time series as a collection of seven variables {declining rapidly, keeping down, declining slowly, horizontal, rising slowly, keeping rising, rising rapidly}, and they were denoted as M = {-3, -2, -1, 0, 1, 2, 3} corresponding to the above-

Algorithm 1: the PAA_ERR method

Input: multivariate time-series data set: MTS; the initial number of segments: k and the threshold of fitting error: threshold.

Output: the representation of time series A using PAA_ERR segmented method.

```
Step 1:
  MTS=(MTS-min(MTS))/(max(MTS)-min(MTS))
                                    // Normalize MTS
Step 2:
  seg_err(i)=Calculate_Seg_Err(MTS) //the fitting error of segmented overall after PAA treatment
Step 3:
  while seg_err(i)>threshold      // the value of fitting error is greater than the threshold value on all the variables
Step 4:
  Divide_Seg(i)                   // refragmenting
Step 5:
  end
```

described mode. The specific numbers was shown in Table 2. This method could reflect the degree of dynamic changing trends. However, in literature [23], the authors used the time series which was completely represented by this model, and this method led to the problem that time series was not sensitive to its stretching. It can be seen clearly in Fig. 2, where the representation modes of these two time series would be identical (calculated by Table 2).

Based on the above cases, this paper combined the representation method of "shape mode" with the tilt angle of time-series. Meanwhile, we extracted various segments of its tilt angle and the shape mode as the segment feature representation from time series which was calculated by PAA_ERR calculation method. The inclination angle had a clear physical meaning, which could reflect the local changing trend of time series, and it reduced much calculation. The value of shape mode could show the trends of time series during a period of time. The combination of "shape mode" and the tilt angle reflected the extent of dynamic changing trends more effectively.

We assume that A [m × n] is a multivariate time series of m-variable n-sections after being segmented, the deformation intensity of each segment is denoted as $T_i$ (i = 1, 2,…,n), $T_i$ = max($y_i$) - min($y_i$) representing the different value between the maximum sample value of each segment and the minimum sample value. $\sum_{j=1}^{n} T_j$ represents the sum of all deformation strength in a variable. Then, the weight of each shape mode segment can be represented by the following formula:

$$\Im_i = \frac{T_i}{\sum_{j=1}^{n} T_j} \tag{5}$$

Where K (i, j) represents the slope of the j-th variable of the i-th segment.

Then multivariate time-series $A_{m*n}$ can be denoted as:

$$\begin{bmatrix} (\alpha_{11},\ \Im_{11}\ p_{11}),\ (\alpha_{12},\ \Im_{12}\ p_{12}),\ldots,\ (\alpha_{1n},\ \Im_{1n}\ p_{1n}) \\ (\alpha_{21},\ \Im_{21}\ p_{21}),\ (\alpha_{22},\ \Im_{22}\ p_{22}),\ldots,\ (\alpha_{2n},\ \Im_{2n}\ p_{2n}) \\ \vdots \\ (\alpha_{m1},\ \Im_{m1}\ p_{m1}),\ (\alpha_{m2},\ \Im_{m2}\ p_{m2}),\ldots,\ (\alpha_{mn},\ \Im_{mn}\ p_{mn}) \end{bmatrix} \tag{6}$$

Which $\alpha$ represents the inclination angle of each segment, $\Im$ is the weight of segmented form, $p$ is the value of segmented shape pattern. And in each variable, the sum weight of the shape characteristics is 1.

## The similarity matching of SA_DTW method

Suppose the shape characteristics representation of two time series A, A ` can be expressed as follows:

$$A = \begin{bmatrix} (\alpha_{11},\ \Im_{11}\ p_{11}),\ (\alpha_{12},\ \Im_{12}\ p_{12}),\ldots,\ (\alpha_{1n},\ \Im_{1n}\ p_{1n}) \\ (\alpha_{21},\ \Im_{21}\ p_{21}),\ (\alpha_{22},\ \Im_{22}\ p_{22}),\ldots,\ (\alpha_{2n},\ \Im_{2n}\ p_{2n}) \\ \vdots \\ (\alpha_{m1},\ \Im_{m1}\ p_{m1}),\ (\alpha_{m2},\ \Im_{m2}\ p_{m2}),\ldots,\ (\alpha_{mn},\ \Im_{mn}\ p_{mn}) \end{bmatrix} = [a_1, a_2, \cdots, a_n] \tag{7}$$

$$A' = \begin{bmatrix} \left(\alpha'_{11},\ \Im'_{11}\ p_{11}\right),\ \left(\alpha'_{12},\ \Im'_{12}\ p'_{12}\right),\ldots,\ \left(\alpha'_{1n},\ \Im'_{1n}\ p'_{1n}\right) \\ \left(\alpha'_{21},\ \Im'_{21}\ p'_{21}\right),\ \left(\alpha'_{22},\ \Im'_{22}\ p'_{22}\right),\ldots,\ \left(\alpha'_{2n},\ \Im'_{2n}\ p'_{2n}\right) \\ \vdots \\ \left(\alpha'_{m1},\ \Im'_{m1}\ p'_{m1}\right),\ \left(\alpha'_{m2},\ \Im'_{m2}\ p'_{m2}\right),\ldots,\ \left(\alpha'_{mn},\ \Im'_{mn}\ p'_{mn}\right) \end{bmatrix} = \left[a'_1, a'_2, \cdots, a'_n\right] \tag{8}$$

The steps of this algorithm are as follows:

Algorithm 2: the algorithm of multivariate time series pattern

Input: (1) time series MTS' (m × v) that dealt by PAA_ERR method, where m is the number of variable of time series and v represents the number of sub-time series. (2) the threshold of morphological patterns distinguish : th

Output: the angle of each segment of multivariate time series: angles, the value of keeping length: pre_length_weights, shape mode value: shape.

(1) for i=1 to m

(2) for j=1 to v

(3) angles[i,j]=calculate_slope(MTS'$_i$)

//the angle of each segmen

(4)pre_length_weights[i,j]=calculate_pre_weight(MTSi) // the value of keeping length

(5)if (K(i-1,j)<−th or −th<K(i-1,j)<th) & K(i,j)<K(i-1,j) then shape(i,j)= −3; end

(6)if K(i-1,j)<−th & K(i,j)= K(i-1,j) then shape(i,j)=−2; end

(7)if K(i-1,j)<−th & K(i-1,j)<K(i,j)<−th then shape(i,j)=−1; end

(8)if K(i-1,j)>th & 0<K(i,j)< K(i-1,j) then shape(i,j)=1; end

(9)if K(i-1,j)>th& K(i,j)= K(i-1,j) then shape(i,j)=2; end

(10)if (K(i-1,j)>th or −th<K(i-1,j)<th) & K(i,j)>K(i-1,j) then shape(i,j)=3; else shape(i,j)=0; end

(11)end

(12)end

**Table 2** Shape mode list of values

| | k(i + 1) < −th | | | −th < k(i + 1) < th | k(i + 1) > th | | |
|---|---|---|---|---|---|---|---|
| ki < −th | Δk < 0 | Δk = 0 | Δk > 0 | 0 | | 3 | |
| | −3 | −2 | −1 | | | | |
| −th < ki < th | −3 | | | 0 | | 3 | |
| ki > th | −3 | | | 0 | Δk < 0 | Δk = 0 | Δk > 0 |
| | | | | | 1 | 2 | 3 |

Where $a_i$, $a_i$ ` (i = 1, 2, …, n) are the column vector of the two time series A, A `. From definition 4, the shape characteristics distance of A, A ` is defined as follows:

$$D_{dtw}(A, A') = D_{base}(a_1, a_1') + min\begin{cases} D_{dtw}(A, A'[2, -]) \\ D_{dtw}(A[2, -], A') \\ D_{dtw}(A[2, -], A'[2, -]) \end{cases} \tag{9}$$

And the base distance is defined as follows:

$$D_{dtw}\left(a_i, a_j'\right) = \begin{cases} \sum_{k=1}^{m}\left[\lambda_k\left|\Im_{ki}p_{ki} - \Im_{kj}'p_{kj}'\right| + \beta_k\left|\alpha_{ki} - \alpha_{ki}'\right|\right], if\,|i-j| \le q \\ \infty, \quad if\,|i-j| > q \end{cases} \tag{10}$$

Where, q takes 10 % of the time series' length. In the k-th variable, $\lambda_k$, $\beta_k$ indicate difference between shape mode and the tilt angle weight values, and

$$\lambda_k + \beta_k = 1 \quad (k = 1, 2, \ldots, m). \tag{11}$$

## Result and analysis

The experiments were performed on Intel (R) Core (TM) i7 CPU, with 2.98 GB memory. The operation system is Windows 7 Ultimate, and the software was MATLAB7.11.0. We tested our implementation with two data sets from http://kdd.ics.uci.edu/databases/: Vicon Physical Action Data Set (VPA) and EEG. The VPA data set was collected from seven male and three female, which reflected human action in both normal conditions and violent scenes; while the EEG data set came from the populations of two distinct: Alcoholic Subjects and Control Subjects. This paper contained two experiments: PAA_ERR and SA_DTW. The PAA_ERR method divided the multivariate time series into several segments, taking compression ratio of multivariate time-series as the main indicator to compare, while SA_DTW method compared time complexity and accuracy of similarity in the algorithm.

### PAA_ERR method

For the above two data sets, PAA_ERR algorithm could maintain the original features of multivariate time series and greatly reduced the amount of data. For example, Fig. 3 showed the waveform of the Headering data by using PAA_ERR algorithm. Since PAA_ERR is based on PAA algorithm, for the compression algorithm of time series, we chose PAA algorithm as control.

In Table 3, initSegNum represented the number of initial segments after PAA_ERR algorithm, and finalSegNum indicated the number of final segments. PAA_ERR used the initial segmentation generated by using PAA, and then, further processed the time series by using fitting error. Therefore, in Table 3, the initSegNum' value and finalSegNum' value for the PAA were always equal. The maxSegErr represented the largest fitting error after

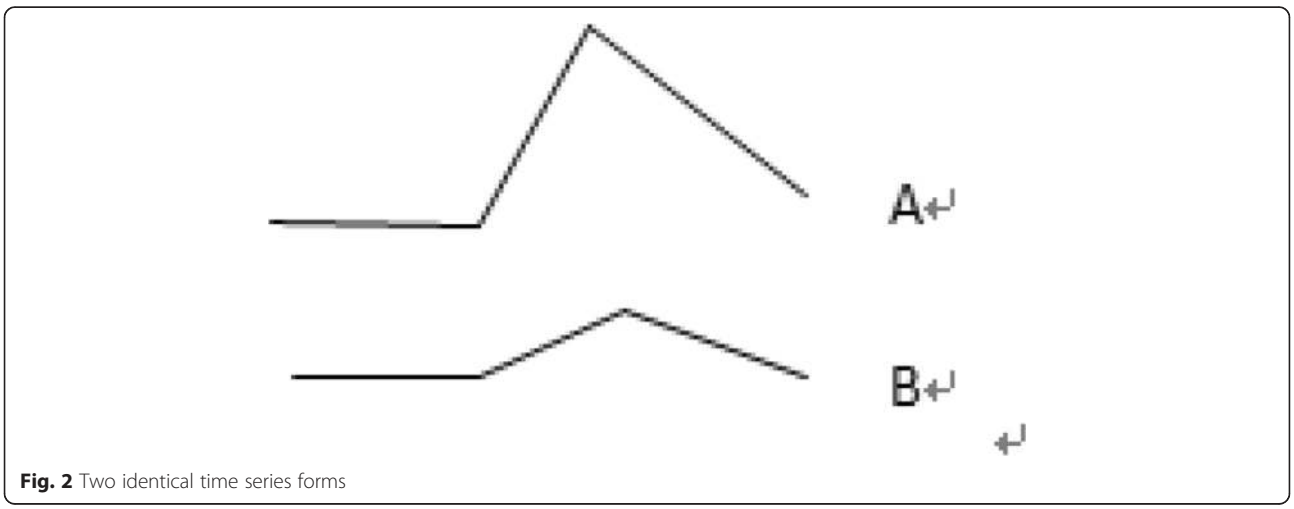


**Fig. 2** Two identical time series forms

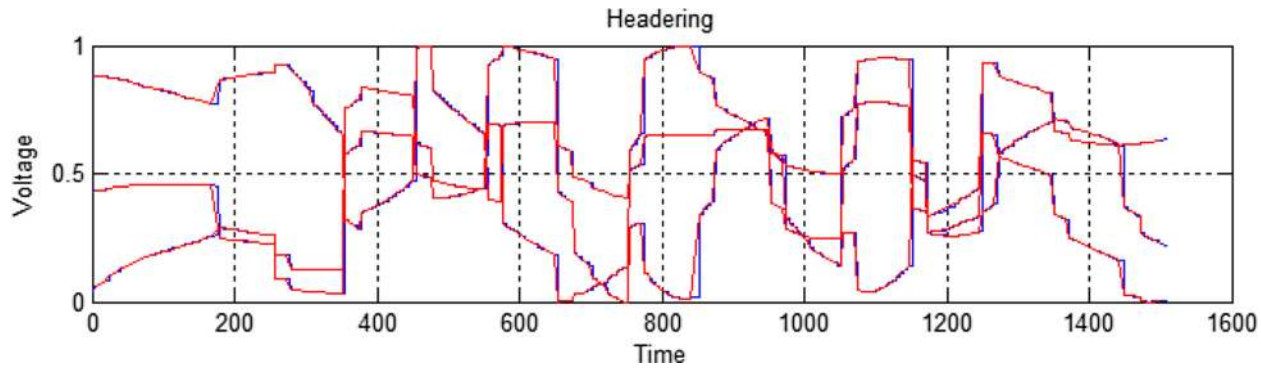


**Fig. 3** The three-variable data of headering's fitting renderings

the segmentation. According to previous algorithm, we knew that fitting error of each segment divided from PAA_ERR algorithm was smaller than the threshold; hence the maxSegErr value was basically equal. The timeSeqErr represented the overall fitting error value of the time series which is the sum of all fitting error value of all segments and variables. Compression ratio is the ratio of the number of time series segments calculated by PAA_ERR algorithm to the length of the original time series.

(1) As can be seen from the above table, multivariate time series calculated by PAA_ERR algorithm could guarantee that every fitting error was far smaller than the PAA algorithm, and the overall fitting error of the time series was also significantly smaller than the PAA direct segments. More importantly, the PAA_ERR algorithm could guarantee small fitting error. It also enabled time series compression rate to reach 90 % and above. So it indicated that PAA_ERR algorithm in multivariate time series piecewise linear representation could get better results by reducing the fitting error.

(2) Figure 3 showed the direct effect from PAA_ERR calculation for the first three-variable data of Headering data when the fitting error threshold was 0.01.

The horizontal axis was a time axis, and the vertical axis was the recorded values after standardization. Blue lines were the original time series and the red ones were the time series after processing algorithms.

Through the above experiment, when the error threshold was set to 0.05, the compression ratio of the original time series could reach 90 %, and the algorithm could effectively preserve the local characteristics and the overall shape trends of the original sequence under the premise of ensuring high compression ratio.

**Table 3** Experimental results between PAA and PAA_ERR method

| | initSegNum | finalSegNum | maxSegErr | timeSeqErr | Compression ratio |
|---|---|---|---|---|---|
| PAA algorithm | 50 | 50 | 0.7556 | 20.1941 | 98.6 % |
| | 100 | 100 | 0.7511 | 19.2898 | 97.3 % |
| | 150 | 150 | 0.5517 | 17.6011 | 96.0 % |
| | 200 | 200 | 0.4393 | 16.7043 | 94.7 % |
| | 250 | 250 | 0.4182 | 15.8199 | 93.3 % |
| | 300 | 300 | 0.3456 | 15.5139 | 92.0 % |
| | 350 | 350 | 0.2931 | 15.6895 | 90.7 % |
| PAA_ERR algorithm | 50 | 219 | 0.05 | 11.5020 | 94.2 % |
| | 100 | 252 | 0.05 | 11.5370 | 93.3 % |
| | 150 | 284 | 0.05 | 11.6466 | 92.4 % |
| | 200 | 312 | 0.05 | 11.3508 | 91.6 % |
| | 250 | 345 | 0.05 | 11.3428 | 90.8 % |
| | 300 | 401 | 0.05 | 11.2568 | 89.3 % |
| | 350 | 453 | 0.05 | 11.4785 | 87.9 % |

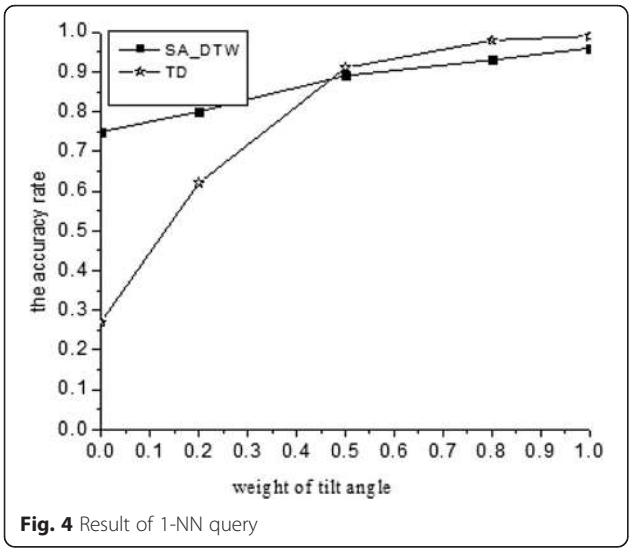


**Fig. 4** Result of 1-NN query

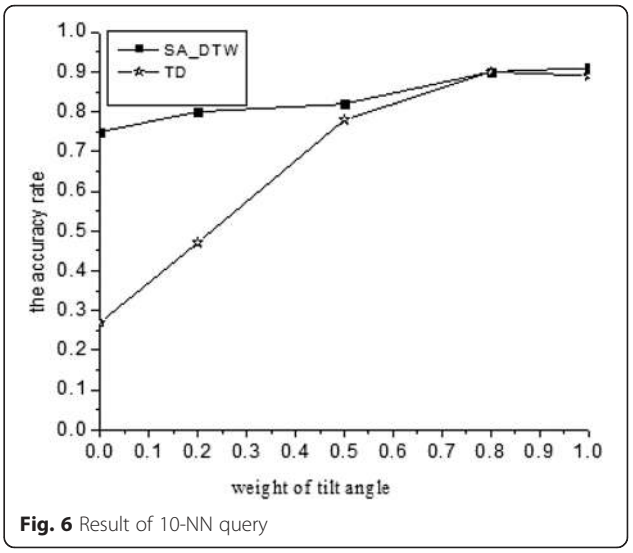


**Fig. 6** Result of 10-NN query

PAA_ERR method used the overall fitting error for the final segment. For those time series of small time span but large sequence trend changes, time segment calculated by PAA_ERR would get very close to the original sequence, and resulted in poor final segment results. Therefore, in the next experiment, we also used a time series with smaller trend change sequence as a test object.

### SA_DTW method

In the experiment, we adopted multivariate time series, Australian Sign Language signs (High Quality) dataset provided by UCI public databases. It used a total of 22 attributes to depict the bending degrees of every finger in right and left hand. And the two hands were depicted with 11 attributes respectively, in which six attributes described the information of hand position and the other five variables used to describe the bending degrees of the thumb, forefinger, middle finger, ring finger and the little finger. The data set contained 95 categories (a category represented a gesture), each category had 27 groups of data, and the sampling rate was 0.01 s, namely each frame data was collected by every 0.01 s. In order to facilitate the comparison of our experiment, we used the TD method, which was also adopted by Li Zhengxin and others in their experiment, as the object of experimental comparison. The entire experiment could be divided into two steps. Step 1: to choose one sequence randomly from the experiment data set as an input instance. Step 2: for the chosen one and remaining sequences, using TD method and SA_DTW method to implement the distance calculation. Then, select the closest one, five, or ten sequences, record the sequence numbers different from the input instance of different categories, and then use the 4 to 7 formula to calculate its accuracy. Repeating the experiment was repeated for 50 times to calculate its accuracy.

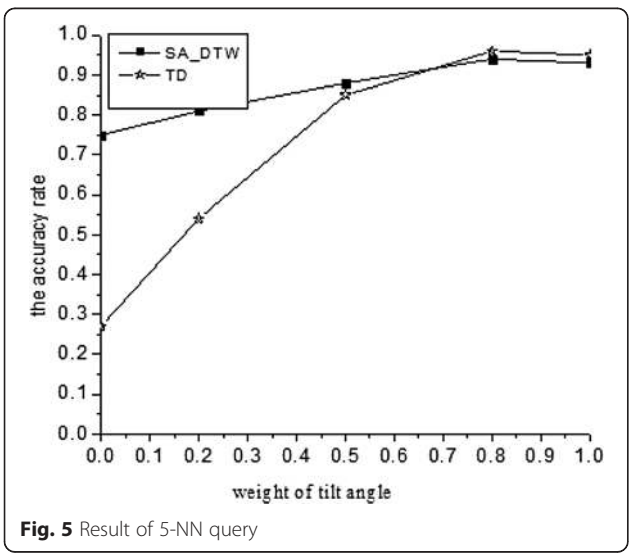


**Fig. 5** Result of 5-NN query

As was shown in Figs. 4, 5 and 6, when the weight of slope angle was changed, the accuracy of TD method and SA_DTW method would be also changed. In the diagram, when the slope angle weight value was 0, the weight of TD method's time span and SA_DTW's shape mode value were 1 respectively, the accuracy of TD method decreased to 20 %, while SA_DTW still had high accuracy. It indicated that the shape weight we proposed in this paper was more suitable for the representation of multivariate time series model. It also can be seen through Table 4 that

**Table 4** Comparison of computation time between SA_DTW and DTW

| Dataset | $T_{SA\_DTW}/T_{TD}$ |
|---|---|
| VPA | 0.6 |
| EEG | 0.75 |

SA_DTW method is better than TD methods on reducing the time complexity.

## Conclusion

This paper presented a new method of similarity search: SA_DTW. It adopted shape mode and tilt angle to represent the feature of each segment. SA_DTW carried out a whole segment on all variables, and reflected the correlation among various variables to some extent. In addition, it supported time series' stretching and bending on the time line. Through the experiments we could see SA_DTW could effectively reduce the time complexity compared with the DTW method. The method that combined shape feature and tilt angle could greatly improve the accuracy of similarity matching.

**Acknowledgments**

This paper is supported by Beijing Natural Science Foundation (NO.4144073), supported by Research Fund of North China University of Technology, supported by National Natural Science Foundation of China (NO. 41471303), supported by training project for outstanding young teachers of North China University of Technology.

**Authors' contributions**

DC provided the idea of the paper, carefully designed the algorithm in the manuscript, reviewed and edited the manuscript. JL performed the experiments and presented performance analysis. Both authors read and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.

Received: 11 December 2015 Accepted: 10 August 2016
Published online: 17 August 2016

**References**

1. Qian-yun MO, Cheng Z (2008) Parallel computing dynamic warping distances for time sequences on the cluster computing systems. Microelectron Comput 25:155–158
2. Khan Z, Anjum A, Soomro K, Tahir MA (2015) Towards cloud based big data analytics for smart future cities. J Cloud Comput 4(1):1
3. Prema V, Uma Rao K (2015) Time series decomposition model for accurate wind speed forecast. Renewables Wind Water Solar 2–18:2015
4. Hai-lin LI (2015) Feature representation of multivariate time series based on correlation among variables. Control Decis 30:441–446
5. Yan-yan Z, Rong-cong X, Xiao-yun C (2006) Time series piecewise linear representation based on slope extract edge point. Comput Sci 33:139–142
6. Chan K, Fu AW (1999) Efficient time series matching by wavelets. In: Richard ST (ed) Proceedings of 15th IEEE International Conference on Data Engineering (ICDE). IEEE Computer Society, Sydney, pp 126–133
7. Korn F, Jagadish H, Faloutsos C (1997) Efficiently supporting ad hoc queries in large datasets of time sequences. In: Joan P (ed) Proceedings of ACM SIGMOD international conference on management of data. Morgan Kaufmann Publisher, Tuescon, pp 289–300
8. Keogh E, Pazzani M (2000) A simple dimensionality reduction technique for fast similarity search in large time series databases. In: Terano T, Liu H, Chen AL (eds) Proceedings of 4th Pacific-Asia conference on knowledge discovery and data mining. Springer-Verlag, Kyoto, pp 122–133
9. Yi B, Faloutsos C (2000) Fast time sequence indexing for arbitrary Lp norms. In: Abbadi AE, Brodie ML, Chakravarthy S et al (eds) Proceedings of the 26th International Conference on Very Large Databases (VLDB). Morgan Kaufmann Publishers, Cairo, pp 385–394
10. Lin J, Keogh E, Londardi S et al (2003) A symbolic representation of time series, with implications for streaming algorithms. In: Mohammed JZ, Charueds AC (eds) Proceedings of the 8th ACM SIGMOD workshop on research (DMKD). ACM Press, San Diego, pp 55–68
11. Bin M, Jin-lai Y (2009) Efficient time series lower bounding technique. Comput Eng Appl 45(11):168–171
12. Agrawal R, Faloutsos C, Swami A (1993) Efficient similarity search in sequence databases. //Proc of the 4th International Conference on Foundations of Data Organization and Algorithms. Chicago, p 69-84.
13. Keogh E (2005) Exact indexing of dynamic time warping[J]. Knowl Inf Syst 7(3):358–386
14. Yi B, Jagadish H, Faloutsos C (1998) Efficient retrieval of similar time sequences under time warping. In: Sipple RS (ed) Proceedings of International Conference of Data Engineering (ICDE). IEEE Computer Society, Orlando, pp 201–208
15. Ristad ES, Yianilos PN (1997) Learning string edit distance. In: Douglas FH (ed) Proceedings of 14th International Conference on Machine Learning (ICML). Morgan Kaufmann Press, Nashville, pp 287–295
16. Chen L, Ng R (2004) On the marriage of Lp-norm and edit distance. In: Nascimento MA, Zsu MT, Kossmann D et al (eds) Proceedings of 30th International Conference on Very Large Databases (VLDB). Morgan Kaufmann Publishers, Toronto, pp 792–801
17. Das G, Gunopulos D, Mannila H (1997) Finding similar time series. In: Komorowski J, Zytkow J (eds) Proceeding of 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD). Springer-Verlag, Bergen, pp 88–100
18. Vlachos M, Kollios G, Gunopulos D (2002) Discovering similar multidimensional trajectories. In: Agrawal R, Dittrich K, Ngu AH (eds) Proceedings of 18thInternational Conference on Data Engineering (ICDE). IEEE Computer Society, San Jose, pp 673–684
19. Vlachos M, Hadjieleftheriou M, Gunopulos D et al (2003) Indexing multidimensional time-series with support for multiple distance measures. In: Getoor L, Senator TE (eds) Proceedings of ACM SIGKDD International Conference on knowledge discovery and data mining. ACM Press, Washington DC, pp 216–225
20. Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series. Working Notes of the Knowledge Discovery in Databases Workshop, Seatle
21. Zhen-xing W, Lai-wan C (2012) Learning causal relations in multivariate time series data[J]. ACM Trans Int Syst Technol 3(4):71–76
22. Zheng-xin L, Feng-ming Z, Ke-wu L (2011) DTW based pattern matching method for multivariate time series. Pattern Recognit Artif Intell 24(3):425–430
23. Hui X, Yun-fa H (2005) Data mining based on segmented time warping distance in time series database. J Comput Res Dev 42(1):72–78
24. Xiao-li D, Cheng-kui G, Zheng-ou W (2007) Research on shape-based time series similarity matching. J Electron Inf Technol 29(5):1228–1231