Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 99 (2015) 1190 – 1202

Procedia Engineering

www.elsevier.com/locate/procedia

“APISAT2014”, 2014 Asia-Pacific International Symposium on Aerospace Technology, APISAT2014

A Novel Data Association Algorithm for Unequal Length Fluctuant Sequence

Xin Guan, Guidong Sun*, Xiao Yi, Qiang Guo

Department of Electronics and Information Engineering, Naval Aeronautical and Astronautical University, Yantai ,Shandong,264001, P.R.China

Abstract

There are quantities of such sensors as radar, ESM, navigator in aerospace areas and the sequence data is the most ordinary data in sensor domain. How to mine the information of these data has attracted a great interest in data mining. But sequence data is easily interfered and produces some fluctuant points. When dealing with these sequences, traditional sequence similarity measurement such as Euclidean distance arises large error, especially for unequal length fluctuant sequence. A novel average weight 1-norm unequal length fluctuant sequence similarity measurement algorithm based on dynamic time warping (DTW) is proposed to solve this problem. It constructs an absolute distance matrix based on DTW firstly, then weight average weight 1-norm and modify it with modifying factor to measure the distance of unequal length fluctuant sequence. It solves the fluctuation sensitivity of maximum distance measurement algorithm. Finally transform distance to similarity as the index of the association, associate the sequence data according to the maximum similarity association rule. Simulation results show the effectiveness of the proposed algorithm when associating unequal length fluctuant sequence, association rate is above 70% and simulate the effect of variation of the sequence length, fluctuant rate and processing time to the proposed algorithm.

Peer-review under responsibility of Chinese Society of Aeronautics and Astronautics (CSAA)

*Keywords:*data association; sequence similarity; fluctuant rate; dynamic time warping (DTW); average weight 1- norm

* Corresponding author. Tel.: +86+13001603600
*E-mail address:*sdwhsgd@gmail.com

Peer-review under responsibility of Chinese Society of Aeronautics and Astronautics (CSAA)
doi:10.1016/j.proeng.2014.12.648

1. Introduction

As a kind of uncertain data, sequence data is the most ordinary data in sensor domain and is the main research object in the field of data mining, it is pervasive across almost all human endeavors, including economic forecasting, medical research, weather forecast, network security, military science and other fields. With the rapid development of information technology and the increasing amount of data, what we have received contains more information, undoubtedly it has entered the age of big data. How to mine the effective information and knowledge hidden in these data has attracted significant attention and extensively research effort in recent years. Sequence data is high dimensional data composed of large data points, the length of it may not be consistent because of data points changing with time, which is called unequal length sequence. How to mining the unequal length sequence data is a key problem in data mining. Sequence similarity measurement may help, it is an important process and basic method in data mining, which can measure the relationship between different objects. There are many definitions to the similarity of unequal sequence, which are inconsistent, especially when the sequence data has some kind of fluctuant point caused by interference. All these make the information fusion of sequence data uncertainty, which encounters many difficulties and challenges in the actual process. Study on the sequence similarity of unequal length, Agrawal, Yi, Faloutsos, Keogh and others have improved sequence similarity query, at present the main existing methods are the discrete Fourier transform(DFT) [1-3], singular value decomposition(SVD)[4], discrete wavelet transform(DWT)[5-8] piecewise aggregate approximation (PAA) [9, 10], dynamic time warping (DTW) [11-17], piecewise linear representation (PLR) [10, 18-20], piece- wise polynomial representation (PPR) [21] etc. These methods can essentially be divided into two categories, one category is based on the Euclidean distance metric, and the other is DTW. Keogh and his partner proved the DTW is optimal in the processing of unequal sequence measurement in [12], so in this paper our study is based on DTW, especially the measurement of the unequal length and fluctuant sequence. Sang-Wook Kim proposed ∞-norm measurement based on DTW in [11], better processing measurement of unequal length smooth sequence, but it will produce serious distortion in the sequence repeated bending section and volatility fluctuant point. Therefore, the average weight 1- norm similarity measurement unequal length fluctuant sequence based on DTW is proposed, this algorithm calculates the elements' distance between two unequal sequence forming absolute distance matrix based on DTW firstly, and then extract the minimum distance absolute distance matrix by rows or columns forming the minimum absolute distance set, instead of using the maximum value in the minimum absolute distance set as the distance measurement of unequal length fluctuant sequence, but summing elements the minimum absolute distance set with average weight 1-norm distance measurement, it can greatly reduce the influence of the fluctuant point. Considering the minimum value of the minimum absolute distance set in each row or column may not be unique, directly summation can cause repeated bending problems, resulting in distance enlarged, so modifying the average weight 1-norm distance measurement with modifying factor to solve the repeated bending problems. Calculate the unequal length fluctuant sequence similarity according to the relation between distance and similarity calculation, finally weighting and fusing these similarities we get the unequal length fluctuant sequence matrix similarity. According to the similarity ranking, we can effectively associate unequal length fluctuant sequence type sensor data by maximum similarity association rules.

This paper is organized as follows: in section 2, introduce and analyze the sequence similarity mining in detail; in section 3, the average weight 1-norm unequal length fluctuant sequence similarity measurement algorithm based on DTW is proposed and deduced in math; in section 4, the simulation analysis is given to prove the effectiveness of the proposed algorithm; finally presents the conclusions.

2. Sequence similarity mining analysis

2.1. Matrix representation of sequence

A sequence can be represented by

$$\boldsymbol{S_i} = (S_{i1}, S_{i2}, \cdots, S_{in}) \tag{1}$$

where S_{ij} is the j th measurement value of the i th sequence, assuming the length of the sequence is $|\boldsymbol{S}_i|$, if $|\boldsymbol{S}_i| \neq |\boldsymbol{S}_j|$,we call they are unequal length sequence.

Definition 1:

One sequence can represent a series of values which the sensor detects at a certain time, we assume that all data point in just one sequence comes from one fixed target, then *m* sequences can represent the measurement data of one target in *m* parameters, it can be defined as a matrix:

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ S_{m1} & \cdots & & S_{mn} \end{bmatrix} \tag{2}$$

We call it sequence matrix, each row in it represents a kind of characteristic parameter of measurement data from one target. Each sequence matrix represents a kind of measurement target.

2.2. The unequal length sequence similarity measurement DTW

The dynamic time warping DTW is a method to measure unequal length sequence, which will respectively launch sequence data along the two directions in the two-dimensional plane. So that we can get a lot of two tuples $c_k = (s_i, q_j)$, calculating the minimum values of the two tuples from the starting point of two tuples to the end and connecting all minimum values of the two tuples to form a bent path, as shown in Figure 1, the bent path is defined as unequal length sequence distance measurement D_{utwi}, D_{utwi} is as follows:

$$D_{utwi}(\boldsymbol{S_i}, \boldsymbol{Q_i}) = D_{base}(first(\boldsymbol{S_i}), first(\boldsymbol{Q_i})) + \min \begin{cases} D_{utwi}(\boldsymbol{S_i}, rest(\boldsymbol{Q_i})) \\ D_{utwi}(rest(\boldsymbol{S_i}), \boldsymbol{Q_i}) \\ D_{utwi}(rest(\boldsymbol{S_i}), rest(\boldsymbol{Q_i})) \end{cases} \tag{3}$$

where D_{base} can be represented by the p-norm, $first(\boldsymbol{S_i})$ and $first(\boldsymbol{Q_i})$ represent the first point of the sequence $\boldsymbol{S_i}$ and $\boldsymbol{Q_i}$ respectively, $rest(\boldsymbol{S_i})$ and $rest(\boldsymbol{Q_i})$ represent the rest point of the sequence $\boldsymbol{S_i}$ and $\boldsymbol{Q_i}$ without the first point respectively.

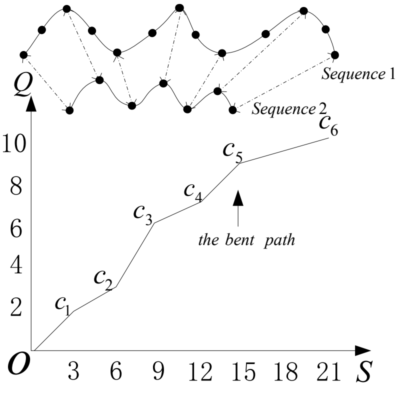

Fig.1 Diagram of DTW

Definition 2:

Based on distance measurement, according the relation between distance and similarity unequal sequence similarity between sequences $\boldsymbol{S}$ and $\boldsymbol{Q}$ can be defined as:

$$Sim_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i)=1-\frac{D_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i)}{D_{utw\max}} \tag{4}$$

where $D_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i)$ represents the distance measurement between the sequence S_i and $\boldsymbol{Q}_i$, $D_{utw\max}$ represents the maximum value in the distance measurement:

$$D_{utw\max}=\max\{D_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i),i=1,\cdots,\dim(\boldsymbol{S})\} \tag{5}$$

where $\dim(\boldsymbol{S})$ represents the number of sequences.

2.3. The meaning of the fluctuant sequence

The detection of characteristics of the sensor target is often interfered by the active or passive factors, resulting in one or some measurement data deviated from the actual data appearing larger fluctuant point as shown in figure 2. The accuracy of the measurement of unequal length sequence shown in Figure 1 by DTW is well, but in the actual process of sequence data, because of some interference some data points are often fluctuant points as shown in Figure 2, here measure the equal length fluctuant sequence by DTW will cause large measurement error. Whether the fluctuant point exists in a sequence is not predictive, one way is to increase the detection process, but it will introduce a new error and increase the processing time. Another method is that propose or improve novel measurement method in the unequal length fluctuant sequence measurement, our proposed algorithm is just in this way to study the relations between unequal length fluctuant sequence to get the similarity and to realize the sequence data association.

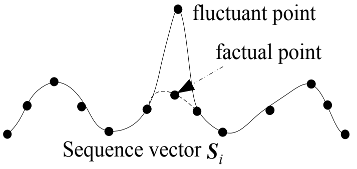

Fig.2 Diagram of fluctuant point

2.4. Sequence association

The association for the query sequence matrix $\boldsymbol{S}^i$ and the reference sequence matrix $\boldsymbol{Q}^j$, given an arbitrary measure of ρ , if

$$\rho(\boldsymbol{S}^i,\boldsymbol{Q}^j)\to 1 \tag{6}$$

or give the threshold ε satisfying

$$\rho(\boldsymbol{S}^i,\boldsymbol{Q}^j)\geq\varepsilon \tag{7}$$

We consider that sequence data matrix $\boldsymbol{S}^i$ is associated with sequence data matrix $\boldsymbol{Q}^j$. In fact, we often use the size of similarity to measure the degree of association. The more similar is the sequence matrix $\boldsymbol{S}$ to the sequence matrix $\boldsymbol{Q}$, the more associated is the sequence data matrix $\boldsymbol{S}$ with the sequence data matrix $\boldsymbol{Q}$.

3. Average weight 1- norm unequal length fluctuant sequence similarity measurement algorithm based on DTW

The 2.2 section introduces the basic concepts of similarity measurement of the unequal length fluctuant sequence, this section proposes an Average weight 1-norm unequal fluctuant sequence similarity measure algorithm based on DTW to deal with the large measurement error which the reference [11] took the maximum measurement when processing unequal length fluctuant sequence.

3.1. Description of algorithm

DTW is a common method of unequal length sequence measurement, but high computational complexity constrains its use. There are a lot of its improved methods, but still inconsistent. This section proposed an algorithm based on DTW to solve the data of fluctuant and repeated bending problem.

We deal with the similarity measurement of unequal length fluctuant sequence S_i and $\boldsymbol{Q}_i$ firstly and then get the similarity measurement of unequal length fluctuant sequence matrix S and $\boldsymbol{Q}$ based on it.

At first, constructing the absolute distance matrix sequence elements with the elements in the sequence S_i and $\boldsymbol{Q}_i$:

$$\boldsymbol{M}_i = \begin{bmatrix} |S_{i1} - Q_{i1}| & |S_{i1} - Q_{i2}| & \cdots & |S_{i1} - Q_{i|\boldsymbol{Q}_i|}| \\ |S_{i2} - Q_{i1}| & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ |S_{i|S_i|} - Q_{i1}| & \cdots & \cdots & |S_{i|\boldsymbol{S}_i|} - Q_{i|\boldsymbol{Q}_i|}| \end{bmatrix} \tag{8}$$

Assuming that $|\boldsymbol{S}_i| < |\boldsymbol{Q}_i|$, let $\boldsymbol{H} = (D_1, D_2, \cdots, D_{|\boldsymbol{H}|})$ denote the best element mappings sets that obtain the minimum absolute distance between S_i and $\boldsymbol{Q}_i$ .Where $|\boldsymbol{S}_i| \le |\boldsymbol{H}| \le |\boldsymbol{S}_i| \cdot |\boldsymbol{Q}_i|$, D_l is the minimum value in each column of $\boldsymbol{M}_i$,

$$D_l = \min\{|S_{im} - Q_{i1}|, |S_{im} - Q_{i2}|, \cdots, |S_{im} - Q_{i|Q_i|}|\}, m = 1, \cdots, |\boldsymbol{S}_i| \tag{9}$$

To describe $\boldsymbol{H}$ in words is that the minimum value for each column in the matrix elements forming $\boldsymbol{H}$, if $|\boldsymbol{S}_i| > |\boldsymbol{Q}_i|$ then minimum value for each row forming $\boldsymbol{H}$.

The reference [11] used the maximum value in $\boldsymbol{H}$ to represent the distance of unequal length fluctuant sequence:

$$D_{ui}^{[11]}(\boldsymbol{S}_i, \boldsymbol{Q}_i) = \max\{\boldsymbol{H}\} = \max\{D_1, D_2, \cdots, D_{|\boldsymbol{H}|}\} \tag{10}$$

Obviously, it did not meet data points' fluctuation in the figure 2, and the error is large. So we introduce the average weight 1-norm to measure the distance of unequal length fluctuant sequence, the average weight 1-norm is to weight the average weight to 1-norm to reduce the effect of the fluctuant points to the distance measurement, the average weight 1-norm distance measurement between the unequal length sequences $\boldsymbol{S}_i$ and $\boldsymbol{Q}_i$ is:

$$D_{ui}(\boldsymbol{S}_i, \boldsymbol{Q}_i) = \sum_{l=1}^{|\boldsymbol{H}|} \frac{1}{|\boldsymbol{H}|} \cdot D_l \tag{11}$$

where $\frac{1}{|\boldsymbol{H}|}$ is the average weight, it can reduce the effect of the fluctuant points to the distance measurement.

Considering that there may be many minimal value when calculating D_l according to equation(9), if directly summing and combining D_l will certainly cause repeated bending problems making the very small distance become large, considering the above problems, our proposed algorithm uses the modifying factors to modify the average 1-norm distance measurement, that is:

$$D_{utwi}(\boldsymbol{S}_i, \boldsymbol{Q}_i) = \xi \cdot D_{ui}(\boldsymbol{S}_i, \boldsymbol{Q}_i) = \xi \cdot \sum_{l=1}^{|\boldsymbol{H}|} \frac{1}{|\boldsymbol{H}|} \cdot D_l \tag{12}$$

where ξ is the modifying factors, it can modify the repeated bending problem of the repeated minimal value , ξ is defined as

$$\xi = \frac{1}{2} \cdot \left(1 + \frac{\max\{|\boldsymbol{S}_i|, |\boldsymbol{Q}_i|\}}{|\boldsymbol{H}|}\right) \tag{13}$$

So the distance measurement between the unequal length sequences $\boldsymbol{S}_i$ and $\boldsymbol{Q}_i$ can be written in DTW canonical form:

$$D_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i)=\xi\cdot\sum_{l=1}^{|\boldsymbol{H}|}\frac{1}{|\boldsymbol{H}|}\cdot\min\begin{cases}D_{base}(first(\boldsymbol{S}_i),first(\boldsymbol{Q}_i))\\D_l(\boldsymbol{S}_i,rest(\boldsymbol{Q}_i))\\D_l(rest(\boldsymbol{S}_i),\boldsymbol{Q}_i)\\D_l(rest(\boldsymbol{S}_i),rest(\boldsymbol{Q}_i))\end{cases}\tag{14}$$

After we have got the distance between the unequal length fluctuant sequence S_i and $\boldsymbol{Q}_i$, we could transform it into the similarity between the unequal length fluctuant sequence S_i and $\boldsymbol{Q}_i$ according to the relation between distance and similarity measurement, we just offer one relation as follows:

$$Sim_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i)=1-\frac{D_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i)}{D_{utw\max}}\tag{15}$$

where $D_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i)$ is the distance measurement between the sequence S_i and $\boldsymbol{Q}_i$, which can be calculated by equation (12), $D_{utw\max}$ is the maximum value in the distance measurement:

$$D_{utw\max}=\max\{D_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i),i=1,\cdots,\dim(\boldsymbol{S})\}\tag{16}$$

In the end, we fuse each sequence in the unequal length fluctuant sequence matrix according to the importance of the sequence in the matrix, the similarity measurement of the unequal length fluctuant sequence matrices S and $\boldsymbol{Q}$ can be weighted as

$$Sim_{utw}(\boldsymbol{S},\boldsymbol{Q})=\frac{1}{\dim(\boldsymbol{S})}\sum_{i=1}^{\dim(\boldsymbol{S})}\lambda_i\cdot sim_{utwi}(\boldsymbol{S}_i,\boldsymbol{Q}_i)\tag{17}$$

where λ_i the important weight of the sequence in the matrix is $\dim(\boldsymbol{S})$ is the number of sequences.

3.2. Algorithm process

The process of the Average weight 1- norm unequal length fluctuant sequence similarity measurement algorithm based on DTW is as follows:

```
Algorithm: Average weight 1-norm unequal length fluctuant sequence similarity measurement algorithm based on
    DTW.
Input: The query sequence matrix group $\{\boldsymbol{S}^i\}$ and reference sequence matrix group $\{\boldsymbol{Q}^j\}$;
Output: The similarity $\{\boldsymbol{Sim}_{utw}^i\}$ between the query sequence matrix group $\{\boldsymbol{S}^i\}$ and reference sequence matrix
    group $\{\boldsymbol{Q}^j\}$;
Step1:If the meaning of sequence in the query sequence matrix group $\{\boldsymbol{S}^i\}$ and reference sequence matrix group
    $\{\boldsymbol{Q}^j\}$ is not corresponding, do Step2, else do Step3;
Step2: Reconstruct the query sequence matrix group $\{\boldsymbol{S}^i\}$ and reference sequence matrix group $\{\boldsymbol{Q}^j\}$ according
    to the physical meaning of the row vector in the sequence data, forming new matrix;
Step3:For a sequence of matrix $\boldsymbol{S}^i$, select the corresponding row sequence $\boldsymbol{S}_k^i$ and $\boldsymbol{Q}^j$ according to the physical
    meaning of matrix forming the comparison of vector group: $\boldsymbol{Com}_{ek}^i=(\boldsymbol{S}_k^i,\boldsymbol{Q}_k^1,\boldsymbol{Q}_k^2,\cdots,\boldsymbol{Q}_k^{num(Q^j)})$, where $k$ is
    the row number in matrix $\boldsymbol{S}^i$, $num(\boldsymbol{Q}^j)$ is the number of matrix in the matrix group;
Step4:Calculate the absolute distance between the sequence $\boldsymbol{S}_k^i$ and $\boldsymbol{Q}_k^j$ according to DTW and form the absolute
    distance matrix group $\{\boldsymbol{M}_i^j\}$;
Step5:If $|\boldsymbol{S}_k^i|<|\boldsymbol{Q}_k^j|$, extract the minimum value in each column of $\{\boldsymbol{M}_i^j\}$ and forming the absolute distance
    set $\{\boldsymbol{H}_k^j\}$, then do Step7, else do step6;
Step6:Extract the minimum value in each row of $\{\boldsymbol{M}_i^j\}$ and forming the absolute distance set $\{\boldsymbol{H}_k^j\}$;
Step7:Calculate the distance between unequal length fluctuant sequence $\boldsymbol{S}_k^i$ and $\boldsymbol{Q}_k^j$ according to equations (11-14)
    and transform it into the similarity between unequal length fluctuant sequence $\boldsymbol{S}_k^i$ and $\boldsymbol{Q}_k^j$ according to
    equations (15), that is $\boldsymbol{Sim}_{utwk}(\boldsymbol{S}_k^i,\boldsymbol{Q}_k^j)$;
Step8: If $k \le \dim(\boldsymbol{S}^i)$, then do Step3-Step7, else do Step9;
Step9:Weight and combine $\boldsymbol{Sim}_{utwk}(\boldsymbol{S}_k^i,\boldsymbol{Q}_k^j)$ according to equation(17) to calculate the similarity between the query
    sequence matrix $\boldsymbol{S}^i$ and each matrix in the reference sequence matrix group $\{\boldsymbol{Q}^j\}$ and forming the
    similarity measurement of the unequal length fluctuant sequence matrices, that is $\boldsymbol{Sim}_{utw}^i$;
Step10:If $i \le num(\boldsymbol{S}^i)$, then do Step3-Step9, else do Step11;
Step11: end.
```

3.3. Association rule

Through the above research we can get unequal length fluctuant sequence matrix similarity measurement $\{\boldsymbol{Sim}_{utw}^i\}$. According to the association method described in section 2.4, sort the similarity measurement $\{\boldsymbol{Sim}_{utw}^i\}$ and associate the unequal length fluctuant sequence matrix according to the largest similarity association rules, we get

$$Sim_{i\max} = \max\{\boldsymbol{Sim}_{utw}^1, \boldsymbol{Sim}_{utw}^2, \cdots, \boldsymbol{Sim}_{utw}^{num(\boldsymbol{Q}^j)}\} \tag{18}$$

Then we associate the sequence data described by each maximum $Sim_{i\max}$, we get the association pairs. Which indicate that the data in unequal length fluctuant sequence matrix and associated data in unequal length fluctuant sequence matrix come from the same target.

Association flow chart of unequal length fluctuant sequence data is given in figure 3. It shows that the sensors detect the targets and get a series of sequence data, in this paper we assume they are unequal length and associated data points. Through preceding processing we get the sequence data matrix of one target going to be associated. What we have done in this paper is that associating these sequence data matrixes to other sequence data matrixes coming from their each target respectively measured by other sensors, whether these target are associated is going to be proved. We associate these data according our proposed Average weight 1-norm unequal length fluctuant sequence similarity measurement algorithm based on DTW and distinguish what sequence data coming from the same target.

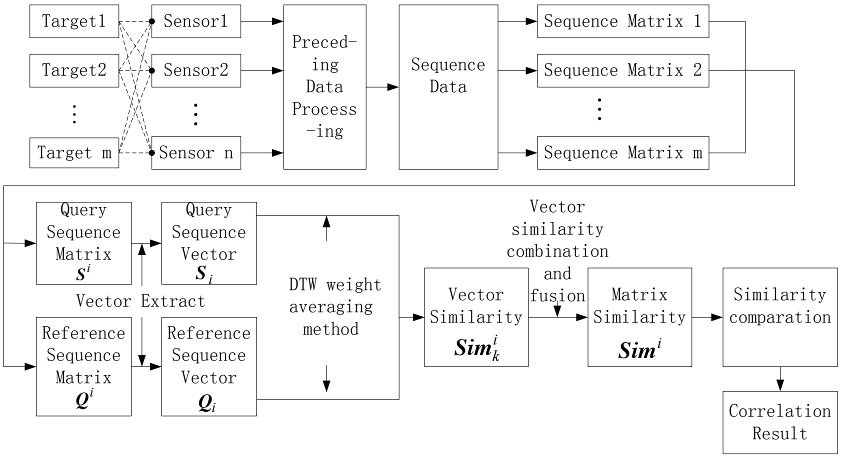
Target1
Target2
Target m
Sensor1
Sensor2
Sensor n
Preced-ing Data Process-ing
Sequence Data
Sequence Matrix 1
Sequence Matrix 2
Sequence Matrix m
Query Sequence Matrix S^i
Query Sequence Vector S_i
Vector Extract
Reference Sequence Matrix Q^i
Reference Sequence Vector Q_i
DTW weight averaging method
Vector similarity combination and fusion
Vector Similarity Sim_k^i
Matrix Similarity Sim^i
Similarity comparation
Correlation Result

Fig.3 Association flow chart of unequal length fluctuant sequence data

4. Simulation and analysis

This section makes two experiments to demonstrate the effectiveness of the proposed algorithm, the first calculates the similarity measurement of the unequal length fluctuant sequence to demonstrate its association effectiveness and the second compared our proposed algorithm to reference [11] with the increase of sequence length and fluctuant rate to analysis their performances.

4.1. The simulation environment

Assuming that sensors at location A detect radar carrier frequency RF, pulse repetition frequency PRF and pulse width PW three kinds of identity information of the target, through the preceding data processing, obtain two unequal length fluctuant sequence data matrices $\boldsymbol{S}^1$ and $\boldsymbol{S}^2$, which is made up of three sequences respectively representing RF, PRF and PW three kinds of parameters, Assuming all data in one sequence coming from the same target. Similarly, measurement data of the sensors at B are represented by four unequal length fluctuant sequence data matrices $\boldsymbol{Q}^1$, $\boldsymbol{Q}^2$, $\boldsymbol{Q}^3$ and $\boldsymbol{Q}^4$, the meaning of the four sequence data matrices is the same with sequence data matrices $\boldsymbol{S}^1$ and $\boldsymbol{S}^2$.What we have to do is that associating sequence data matrices $\boldsymbol{S}^1$ and $\boldsymbol{S}^2$ with which the sequence data matrices $\boldsymbol{Q}^1$, $\boldsymbol{Q}^2$, $\boldsymbol{Q}^3$ and $\boldsymbol{Q}^4$.

Firstly describe generation of simulation data before the simulation, sequence data can be produced by

$$data = a + \alpha \cdot b \qquad (19)$$

where a are the discrete sequence data values obeyed uniform distribution, b are the discrete sequences sequence data values obeyed Gauss distribution, α is the standard deviation for the Gauss distribution, can be used to describe the measurement error. Distribution range of measurement value and measurement error can be as shown in table 1. Assuming that they are all normalized.

Table 1 The simulation data table

data	$a \sim U(start, end)$			$b \sim N$	α
	RF	PRF	PW	(S,T)	
$\boldsymbol{Q}^1$	(8.9, 9.1)	(18.9, 21.1)	(6.5,7.5)	(0,1)	0.5
$\boldsymbol{Q}^2$	(8.5, 9.5)	(15.8, 24.2)	(5.2, 8.8)	(0,1)	0.5
$\boldsymbol{Q}^3$	(5.9, 6.1)	(8.9, 11.1)	(4.5, 5.5)	(0,1)	0.5
$\boldsymbol{Q}^4$	(4.9, 7.1)	(7.9, 12.1)	(3.4, 6.6)	(0,1)	0.5
$\boldsymbol{S}^1$	(8.8, 9.2)	(19.9, 20.1)	(6.1,7.9)	(0,1)	0.5
$\boldsymbol{S}^2$	(5.5, 6.5)	(9.5, 10.5)	(4.7, 5.3)	(0,1)	0.5

4.2. The simulation experiment

4.2.1 Unequal length fluctuant sequence simulation experiment

Sensor 100 measurement cycle forming the targets' measurement unequal length fluctuant sequence matrix of 100 lengths at A, and 200 lengths at B. In order to highlight the contrast effect of fluctuant data and repeated bending, we increase measurement error among 20 lengths data of the parameter PF in the sequence randomly, the specific data values are in table 1.

We produce simulation data of 2D alignment analysis results on macroscopic as shown in Figure 4, including the fluctuant data and repeated bending data points.

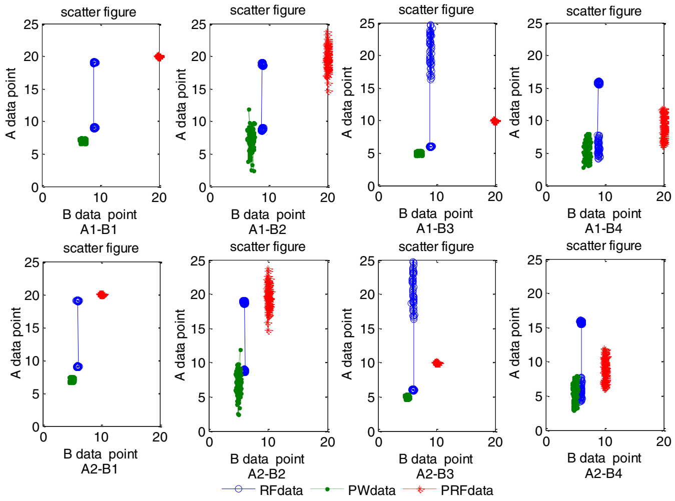

Fig.4 Two-dimensional comparison chart of unequal length sequence simulation data

The abscissa data in the figure 4 is unequal length fluctuant sequence matrix data from sensor at location B and the ordinate data is unequal length fluctuant sequence matrix data from sensor at location A. Figure A1-B1 to B4 is clustering figure of sequence data matrix S^1 from sensor at location A and sequence data matrix Q^1 to Q^4 from sensor at location B. Figure A2-B1 to B4 is clustering figure of sequence data matrix S^2 from sensor at location A and sequence data matrix Q^1 to Q^4 from sensor at location B, the jump of the parameter PF is because of what we have increased the measurement to reflect the fluctuant points and it is what we want to emphasize and simulate. It shows that the effect of figure A1-B1, A1-B3, A2-B1 and A2-B3 is better than others in clustering. Besides the slope of figure A1-B1and A2-B3 is almost 1 in value. It proves they are very similar. The concrete association results must be tested with algorithm and we offer the simulation of algorithm.

According to the proposed algorithm, we first calculate the absolute value matrix between the described target RF, PRF and PW three kinds of sequence measure data at A and measured data at B according to equations (8, 9). Then we can get the distance measurement between data at A and B and according to equation (11-14) and transform it into the sequence similarity measurement of unequal length fluctuant data at A and B according to equations (15, 16). In the end, according to the equation (17), weighting and combining the gotten unequal length fluctuant sequence similarity we get the similarity of sequence matrix data between A and B. The similarity results are as shown in table 2.

Table 2 Similarity degree table between measurement data and database data

Similarity	Q^1	Q^2	Q^3	Q^4
S^1	0.7461	0.4707	0.2435	0.0309
S^2	0.1783	0	0.6010	0.2467

Table 2 shows that, the similarity of measured data of the described goal 1 at location A and measured data of the described goal 1 at location B is the largest, similarity reaches more than 70%. The similarity of measured data of the described goal 2 at A and measured data of the described goal 3 at B is the largest, similarity reached more than 60%. According to the maximum similarity association criterion, we can judge that measured datas of described goals 1 and 2 at location A are associated with measured data of described goals 1 and 3 at location B.

4.2.2 Algorithm performance analysis

This experiment analyzes performance of our proposed unequal length fluctuant similarity measurement algorithm compared with the reference [11]. In the experiment above, if we do not use in accordance with the equations (11-14) to calculate the distance instead the maximum distance according to equation (10) in reference [11], then we would get another similarity between datas compared with our proposed similarity algorithm, the results are shown in table 3.

Table 3 Similarity degree table between two algorithms

Similarity		Q^1	Q^2	Q^3	Q^4
proposed algorithm	S^1	0.7491	0.4668	0.2461	0.0231
	S^2	0.1818	0.0039	0.6160	0.2589
reference 11algorithm	S^1	0.5384	0.4042	0.1150	0.0852
	S^2	0.1541	0.0098	0.4981	0.4707

Table 3 shows that, the effect of our proposed algorithm is better than reference [11], mainly because the reference [11] takes the maximum distance as the distance between datas ignoring some data points are fluctuant making the distance increased. It is only the partial distance which resulting in that this distance cannot reflect the whole sequence’s distances producing large error while our proposed algorithm can reduce the error with 1-norm method and normalized, getting satisfactory similarity result.

The sequence length has an important role to sequence similarity, it is necessary to measure the effect of sequence similarity with variation of the sequence length. The sequence length at location A is set to 0, 50,100, step 50 to 500, Similarly, the sequence length at location B are twice the length of location A, Simulate our proposed algorithm and reference [11] under both our simulation data from equation (19) and simulation data in reference [11] (produced by rand walk) and record the changes of similarity as shown in figure 5.

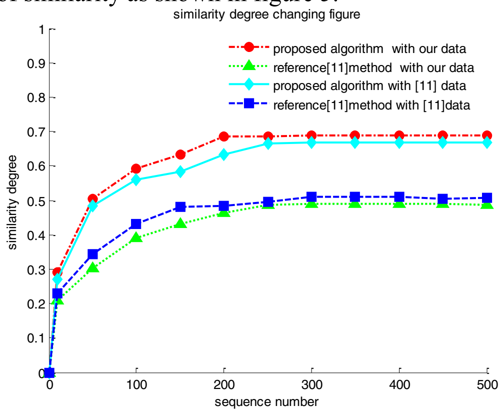

Fig.5 Changing figure of algorithm similarity degree as length

Figure 5 shows that, with the increase of the sequence length, both two similarities under two kinds of simulation data grow very quickly at first, but with the sequence length increases to a certain value, similarities are no longer increase with the increase of the sequence length; instead stay stable in the vicinity of a certain value. It shows that increasing sequence length in a certain range can improve the similarity, but when the similarity increases to the limit, increasing the sequence length is invalid. The reason for this change is that when the sequence length is short the sequence measurement error is large. It will inevitably destroy the similarity. With the increase of the sequence length, the measurement error reduces, similarity increases, but to a certain limit, it will not increase, here the sequence length at this time is the optimal length. So we should try to select the appropriate sequence length to measure the sequence similarity, otherwise the result error may be larger. Comparison between the calculated results of the two algorithms show that, our proposed algorithm in the calculation of similarity is superior to reference [11] algorithm, mainly because under the condition of a fixed fluctuant rate when length increases, the reference [11] algorithm can not deal with fluctuant point well, but our proposed algorithm due to the superiority in the treatment of fluctuant data, can get the appropriate similarity measurement.

Since our proposed algorithm is mainly used for processing the fluctuant point, so it is necessary to simulate the algorithm in the condition of different fluctuant rates. We define the fluctuant rate as follows:

$$saltatorial\ rate = \frac{the\ number\ of\ saltatorial\ points}{the\ number\ of\ sequence\ data\ points} \quad （20）$$

In this experiment, fluctuant rate of the sequence point at location A will be from 0 by step 0.05 to 1, recording the variations of similarities of our proposed algorithm and the reference [11] under two kinds of simulation data as shown in figure 6.

Figure 6 shows that, with the increase of the fluctuant rate at location A, both the similarity obtained by the two algorithms under two kinds of simulation data are all overall downward trend, finally stay stable in the vicinity of a fixed value, at this time the sequence similarity is weak. Comparison between two algorithms shows that the similarity obtained by the reference [11] algorithm has declined sharply at the increase of fluctuant rate, and the similarity obtained by our proposed algorithm is stable at a higher value, only when the fluctuant rate reaches about 25%, it will decrease significantly. It proved that our proposed algorithm average the fluctuant point effectively weaken the influence, while the [11] algorithm can not weaken fluctuant points' effect and lead to dramatic changes in similarity.

Because of the high calculation of DTW, which constrains its use, but we must use it because of its high precision. It is essential to discuss the processing time of our proposed algorithm and compared with the algorithm in reference [11].The simulation environment is the same as previously, the processing of two is as shown in figure 7.

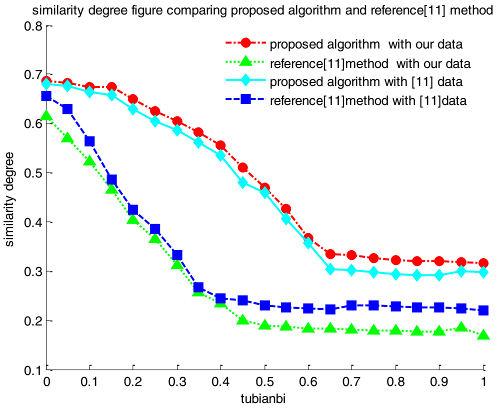

Fig.6 Comparison chart of two unequal length sequence algorithm similarity degree as fluctuant rate

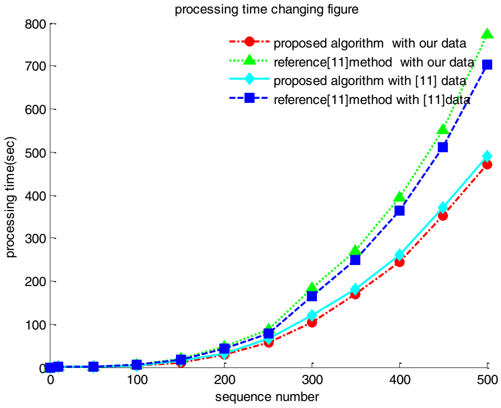

Fig.7 Comparison chart of two unequal length sequence algorithm processing time as length

Figure 7 shows that, with the increase of the sequence length, both the processing time of both two algorithms under two kinds of simulation data increase, especially when the sequence length increases to a relatively large, the processing time increases sharply. Compared with two algorithms, the processing time of our proposed algorithm is less than the algorithm in reference [11]. This is because our proposed algorithm use normalized 1-norm to measure the minimum absolute distance sets instead of the maximal distance measurement which must query all the sets and increase the processing time, especially when the sequence is very large, the processing time will increase quickly.

5. Conclusion

(1)Considering the fluctuant points of the sequence data interfered, we propose an average weight 1- norm unequal length fluctuant sequence similarity measurement algorithm based on DTW .This algorithm uses average weight 1-norm effectively weakening the effect of the fluctuant points and gets satisfactory similarity which reaches more than 60%. Combined with the association principle of maximum similarity we can easily associate unequal length fluctuant sequence data.

(2)Compared similarity obtained with the reference [11] algorithm in this paper, similarity obtained by our proposed algorithm is much larger. When data is fluctuant, what the maximum distance method measure is the fluctuant points' distance instead of the whole sequence's distance, which caused the similarity error of the entire sequence larger, but average weight 1- norm use 1- norm first and then weight average weight effectively weaken the fluctuant points avoiding reference [11]'s shortcoming.

(3)The similarity obtained by algorithm increase to a fixed maximum value with the sequence length becomes larger, which corresponds the optimal length. With the increase of the fluctuant rate, the effect of our proposed algorithm is better than reference [11] algorithm. It proves that our proposed algorithm can be used for unequal length fluctuant sequence similarity measurement well within a certain range of the fluctuant rate.

Acknowledgements

We would like to thank all the donors of the research institute of information fusion.

References

[1] AgrawalR, FaloutsosC, SwamiA, Efficient Similarity Search in Sequence Databases[C]//Proc. 4th Int. l Conf Foundations of Data Organization and Algorithms, Chicago, IL, Oct, 1993: 69-84.

[2] RafieiD, MendelzonAO.Querying time series data based on similarity [J]. IEEE Trans on Knowledge and Data Engineering, 2000, 12(5): 675-693.

[3] Wang Chang-Zhou, Wang Xiao-yang. Multilevel filtering for high dimensional nearest neighbor search[C]// Proc of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Dallas: ACM Press, 2000: 37-43.

[4] Korn, F., Jagadish, H.V., Faloutsos, C.: Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences[C]//Proceedings of ACM SIGMOD International Conference on Management of Data (1997) 289-300.

[5] HuhtallaY, Krkkinen J, ToivonenH.Mining for similarities in aligned time series using wavelets[C]//Proc of Data Mining and Knowledge Discovery: Theory, Tools, and Technology. Orlando: [s.n.], 1999:150-160.

[6] StruzikzR, SiebesA. Measuring time series similarity through large singular features revealed with wavelet transformation[C]//Proc of the International Workshop on Database and Expert Systems Application. Florence: IEEE Computer Society Press, 199:162-166.

[7] PopivanovI, MlitterrJ. Similarity search over time series data using wavelets[C]//Proc of the 18th International Conference on Data Engineering. San Jose: IEEE Computer Society Press, 2002:212-221.

[8] Zhang Haiqin, Cai Qingsheng. Time series similarity Querying based on wavelets [J]. Computer Journal, 2003, 26(3): 373-37. (in Chinese)

[9] KeoghE. Data mining and machine learning in time series database[C]//Proc of the 5th Industrial Conference on Data Mining (ICDM).Leipzig: [s.n.], 2005.

[10] KeoghE, ChakrabartiK, PazzaniM, etal. Dimensionality reduction for fast similarity search in large time series databases [J]. Journal of Know ledge and Information Systems, 2001, 3 (3):263-286.

[11] Sang-Wook Kim, Sanghyun Park, Wealey W.Chu. An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases [C]// Proceedings 17th International Conference on Data Engineering. Heidelberg, 2001:607-614.

[12] Thanawin Rakthanmanon, Bilson Campana, KeoghE, etal. Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping[C]// Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. Beijing, 2012:262-270

[13] AgrawalR, LinkI, SawhneyhS, etal. Fast similarity search in the presence of noise, scaling, and translation in time series databases[C]//Proc of the 21st Conference on Very Large Databases. San Francisco: Morgan Kaufmann, 1995:490-501.

[14] KeoghE. Exact indexing of dynamic time warping[C]//Proc of the 28th International Conference on Very Large Databases. Hong Kong: Morgan Kaufmann, 2002:406-417.

[15] RathtM, ManmathaR. Lower bounding of dynamic time warping distances for multivariate time series, Technical ReportMM-40[R]. Amherst: Center for Intelligent Information Retrieval Technical Report, University of Massachusetts, 2003.

[16] ChiusS, KeoghE, HartD, PazzaniM. Iterative deepening dynamic time warping for time series[C]//Proc of the 2nd SIAM International Conference on Data Mining. Baltimore: SIAM Press, 2002:148-156.

[17] KeoghE, PazzaniM. Derivative dynamic time warping [C]//Proc of the 1st SIAM International Conference on Data Mining. Chicago: SIAM Press, 2001:209-211.

[18] KeoghE, ChakrabartiK, PazzaniM, etal. Locally adaptive dimensionality reduction for indexing large time series databases [J]. ACM Transactions on Database Systems, 2002, 27 (2):188-228.

[19] GeX, SmythP. Deformable markov model templates for time series pattern matching[C] //Proc of the 6th ACM SIGKDD Intl Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000: 81-90.

[20] Michail Vlachos, Marios Hadjieleftheriou, Dimitrios Gunopulos, KeoghE.Indexing Multidimensional Time-Series[J] The Vldb Journal - VLDB ,2006,15(1):1-20.

[21]BerndtD, CliffordJ. Using dynamic time warping to find patterns in time series[C]//AAAI94 Workshop on Knowledge Discovery in Databases. Seattle, 1994.