International Journal on Data Science and Technology
2021; 7(3): 54-61
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20210703.12
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Morphological Similarity Clustering and Its Applications in Anomaly Detection of Time Series

Hu Shaolin[1, *], Huang Xiaomin[2], Su Naiqian[1], Wang Shihua[1]

[1]Institute of Intelligent Perception and System Safety, Guangdong University of Petrochemical Technology, Maoming, China
[2]School of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China

Email address:
hfkth@gdupt.126.com (Hu Shaolin)
[*]Corresponding author

To cite this article:
Hu Shaolin, Huang Xiaomin, Su Naiqian, Wang Shihua. Morphological Similarity Clustering and Its Applications in Anomaly Detection of Time Series. *International Journal on Data Science and Technology*. Vol. 7, No. 3, 2021, pp. 54-61. doi: 10.11648/j.ijdst.20210703.12

Received: August 1, 2021; **Accepted**: August 16, 2021; **Published**: August 27, 2021

Abstract: Time series data clustering is an important branch and difficult topic in the field of data clustering. In this paper, the definition of temporal data morphological similarity is proposed, a set of affine invariant morphological similarity measurement methods of time series data is established, and a morphological clustering algorithm based on morphological similarity measurement is developed. Using morphological similarity measurement of time series data, two groups of abnormal change detection algorithms for time series data are established, which can be used to detect the morphological consistency of different periodical sampling series in the same time series and the morphological consistency among several time series in the same period. Based on these algorithms stated above, the multiple monitoring algorithms are proposed, which can be used to monitor states of many kinds of industry process. The effectiveness of the methods and algorithms is verified with theoretical deduction and simulation results. Simulation results show that these algorithms are very valuable for mining, clustering, modeling, statistical learning of multi-source time series data, as well as the detection and diagnosis of abnormal process changes.

Keywords: Data Clustering, Time Series, Change Detection

1. Introduction

A time series is a series of data that occur over different sampling times. Time series data exist in many kinds of areas, such as the process control, the target motion navigation, the climate change, the stock market and other areas [1]. In most of the complex system control engineering fields, some of the time series from different sensors are similar in morphological variations. In order to monitor the operational status of complex systems, it is necessary to detect the abnormal changes in time series of data from different sources. Under the normal circumstances, the monitoring results should be similar for different time series with similar morphologies. Monitoring a large number of time series with similar shapes can result in large cost overhead.

All of the morphological similar sequences are effectively clustered into one class, which greatly facilitates the monitoring of complex system operation process under the condition of big data, and also helps to find the similar changes in the abnormal sequence, so as to improve the reliability of monitoring.

Morphological clustering is a useful tool for discovering knowledge from a large number of time series data. The difficulty of time series morphological clustering is how to measure the morphological similarity of different time series.

Choosing an appropriate similarity measure is very important for us to cluster multi-source time series data. Some means were considered in research of time series data mining in recent years [2]. Aghabozorgi et al. (2015) indicated that the dynamic time warping (DTW) [3, 4] is the most widely used measures [5]. DTW was used as an important tool for time series analysis, which was one of the important research topics [6].

However, the DTW algorithm has a few serious dis-advantages, such as the computational complexity of $O(N^2)$, which limits its practical use to the analysis of time series big data. Zhou Keyi, et al (2018) pointed out that the DTW

algorithm lacks time shift invariance and spatial mobility invariance [7]. In other words, when a distance based similarity measure (such as DTW [8]) is used to measure two different time series that are visually similar in shape, even if one of the time series is a fixed value offset sequence of the other time series, the judgment given may be a misjudgment with different shapes.

In the field of practical engineering, especially in the process monitoring, we are more concerned with changing patterns than with magnitude. For example, when the attitude control state of the spacecraft is monitored, the temperature variation curves of different gyroscopes are different, but the shapes show similar variation characteristics. When testing whether the gyroscope is working normally, we pay attention not to the single temperature change amplitude, but to whether the change patterns of multiple temperature curves are consistent. Using these inconsistencies, the gyroscopes with abnormal changes can be found.

In order to overcome the shortcomings of some similarity measure existed for time series and clustering algorithm lacking the affine invariance in spatiotemporal space, a new similarity measure method is proposed in section 2. In section 3, a multi-source time series clustering algorithm based on morphological similarity is established; in section 4, two groups of anomaly detection algorithms based on morphological similarity measurement and morphological clustering algorithm are constructed for two different types of practical problems. In section 5, the simulation calculation and result analysis of the algorithms given in this paper are given. The results further show that the algorithms proposed are effective.

2. Morphological Similarity Measure of Time Series

In this section, a new approach is proposed to measure morphological similarity of different time series.

2.1. Morphological Similarity of Time Series

The similarity of time series means that two time series have similar characteristics in a sense. There are many definitions of similarity, such as time-varying similarity, shape similarity and model similarity [9, 10], etc.

This paper focuses on the morphological similarity and how to judge the morphological similarity of time series.

Definition 1. If a time series moves to the left or right, or moves up or down, or stretches or compresses in amplitude, it can be basically consistent or coincides with another time series, then two time series are considered to be similar in form, or called similar in form.

Obviously, the above definition can ensure that the morphological similarity has time-shifting invariance and spatial mobility invariance, as well as the invariance of amplitude stretching changes and amplitude compression changes.

Undoubtedly, this definition is reasonable and appropriate. For example, a time series with a sine wave change, whether it is shifted left, right, up, down, stretched, compressed, etc., we still think it as a sine wave curve.

2.2. Model Representation of Morphological Similarity

Considering that for two time series, the time-lapse similarity is relatively easy to deal with. Therefore, this section mainly considers the similarity of amplitude change, including the similarity of amplitude translation, stretching and compression in time series [11].

Theorem 1: Two series $S_j = \{y_j(t_i) : t_i = t_0 + ih, i = 1,2,3,\cdots\}$ $(j = 1,2)$ are morphologically similar if and only if there are two unknown constants a and b which satisfy the following formula

$$y_1(t_i) = a + by_2(t_i) \quad (t_i = t_0 + ih, i = 1,2,3,\cdots) \tag{1}$$

in the case that these two time series data do not contain any stochastic errors.

Proof: It is obvious that one of the time series can be merged with the other time series by up, down, and amplitude scaling, if these two series satisfy the formula (1), The reverse form is also true.

Theorem 1 tells us that if two time series are similar in shape, one of them must be overlapped with the other after translation and amplitude compression or stretching. A new time series formed by time axis translation and amplitude compression or stretching is similar to that before transformation. In other words, the morphological features of time series are affine invariant [12].

Considering that the actual sampling time series inevitably contains errors, two morphologically similar time series can be modeled as

$$y_1(t_i) = a + by_2(t_i) + \varepsilon(t_i) \quad (t_i = t_0 + ih, i = 1,2,3,\cdots) \tag{2}$$

where the parameters a and b are unknown constants, the error series $\{\varepsilon(t_i)\}$ is a zero mean stationary sequence.

2.3. Measure of Morphological Similarity

Mathematical statistics tell us that for any two time series $\{S_1, S_2\}$ which satisfy the relation (2), it is possible to find constants (a,b) to minimize the difference between time series S_1 and S_2 . In fact, using the least squares method,

$$f(a,b) = \sum_t \{y_1(t) - (a + by_2(t))\}^2 \xrightarrow{a,b} \min \tag{3}$$

these two parameters can be determined as follows

$$\left.\begin{aligned} \hat{a} &= \frac{\hat{E}y_1\hat{E}y_2^2 - \hat{E}y_2\hat{E}(y_1y_2)}{\hat{E}y_2^2 - (\hat{E}y_2)^2} \\ \hat{b} &= \frac{\hat{E}(y_1y_2) - \hat{E}y_1\hat{E}y_2}{\hat{E}y_2^2 - (\hat{E}y_2)^2} \end{aligned}\right\} \tag{4}$$

where the operator $\hat{E}$ is the mean operator to calculate the mean.

For any sample segments of these two long time series S_1 and S_2, the estimators (4) can be expressed in matrix form as follows

$$\begin{pmatrix} \hat{a}_s \\ \hat{b}_s \end{pmatrix} = \begin{pmatrix} n-s+1 & \sum_{i=s}^{n} y_1(t_i) \\ \sum_{i=s}^{n} y_1(t_i) & \sum_{i=s}^{n} y_1^2(t_i) \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=s}^{n} y_2(t_i) \\ \sum_{i=s}^{n} y_1(t_i) y_2(t_i) \end{pmatrix} \quad (5)$$

where, t_s and t_n denote the start time and end time respectively.

Using equation (5) to construct a similarity measure [13] for two time series segments

$$D_s(y_2, y_1) = \sum_{i=s}^{n} (y_2(t_i) - (\hat{a}_s + \hat{b}_s y_1(t_i))^2 \quad (6)$$

In the form of vector and matrix, equation (6) can be expressed as follows

$$D_s(y_2, y_1) = \begin{pmatrix} y_2(t_s) \\ \vdots \\ y_2(t_s) \end{pmatrix}^T (I - H_s^n) \begin{pmatrix} y_2(t_s) \\ \vdots \\ y_2(t_s) \end{pmatrix} \quad (7)$$

where, the matrix

$$H_s^n = X(X^T X)^{-1} X^T, \quad X = \begin{pmatrix} 1 & \cdots & 1 \\ y_1(t_s) & \cdots & y_1(t_n) \end{pmatrix}^T$$

Obviously, the similarity measure D_s is approximately equal to zero if the segment $\{y_1(t_s), \cdots, y_1(t_n)\}$ from S_1 are similar morphologically to the segment $\{y_2(t_s), \cdots, y_2(t_n)\}$ from S_2.

Theorem 2: The measure $D_s(y_2, y_1)$ of morphological similarity between two time series $\{y_1(t_i) \mid i = s, \cdots, n\}$ and $\{y_2(t_i) \mid i = s, \cdots, n\}$ is affine invariant about the time series $\{y_1(t_i) \mid i = s, \cdots, n\}$, which means that for any constants a and b, the measure D_s satisfies the following expression

$$D_s(y_2, a + by_1) = D_s(y_2, y_1) \quad (8)$$

Proof: An affine transformation, also known as affine mapping, refers to a vector space undergoing a linear transformation and a translation, transforming into another vector space. For any constant a and b, the affine transformation of the time series $\{y_1(t_i)\}$ is $\tilde{y}(t_i) = a + by_1(t_i)$. So, we have

$$D_s(y_2, \tilde{y}) = \begin{pmatrix} y_2(t_s) \\ \vdots \\ y_2(t_n) \end{pmatrix}^T (I - \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T) \begin{pmatrix} y_2(t_s) \\ \vdots \\ y_2(t_n) \end{pmatrix} \quad (9)$$

where $\tilde{X} = \begin{pmatrix} 1 & \cdots & 1 \\ \tilde{y}(t_s) & \cdots & \tilde{y}(t_n) \end{pmatrix}^T$. So, we have

$$(\tilde{X}^T \tilde{X})^{-1} = \begin{pmatrix} 1 & \tilde{y}(t_s) \\ \vdots & \vdots \\ 1 & \tilde{y}(t_n) \end{pmatrix}^T \begin{pmatrix} 1 & \tilde{y}(t_s) \\ \vdots & \vdots \\ 1 & \tilde{y}(t_n) \end{pmatrix}^{-1} = \frac{1}{n-s+1} \begin{pmatrix} 1 & \hat{E}\tilde{y} \\ \hat{E}\tilde{y} & \hat{E}\tilde{y}^2 \end{pmatrix}^{-1}$$

Namely,

$$(\tilde{X}^T \tilde{X})^{-1} = \frac{1}{(n\text{-s}+1)(\hat{E}\tilde{y}^2 - (\hat{E}\tilde{y})^2)} \begin{pmatrix} \hat{E}\tilde{y}^2 & -\hat{E}\tilde{y} \\ -\hat{E}\tilde{y} & 1 \end{pmatrix}$$

Using the notation

$$W_y = (\hat{E}y_2)^2 \hat{E}\tilde{y}^2 - 2\hat{E}y_2 \hat{E}\tilde{y} \hat{E}(y_2\tilde{y}) + [\hat{E}(y_2\tilde{y})]^2]\}$$

then we have

$$D_s(y_2, \tilde{y}) = (\text{n-s}+1)\left\{\hat{E}y_2^2 - \frac{W_y}{\hat{E}\tilde{y}^2 - (\hat{E}\tilde{y})^2}\right\} \quad (10)$$

Inserting the affine transformation $\tilde{y}(t_i) = a + by_1(t_i)$ into the formula (9), we have

$$\begin{aligned} \hat{E}\tilde{y} = \hat{E}(a + by_1) &= \frac{1}{n-s+1} \sum_{i=s}^{n} (a + by_1(t_i)) \\ &= a + \frac{1}{n-s+1} \sum_{i=s}^{n} by_1(t_i) = a + b\hat{E}y_1 \end{aligned} \quad (11)$$

$$\begin{aligned} \hat{E}\tilde{y}^2 = \hat{E}(a + by_1)^2 &= \frac{1}{n-s+1} \sum_{i=s}^{n} (a + by_1(t_i))^2 \\ &= a^2 + 2ab \frac{1}{n-s+1} \sum_{i=s}^{n} y_1(t_i) + \frac{1}{n-s+1} \sum_{i=s}^{n} b^2 y_1^2(t_i) \\ &= a^2 + 2ab\hat{E}y_1 + b^2 \hat{E}y_1^2 \end{aligned} \quad (12)$$

$$D_s(y_2, \tilde{y}) = Y_2^T Y_2 - \frac{\text{n-s}+1}{\hat{E}(a+by_1)^2 - (\hat{E}(a+by_1))^2} \{(\hat{E}y_2)^2 \hat{E}(a+by_1)^2 - 2\hat{E}y_2 \hat{E}(a+by_1)\hat{E}(y_2(a+by_1)) + (\hat{E}y_2(a+by_1))^2\}$$

$$= (n-s+1)\{\hat{E}y_2^2 - \frac{\text{n-s}+1}{b^2[\hat{E}y_1^2 - (\hat{E}y_1)^2]} \{b^2(\hat{E}y_2)^2 \hat{E}y_1^2 - 2b^2 \hat{E}y_1 \hat{E}y_2 \hat{E}(y_1 y_2) + b^2[\hat{E}(y_1 y_2)]^2\}\}$$

So, we get

$$D_s(y_2,\tilde{y}) = D_s(y_2,y_1) \tag{13}$$

This formula (12) shows that the similarity measure D_s is affine invariant on $\{y_1(t_i) \mid i = s,\cdots,n\}$.

Theorem 2 shows that the similarity measure D defined in formula (6) is invariant to affine transformation and will not change the size of the measure due to the translation or expansion of the time series. Therefore, it is appropriate to use the measure D to measure the degree of morphological similarity between two time series, which is called morphological similarity measure MSM.

Furthermore, through the comparison of measurement values, we can filter out all time series with similar morphology from a large number of time series sets, which is of great significance for time series data mining, morphological analysis and mutation detection.

3. MSM Based Clustering of Multi-source Time Series

In the field of industrial production and large-scale engineering, in order to monitor the production process or equipment working conditions, it is very necessary to accurately judge how the measurement data from different sensing channels change with the advancement of time, for example, it is a fixed value change, or it changes with time. In this regard, using the above similarity measure, we can consider establishing a big data clustering algorithm based on morphological "seed".

Step 1. Set up several typical morphological patterns or seeds, such as constant values, monotonically increasing or monotonically decreasing, sine waves, cosine waves, quadratic parabola, etc, build a collection set of seed forms or patterns that can be expanded. In other words, it is assumed that there are k typical morphological sequences in the seed selection set $\{\overline{S}_1,\cdots,\overline{S}_k\}$.

Step 2. Take any time series y in the time series big data set B_s, and calculate the median similarity between y and the seed set $\overline{S}_i$

$$D_{\text{med}}(y \mid \overline{S}_i) = \underset{y_i \in \overline{S}_i}{med}\{D_s(y,y_i)\} \tag{13}$$

Step 3. Calculate the median of similarity measure for all sequence pairs in any two subsets $\overline{S}_i$ and $\overline{S}_j$

$$D_{\text{med}}(\overline{S}_i,\overline{S}_j) = \underset{y_i \in \overline{S}_i, y_{ji} \in \overline{S}_j}{med}\{D_s(y_i,y_j)\} \tag{14}$$

Step 4. The maximum and minimum values of the median series of similarity measure for all two seed sets $\overline{S}_i$ and $\overline{S}_j$ from all sequence pairs are calculated as follows

$$\begin{cases} M_D = \max\{D_{\text{med}}(\overline{S}_i,\overline{S}_j) \mid i,j = 1,2,\cdots,k\} \\ L_D = \min\{D_{\text{med}}(\overline{S}_i,\overline{S}_j) \mid i,j = 1,2,\cdots,k\} \end{cases} \tag{15}$$

Step 5. The minimum value of similarity median series of time series between y and $\{\overline{S}_1,\cdots,\overline{S}_k\}$

$$K_S(y) = \min\{D_{\text{med}}(y \mid \overline{S}_j) \mid j = 1,2,\cdots,k\} \tag{16}$$

Step 6. Comparison: if $K_S(y) \geq M_D$, there must be two seed sets $\overline{S}_{i0}$ and $\overline{S}_{j0}$ which satisfy the following relationship $D_{\text{med}}(\overline{S}_{i0},\overline{S}_{j0}) = L_D$. To merge the nearest seed set $\overline{S}_{i0}$ and $\overline{S}_{j0}$ and make it a new set $\overline{S}_{i0}$, namely $\overline{S}_{i0} \bigcup \overline{S}_{j0} \Rightarrow \overline{S}_{i0}$, and to construct seed set $\overline{S}_{j0} = \{y\}$ so as to update the seed collection. If $K_S(y) < M_D$, there must be at least one seed set $\overline{S}_{j0}$ satisfying $K_S(y) = D_{\text{med}}(y \mid \overline{S}_{j0})$, then the time series y is incorporated into the seed set $\overline{S}_{j0}$, namely $\overline{S}_{j0} \bigcup \{y\} \Rightarrow \overline{S}_{ji0}$.

Step 7. Return to step 2 and repeat steps 2-6 until all sequences are sorted.

The clustering process stated above can be intuitively represented by flow chart, as shown in Figure 1. Since the core algorithm adopts median operator instead of simple average, the above-mentioned time series clustering algorithm based on affine invariant morphological similarity MSM has good fault tolerance to avoid classification errors caused by outlier data.

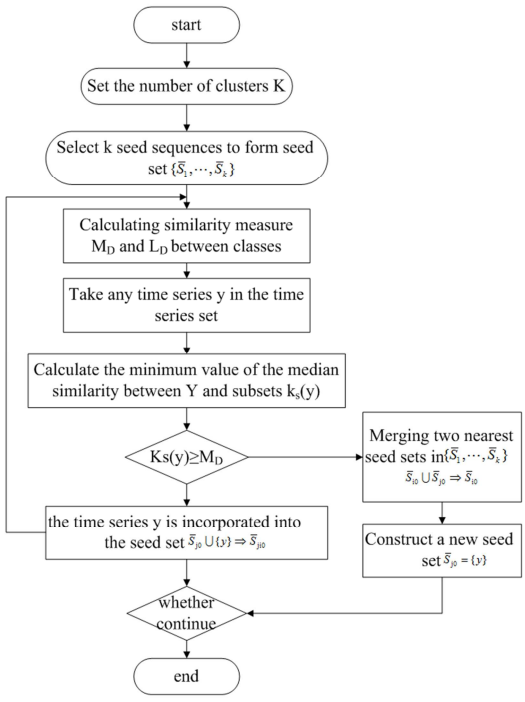


Figure 1. *Morphological clustering algorithm of time series data based on MSM.*

4. MSM Based Anomaly Detection of Time Series

MSM is very suitable for the detection of abnormal changes in complex processes such as process industry [14], including the detection of abnormal changes in different periods of the same process with periodic characteristics, the detection of abnormal morphology of the same type of equipment under different working conditions, and the detection of abnormal changes of monitoring data in different parts of complex structures.

4.1. Detection of Abnormal Changes in Different Periods with Periodic Components

In many practical fields, such as ethylene cracking, we usually encounter such a situation: a chemical plant or system has a periodic continuous operation, which makes the measured data contain the periodic component of the threshold, for example, the temperature data of ethylene cracking furnace, whether the sampling data of different periods are similar or not is often used as an important basis to judge the abnormal working condition of the system.

In order to simplify the description, the sampling data is abbreviated as $\{y(t_i): t_i = t_0 + ih\}$, where t_0 is the starting sampling time, $h = \frac{T}{N}$ is the sampling interval, T is the process cycle of each batch in the production process, and N is the sampling data points in each cycle. The sampling data sequence $\{y(t_i)\}$ of the long-time running process of the detected object is divided into time series data segments with N sampling points in each segment

$$S_j = \{\tilde{y}_j(t_i): \tilde{y}_j(t_i) = y(t_0 + (j-1)T + ih))\} \quad (17)$$

Reference formula (7), cycle to calculate the MSM between each adjacent segment S_{j+1} and S_j

$$D_j = \begin{pmatrix} \tilde{y}_{j+1}(t_1) \\ \vdots \\ \tilde{y}_{j+1}(t_N) \end{pmatrix}^T (I - X(X^T X)^{-1} X^T) \begin{pmatrix} \tilde{y}_{j+1}(t_1) \\ \vdots \\ \tilde{y}_{j+1}(t_N) \end{pmatrix} \quad (18)$$

Where $X = \begin{pmatrix} 1 & \cdots & 1 \\ \tilde{y}_j(t_1) & \cdots & \tilde{y}_j(t_N) \end{pmatrix}^T$.

If the segments of the time series with periodic components shown in equation (17) are similar in shape, then the MSM sequence $\{D_j, j = 1,2,...\}$ shown in equation (18) is a non negative first-order weakly stationary sequence. Take the default value

$$C = 1.5 med\{D_j, j = 1,2,..., N\} \quad (19)$$

Construct a MSM based temporal data mutation detection corridor, as shown in Figure 2:

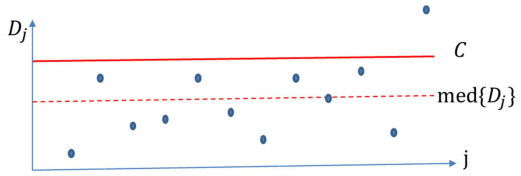


Figure 2. *Anomaly detection corridor based on morphological similarity.*

Detection method: In scatter figure 2, if there is a point D_{jo} over the upper wall of the detection corridor, that is $D_{jo} > C$, then it can be judged that the change shape of time series data fragment S_{jo+1} is abnormal. Further, if D_{jo+1} and later points still return to the corridor, then it can be determined that the change mode of time series data starts from segment S_{jo+1}, That is before abnormal morphologies occur S_{jo} and after S_{jo+1} are two different modes.

4.2. Abnormal Change Detection of Different Objects with Similar Features

For large production units such as cracking furnace, the key characteristic quantity (such as furnace tube temperature) usually requires multiple sets of sensor equipment to collect data from different parts to reflect the changes of different parts. Although the amplitude and speed of sampling data of different parts may be different, they all contain the working state of corresponding parts. Using proper fusion or comparison of multi-source sampling time series, it is helpful to identify and judge whether the working state of corresponding parts is normal.

For convenience of description, it is assumed that there are M-group sensing devices participating in data acquisition of complex device, and the sampling data obtained by these M-group sensing devices are abbreviated as time series respectively $S_j = \{y_j(t_i): t_i = t_0 + ih\}$ $(j = 1,2,...,M)$, where t_0 is the initial sampling time and h is the sampling time interval.

Setting the number of clusters k (k<M/3) and using the MSM-based time series data morphological clustering algorithm stated in Figure 1, for M sampling time series data segments

$$\begin{cases} S_{1,n} = \{y_1(t_i): t_i = t_0 + ih, \ \ i = 1,...,n\} \\ \vdots \\ S_{M,n} = \{y_M(t_i): t_i = t_0 + ih, \ \ i = 1,...,n\} \end{cases} \quad (20)$$

the morphological clustering is shown in Figure 3

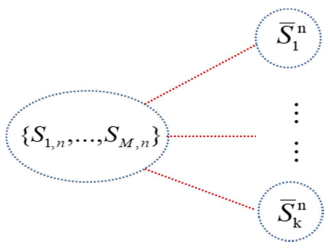


Figure 3. *Schematic diagram of morphological clustering based on MSM.*

If the process continues from time t_n to t_N, correspondingly, the sample time series data grows segments

$$\begin{cases} S_{1,N} = \{y_1(t_i) : t_i = t_0 + ih, \ \ i = 1,...,N\} \\ \vdots \\ S_{M,N} = \{y_M(t_i) : t_i = t_0 + ih, \ \ i = 1,...,N\} \end{cases} \tag{21}$$

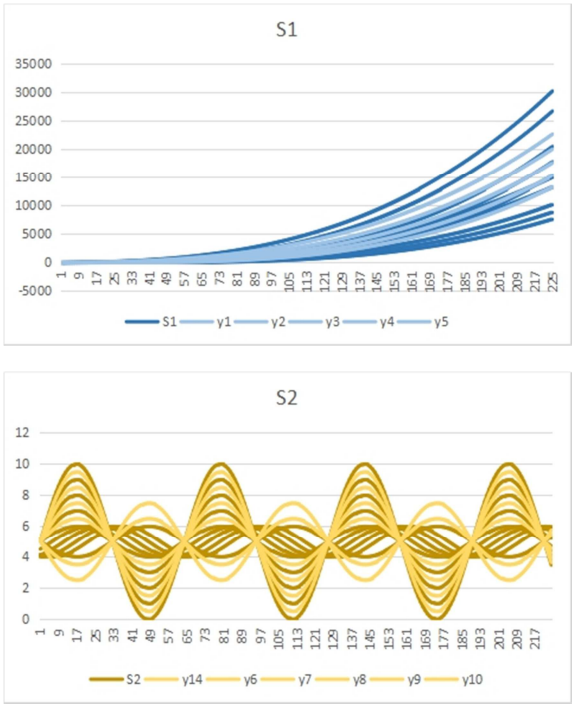
Similarly, M time series with growing segments (21) are clustered again to obtain clustering $\{\overline{S}_1^N, \cdots, \overline{S}_k^N\}$.

Detection method: If the two clustering results are different, that is $\{\overline{S}_1^n, \cdots, \overline{S}_k^n\} \neq \{\overline{S}_1^N, \cdots, \overline{S}_k^N\}$, Then it can be judged that the system has abnormal state change from time t_n to time t_N.

5. Simulation Calculation and Result Analysis

5.1. Clustering Example

In order to verify the validity and feasibility of the time series clustering algorithm proposed in this article, Setting the number of clusters to 3, three typical morphological sequences of sine function, cosine function, and third-order polynomial are selected to form the seed set, recorded as $\{\overline{S}_1, \overline{S}_2, \overline{S}_3\}$ with time shifting and changing the amplitude and other forms of change, time series data are generated for the three typical morphological sequences, which are recorded as $\{y_1,...,y_5\}$、$\{y_6,...,y_{10}\}$、$\{y_{11},...,y_{15}\}$ constitute time series data set $\{y\}$. Use this algorithm to cluster time series data sets, the cluster results are shown in Figure 4.



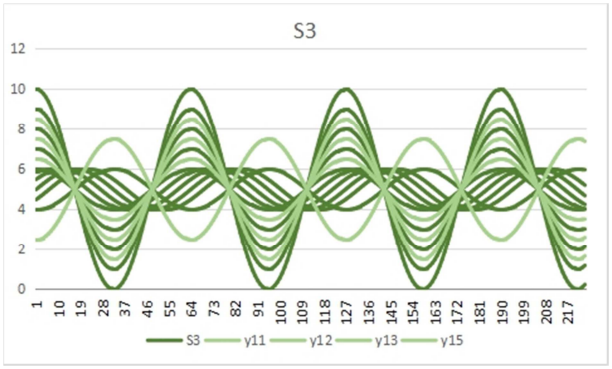


Figure 4. *Clustering results of three kinds of time series.*

5. Example of Mutation Detection

As a simulation analysis, in order to verify the ability of the monitoring algorithm to detect the occurrence of time series, a group of time series data is spliced, as shown in Figure 5.

The first half (the first 100 sampling points) is the sine change data, and the second half (the last 100 sampling points) is the cosine data. Set n = 5, and its timing data anomaly detection channel based on MSM is shown in Figure 6.

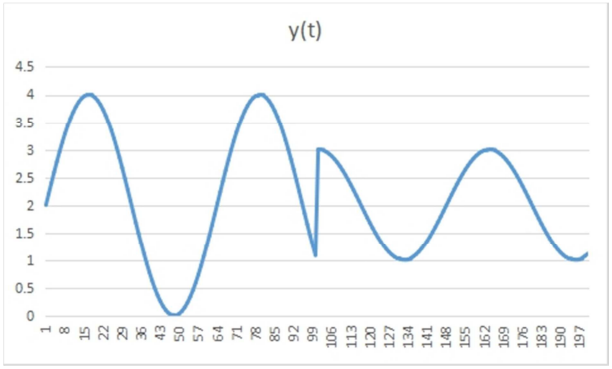


Figure 5. *Splicing time series y(t).*

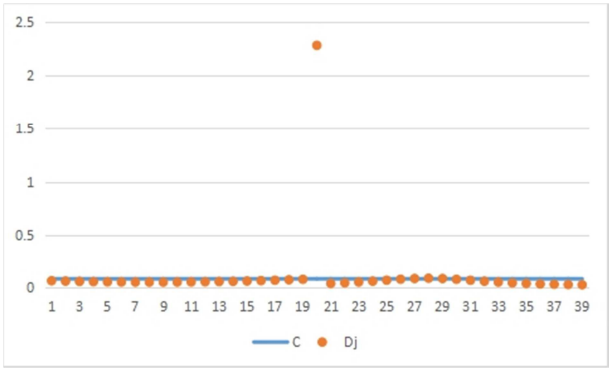


Figure 6. *Anomaly detection corridor.*

According to the anomaly detection corridor in Figure 6, the first half and the second half of the time series belong to two different modes of change, and the position of abnormal sequence data segment is consistent with the actual situation in Figure 5. It can be proved that the proposed method is effective.

Method evaluation

As a clustering algorithm, we need to consider whether the members of the class are reasonable, that is, the rationality of the clustering results. The clustering accuracy index is introduced to evaluate the rationality of the algorithm clustering results [15]. For the time series set

$$S = \{y_1, ...y_i, ..., y_n\} \ (i = 1, 2, ..., n)$$

according to different clustering division, it can be expressed as

$$S = \{\overline{S}_1, ..., \overline{S}_r, ..., \overline{S}_w\} \ (r = 1, 2, ..., w)$$

where $\overline{S}_r$ represents the time series contained in the r-th cluster, $|\overline{S}_r|$ represents the number of time series contained in cluster $\overline{S}_r$; According to the different categories of time series, $\overline{S}_r$ can be expressed as

$$\overline{S}_r = \{y_{1r}, ...y_{ir}, ..., y_{gr}\}$$

where $|y_{ir}|$ is the number of time series in cluster $\overline{S}_r$ that belongs to class i, then the clustering accuracy is shown in equation (22):

$$Accuracy = \sum_{r=1}^{w} \frac{|\overline{S}_r|}{n} \times \max_{i=1,2,...,g} \frac{|y_{ir}|}{|\overline{S}_r|} \tag{22}$$

Combined with the clustering results of three kinds of time series in Figure 4, the accuracy of clustering algorithm is 0.9476, prove that the algorithm has good clustering effect, solve the problems of the existed time series similarity measure and clustering algorithm lacking spatiotemporal invariance.

6. Conclusion

In the process of dynamic system operation, the typical form of sensor sampling data is time series, also known as sampling data sequence with time variation, or data sequence with time sequence. Time series data clustering is an important branch of data clustering research, which is of great significance for statistical learning and data knowledge discovery.

In this paper, the definition and measurement method of temporal data sequence morphological similarity are systematically proposed, based on the proof of affine invariance of similarity measure, a morphological clustering algorithm based on morphological similarity measure is established. Based on the morphological similarity measurement and morphological clustering algorithm of time-series data, this paper focuses on the actual needs of process industry process monitoring and establishes two groups of abnormal change detection algorithms, including morphological consistency detection in different periods of the same process and anomaly detection algorithm for different objects with similar features at multiple monitoring points. The results stated above can be used for mining, clustering, modeling, statistical learning of multi-source time series data, as well as the detection and diagnosis of abnormal changes in process industries; also can be applied to different fields such as safety management and control of spacecraft in orbit, has potential technical value.

Acknowledgements

This paper is supported by the Nature Science Foundation of China (61973094), the Maoming Nature Science Foundation (2020S004), and the Guangdong Basic and Applied Basic Research Fund Project (2020B1515310003).

References

[1] Hai-Lin L I, Chong-Hui G. Survey of feature representations and similarity measurements in time series data mining [J]. Application Research of Computers, 2013.

[2] Liao T W. Clustering of time series data—a survey [J]. Pattern Recognition, 2005, 38 (11): 1857-1874.

[3] Hsu C J, Huang K S, Yang C B, et al. Flexible Dynamic Time Warping for Time Series Classification [J]. Procedia Computer ence, 2015, 51: 2838-2842.

[4] Mao H B, Wu H S, Li Z X, et al. Research on similarity measurement methods for multivariate time series [J]. Control and Decision, 2011, 26 (4): 565-570.

[5] Aghabozorgi S, Shirkhorshidi A S, Wah T Y. Time-series clustering – A decade review [J]. Information Systems, 2015, 53 (C): 16-38.

[6] Müller, Meinard. Information Retrieval for Music and Motion [J]. 2007.

[7] Ke, Yi, Zhou. An improved morphological weighted dynamic similarity measurement algorithm for time series data [J]. International Journal of Intelligent Computing & Cybernetics, 2018.

[8] Keogh E, Ratanamahatana C A. Exact indexing of dynamic time warping [J]. Knowledge & Information Systems, 2005, 7 (3): 358-386.

[9] Balasubramaniyan R, Huellermeier E, Weskamp N, et al. Clustering of gene expression data using a local shape-based similarity measure [J]. Bioinformatics, 2005, 21 (7): 1069-1077.

[10] Shumway R, Stoffer D. Time series analysis and its applications with R examples. New York: Springer, 2009.

[11] Nguyen H. A New Similarity Measure for Intuitionistic Fuzzy Sets [C]// Asian Conference on Intelligent Information and Database Systems. Springer, Berlin, Heidelberg, 2016.

[12] LOWE, D. G. Distinctive Image Feature from Scale-Invariant Key points [J]. International Journal of Computer Vision, 2004.

[13] Keogh E, Chu S, Hart D, et al. Segmenting Time Series: A Survey and Novel Approach [M]. 2003.

[14] Hanlon B, Forbes C. Model Selection Criteria for Segmented Time Series from a Bayesian Approach to Information Compression [J]. Monash Econometrics and Business Stats Working Papers, 2002.

[15] Izakian H, Pedrycz W, Jamal I. Fuzzy clustering of time series data using dynamic time warping distance [J]. Engineering Applications of Artificial Intelligence, 2015, 39 (mar.): 235-244.