ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

Research on Improving Time Series Similarity Based on Segmented Local Representation

Jing Zhou*, Mingwei Li

Faculty of Science, Northeastern University, Shenyang, 110819, China
**Corresponding author*

Abstract: *The similarity measurement of time series has important application value in multiple fields. Among them, the approximate representation and similarity measurement of sequences are the key to solving similarity search. The main purpose of this study is to solve the problem that traditional similarity measures cannot capture the similarity between long time series well. By using segmented local representation and the loss function based on local representation of time series, the loss distance is proposed as an indicator for similarity measurement. The method proposed in this paper is to match the traditional point-to-point according to the time scale to the local representation matching between the current segments according to the morphological characteristics of the time series, which saves the time cost and improves the efficiency of the model.*

Keywords: *Segmented local representation; time series similarity; Loss distance*

1. Introduction

A time series is a series of observations recorded in an orderly manner based on time as a node, which can also be understood as an ordered sequence. The similarity search of time series has always been a research hotspot in Time Series Data Mining (TSDM). The research on similarity measurement of time series not only has profound theoretical significance but also has important practical application value in multiple fields such as mathematical statistics, economic finance, machine learning, medicine, meteorology, etc. If a certain method can be found to process these data reasonably and efficiently, discovering their internal interrelationships, the potential information in sequence data can be fully mined, greatly improving the practical value of time series data.

In recent years, the combination of symbolization and morphological features in time series similarity search has attracted extensive research and achieved certain research results. For example, literature [1] and others proposed a similarity measurement method based on curvature distance, which uses the curvatures of fitted curves of time series at different times to form curvature sequences and defines the curvature distance of the sequence as the overall similarity of the sequence. Reference [2] proposed a method that comprehensively considers the mean and slope of the sequence, and implements symbolic transformation through discourse, which can reflect more data information of the original sequence and objectively reflect the trend of time series changes. References [3] and [4] also used symbolic methods and morphological features of sequences to perform similarity searches on time series, and achieved certain search results. Keogh et al. [5] defined similarity based on compression and demonstrated its superior performance in clustering, classification, anomaly detection, and other work. Keogh[6]applied precise indexing based on DTW to time series mining. Gorecki et al. [7] considered the problem of local morphological feature processing and proposed a method of using dynamic time warping and differential dynamic time warping to obtain the distance of multivariate time series, and then combining the metrics with parameters. Keogh et al. [8] proposed an extension method called DDTW, select an appropriate calculation method based on the adjacent information of a certain point in the time series to construct a new time series, achieving the goal of no longer being sensitive to outliers in the new sequence. Nakamura et al. [9] first proposed the Shape Similarity Based Angular Measure Method (AMSS) in the literature, which is not based on a single point in the sequence, but on a vector representing the sequence. Wang Da proposed the concept of pattern distance in reference [10]. Pattern distance can overcome the shortcomings of Euclidean distance and dynamic time curvature distance, and can effectively calculate the similarity in the trend of changes between sequences. It is a novel measurement method. Reference [11] proposes a composite metric algorithm. This composite distance algorithm not only inherits the advantages of Minkowski distance, but also effectively describes the

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

morphological differences between sequences, that is, the differences in the morphological and statistical features of time series.

2. Measurement of similarity in segmented local representations

Similarity can fully reflect the magnitude of the difference between two objects, and similarity measurement in time series is the main method to determine the relationship between two sequences. The distance between sequences is determined by the correlation coefficient between sequences.This article briefly notes this method as MSSLR.

2.1. Local representation

$S_1 = \{s_{11}, s_{12}, \ldots, s_{1n}\}$, $S_2 = \{s_{21}, s_{22}, \ldots, s_{2m}\}$ ordered sequences of length n and m, respectively, $\|A\|_F = \sqrt{\sum_i^n \sum_j^m a_{ij}^2}$ A^T representing the transposition of matrix A, Frobenius norm for matrix A, and it exists $\|A\|_F^2 = tr(AA^T)$.

Therefore, based on reviewing literature on time series similarity, this article makes the following definition:

Definition 1: For two time series S_1 and S_2, if $S_2 \propto \omega S_1$, it is said that S_1 and S_2 are related, and similar in shape. $\propto$, a mathematical symbol that represents being directly proportional to a quantity.

Definition 2: $\{u_1, u_2, \ldots, u_{Nx}\}$, $\{v_1, v_2, \ldots, v_{Ny}\}$,The subsequence sets of sequence S_1 and sequence S_2, respectively.

If $N_x = N_y$ (N_x, N_y represent the number of subsequences in S_1 and S_2, respectively) and for $\forall u_i$, a vj can be found to construct a representation of $u_i = \omega_{ij} v_j + b_{ij}$, and ui and vj are one-to-one corresponding, that is, they satisfy a bijective relationship, then S_1 and S_2 are mutually locally representable.

If $N_x < N_y$, if a subset can be found in the set of S_2 subsequences, so that this subset and the set of S_1 subsequences meet the above conditions, then S_1 and S_2 are said to be mutually locally representable subsequences.

We understand that greedy algorithm is an algorithm that takes the best or optimal choice in the current state at each step of selection, hoping to result in a globally best or optimal result. Therefore, from the perspective of greedy algorithm, if two sequences are mutually localizable and ly represented, it can be considered that these two sequences are also similar globally.

2.2. Segmentation of subsequences

To a certain extent, morphological features can reflect the basic characteristics of a time series with a certain length. In reality, time series often span longer time periods, and as time increases, the characteristics of the series will change, which is not conducive to analysis and measurement. For the convenience of subsequent research work, the time series is segmented. By segmenting subsequences, we can transform complex time series into segmented sequences.

This article obtains the subsequence set of sequences through sliding windows. Here, we take the width of the sliding window as p and the step size of sliding as s. Therefore, for any sequence of length n, we can obtain a subsequence of length p. The process of extracting subsequences through window sliding is shown in Figure 1:

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

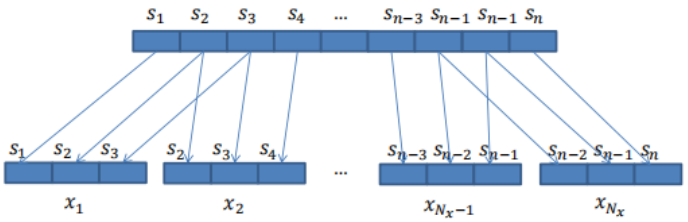

Figure 1: Segmentation diagram

Assuming there are two existing sequences S_1 and S_2, obtain n and m subsequences respectively through the above steps

Order

$$X = (x_1, x_2, \ldots, x_N)^T = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{13} & \cdots & s_{1(p+1)} \\ \cdots & \cdots & \cdots & \cdots \\ s_{1(n-p)} & s_{1(n-p+1)} & \cdots & s_{1n} \end{pmatrix}$$

Remembering their subsequence matrices as $X = (x_{ij}) \in \Re^{n \times p}, Y = (y_{ij}) \in \Re^{m \times p}$

Therefore, the objective functions for the local representations of sequences S_1 and S_2 can be obtained as

$$f_1 = \min_{W,b} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| x_i - w_{ij} y_j - b_{ij} 1_p \right\|_2^2$$

Where x_i and y_j are the ith row of matrix X and the jth row of matrix Y, respectively, $W = (w_{ij}) \in \Re^{n \times m}$ is a local coefficient matrix, $b = (b_{ij}) \in \Re^{n \times m}$ is the local deviation matrix, 1_p is elements that are all 1's 1 × P row vector.

2.3. Measurement of similarity

Due to the sensitivity of representation to outliers, if there are indeed some mutation points in the sequence, it is easy to affect the measurement of similarity. Therefore, this article aims to obtain local coefficients with stronger robustness. In regression analysis, some penalties can usually be applied to the coefficients, such as ℓ1 regularization and ℓ2 regularization, to constrain the values of the coefficients not to be overly exaggerated. Generally speaking, ℓ1 regularization is more easily used to obtain sparse matrices and is therefore more commonly used for feature selection; And ℓ2 regularization can shrink coefficients with larger values to a relatively stable range, thereby enhancing the model's generalization ability. As a generalization of ℓ2 regularization in matrix regularization, ℓF regularization has similar properties. We add ℓF regularization to the objective function and obtain a locally representation of the objective function with regularization as follows

$$f_2 = \min_{W,b} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| x_i - w_{ij} y_j - b_{ij} 1_p \right\|_2^2 + \alpha \|W\|_F^2$$

By optimizing the above equation, we can obtain the parameter matrix W and parameter matrix b, which can then calculate the loss of local representation. The loss of local representation can be

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

divided into two parts, One is distance loss $D_1 = \left(\left|b_{ij}\right|\right) \in \Re^{n\times m}$, The difference in spatial location between the two; The other is shape loss $D_2 = \left(\left|x_i - w_{ij}y_j - b_{ij}1_p\right|\right) \in \Re^{n\times m}$, That is, the residual of the local representation.

Based on the background situation, sometimes we have higher requirements for shape similarity, while sometimes we may pay more attention to the differences in spatial positions between two sequences. In order to adjust the contradiction of demand, the adjustment factor ρ is introduced to obtain the weighted loss matrix L, where $L = \rho D_1 + (1-\rho)D_2$. Normalization of datasets, especially non- regression datasets, is a necessary step because normalization can not only eliminate dimensionality and unify data on the same scale, but also improve the convergence speed of the model. Here, in order to reduce the impact of high or low value levels, we normalize the distance loss and shape loss before calculating the loss matrix.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

We use a loss matrix as an indicator to measure the similarity between two sequences, which minimizes the sum of losses in the local representation when the subsequences of the two sequences are used at most once. The smaller the value of this loss, the more similar the two sequences will be; The reverse is also true. This problem can be transformed into the following planning problem, which can be solved using the Hungarian algorithm. Here, we define the norm of the loss matrix as the loss distance, which is the same as the Euclidean distance equidistance we understand, that is, the smaller the distance, the greater the similarity, and vice versa.

$$\min \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}L_{ij}$$
$$s.t. \quad C_{ij} \in \{0,1\}, \forall i = 1,2,\ldots,n;\ j = 1,2,\ldots,m$$

2.4. Optimization of loss function

$$f(W,\ b) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\|x_i - w_{ij}y_j - b_{ij}1_p\right\|_2^2 + \alpha\|W\|_F^2$$

$$f(W,\ b) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\|x_i - w_{ij}y_j - b_{ij}1_p\right\|_2^2 + \alpha\|W\|_F^2$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_i - w_{ij}y_j - b_{ij}1_p\right)\left(x_i - w_{ij}y_j - b_{ij}1_p\right)^T + \alpha\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}^2$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\alpha + y_i y_i^T\right)w_{ij}^2 + p\sum_{i=1}^{n}\sum_{j=1}^{m} b_{ij}^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{m} y_j 1_p{}^T w_{ij}b_{ij} - 2\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j^T w_{ij}$$
$$-2\sum_{i=1}^{n}\sum_{j=1}^{m} x_i 1_P{}^T b_{ij} + n\sum_{i=1}^{n} x_i x_i^T$$

We obtain the optimal solution for $f(W,b)$ by taking partial derivatives

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

Order $\frac{\partial f(W,\ b)}{\partial w_{ij}} = 0,$ obtain $\left(\alpha + y_j y_j^T\right) = x_i y_j^T - y_j 1_p{}^T b_{ij}$

Order $\frac{\partial f(W,\ b)}{\partial b_{ij}} = 0,$ obtain $pb_{ij} = x_i 1_p{}^T - y_j 1_p{}^T w_{ij}$

By combining two equations, we obtain

$$\begin{cases} w_{ij} = \dfrac{x_i y_j^T - \dfrac{1}{p} y_j 1_p{}^T x_i 1_p{}^T}{\alpha + y_j y_j^T - \dfrac{1}{p} y_j 1_p{}^T x_i 1_p{}^T} \\ b_{ij} = \dfrac{1}{p} x_i 1_p{}^T - \dfrac{1}{p} y_j 1_p{}^T w_{ij} \end{cases}$$

For a more intuitive expression, we remember

$$\overline{x_i} = \frac{1}{p}\sum_{j=1}^{p} x_{ij}, i = 1,\ldots,\ n. \quad \overline{y_j} = \frac{1}{p}\sum_{i=1}^{p} y_{ij}, j = 1,\ldots,\ m.$$

Order $\overline{X} = \left(\overline{x_1},\ldots,\overline{x_n}\right)^T, \overline{Y} = \left(\overline{y_1},\ldots,\overline{y_m}\right)^T$ The row mean matrix of matrix X and matrix Y, respectively. The dispersion matrices of matrices X and Y are σ_{ij}^X和σ_{ij}^Y , The covariance matrix of X and Y is σ_{ij}^{XY} ,

Among them, $\sigma_{ij}^X = \sum_{j=1}^{p}\left(x_{ij} - \overline{x_i}\right)^2, \quad \sigma_{ij}^Y = \sum_{j=1}^{p}\left(y_{ij} - \overline{y_j}\right)^2, \quad \sigma_{ij}^{XY} = \sum_{k=1}^{p} x_{ik} y_{ik} - \frac{1}{p}\sum_{k=1}^{p} x_{ik}\sum_{k=1}^{p} y_{jk}$

Therefore, the above equation can be organized as follows

$$\begin{cases} w_{ij} = \dfrac{\sigma_{ij}^{XY}}{\alpha + \sigma_{ij}^Y} \\ b_{ij} = \overline{x_i} - \overline{y_j} w_{ij} \end{cases}$$

2.5. Algorithm implementation

The algorithm pseudocode is shown in Algorithm 1.

Algorithm 1: Algorithm based on similarity of morphological features

Input	Time series $S_1,\ S_2$
Output	Loss distance L ,Similarity matching result graph
1	Enter two time series $S_1,\ S_2$
2	Sliding Window Segments Time Series into Subsequence Sets;
3	Construct subsequence matrix X Y
4	$f_2 = \min_{W,b}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\|x_i - w_{ij} y_j - b_{ij} 1_p\right\|_2^2 + \alpha\left\|W\right\|_F^2$ Objective function
5	For i to n
6	For j in m

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

7	Local coefficient matrix W $w_{ij} = \frac{\sigma_{ij}^{XY}}{\alpha + \sigma_{ij}^{Y}}$
8	Local deviation matrixb b $b_{ij} = \overline{x_i} - \overline{y_j} w_{ij}$
9	return W, b
10	End
11	distance = (distance - Min_dist) / (Max_dist - Min_dist);
12	$L = \rho \left\| b_{ij} \right\| + (1-\rho) \left\| x_i - w_{ij} y_j - b_{ij} 1_p \right\|$
13	Minimizing the sum of losses in local representations $\min \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij} L_{ij}$
14	End While
15	Plt show
16	Loss distance L=$\|L\|_F$

3. Simulation experiment

3.1. Visualization of experimental results

Using visualization of experimental results to further illustrate the similarity measure proposed in this article.

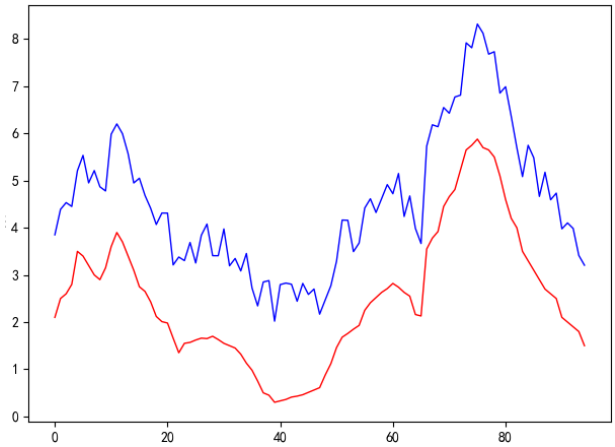

Figure 2: Visualization of Time Series Data

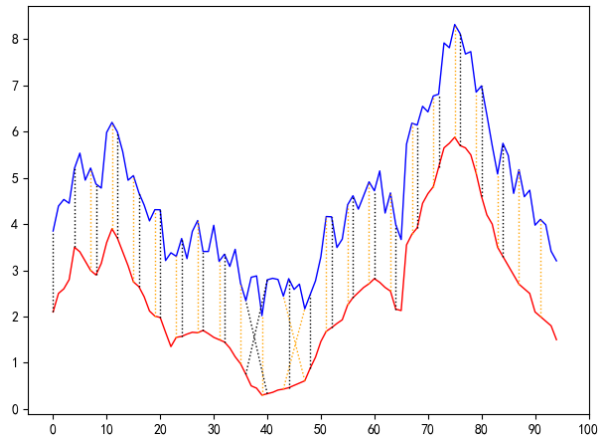

Figure 3: Visual image of matching effect

The horizontal axis represents time, and the vertical axis represents numerical values.We can see

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

from Figures 2 and 3 that the segmented time series segments are not strictly matched according to the corresponding time points. Instead, one segment is traversed through all segments of another time series, and the matching result with the smallest loss distance is selected for presentation. The similarity matching principle proposed in this article can be more clearly seen in the matching result graph during the 30-50 time period.

The loss distance obtained from the two experimental time series in the figure is 2.753, and the loss distance value is relatively small. From the graph, it can be seen that the similarity between the two time series in the experimental data is high, which is consistent with the principle that the smaller the loss distance, the greater the similarity between the two time series. Further verification of the feasibility and effectiveness of the proposed method in this article.

3.2. K-means clustering based on improved similarity measure

Clustering algorithms, as an important component of temporal clustering, are one of the popular research theories in machine learning today. Unlike classification, clustering is an unsupervised learning method that does not require labels for model training. Based on similarity, data is divided into different clusters, ultimately achieving high data similarity within the same cluster and low data similarity between different clusters.

3.2.1. Algorithm

The algorithm pseudocode is as Algorithm 2

Algorithm 2: Clustering Algorithm Improved Based on Loss Distance

Input	Cluster Dataset X , Nearest Neighbor Number K , Number of clusters k , Iterations N
Output	Cluster results C
1	Initialize k clusters $C_1 \leftarrow \phi, C_2 \leftarrow \phi, \cdots, C_k \leftarrow \phi$;
2	Randomly select k points from X as the initial clustering center. $x_{11}, x_{12}, \cdots x_{1k}$
3	$C_1 \leftarrow x_{11}, C_2 \leftarrow x_{12}, \cdots, C_k \leftarrow x_{1k}$;
4	While($i < N$)
5	For j=1 to k
6	Calculate loss matrix $L = \rho\left\|b_{ij}\right\| + (1-\rho)\left\|x_i - w_{ij} y_j - b_{ij} 1_p\right\|$
7	The degree of dispersion is calculated by the Loss distance;
8	End For
9	For each $x \in X$
10	Calculate the distance from x to each cluster center $d^R(x, x_{ij})$
11	$q = \arg\min d^R(x, x_{ij})$;
12	$C_q \leftarrow x$;
13	End
14	$i++$;
15	For j=1 to k
16	$x_{ij} = \frac{\sum_{o \in C_j} o}{\left\|C_j\right\|}$ Update center by averaging of objects within the cluster
17	End For
18	End While
19	$S = \{C_1, C_2, \cdots, C_k\}$

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

3.2.2. Data preprocessing

Due to the large amount of time series data and the presence of noise, directly measuring the similarity of the sequence not only requires a large amount of computation, but also may result in inaccurate measurement results. In order to improve computational efficiency and the robustness of the model and make it easy to compare, this article first preprocesses the experimental data.

Data normalization. Data normalization is a common operation in the preprocessing process of time series, aimed at promoting meaningful comparisons between them. The calculation is as follows

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This article uses interpolation to fill in missing values based on their adjacent data points and remove outliers. Anomalies in the description of the time series exhibit significant fluctuations, usually greater than 2 standard deviations, meaning that the deviation between the anomaly and the average is the largest. Generally speaking, the ratio of outliers in a time series is less than 5%. Therefore, for a simpler calculation, we delete the top 5% data with the highest deviation from the average value, and then perform a complete operation on it using a method consistent with filling in missing values. This way, extreme anomalies (usually huge peaks or valleys) are removed and replaced with their adjacent normal observations. Note that the size of the removal rate is related to the anomaly rate of the dataset and can be adjusted. For different data anomaly rates, it should be possible to cover anomalies caused by extreme values. Afterwards, we use moving average for noise reduction processing.The formula for interpolation method is

$$Y = Y_1 + (Y_2 - Y_1)\frac{X - X_1}{X_2 - X_1}$$

Specifically, for a time series sliding window T with a sliding window length of W, step=1, for each point x_t, The corresponding points on the baseline are represented as x*. It is the average of vector $(x_{t-W+1}, \cdots, x_t)$ .The difference between x_t and x* is called residual. The baseline B and residual R can be calculated as:

$$T = (x_1, x_2, \ldots, x_m)$$

$$x_t^* = \frac{1}{W}\sum_{i=1}^{W} x_{t-i+1}$$

$$B = (x_W^*, x_{W+1}^*, \ldots, x_m^*)$$

$$R = (x_W - x_W^*, \ldots, x_m - x_m^*)$$

Our algorithm extracts the baseline while preserving its basic shape while removing most anomalies and noise. The residual contains random noise and is not considered in similarity measurement and clustering. This baseline is used as input for our similarity measurement algorithm and subsequent clustering algorithms.

3.2.3. Experiments and Indicator Analysis

Replace the Euclidean distance in classical k-means clustering with the sum of the loss distances proposed in the previous chapter. This article uses electrocardiogram related data from the UCR dataset. In order to demonstrate the performance of the proposed method in time series clustering applications, this method is experimentally compared with ordinary k-means clustering to measure the clustering effect.

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

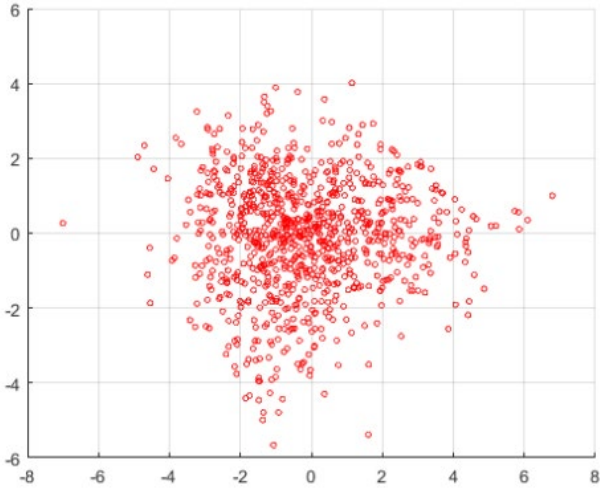

Figure 4: Visualization of raw data

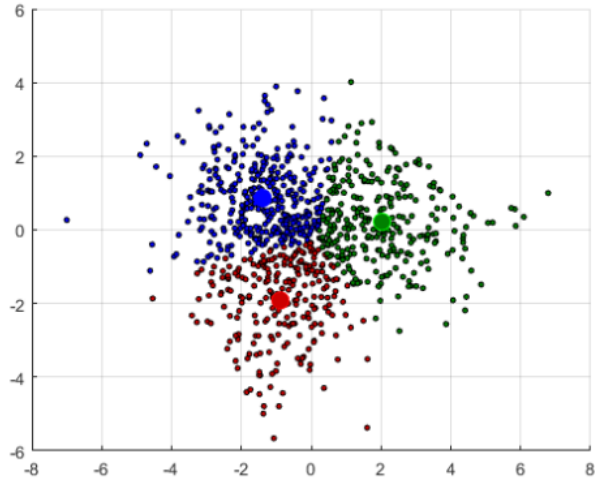

Figure 5: Visual image of clustering effect

Figure 4 shows the visualization of the original data, while Figure 5 shows the clustering effect obtained using the algorithm proposed in this paper. Then we use comparative analysis of indicators to demonstrate the feasibility and effectiveness of the MSSLR method.

(1)Calinski-Harabaz index

The CH indicator is the ratio of the separation and compactness of a dataset, measured by the sum of the squares of the distances between various center points and the center points of the dataset, and the compactness of the data is measured by the sum of the squares of the distances between each point within the class and its class center. The better the clustering effect, the larger the inter class gap and the smaller the intra class gap, that is, the closer the class itself, the more dispersed the inter class. Therefore, the larger the CH index value, the better the clustering effect.

The essence of the Calinski Harabasz index is the ratio of inter cluster distance to intra cluster distance, and the overall calculation process is similar to the variance calculation method, so it is also called the variance ratio criterion. Aggregate the dataset X with a capacity of N into K classes, and measure the compactness within the class (intra class distance) by calculating the sum of the square distances between each point within the class and the center of the class. Measure the separation of the dataset (inter class distance) by calculating the sum of the square distances between each center point of the class and the center point of the dataset

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

The calculation formula for CH index is $s = \frac{tr(B_k)(N-K)}{tr(W_k)(K-1)}$

The CH indicators obtained by the two methods are 0.752 and 0.836, respectively. The larger the CH indicator value, the better the clustering effect. It can be concluded that using MSSLR clustering is better than using kmeans clustering with Euclidean distance.

(2)Precision and Recall

Method 1: Euclidean distance is directly used for k-means clustering

Method 2: Losing distance for k-means clustering

Table 1: Average Precision and Recall

	Method 1	Method 2
Average precision	0.4837	0.5514
Average recall rate	0.6042	0.6121

From the Table 1, it can be seen that compared with Method 1, although the average precision and recall of Method 2 have increased slightly, they have overall improved. Method 1: Euclidean distance directly calculates point-to-point time series without segmented representation, resulting in high computational complexity and susceptibility to noise; The method proposed in this article first segments the time series and reduces the data dimension. It represents time series segments with the same trend, which conforms to the changing trend of the sequence and reduces the impact of noise. Therefore, the clustering results are good. Table reflects that the clustering accuracy of this method has also improved.

4. Summary and Outlook

This article first introduces the concept of time series, the research background, significance, current status, and applications of time series similarity measurement technology in real life. The similarity measurement of time series is the foundation of time series clustering. Analyzing time series data and finding potential patterns have a significant impact on people's production and life. Time series usually exist in a high-dimensional and variable form, making it difficult to intuitively measure similarity. Existing similarity measures have limitations. Therefore, this paper proposes a new method for improving similarity measurement based on time series morphology, which is Measurement of similarity in segmented local representations (MSSLR).

The method proposed in this article first performs sliding window segmentation on the time series to obtain two construction matrices of the subsequence. By minimizing the sum of local representation losses of the two matrices, the similarity matching results of the two sequences are obtained. Afterwards, in order to enhance the robustness of the locally represented coefficients, regularization was added to the objective function, ultimately obtaining the loss distance for measuring similarity. Compared to the kmeans clustering using Euclidean distance, the kmeans clustering using the loss distance in this paper has higher clustering index results, with a certain range of improvement in precision and recall, proving the feasibility and effectiveness of the proposed pie wise local linear representations a measure of similarity.

In today's society, with the continuous deepening of research on similarity measurement algorithms for sequences, people have generated more accurate similarity measurement algorithms based on different theories. How to better segment data by combining new knowledge domains with feature points or key points of sequences based on research, and finally, how to use models for data query and prediction processes while considering time factors, is a topic that needs further research.

References

[1] Liu Boning, Zhang Jianye, Zhang Peng, et al. Time series similarity search method based on curvature distance [J]. Journal of Electronics and Information Technology, 2012, 34 (9): 2200-2207
[2] Li Hailin, Guo Chonghui. Symbolic Aggregation Approximation Method for Time Series Based on Morphological Features [J]. Pattern Recognition and Artificial Intelligence, 2011,24 (5): 665-672
[3] Liu Fen, Guo Gongde. A composites measurement method for time series similarity based on symbolic aggregation approximation [J]. Computer Application, 2013,33 (1): 192-198

ISSN 2616-5775 Vol. 7, Issue 4: 45-55, DOI: 10.25236/AJCIS.2024.070407

[4] Xiao Rui, Liu Guohua. Trend based similarity measurement and clustering research for time series [J]. Computer Application Research, 2014,31 (9): 2600-2605

[5] Keogh E, Lonardi S, Ratanamahatana C A. Towards parameter-free date mining[C].Proceedings of the 10th ACM SIGKDD international conference on Konwledge discovery and date mining. ACM, 2004: 206-215

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

[7] Gorecki T, Luczak M. Using derivatives in time series classification [J]. Data Mining and Knowledge Discovery, 2013, 26(26): 310-331.

[8] Keogh E J, Pazzani M J. Derivative dynamic time warping[C]//Proceedings of the 1st SIAM International Conference on Data Mining, Chicago, Apr 5-7, 2001. Philadelphia: SIAM, 2001: 1-11

[9] Nakamura T, Taki K, Nomiya H, et al. A Shape-based Similarity Measure for Time Series Data with Ensemble Learning[J]. Pattern Analysis and Applications, 2013, 16(4): 535-548.

[10] Wang Da, Rong Gang. Pattern Distance of Time Series [J]. Journal of Zhejiang University (Engineering Edition), 2004,38 (7): 795-798

[11] Xie Fuding, Li Ying, Sun Yan, et al. Improved Symbolic Time Series Processing Method [J]. Computer Engineering and Design, 2012,33 (10): 3950-3953