Research Article

Fuzzy Clustering Method Based on Improved Weighted Distance

Honghan Bei ^{1,2}, Yingchao Mao ³, Wenyang Wang ^{1,2} and Xu Zhang ³

¹School of Maritime Economics and Management, Dalian Maritime University, Dalian 116026, China

²Collaborative Innovation Center for Transport Studies, Dalian Maritime University, Dalian 116026, China

³School of Mathematical Science, Dalian University of Technology, Dalian 116026, China

Correspondence should be addressed to Wenyang Wang; wangwenyang@dmlu.edu.cn and Xu Zhang; xuzhang@dlut.edu.cn

Received 30 December 2020; Revised 24 January 2021; Accepted 28 February 2021; Published 8 March 2021

Academic Editor: Ant nio M. Lopes

Copyright © 2021 Honghan Bei et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As an essential data processing technology, cluster analysis has been widely used in various fields. In clustering, it is necessary to select appropriate measures to evaluate the similarity in the data. In this paper, firstly, a cluster center selection method based on the grey relational degree is proposed to solve the problem of sensitivity in initial cluster center selection. Secondly, combining the advantages of Euclidean distance, DTW distance, and SPDTW distance, a weighted distance measurement based on three kinds of reach is proposed. Then, it is applied to Fuzzy C-MeDOIDS and Fuzzy C-means hybrid clustering technology. Numerical experiments are carried out with the UCI datasets. The experimental results show that the accuracy of the clustering results is significantly improved by using the clustering method proposed in this paper. Besides, the method proposed in this paper is applied to the MUSIC INTO EMOTIONS and YEAST datasets. The clustering results show that the algorithm proposed in this paper can also achieve a better clustering effect when dealing with practical problems.

1. Introduction

Clustering analysis, clustering for short, is the process of subdividing data objects into subsets. Each subset is a cluster, and items in the same cluster have a high similarity, while objects in different clusters have a relatively high divergence. Clustering algorithms are widely used in engineering (such as artificial intelligence and machine learning), biomedical (such as blood pressure and electrocardiogram measurement), biometric data (facial recognition image data), sentiment analysis (sentiment classification scheme), meteorological data, physical particle tracking, and other fields [1].

Data sequences are characterized by nonlinearity, high dimension, complexity, and redundancy. Although complex, they often contain the evolution law of things [2]. To find these potential rules, the data mining algorithm plays a vital role in data analysis, the clustering of which is one of the critical technologies. Three factors directly affect data clustering: correct acquisition of clustering number, efficient and excellent clustering algorithm, and distance function determining the similarity between data points.

K-means is a classical clustering algorithm based on division. The main idea is to find out K clustering centers, divide the dataset into K classes, and make each data sequence belong to a specific class. However, it is difficult to determine the number of clustering K in this algorithm, and the unsuitable K is not worth the ideal clustering result. Therefore, Gupta and Srivastava proposed a document-based K-means algorithm that automatically selects K and refined clusters, thus solving cluster number selection problems [3]. Onan and Toolu have made generous contributions to the study of clustering. Onan, A. presents a consensus clustering-based undersampling approach to imbalanced learning. In this scheme, five supervised learning methods (namely, k -means, k -modes, k -means++, self-organizing maps, and DIANA algorithm), three ensemble learner methods, five clustering algorithms, and their combinations are adopted. In the clustering stage, Onan, A. proposes an improved clustering integration framework, which combines the traditional clustering methods (i.e., K-means, K-modes, K-means++, self-organizing mapping, and Diana algorithm) through iterative voting consistency. Onan and Toolu present a document-clustering model on

MOOC discussion forum posts. In their study, four word-embedding schemes, four weighting functions, and four clustering algorithms (i.e., K-means, K-means++, self-organizing maps, and divisive analysis clustering algorithm) are used for document clustering [4]. Onanand Toolu augment the randomized seeding technique to overcome poor initialization of medoids in the PAM algorithm. The proposed approach (PAM++) is compared with other partitioning clustering algorithms, such as K-means and K-means++ on text document clustering benchmarks, and evaluated in terms of F-measure. However, in practical applications, data sequences sometimes belong to multiple classes rather than just one class. To deal with this problem, fuzzy clustering methods, such as the FCM algorithm and the FBSA algorithm, came into being. Li et al. [5] summarize data mining research based on fuzziness theory, expound the research results of fuzzy clustering methods in recent years, and forecast the research prospect. The most representative fuzzy clustering algorithm is the fuzzy C-mean (FCM) algorithm. The fuzzy C-mean algorithm uses fuzzy theory to describe the data. This algorithm is a flexible partition that can obtain much clustering information and reflect the actual distribution of the samples more accurately. However, at the same time, this algorithm is sensitive to the selection of the initial clustering center point and prone to fall into the problem of an optimal local solution. To make up for the shortcomings of the algorithm and improve the accuracy of the algorithm, researchers always put forward improved clustering methods. Zhu et al. [6] proposed an improved fuzzy C-means algorithm combining genetic algorithm, particle swarm optimization, and FCM clustering algorithm. The enhanced algorithm optimized the performance of the clustering algorithm to some extent and reduced the dependence of the algorithm on the initial center point. Zhang and Wang [7] proposed a new method for cluster center selection, which took the midpoints of the two closest sample points as cluster center points for cluster analysis. Bharill and Tiwari [8] introduced a random sampling iterative optimized fuzzy C-mean (RSIO-FCM) clustering algorithm, which divided big data into different subsets, formed effective clustering, and eliminated the problem of overlapping clustering centers. Bulut et al. [9] have developed an efficient clustering scheme for microarray gene expression data. The algorithm utilizes the feature selection algorithm to overcome the high-dimensionality problem encountered in the bioinformatics domain. The clustering quality of the ant-based clustering algorithm is enhanced with the use of fuzzy c-means algorithm and heaps merging heuristic. Based on the improved fuzzy C-means method, Liu et al. [10] used the chaotic quantum particle swarm optimization algorithm to generate initialization and global optimal clustering center. They then applied the improved algorithm to the clustering problem of time series datasets, achieving a good clustering effect.

Besides, in clustering, different distance functions often have a significant impact on the clustering results. Euclidean distance is a widely used distance function, but Euclidean distance is difficult to solve the linear drift problem. Rammal et al. studied infrared spectrum clustering and

experimentally verified that the experimental results using L1 or Manhattan distance were superior to the Euclidean and Cheyne distances [11]. Petitjean et al. [12] proposed the average value method based on DTW. Izakian et al. [13] applied the average value method of Petitjean et al. to the FCMDD algorithm and achieved satisfactory results. Based on limited DTW, Kathiresan and Sumathi proposed an improved FCM algorithm [14], and the experimental results showed that the improved algorithm was more accurate.

In this paper, a weighted distance method based on a fuzzy model is proposed to cluster different sequences. We first determined the initial clustering center based on the grey relational degree method, then applied the weighted distance of Euclidean distance, DTW, and SPDTW to the FCM algorithm, and carried out the optimal clustering for the sequence through the particle swarm optimization algorithm. Through numerical experiments, different datasets are clustered, and the experimental results show that the improved algorithm proposed in this paper is more accurate than the existing literature.

The rest of the paper is organized as follows: Section 2 introduces the basic principles of grey relational degree, fuzzy C-means clustering, and distance function. Section 3 introduces the implementation of the improved fuzzy C-means clustering method. In Section 4, the improved clustering method is applied to different datasets, and numerical experiments are carried out to verify the effectiveness of the method through comparison. Besides, the method proposed in this paper is applied to the self-care (SCADI) problems of children with physical and motor impairments. The clustering results show that the algorithm proposed in this paper can also achieve a better clustering effect when dealing with practical problems. Section 5 gives the summary.

2. Basic Knowledge

This section introduces the basic concepts and principles of grey relational degree, fuzzy C-means clustering, and distance function.

2.1. Grey Relational Degree Analysis. The basic ideas of grey relation analysis method are as follows: according to the comparative sequence, set up the constituent of a family of curves and the reference sequence curve to determine the geometric similarity degree between the reference sequence and compared sequence set, and the correlations between the comparative sequence consisting of the curve and constituent of the reference sequence curve geometry are similar, and the correlation is more significant. The comprehensive evaluation method using grey correlation analysis is as follows.

Assuming that the evaluation problem is composed of n objects and m indicators, x_{ij} represents the index value of the i object and the j index, and the original evaluation matrix is $X = (X_1, X_2, \dots, X_n)^T$, in which $X_i = (x_{i1}, x_{i2}, \dots, x_{im})$, $i = 1, 2, \dots, n$. According to the purpose of evaluation and index situation, the reference

number is set as $X_0 = (x_{01}, x_{02}, \dots, x_{0j}, \dots, x_{0m})$. We begin by standardizing the index data (adopting different standardized processing strategies for different index types). The data sequence after standardized processing is recorded as $Z = (Z_1, Z_2, \dots, Z_n)^T$, in which $Z_i = (z_{i1}, z_{i2}, \dots, z_{im})$, $i = 1, 2, \dots, n$. After standardized treatment, the reference number column is $Z_0 = (z_{01}, z_{02}, \dots, z_{0j}, \dots, z_{0m})$. The absolute difference

between the index sequence (comparison sequence) of each evaluated object and the corresponding element of the reference sequence is calculated one by one, $|z_{ij} - z_{0j}|$, $i = 1, 2, \dots, n$, to confirm $\min_{i=1}^n \{\min_{j=1}^m (|z_{ij} - z_{0j}|)\}$ and $\max_{i=1}^n \{\max_{j=1}^m (|z_{ij} - z_{0j}|)\}$.

Then, the correlation coefficient between the corresponding elements of each comparison sequence and the reference sequence is calculated:

$$\zeta_{ij} = \frac{\min_{i=1}^n \left\{ \min_{j=1}^m (|z_{ij} - z_{0j}|) \right\} + \eta \cdot \max_{i=1}^n \left\{ \max_{j=1}^m (|z_{ij} - z_{0j}|) \right\}}{|z_{ij} - z_{0j}| + \eta \cdot \max_{i=1}^n \left\{ \max_{j=1}^m (|z_{ij} - z_{0j}|) \right\}}, \quad (1)$$

where η is the resolution ratio, the value of which is between (0, 1). The smaller the value of η is, the greater the difference between correlation coefficients is, and the stronger the resolution ability is. Usually, $\eta = 0.5$.

For each evaluation object, the mean value of the correlation coefficient of the corresponding element of each index and the reference sequence is calculated, respectively, to reflect the correlation degree of each evaluation object and the reference sequence, denoted as

$$\begin{aligned} Y &= (y_{01}, \dots, y_{0i}, \dots, y_{0n})^T, \\ y_{0i} &= \frac{1}{m} \sum_{j=1}^m \zeta_{ij}(j), \\ i &= 1, 2, \dots, n; \\ j &= 1, 2, \dots, m. \end{aligned} \quad (2)$$

2.2. Fuzzy C-Means Clustering. Fuzzy c-means (FCM) clustering uses membership to determine that each data point belongs to a certain degree of clustering of a fuzzy clustering algorithm. Its core idea is based on the fuzzy membership degree matrix obtained. The membership degree of each data sample's power and the distance between every center-weighted clustering define the objective function. The fuzzy C-means clustering criterion is to find the membership matrix and the clustering center to minimize the objective function.

Given dataset $X = (x_1, x_2, \dots, x_n)$, fuzzy clustering is the division of samples into classes ($2 \leq c \leq n$); denote $V = (v_1, v_2, \dots, v_c)$ as c center of clustering, $U = \{u_{ji}\}$ is the degree of membership of i th data base x_i corresponding to the j th sort, and then the clustering loss function based on membership function is

$$J(U, V) = \sum_{i=1}^n \sum_{j=1}^c u_{ji}^m d^2(x_i, v_j), \quad \text{s.t. } 0 \leq u_{ji} \leq 1, \quad \sum_{j=1}^c u_{ji} = 1, \quad (3)$$

where d is the distance measurement, and m is the fuzzy weighted index, also known as a smoothing factor, which controls the degree of fuzziness between fuzzy classes.

Let loss function $J(U, V)$ be the partial derivative v_j and u_{ji} is 0, so the necessary conditions for obtaining the formula minimum (3) are

$$v_j = \frac{\sum_{i=1}^n (u_{ji})^m x_i}{\sum_{i=1}^n (u_{ji})^m}, \quad (4)$$

$$u_{ji} = \frac{1}{\sum_{k=1}^c (d(x_i, v_j)/d(x_i, v_k))^{(2/(m-1))}}. \quad (5)$$

Therefore, the membership matrix and clustering center are calculated as (4) and (5). By optimizing the objective function to minimize it, each sample point's final membership matrix to all class centers is obtained, thus determining the optimal classification of all sample data.

In this paper, we also mentioned the fuzzy C center point (FCMdd) clustering algorithm. FCMdd algorithm, like the FCM algorithm, uses (3) as the objective function and updates the clustering center and membership matrix through formulas (4) and (5). However, they have different choices for the initial clustering center. FCM uses the weighted average value of data or randomly generates a set of data as the initial clustering center, while FCMdd selects the existing data in the dataset as the initial center.

2.3. Distance Measurement

2.3.1. Euclidean Distance. In many clustering algorithms, the Euclidean distance formula is used as the basis for clustering. The Euclidean distance (ED) formula can be expressed as

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}. \quad (6)$$

In this paper, the weighted Euclidean distance [15] will be adopted, which can be expressed as

$$d_{ED} = d(x_i, x_j) = \sqrt{\sum_{k=1}^m w_k (x_{ik} - x_{jk})^2}, \quad (7)$$

where w_k is the weight. In the present paper, the weight is determined in the following form:

$$w_k = \frac{v_k}{\sum_{k=1}^m v_k}, \quad (8)$$

$$v_k = \frac{\sigma_k}{\bar{x}_k},$$

where σ_k is the standard deviation of the index k and $\bar{x}_k$ is the average of the number k .

2.3.2. Dynamic Time Structuring (DTW). DTW is based on the idea of dynamic programming (DP). Given two discrete sequences, DTW can measure how similar the two sequences are, or how far apart they are. If two feature sequences X and Y have i frames and j frames, respectively, it is assumed that the distance of every two frames between X and Y is calculated to obtain the difference matrix $D = d_{ij}$ with i rows and j columns, and the element d_{ij} is expressed as

$$d_{ij} = d(i, j) = \text{Dist}(X_i, Y_j), \quad (9)$$

where $\text{Dist}(X_i, Y_j)$ is the deviation between X_i and Y_j .

The calculation formula of DTW is

$$\begin{aligned} \text{DTW}(X_j, Y_j) = \min & (\text{DTW}(X_{i-1}, Y_{j-1}), \text{DTW}(X_{i-1}, Y_j), \\ & \cdot \text{DTW}(X_i, Y_{j-1})) + d(i, j), \end{aligned} \quad (10)$$

where $\text{DTW}(X_i, Y_j)$ is the summation of the minimum differences between frame i of X sequence and frame j of Y sequence.

Also, the SPDTW algorithm is an improvement on DTW. The calculation formula of the SPDTW algorithm is as follows:

$$\begin{aligned} \text{SPDTW}(X_i, Y_j) &= \text{DTW}(X'_i, Y'_j), \\ X'_i &= X_{i-1} - X_i, \quad i = 1, 2, \dots, n-1. \end{aligned} \quad (11)$$

Compared with the DTW algorithm, the introduction of SPDTW will reduce computation and improve the clustering accuracy to a certain extent.

3. Improved Fuzzy Clustering Method

3.1. Selection of Initial Cluster Centers. In general fuzzy clustering algorithms, the selection of the initial clustering center is usually randomly generated. However, since FCM and FCMdd clustering algorithms are susceptible to selecting the initial clustering center, the randomly generated clustering center may not have a good clustering effect. Therefore, this paper proposes an initial cluster center selection method based on the highest density and grey

relational degree principle to reduce the dependence of the algorithm on the initial center point.

In the clustering analysis of datasets, the clustering center must be in the area with a high distribution density of sample points. If the region centered on a sequence contains the most samples, the sample is the densest sequence. Therefore, we can first find the first initial clustering center in the region with the densest data distribution, and the distance between sample points located in the region with dense data distribution is small, and the mean distance of sample points in the dense region is also small. If an appropriate distance mean threshold is given, the distance and the mean distance between the initial cluster center in the dense region and other sample points in the region are generally smaller than the threshold. In this paper, the distance threshold used is determined by the average distance of all data samples and expressed as follows:

$$d_{\text{avg}} = \frac{1}{C_n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n d(x_i, x_j), \quad (12)$$

where $d(x_i, x_j)$ represents the distance between data x_i and x_j and n represents the number of all sample points to be clustered.

The clustering center selection algorithm based on the principle of highest density is as follows:

Step 1. Calculate the average distance value of all sample sequences through formula (12) and use it as the threshold value in the above analysis;

Step 2. Draw a circle with the radius of d_{avg} and the sample $i (1 \leq i \leq n)$ as the center. Calculate the Euclidean distance between the first sample and other samples. If $d < d_{\text{avg}}$, then the sample is placed in the circle with the center of the sample i . Calculate the number of samples contained in the circle with the center of the sample i , and write it as n_i ;

Step 3. The corresponding sequence $\max\{n_i, 1 \leq i \leq n\}$ is obtained as the sequence i with the highest density, which denoted C_1 as the first initial clustering center.

In the clustering analysis, we need to determine C clustering centers. The above algorithm has given the first initial clustering center, and the selection of the remaining initial centers is given based on the grey relational degree method. The more similar the geometric shapes of the curve formed by the comparison sequence and the curve formed by the reference sequence are, the greater the correlation degree will be; otherwise, the smaller the correlation degree will be. The clustering center selection algorithm based on the grey relational degree is as follows:

Step 1. Take the first determined initial clustering center C_1 as the reference sequence and remove the sequence C_1 from the dataset sequence of the clustering as the evaluation matrix.

Step 2. According to the grey relational degree calculation formula (1), the correlation coefficient ζ_{ij} of corresponding elements between the index sequence of

each evaluated object (comparison sequence) and the reference sequence is calculated, and the correlation degree of each evaluation index sequence is $\gamma_{0i} = (1/m) \sum_j \zeta_{ij}(j)$, in which the corresponding sequence with the lowest correlation degree is denoted as the second cluster center, denoted as C_2 .

Step 3. Take the second cluster center C_2 as the reference sequence, remove the dataset C_1 , C_2 of the cluster, and repeat Step 2 as the new evaluation matrix C_3 . Repeat the above steps and get C as the initial cluster center.

3.2. Determination of Weighted Distance. In the process of clustering, different distance functions often have a great influence on the clustering results. Euclidean distance and DTW distance are widely used distance functions for data series clustering.

The Euclidean distance has translation and rotation invariance. The use of the Euclidean distance as the objective function has intuitive significance and is easy to calculate. However, the Euclidean distance only considers the direct geographical distance between data and ignores the other valuable information, and it cannot solve the problem of linear drift.

Dynamic time structuring distance (DTW) can well support time axis bending by stretching and compressing sequences. It can measure nonisometric sequences, flexibly match time series data, and determine the similarity between any two sequences by stretching or compressing data segments. However, traditional DTW algorithms are often limited by their high time and space complexity and are affected to some extent in practical applications.

SPDTW is the DTW distance after differentiating the data sequence. Pure SPDTW is not very useful as a general distance metric. For example, a classification algorithm uses

the nearest neighbor method, which tends to be less effective than DTW on a large database of time series. However, SPDTW can attenuate sequence fluctuations in many datasets and provide good results. This case is often used in conjunction with other distances.

To overcome the disadvantages of different distance measurements, we combine the three kinds of distance measurement advantages. There may be a larger dataset dependent on a certain distance or a different degree depending on the distance measure problem (see Section 4). This paper proposes a new method of weighted distance measurement, and it can be expressed as a linear combination of the form:

$$\text{dist}(X_1, X_2) = a \cdot \text{dist}_1(X_1, X_2) + b \cdot \text{dist}_2(X_1, X_2) + c \cdot \text{dist}_3(X_1, X_2), \quad (13)$$

where $a, b, c \in [0, 1]$, $a + b + c = 1$.

Based on the measurement method introduced in this article, we make the following transformation:

$$\begin{aligned} a &= x(1), \\ b &= x(2), \\ c &= 1 - (x(1) + x(2)), \\ x(1) &\in [0, 1], \\ x(2) &\in [0, 1], \\ (x(1) + x(2)) &\in [0, 1], \\ \text{dist}_1 &= \text{DTW}, \\ \text{dist}_2 &= \text{SPDTW}, \\ \text{dist}_3 &= d_{\text{ED}}. \end{aligned} \quad (14)$$

Thus, a new weighted distance measurement is obtained as follows:

$$\begin{aligned} \text{newdist}(X_1, X_2) &= x(1)\text{DTW}(X_1, X_2) + x(2)\text{SPDTW}(X_1, X_2) \\ &+ (1 - x(1) - x(2))d_{\text{ED}}(X_1, X_2), \quad \text{s.t. } x(1) \in [0, 1], x(2) \in [0, 1], (x(1) + x(2)) \in [0, 1]. \end{aligned} \quad (15)$$

It can be seen from equations (14) and (15) that the selection of distance directly affects the selection of membership matrix and clustering center, and from equation (15), we can see that the selection of parameter sum determines the value of distance. The Cross-Validation method is used here for the parameters' tune. Therefore, let $x = (x(1), x(2))$, so the value of x would directly affect the quality of the clustering result. To select the appropriate value of x , the particle swarm optimization algorithm of x is used in this paper for optimization so as to obtain the optimal clustering effect.

In particle swarm optimization (PSO), each optimization problem's solution is regarded as a particle in the search space. The initial population is first generated, and the objective function determines a fitness value. Each particle will travel in solution space, and its velocity will determine

its direction and distance. In each iteration, the particle updates itself by tracking two "extremes": the one found by the particle itself, known as the individual best (pbest), and the one so far found by the entire population, known as the global best (gbest).

In the particle swarm algorithm, particle velocity and position updating formulas are, respectively,

$$\begin{aligned} v_n^{k+1} &= \omega v_n^k + c_1 r_1 (p_n^k - x_n^k) + c_2 r_2 (g - x_n^k), \\ x_n^{k+1} &= x_n^k + v_n^{k+1}, \end{aligned} \quad (16)$$

where $v_n^k = (v_{n,1}^k, v_{n,2}^k, v_{n,3}^k, \dots, v_{n,m}^k)$ is the particle velocity of the particle n in the iteration k , ω is the inertia weight, p_n^k is the best position of the particle n in the first k iteration, g is the best position in the iteration population k , c_1, c_2 are the

acceleration constants, and r_1, r_2 are the random numbers between 0 and 1.

In this paper, the particle swarm optimization algorithm is used to find the optimal value of particles, to obtain the optimal fitness value, namely, the optimal clustering effect. PSO algorithm retains the global search strategy based on population and adopts the speed-displacement model, which is easy to operate and can find the optimal solution in fewer iterations.

3.3. Mixed Fuzzy Clustering. With the continuous development and progress of the data information age, traditional FCM clustering technology can no longer meet people's demand for data processing. In this section, we present a hybrid technique that takes advantage of the FCM and FCMdd techniques.

In this method, for the data sequence of $n \times m$, the FCMdd method firstly clusters the data into r ($r \geq c$) classes, where c is the number of final clustering, and then obtains the cluster center $V_{r \times 1}$ and partition matrix $U_{r \times n}$ with r rows and n columns. In fact, the FCMdd technique is used here to transform the data, and usually $r \leq n$, the transformed dataset is located in a smaller search space, and $U_{r \times n}$ is treated as a new dataset with r features and n objects. In the next step of this algorithm, the new weighted distance proposed in this paper is used as the distance function in the FCM algorithm to cluster the dataset U into class c , so as to obtain the final membership matrix U_0 and clustering center V_0 , and analyze the clustering results to complete the clustering analysis of multiple sequences.

To quantify the clustering methods in cluster analysis, precision is considered as an appropriate evaluation standard. In this paper, the method in Izakian et al. [10] is adopted to determine the clustering accuracy, and the calculation formula of the accuracy is

$$P = \sum_{j=1}^k \frac{|C_j|}{N} \times \max_{i=1,2,\dots,c} p(K_i, C_j), \quad (17)$$

$$p(K_i, C_j) = \frac{|K_i \cap C_j|}{|C_j|},$$

where c is the number of clustering clusters, k is the number of known classes, N is the number of dataset sequences, K_i is the data sequence belonging to class i (known sequences), C_j is the data sequence belonging to class j , and $|\cdot|$ represents the number of elements in the set. To sum up, the overall scheme of the hybrid fuzzy clustering method proposed in this paper is as follows:

Step 1. For the sequence to be clustered, we determine the initial clustering center based on the grey relational degree, divide the data into r ($c \leq r \leq n$) classes by FCMdd algorithm and DTW distance, determine the membership matrix and clustering center, and then transform the dataset into a smaller search space.

Step 2. The FCMdd algorithm needs to determine the membership degree matrix clustering as a new dataset

for the secondary clustering by the FCM algorithm. It requires the presented weighted distance as the FCM algorithm's function through the weight of the particle swarm optimization algorithm to optimize the weighted distance. It obtains the final membership degree matrix and cluster centers. The optimal weight is obtained as the best clustering accuracy.

Figure 1 shows the flow chart of the above process.

4. Experimental Analysis

To evaluate the performance of the algorithm, datasets from the UCI machine learning database are used in this paper for experiments. Datasets with different numbers, different sequence lengths, and categories were selected for the experiment. Table 1 shows each experimental dataset's overall characteristic distribution information, including the dimension and number of data samples in each dataset, the number of clusters, and other indicators.

The main parameter settings of the algorithm are listed as follows:

- (1) In particle swarm optimization, the acceleration constants c_1, c_2 are both 1.5
- (2) The initial inertia weight ω is 1
- (3) The population size is 10

In both the FCM and the FCMdd algorithms, the fuzzy coefficient m is set as 2, the iteration termination condition is $|P(\text{iter} + 1) - P(\text{iter})| < 1.0e - 5$, and the maximum iteration number is set as 100. For hybrid technology, the number of clusters considered in the first step of the algorithm may substantially impact the performance of the algorithm. The optimal value of this parameter is determined by the internal structure of the clustered dataset. To simplify the calculation, take $r = \sqrt{n}$ (n is the number of datasets to be clustered).

To eliminate the contingency of the experimental results and improve the accuracy of the results, this paper conducted 10 experiments for each dataset, and the final result was the average value of the data recorded in the 10 experiments.

Table 2 shows the comparison between the clustering accuracy obtained by using the Euclidean distance only (with weights), DTW distance only, SPDTW distance only, and the accuracy obtained by using the weighted distance proposed in this paper. Besides, the comparison results of clustering accuracy in the existing literature are also given. We see that the clustering method described in this paper determined by weighted distance has higher clustering accuracy and better clustering effect.

Luczak [16] adopts the double-layer fuzzy clustering method and uses the weighted distance of DTW and differential form DTW as distance measurement to cluster the data. Huang et al. [17] developed a new objective function based on the k-means algorithm to cluster time series data by extracting the hidden smooth subspace. In Yu et al. [18], based on the best U-Shapelet method, combined with I index evaluation and multiple top-k query technology, the subsequence of the best dataset, namely, the U-Shapelets set,

FIGURE 1: Flow chart of the fuzzy clustering method.

TABLE 1: The summary of the characteristics of the datasets.

Dataset	Sample size	Dimension	Categories
SonyAIBORobotSurface	621	70	2
SonyAIBORobotSurfaceII	980	65	2
OliveOil	60	570	4
CBF	930	128	3
ECG200	200	96	2
Synthetic	600	60	6
Lighting2	121	637	2
Yoga	3300	426	2
Beef	60	470	5
Gisette	13500	5000	2
Arcene	900	10000	2

TABLE 2: Comparison of clustering results.

Dataset	WED	DTW	SPDTW	Literature	Newdist
SonyAIBORobotSurface	0.6055	0.9452	0.9662	0.5620 [13]	0.9452
SonyAIBORobotSurfaceII	0.6168	0.6173	0.6173	0.6320 [13]	0.8071
OliveOil	0.5313	0.6833	0.5167	0.7764 [15]	0.8833
CBF	0.3368	0.3333	0.3543	0.5810 [14]	0.6441
ECG200	0.6750	0.6950	0.6700	0.5952 [15]	0.7850
Synthetic	0.2172	0.2500	0.2094	0.6670 [13]	0.7117
Lighting2	0.6116	0.6116	0.6033	0.4993 [15]	0.6033
Yoga	0.5366	0.5367	0.5364	0.5100 [16]	0.5433
Beef	0.3750	0.4692	0.4833	0.3170 [13]	0.5000
Gisette	0.3322	0.3401	0.3918	*	0.4392
Arcene	0.4021	0.4881	0.4701	*	0.5801

was selected, and the distance between each U-Shapelet in the U-Shapelets set and all sequences in the dataset was calculated to form the distance matrix Dist. The traditional K-MESNS clustering method is used to cluster the distance matrix to obtain the final clustering result. Ferreira and Zhao [19] establish a network in which each sequence is represented by a vertex and each vertex is connected to the most similar vertex by DTW. A hierarchical clustering is proposed, which uses DTW based functions to measure the similarity between clusters and iteratively calculates the most similar clustering.

Table 3 shows the weighted distance obtained when different data cluster classes have the highest accuracy, namely, the optimal weight. It can be seen from the table that different datasets have different degrees of dependence on different distances. Some datasets are more dependent on a certain distance, such as the Yoga dataset, which is more dependent on Euclidean distance. Some datasets are

dependent on the three distances to varying degrees. For example, synthetic datasets are subject to similar dependence on the three distances. Combined with the data in Tables 2 and 3, the effectiveness and accuracy of the weighted distance method proposed in Section 3 of this paper are verified through multiple experiments. We believe that the weighted distance method proposed in this paper can get better clustering results when dealing with clustering problems of different data.

Figure 2 shows the cluster center scatter diagram determined after secondary clustering of CBF, ECG200, OliveOil, and Yoga. Points with different colors in the figure represent clustering centers of different categories, where the x-coordinate represents the feature n of the dataset, and the y-coordinate represents the value corresponding to its feature. It can be seen from the figure that the determined cluster centers are of great difference and produce good clustering results. Therefore, the cluster center selection

TABLE 3: The optimal weights.

Datasets weights	SonyAIBORobotSurface	SonyAIBORoboSurfaceII	OliveOil	CBF	ECG200	Synthetic	Lighting2	Yoga	Beef	Gisette	Arcene
X_1	0.283	0.367	0.344	0.407	0.063	0.460	0.372	0.073	0.498	0.377	0.405
X_2	0.135	0.332	0.175	0.453	0.150	0.310	0.196	0.012	0.039	0.026	0.031

FIGURE 2: The trend of the clustering center. (a) CBF. (b) ECG200. (c) Olive oil. (d) Yoga.

method based on the combination of maximum density and grey relational degree proposed in this paper has a good effect.

To further verify the feasibility of the clustering algorithm in this paper, cluster analysis was carried out for the MUSIC INTO EMOTIONS problem, the MUSIC INTO EMOTIONS standard dataset [20]. Experiments are conducted on 593 songs with 6 clusters of music emotions based on the Tellegen-Watson-Clark model (six classes: Classical, Reggae, Rock, Pop, Hip-Hop, Techno, and Jazz).

Results of the MUSIC INTO EMOTIONS problem are presented in Table 4, where the clustering of the FCM is using a classical fuzzy clustering algorithm to get the clustering accuracy, Hybrid-ED, Hybrid DTW, and Hybrid SPDTW, respectively, the Euclidean distance, and DTW

TABLE 4: Comparison of MUSIC into EMOTIONS data accuracy.

FCM	0.8112
Hybrid-ED	0.6523
Hybrid-DTW	0.7051
Hybrid-SPDTW	0.6774
Hybrid-newdist	0.9027

distance, and only SPDTW distance is applied to present the FCMdd and FCM clustering technology in precision, using the proposed Hybrid-Newdist weighted distance metrics as Hybrid clustering technology to measure the distance of the results obtained.

As can be seen from the comparison results in Table 4, the hybrid fuzzy clustering technique based on weighted

TABLE 5: Comparison of YEAST data accuracy.

FCM	0.9044
Hybrid-ED	0.8791
Hybrid-DTW	0.8827
Hybrid-SPDTW	0.8518
Hybrid-newdist	0.9360

distance proposed in this paper also achieves good results when dealing with practical problems, thus verifying the effectiveness of the algorithm.

To enhance the advantages of our proposed model, another real data YEAST is applied [21], which is a mixed distributed dataset. The Yeast dataset is formed by micro-array expression data and phylogenetic profiles with 1500 genes in the learning set and 917 in the test set. The input dimension is 103. Each gene is associated with a set of functional classes whose maximum size can be potentially more than 190. The same experiments are repeated, and the results are listed in Table 5, which provides another convincing support for our model. Hybrid-Newdist provides significantly better results than other traditional algorithms.

5. Conclusion and Discussion

Based on the research content and numerical analysis in this paper, the conclusions are as follows:

- (1) The initial clustering center determined by the method based on the principle of maximum density and the grey relational degree proposed in this paper can reduce the dependence of the algorithm on the initial center point and avoid falling into the problem of an optimal local solution.
- (2) In this paper, double-layer fuzzy clustering is adopted to realize data transformation, which improves the algorithm's running speed to a certain extent.
- (3) The weighted distance measurement proposed in this paper combines the advantages of Euclidean distance, DTW distance, and SPDTW distance. We use particle swarm optimization to improve clustering accuracy and make the clustering results more accurate.
- (4) Through numerical experiments on the part of UCR datasets, the clustering accuracy obtained is improved to different degrees than the existing literature.

By analyzing the data of the two real problems, the comparison results show that the clustering accuracy of the algorithm proposed in this paper is almost 7.1% higher than that of the classical FCM algorithm, which indicates that the algorithm proposed in this paper can also achieve a better clustering effect when dealing with practical problems.

The most significant limitation of our study is that the multiclass problem cannot be solved well. We see that all the binary-class datasets obtain convincing results with our proposed algorithms from the numeric research. However,

the multiclass datasets perform poorly. In the future, a further study on the multiclass datasets is needed. Also, the computational complexity of our proposed algorithm is larger than the traditional ones. It needs longer computational time and requires us to run a more complex programming code. A more efficient algorithm is another goal of our future research.

Data Availability

The “SonyAIBORobotSurface,” “SonyAIBORobotSurfaceII,” “OliveOil,” “CBF,” “ECG200,” “Synthetic,” “Lighting2,” “Yoga,” and “Beef” data used to support the findings of this study have been deposited in the UCI machine learning database repository (<https://archive.ics.uci.edu/ml/index.php>). The MUSIC INTO EMOTIONS data used to support the findings of this study have been deposited in the SourceForge repository (DOI: 10.1186/1687-4722-2011-426793). The YEAST data used to support the findings of this study have been deposited in the SourceForge repository (DOI: <https://doi.org/10.7551/mitpress/1120.003.0092>).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank the editors Lopes Ant nio M. and Guangming Xie. This work was supported by the National Key Research and Development Program of China, under No. 2019YFB1600405; National Natural Science Foundation of China, under No. 71774018; China Postdoctoral Science Foundation, under No. 2015M580128; and Natural Science Foundation of Liaoning Province, under No. 2020-HYLH-20.

References

- [1] A. Onan, “Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering,” *IEEE Access*, vol. 7, pp. 145614–145633, 2019.
- [2] J. E. H. Davidson, D. B. Stephenson, and A. A. Turasie, “Time series modeling of paleoclimate data,” *Environmetrics*, vol. 27, no. 1, pp. 55–65, 2016.
- [3] H. Gupta and R. Srivastava, “K-means based document clustering with automatic “K” selection and cluster refinement,” *International Journal of Computer Science and Mobile Applications*, vol. 2, no. 5, pp. 7–13, 2014.
- [4] A. Onan and M. A. Toolu, “Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts,” *Computer Applications in Engineering Education*, vol. 1–15, 2020.
- [5] G. Li, B. Yang, Y. Liu, and D. Cao, “Survey of data mining based on fuzzy set theory,” *Computer Engineering and Design*, vol. 32, no. 12, 2011, in Chinese.
- [6] K. Zhu, L. Li, and H. Guo, “An improved FCM algorithm integrating GA and PSO,” *Journal of System & Management*, vol. 20, no. 6, pp. 91–96, 2011, in Chinese.

- [7] H. Zhang and J. Wang, "Improved fuzzy C means clustering algorithm based on selecting initial clustering centers," *Computer Science*, vol. 36, no. 6, pp. 206–209, 2009.
- [8] N. Bharill and A. Tiwari, "Handling big data with fuzzy based classification approach," *Advance Trends in Soft Computing*, vol. 312, pp. 219–227, 2014.
- [9] H. Bulut, A. Onan, and S. Korukoğlu, "An improved ant-based algorithm based on heaps merging and fuzzy c-means for clustering cancer gene expression data," *Sadhana*, vol. 45, no. 1, 2020.
- [10] H. Izakian, W. Pedrycz, I. Jamal, X. Yue, and D. Zhang, "Fuzzy clustering of time series data using dynamic time warping distance," *Engineering Applications of Artificial Intelligence*, vol. 39, pp. 235–244, 2015.
- [11] A. Rammal, E. Perrin, V. Vrabie et al., "Optimal preprocessing and FCM clustering of MIR, NIR and combined MIR-NIR spectra for classification of maize roots," in *Proceedings of the 3rd International Conference on E-Technologies and Networks for Development*, Beirut, Lebanon, 2014.
- [12] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, no. 3, pp. 678–693, 2011.
- [13] H. Izakian, W. Pedrycz, and I. Jamal, "Fuzzy clustering of time series data using dynamic time warping distance," *Engineering Applications of Artificial Intelligence*, vol. 39, pp. 235–244, 2015.
- [14] V. Kathiresan and P. Sumathi, "A new fuzzy c means clustering algorithm based on constrained dynamic time warping distance measure," *Journal of Theoretical & Applied Information Technology*, vol. 67, no. 1, pp. 212–219, 2014.
- [15] L. Zhu, "Improvement of weighted Euclidean distance clustering algorithm," *Computer & Digital Engineering*, vol. 3, no. 44, pp. 421–424, 2016, in Chinese.
- [16] M. Luczak, "Hierarchical clustering of time series data with parametric derivative dynamic time warping," *Expert Systems With Applications*, vol. 62, pp. 116–130, 2016.
- [17] X. Huang, Y. Ye, L. Xiong, R. Y. K. Lau, N. Jiang, and S. Wang, "Time series k-means: a new k-means type smooth subspace clustering for time series data," *Information Sciences*, vol. 367–368, pp. 1–13, 2016.
- [18] S. Yu, Q. Yan, and X. Yan, "Clustering algorithm of time series with optimal U-shapelets," *Journal of Computer Applications*, vol. 37, no. 8, pp. 2349–2356, 2017, in Chinese.
- [19] L. N. Ferreira and L. Zhao, "Time series clustering via community detection in networks," *Information Sciences*, vol. 326, pp. 227–242, 2015.
- [20] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multilabel classification of music into emotions," in *Proceedings of the 2008 International Conference on Music Information Retrieval (ISMIR 2008)*, pp. 325–330, Philadelphia, PA, USA, 2008.
- [21] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in Neural Information Processing Systems*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., Vol. 14, MIT Press, Cambridge, MA, USA, 2001.