IJLAI Transactions on Science and Engineering (ISSN: 3006-3418) 2024, vol. 2, no. 3, pp: 26–35

Big Data Clustering Optimization Based on Intuitionistic Fuzzy Set Distance and Particle Swarm Optimization for Wireless Sensor Networks

Ye Li[1], Tianbao Shang[1], and Shengxiao Gao[1]

Software College, Shenyang Normal University
Shenyang 110034 China
liye@163.com
Received June. 29, 2024; Revised and Accepted July. 12, 2024

Abstract. Big data clustering plays an important role in the field of data processing in wireless sensor networks. However, there are some problems such as poor clustering effect and low Jaccard coefficient. This paper proposes a novel big data clustering optimization method based on intuitionistic fuzzy set distance and particle swarm optimization for wireless sensor networks. This method combines principal component analysis method and information entropy dimensionality reduction to process big data and reduce the time required for data clustering. A new distance measurement method of intuitionistic fuzzy sets is defined, which not only considers membership and non-membership information, but also considers the allocation of hesitancy to membership and non-membership, thereby indirectly introducing hesitancy into intuitionistic fuzzy set distance. The intuitionistic fuzzy kernel clustering algorithm is used to cluster big data, and particle swarm optimization is introduced to optimize the intuitionistic fuzzy kernel clustering method. The optimized algorithm is used to obtain the optimization results of wireless sensor network big data clustering, and the big data clustering is realized. Simulation results show that the proposed method has good clustering effect by comparing with other state-of-the-art clustering methods.

Keywords: Big data clustering, Intuitionistic fuzzy set distance, Particle swarm optimization, Wireless sensor networks.

1. Introduction

With the continuous development of information technology and communication technology, wireless sensor networks gradually increase their structural complexity and scale, and the big data in the network increases. Big data in wireless sensor networks is characterized by heterogeneity, diversity and complexity. The value of big data is high and it plays an important role in scientific research, economy and society. Big data clustering is the basic content and key point of big data analysis [1-3]. In the field of data mining, big data clustering has become an important research topic at home and abroad, which can provide a basis for people to know and understand things. In this context, it is of great practical significance to study the clustering optimization method of big data in wireless sensor networks.

In reference [4], big data was input into the convolutional neural network to obtain the coarse features of the data, and the coarse features of the data were obtained by training the rough features of the data through the hole convolution, and the data was input into the capsule network to complete the big data clustering. The Jaccard coefficient of clustering results of this method was low, and the data clustering effect was not good. The method in reference [5] obtained the non-boundary point set and boundary point set of network big data on the basis of supporting the concept of k outlier, and used the SMOTE algorithm and the distance-based undersampling algorithm respectively to perform clustering processing on the above point set to achieve big data classification. However, the average entropy of this method was high and the clustering accuracy was low. Reference [6] run K-means clustering efficiently in a distributed scheme based on mobile machine learning to process big data clustering on the network. It constructed a big data clustering method through K-means clustering technology of neural processor, but the clustering effect of this method was poor.

Since fuzzy sets can depict the fuzzy nature of objective things well, problems such as multi-attribute decision making, pattern recognition and classification based on fuzzy sets have been extensively studied [7,8]. Zadeh fuzzy set describes fuzzy concepts with ambiguous extension through membership degree. Because the fuzzy set can only describe the uncertainty of things through membership degree and reflect the information of yes or no, it is difficult to describe the fuzziness of things comprehensively. Alkan et al. [9] proposed the concept of intuitionistic fuzzy set on the basis of Zadeh fuzzy set theory. The fuzzy set was represented by membership function and non-membership function, which could express the information of membership degree, non-membership degree and hesitation degree at the same time. Compared with the traditional Zadeh fuzzy set, it is more flexible and

authentic in dealing with the fuzzy uncertainty of concepts. Intuitionistic fuzzy set theory has developed rapidly and has become a research hotspot. Research achievements on the application of intuitionistic fuzzy sets are mainly concentrated in the fields of multi-attribute decision making, pattern recognition, fuzzy optimization and fault diagnosis [10-12].

According to the characteristics of intuitionistic fuzzy set, intuitionistic fuzzy set can not only describe the uncertainty degree of fuzzy information, but also describe its unknown degree, in which membership degree and non-membership degree reflect the uncertainty degree of information, and hesitation degree reflects the unknown degree of information [13-15]. In practice, it often encounters the problem of comparing two or more fuzzy concepts. One of the methods is to compare the distance between fuzzy concepts. Since intuitionistic fuzzy sets express fuzzy concepts through membership degree, non-membership degree and hesitation degree, this paper mainly studies the distance measurement problem between intuitionistic fuzzy sets. However, in the definition of distance measure between fuzzy sets or intuitionistic fuzzy sets, the distance measure is expressed by membership degree, non-membership degree and/or hesitancy degree, instead of extending the distance measure definition between sets to the distance measure definition between fuzzy sets or intuitionistic fuzzy sets. Therefore, it should be made clear here that the distance measurement between intuitionistic fuzzy sets in this paper is expressed for two intuitionistic fuzzy sets in the same domain by their membership degree, non-membership degree and/or hesitation degree, which is different from the usual distance definition between two sets, but the expression of intuitionistic fuzzy set distance is still used in this paper.

The existing methods on distance measurement of intuitionistic fuzzy sets are mainly divided into two aspects:

1. Intuitionistic fuzzy set distance based on Hamming distance and Euclidean distance. Based on the distance of Zadeh fuzzy set, Zhou et al. [16] proposed the intuitionistic fuzzy set Hamming distance and Euclidean distance, which only considered membership degree and non-membership degree. Alsattar et al. [17] generalized Atanassov distance by introducing weights and constructed several intuitionistic fuzzy set distances, but these distances did not take hesitation into account. For this reason, Ngan et al. [18] introduced hesitation degree into Atanassov method to improve it. Later, references [19,20] extended the method of reference [18].
2. Intuitionistic fuzzy set distance based on Hausdorff metric. For example, references [21,22] defined the distance of intuitional fuzzy sets based on the Hausdorff metric. Later, reference [23] revised the method in references [22]. However, hesitancy was not considered in these distances. To this end, the hesitancy degree was introduced in reference [24] on the basis of reference [23], and the weight was introduced into the method in reference [25].

Although the above method plays a certain role in measuring the distance of intuitionistic fuzzy sets, and has been applied in practical problems, it still has some shortcomings in distinguishing the distance between some intuitionistic fuzzy sets. Therefore, based on the conditions of distance measurement and intuitionistic fuzzy set distance measurement, this paper analyzes the shortcomings of the existing intuitionistic fuzzy set distance measurement. Secondly, since both membership degree and non-membership degree reflect the uncertainty degree of information, and membership degree (non-membership degree) clearly describes the degree to which an element belongs to (does not belong to) an uncertain object, they have the same status and role in intuitionistic fuzzy sets. The degree of hesitation reflects the unknown degree of information, which can be interpreted as an ambiguous degree of decision maker's affirmation or negation of an uncertain object. Therefore, there may be a degree of partial affirmation and a degree of partial negation in the degree of hesitation. The degree of hesitation may be partially or fully converted to membership (or non-membership) if the uncertain object is further explained or explained. Therefore, the degree of hesitation may have a certain degree of allocation to the degree of membership and non-membership. In order to solve the problems in the above methods, a clustering optimization method for big data of wireless sensor networks based on particle swarm optimization is proposed. This method mainly introduces particle swarm optimization algorithm, and combines principal component analysis and information entropy to further optimize the clustering effect of big data and achieve efficient clustering of big data in wireless sensor networks.

2. Big Data Dimension Reduction Processing

The high dimension of big data in wireless sensor networks increases the difficulty of clustering. Therefore, it is necessary to pre-process big data in wireless sensor networks, that is, dimensionality reduction processing data. The dimensionality reduction process is a clustering optimization method for big data of wireless sensor networks based on particle swarm optimization. The concept of information entropy is introduced into the principal component analysis method [26-28] to carry out dimensionality reduction processing for big data of wireless sensor networks.

2.1. Information Entropy

The data transmitted by the data source in the wireless sensor network has m values $X = s_1, s_2, \cdots, s_m$, the probability of each value is expressed by $(a_1, a_2, \cdots, a_m)$, and there is $\sum_{i=1}^{m} a_i = 1$. Information entropy J describes the average value of data uncertainty $-\log a_i$, expressed as follows:

$$J = -\sum_{i=1}^{m} a_i \log a_i. \tag{1}$$

The smaller the information entropy, the less information exists in the data; conversely, the larger the information entropy, the more information exists in the data. Therefore, in the process of data dimensionality reduction, the data with high information entropy should be retained.

2.2. Principal Component Analysis (PCA)

The process of PCA for data as follows:

1. In the wireless sensor network, the m sample data is observed for n times, and the observation matrix x_{ij} is established according to the observation value X:

$$X = [x_{ij}]_{n \times m}. \tag{2}$$

2. Calculate the mean $\bar{x}_j$ and standard deviation s_j of the data by the following formula.

$$\bar{x}_j = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}}{m}. \tag{3}$$

$$s_j = \sqrt{\frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_j)^2}{m}}. \tag{4}$$

Standardized processing of wireless sensor network data to obtain a standard value $\bar{x}_{ij}$:

$$\bar{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}. \tag{5}$$

The standardized observation matrix $\tilde{x}_{ij}$ of wireless sensor networks is established using the standard value $\tilde{X}$:

$$\tilde{X} = [\tilde{x}_{ij}]_{n \times m}. \tag{6}$$

3. Establish the correlation matrix E of the data:

$$E = \frac{\tilde{X}^T \tilde{X}}{m} = [e_{ij}]_{n \times m}. \tag{7}$$

Where e_{ij} represents the elements present in the correlation matrix.

4. Matrix E is decomposed by the following process:
A. Let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m \geq 0$ represent the eigenvalue of matrix E, and calculate the contribution rate V_j of the eigenvalue by the following formula:

$$V_j = \frac{\mu_j}{\mu_1 + \mu_2 + \cdots + \mu_m}. \tag{8}$$

B. Select the eigenvalues of $V_j > 80\%$ as the principal components of the big data of wireless sensor networks, and use $r_1, r_2, \cdots, r_m$ to represent the eigenvectors of the eigenvalues.
C. The first a eigenvectors in $r_1, r_2, \cdots, r_m$ are selected to establish the principal component load array $I_{m \times a} = (r_1, r_2, \cdots, r_m)$.

5. The principal components of wireless sensor network data are obtained.

2.3. PCA Based on Information Entropy

Combined with information entropy and principal component analysis [29], dimensionality reduction of wireless sensor networks is carried out. The specific process is as follows:

1. Set the threshold of information entropy ε, compare J and ε of the data, screen the data features, calculate the attribute r_i, and the corresponding information entropy $J(r_i)$. When $J(r_i) > r_i$, r_i is stored in the set S.
2. Centralized processing data matrix to obtain $X_{n\times m}$.
3. Calculate the covariance matrix Cov between the data.
4. The eigenvector and eigenvalue corresponding to Cov are obtained.
5. The first l eigenvector with large eigenvalue of wireless sensor network data is selected, and the eigenvector matrix $B_{n\times 1}$ of the data is established.
6. Dimensionality reduction result U of wireless sensor network is obtained.

$$U = B^T X. \tag{9}$$

At this point, dimensionality reduction results are output through equation (9) to complete dimensionality reduction processing of big data, which lays a foundation for the introduction of particle swarm optimization algorithm and the construction of big data clustering in wireless sensor networks.

3. Particle Swarm Optimization (PSO) for Big Data Clustering Method

After dimensionality reduction of big data, the clustering optimization algorithm of big data of wireless sensor network based on particle swarm optimization algorithm adopts particle swarm optimization algorithm to optimize the clustering center of intuitionistic fuzzy kernel clustering algorithm, and completes the clustering optimization of big data of wireless sensor network by using the optimized algorithm.

Let $X = x_1, x_2, \cdots, x_n$ represent the particle population, which is composed of n particles existing in the M-dimensional space. In the particle swarm algorithm, let $x_{id}(t)$ represent the corresponding position of the population at the current moment. $v_{id}(t)$ represents the velocity corresponding to the population at the current moment, the inertia factor ξ is set, and the position $x_i = x_{i1}, x_{i2}, \cdots, x_{in}$ and velocity $v_i = v_{i1}, v_{i2}, \cdots, v_{in}$ of the $i-th$ particle existing in the population in the optimization process are updated by equations (10) and (11):

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1). \tag{10}$$

$$v_{id}(t+1) = \xi v_{id}(t) + c_1 r_1 x_{id}(t) + c_2 r_2 x_{id}(t). \tag{11}$$

Where t represents the number of iterations of the population. c_1 and c_2 represent the acceleration constant. $x_{id}(t+1)$ and $v_{id}(t+1)$ represent the new position and velocity obtained by the particle after the update of the above formula. $r_1, r_2 \in [0, 1]$ is a random number.

Particle swarm optimization algorithm has strong convergence speed and global search ability [30]. The clustering optimization method of big data of wireless sensor networks based on particle swarm optimization algorithm takes advantage of this feature to optimize the intuitive fuzzy kernel clustering algorithm and improve the clustering efficiency of the clustering optimization method of big data of wireless sensor networks based on particle swarm optimization algorithm.

$X = x_1, x_2, \cdots, x_n$ is used to represent the data sample space, and the data cluster center is represented by particles, forming a set $V = v_1, v_2, \cdots, v_n$. Set the fitness function $g(x_i)$ of the PSO algorithm:

$$g(x_i) = [K_{km}(I_{kv}, I_{k\eta}, A) + 1]^{-1}. \tag{12}$$

Where $K_{km}(I_{kv}, I_{k\eta}, A)$ represents intuitionistic fuzzy kernel. I_{kv} represents the membership matrix. $I_{k\eta}$ represents a non-membership matrix. A represents the clustering result when the optimal solution is output.

3.1. Zadeh Fuzzy Set Distance

In order to simplify the expression of the distance formula of intuitionistic fuzzy sets, this paper uses the following abbreviated notation, that is, for any two intuitionistic fuzzy sets A and B on the domain X, let $\Delta_{\mu}^{AB}(i) = \mu_A(x_i) - \mu_B(x_i)$, $\Delta_{v}^{AB}(i) = v_A(x_i) - v_B(x_i)$, $\Delta_{\pi}^{AB}(i) = \pi_A(x_i) - \pi_B(x_i)$ represent the difference of membership degree, the difference of non-membership degree and the difference of hesitation degree of two intuitionistic

fuzzy sets A and B respectively. In addition, in order to facilitate subsequent formula is concise, $\Delta_\mu(i) = \Delta_\mu^{AB}(i)$, $\Delta_v(i) = \Delta_v^{AB}(i)$, $\Delta_\pi(i) = \Delta_\pi^{AB}(i)$.

Based on the distance of Zadeh fuzzy sets [31], Atanassov gives the distance metric of intuitive fuzzy sets. The distance measure between any two Zadeh fuzzy sets A and B on the domain X is first given. Zadeh fuzzy set Hamming distance HD_{FS} and standardized Hamming distance NHD_{FS} are respectively expressed as:

$$HD_{FS}(A,B) = \sum_{i=1}^{n} |\Delta_\mu(i)|, NHD_{FS}(A,B) = \frac{1}{n}\sum_{i=1}^{n} |\Delta_\mu(i)|. \tag{13}$$

Zadeh fuzzy set Euclidean distance ED_{FS} and standardized Euclidean distance NED_{FS} are respectively expressed as:

$$ED_{FS}(A,B) = \sqrt{\sum_{i=1}^{n} (\Delta_\mu(i))^2}. \tag{14}$$

$$NED_{FS}(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\Delta_\mu(i))^2}. \tag{15}$$

In the above formula, only the membership difference of Zadeh fuzzy set A and B is considered, but the non-membership difference is not considered. According to the relation between membership degree and non-membership degree of Zadeh fuzzy set, i.e.,

$$\mu_A(x_i) = 1 - v_A(x_i). \tag{16}$$

$$\mu_B(x_i) = 1 - v_B(x_i). \tag{17}$$

The difference of the non-membership degree of Zadeh fuzzy set A and B can also be introduced into the corresponding distance metric formula, and thus,

$$HD'_{FS}(A,B) = \sum_{i=1}^{n} [|\Delta_\mu(i)| + |\Delta_v(i)|]. \tag{18}$$

$$NHD'_{FS}(A,B) = \frac{1}{n}\sum_{i=1}^{n} [|\Delta_\mu(i)| + |\Delta_v(i)|]. \tag{19}$$

$$ED'_{FS}(A,B) = \sqrt{\sum_{i=1}^{n} [(|\Delta_\mu(i))^2| + |(\Delta_v(i))^2|]}. \tag{20}$$

$$NED'_{FS}(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} [(|\Delta_\mu(i))^2| + |(\Delta_v(i))^2|]}. \tag{21}$$

3.2. Big Data Clustering Optimization

The specific process of big data clustering optimization in wireless sensor networks is as follows:

1. Initialization parameters, including maximum speed v_{max}, termination threshold ϕ, inertia factor ξ, maximum number of iterations Y_{max}, constant c_1, c_2, population size z.
2. Initializing the particle swarm and select the initial population $V_1, V_2, \cdots, V_z$ of the algorithm. The set $v_1, v_2, \cdots, v_v$ composed of cluster centers of data can be represented by particle V_i.
3. Divide the non-membership matrix $I_{k\eta}$ and membership matrix I_{kv} of wireless sensor network big data, and set $F_{GK}(x_j, a_i)$ represents the intuitive fuzzy Euclidean distance between data cluster center a_i and data x_j [14-16]. When the value of $F_{GK}(x_j, a_i)$ is not zero, the following formula exists:

$$v_{ij} = \sum_{k=1}^{c}\sum_{t=1}^{m} [F_{GK}(x_j, a_i)/F_{GK}(x_t, a_k)]^{2/(m-1)}{}^{-1}. \tag{22}$$

$$\eta_{ij} = 1 - \sum_{k=1}^{c}\sum_{t=1}^{m}[F_{GK}(x_j, a_i)/F_{GK}(x_t, a_k)]^{2/(m-1)}{}^{-1}. \quad (23)$$

In the formula, $v_{ij}^{(b)}$ and $\eta_{ij}^{(b)}$ represent Gaussian kernel and fuzzy kernel respectively. Let the value of $F_{GK}(x_j, a_i)$ be zero, there is the following formula:

$$v_{ij} = 1, \eta_{ij} = 0, i = k; v_{ij} = 0, \eta_{ij} = 1, i \neq k \quad (24)$$

4. Calculate the particle's $g(x_i)$ on the basis of equation (22).
5. Let $A_{id}(t)$ represent the optimal value obtained by the particle in the optimization process, for $A_{id}(t)$, $g(x_i)$ expansion judgment. When the optimal value $A_{id}(t)$ is better than $g(x_i)$, $A_{id}(t)$ is taken as the new position of the particle in the population. Let $V_{gd}(t)$ represent the velocity of the optimal value obtained by the particle swarm in the optimization process, expand the judgment of $V_{gd}(t)$ and $g(x_i)$, when $V_{gd}(t)$ is better than $g(x_i)$, $V_{gd}(t)$ is taken as the new velocity of the particle swarm.
6. The speed and position of particles in the population are updated, and the updated particles are used to form the next generation population of the algorithm.
7. When the number of iterative updates is $t = t + 1$, whether the algorithm meets the termination condition is judged. If so, the optimal solution of the algorithm at this time is output, and the clustering result A of wireless sensor network big data is obtained. If the termination condition is not met, return to step 3.
8. The non-membership matrix $I_{k\eta}$ and membership matrix I_{kv} of wireless sensor network big data are redivided.
9. Set parameters av_{ij}, $a\eta_{ij}$, $a\pi_i$, and update the clustering result A of big data of wireless sensor network using the above parameters.
10. Set the termination threshold ϕ of PSO, when the number of iteration updates is $t = t+1$, if $||A^{t+1} - A^t|| \geq \phi$, return to step 8, if $||A^{t+1} - A^t|| < \phi$, the clustering optimization result A of big data of wireless sensor network is output.

At this point, the design of big data clustering optimization method for wireless sensor network is completed, and the dimensionality reduction processing of wireless sensor network is realized by combining information entropy and principal component analysis method. Finally, particle swarm optimization algorithm is introduced to realize big data clustering.

4. Experimental Comparison and Result Analysis

4.1. Experimental Environment

In order to fully verify the clustering effect of the fast clustering method of structured big data in the parallel processing network designed in this paper, the clustering experiment of semi-structured data will be carried out through the Hadoop experimental platform. In the experiment, seven computers with the same configuration are used to form a Spark cluster, among which one computer serves as the primary control node and the remaining six computers serve as computing nodes. Each node is deployed on the LAN in the laboratory. The related hardware and software configurations are shown in Table 1.

Table 1. Experimental environment hardware and software configuration parameters

Parameter	Size
CPU	Intel Core i7 12700H
Hard disk	2T
Memory	32GB
Network environment	200M LAN
Operating system	Ubuntul16.04

It is known that this simulation experiment adopts 1 main control node and 6 compute nodes to deploy Spark cluster, but the experimental resources are limited, and the main control node is also used as the compute node in this experiment. In this simulation experiment, all the Spark cluster nodes are integrated into the Hadoop platform, and then software installation and testing are carried out through Java language. After the experimental environment is detected correctly, this semi-structured data clustering experiment can be carried out.

4.2. Experimental Index

On the basis of the above experimental environment, the design method in this paper was used as the experimental group, and the methods in reference [32] and reference [33] were used as the control group, and then the simulation experiment was carried out in two stages. In the first stage, based on a published semi-structured data set as experimental data, the experimental group and the control group were respectively used for cluster mining of experimental data, and then the data clustering results under different methods were compared and analyzed. In the second stage, based on the artificially generated semi-structured data set as the experimental data, cluster mining was also done for the data by the experimental group and the control group, and the clustering results of different methods were compared and analyzed. Different data sets were used for cluster analysis in this experiment, the main purpose of which is to verify the applicability of the design method in semi-structured big data clustering and avoid the chance of experimental data affecting the accuracy of experimental results. Table 2 shows the specific distribution of the experimental data sets.

Table 2. Experimental semi-structured data set distribution

Data set	Sample number	Number of categories	Type
Public	300	4	Log file, XML, JSON, email
Manual data set	200	3	XML, JSON, email

At the same time, this simulation experiment takes the efficiency of cluster mining as the experimental index, that is, in the whole mining process of semi-structured data under different methods, the clustering effect of the data is checked every once in a while, so as to measure the clustering efficiency of the experimental group method and the control group method according to the clustering quality of the data.

4.3. Analysis of Experimental Results

Considering the actual situation of semi-structured big data cluster mining, the mining time of two different types of semi-structured data cluster in this simulation experiment is controlled to 30s, and the data clustering effect is checked every 10s during the experiment. After the experiment, the cluster mining results of semi-structured data in the experimental group and the control group were collected and sorted out, as shown in Figure 1 and Figure 2.

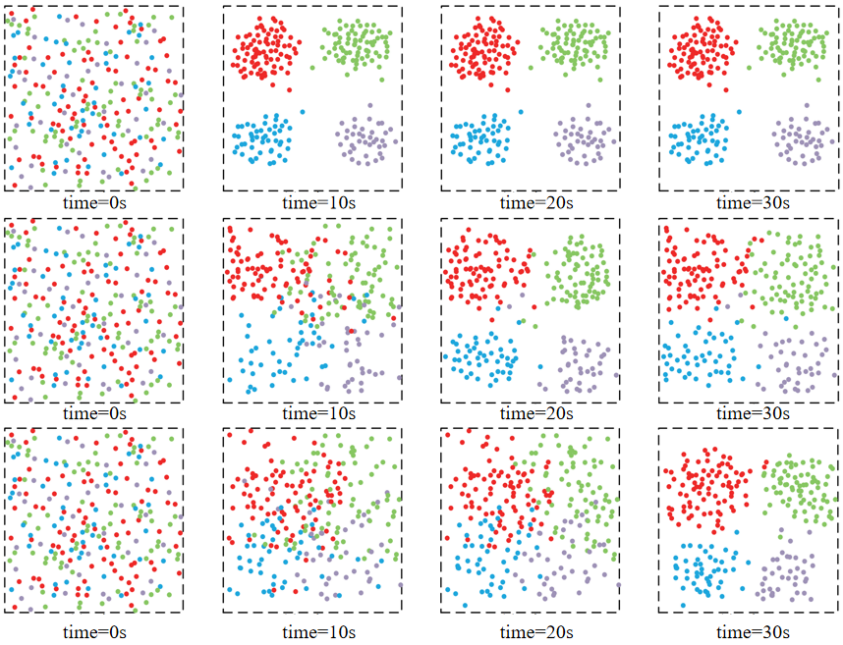

Fig. 1. Comparison of clustering results in public data sets. The first row is the proposed method, the second row is the reference [32] method, and the third row is the reference [33] method

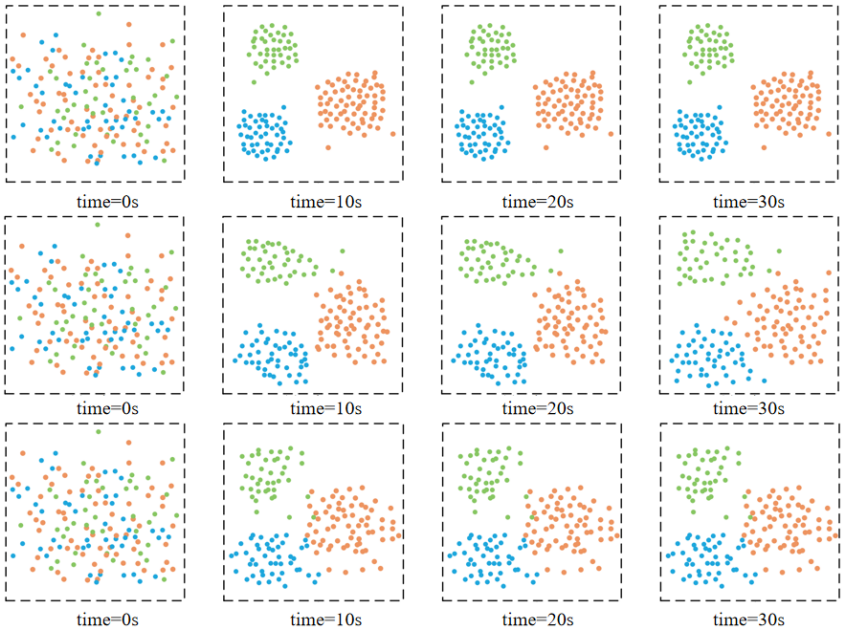

Fig. 2. Comparison of clustering results in manual data sets. The first row is the proposed method, the second row is the reference [32] method, and the third row is the reference [33] method

It can be seen from Figure 1 and Figure 2 that under the artificial semi-structured data set, both the design method in this paper and the method in reference [32] and reference [33] can realize data clustering within 10s. However, compared with the control group method, the semi-structured data clustering results under the design method in this paper are better, and there is no change over time, that is, the stability of the semi-structured data clustering results under the design method in this paper is good. Under a certain public semi-structured data set, with the increase of local clusters to be clustered, the methods in reference [32] and [33] in the control group not only cannot complete the data clustering within 10s, but also the methods in reference [32] cannot guarantee the stability of the clustering results. However, the method designed in this paper can still complete the data clustering within 10s. The clustering quality has been superior, and the clustering results are still relatively stable. This is mainly because the MapReduce framework under parallel processing network is introduced in this paper to carry out parallel clustering of semi-structured data, so the cluster mining speed of the design method will not be affected by the scale of the data set to be clustered. It can be shown that the fast clustering method of semi-structured big data under parallel processing network designed in this paper is reasonable and correct, and can ensure the stability of clustering results on the basis of improving the efficiency of cluster mining of semi-structured big data.

5. Conclusion

With the advent of the era of big data, semi-structured big data is becoming more and more important in the Internet environment. However, the rapid development of computer and other technologies makes the scale of big data continuously expand and the structure gradually complex. Traditional data cluster analysis methods have been unable to meet the mining needs of structured big data in the era of big data. Therefore, this paper proposes a fast clustering method for semi-structured big data under parallel processing network. First of all, large data packets in Linux and Windows network environment are captured. Since the captured data belongs to multi-source heterogeneous data and the quality is poor, a series of pre-processing of original big data is needed to improve the quality of big data. Then cluster analysis is carried out on the prepared multi-source heterogeneous data under the parallel processing network. This paper uses MapReduce framework to improve the conventional K-means clustering algorithm and form a parallel clustering algorithm, so as to complete the fast clustering mining of semi-structured big data. Finally, through the simulation and comparison experiment results, it is verified that the semi-structured big data under the design method in this paper has a good cluster mining efficiency, and the data mining clustering results are very stable, which can provide theoretical reference for actual big data analysis.

6. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. 2021 Scientific Research Funding Project of Liaoning Provincial Education Department (Research and implementation of university scientific research information platform serving the transformation of achievements).

References

1. Ikegwu A C, Nweke H F, Anikwe C V. Recent trends in computational intelligence for educational big data analysis[J]. Iran Journal of Computer Science, 2024, 7(1): 103-129.
2. Fanelli S, Pratici L, Salvatore F P, et al. Big data analysis for decision-making processes: challenges and opportunities for the management of health-care organizations[J]. Management Research Review, 2023, 46(3): 369-389.
3. Manikandan N, Tadiboina S N, Khan M S, et al. Automation of Smart Home for the Wellbeing of Elders Using Empirical Big Data Analysis[C]//2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2023: 1164-1168.
4. Manikandan N, Tadiboina S N, Khan M S, et al. Automation of Smart Home for the Wellbeing of Elders Using Empirical Big Data Analysis[C]//2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2023: 1164-1168.
5. Widiasanti I, Zanuara A A F, Maulidina D F, et al. Implementation of Big Data in the Zoom and Google Classroom Applications as Online Learning Media[J]. Indonesian Journal of Education and Mathematical Science, 2023, 4(2): 86-92.
6. Sun H. Construction of integration path of management accounting and financial accounting based on big data analysis[J]. Optik, 2023, 272: 170321.
7. Irinyi L, Roper M, Malik R, et al. In silico environmental sampling of emerging fungal pathogens via big data analysis[J]. Fungal Ecology, 2023, 62: 101212.
8. Shrivastava A, Nayak C K, Dilip R, et al. Automatic robotic system design and development for vertical hydroponic farming using IoT and big data analysis[J]. Materials Today: Proceedings, 2023, 80: 3546-3553.
9. Alkan N, Kahraman C. Continuous intuitionistic fuzzy sets (CINFUS) and their AHP&TOPSIS extension: Research proposals evaluation for grant funding[J]. Applied Soft Computing, 2023: 110579.
10. Gohain B, Chutia R, Dutta P. A distance measure for optimistic viewpoint of the information in interval-valued intuitionistic fuzzy sets and its applications[J]. Engineering Applications of Artificial Intelligence, 2023, 119: 105747.
11. Yu J, Lu Z, Yin S, et al. News recommendation model based on encoder graph neural network and bat optimization in online social multimedia art education[J]. Computer Science and Information Systems, 2024. doi: 10.2298/CSIS231225025Y.
12. Yin S, Li H, Sun Y, et al. Data Visualization Analysis Based on Explainable Artificial Intelligence: A Survey[J]. IJLAI Transactions on Science and Engineering, 2024, 2(2): 13-20.
13. Krishankumar R, Ravichandran K S, Aggarwal M, et al. An improved entropy function for the intuitionistic fuzzy sets with application to cloud vendor selection[J]. Decision Analytics Journal, 2023, 7: 100262.
14. Gogoi S, Gohain B, Chutia R. Distance measures on intuitionistic fuzzy sets based on cross-information dissimilarity and their diverse applications[J]. Artificial Intelligence Review, 2023, 56(Suppl 3): 3471-3514.
15. Acharya A, Mahata A, Sil N, et al. A prey-refuge harvesting model using intuitionistic fuzzy sets[J]. Decision Analytics Journal, 2023, 8: 100308.
16. Zhou Y, Ejegwa P A, Johnny S E. Generalized similarity operator for intuitionistic fuzzy sets and its applications based on recognition principle and multiple criteria decision making technique[J]. International Journal of Computational Intelligence Systems, 2023, 16(1): 85.
17. Alsattar H A, Mourad N, Zaidan A A, et al. Developing IoT sustainable real-time monitoring devices for food supply chain systems based on climate change using circular intuitionistic fuzzy set[J]. IEEE Internet of Things Journal, 2023.
18. Ngan S C. An extension framework for creating operators and functions for intuitionistic fuzzy sets[J]. Information Sciences, 2024, 666: 120336.
19. Patel A, Jana S, Mahanta J. Construction of similarity measure for intuitionistic fuzzy sets and its application in face recognition and software quality evaluation[J]. Expert Systems with Applications, 2024, 237: 121491.
20. Yazdi M, Kabir S, Kumar M, et al. Reliability analysis of process systems using intuitionistic fuzzy set theory[M]//Advances in Reliability, Failure and Risk Analysis. Singapore: Springer Nature Singapore, 2023: 215-250.
21. Ali Z, Emam W, Mahmood T, et al. Archimedean Heronian mean operators based on complex intuitionistic fuzzy sets and their applications in decision-making problems[J]. Heliyon, 2024, 10(3).
22. Bharati P, Acharya A, Mahata A, et al. A two-compartment drug concentration model using intuitionistic fuzzy sets[J]. Decision Analytics Journal, 2024, 10: 100386.
23. Joshi B P, Joshi N, Gegov A. TOPSIS based Renewable-Energy-Source-Selection using Moderator Intuitionistic Fuzzy Set[J]. International Journal of Mathematical, Engineering and Management Sciences, 2023, 8(5): 979-990.
24. Malik S C, Raj M, Thakur R. Weighted correlation coefficient measure for intuitionistic fuzzy set based on cosine entropy measure[J]. International Journal of Information Technology, 2023, 15(7): 3449-3461.

25. Zhou Y, Zhang X, Chen Y, et al. A water-land-energy-carbon nexus evaluation of agricultural sustainability under multiple uncertainties: The application of a multi-attribute group decision method determined by an interval-valued intuitionistic fuzzy set[J]. Expert Systems with Applications, 2024, 242: 122833.
26. Yin S, Li H, Laghari A A, et al. An anomaly detection model based on deep auto-encoder and capsule graph convolution via sparrow search algorithm in 6G internet-of-everything[J]. IEEE Internet of Things Journal, 2024. doi: 10.1109/JIOT.2024.3353337.
27. Ikotun A M, Ezugwu A E, Abualigah L, et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data[J]. Information Sciences, 2023, 622: 178-210.
28. Arulananth T S, Balaji L, Baskar M, et al. PCA based dimensional data reduction and segmentation for DICOM images[J]. Neural Processing Letters, 2023, 55(1): 3-17.
29. Akbar M, Ahmad I, Mirza M, et al. Enhanced authentication for de-duplication of big data on cloud storage system using machine learning approach[J]. Cluster Computing, 2024, 27(3): 3683-3702.
30. Liu T, Yin S. An improved particle swarm optimization algorithm used for BP neural network and multimedia course-ware evaluation[J]. Multimedia Tools & Applications, 76(9):11961-11974, 2017.
31. Mashchenko S O. On a value of a matrix game with fuzzy sets of player strategies[J]. Fuzzy Sets and Systems, 2024, 477: 108798.
32. Fazel E, Najafabadi H E, Rezaei M, et al. Unlocking the power of mist computing through clustering techniques in IoT networks[J]. Internet of Things, 2023, 22: 100710.
33. Jiang Y, Yang G, Li H, et al. Knowledge driven approach for smart bridge maintenance using big data mining[J]. Automation in Construction, 2023, 146: 104673.

Biography

Ye Li is with the Software College, Shenyang Normal University. Research direction is IoT, Information systems, Computer application and AI.
Tianbao Shang is with the Software College, Shenyang Normal University. Research direction is IoT, Computer application and AI.
Shengxiao Gao is with the Software College, Shenyang Normal University. Research direction is IoT, Computer application and AI.