Cloud Computation-Based Clustering Method for Nonlinear Complex Attribute Big Data

Yanfei Lv

Abstract—**Traditional big data clustering methods for complex attributes have problems of poor clustering results and long clustering running time. Therefore, this paper proposes a nonlinear big data clustering method for complex attributes based on cloud computing. Based on cloud computing environment, the storage framework of big data is constructed to extract the characteristics of information flow model of big database. Multiple regression analysis method is used to obtain the output update rules of big data clustering with nonlinear complex attributes, and achieve balanced configuration of big data transmission channels. Hash function of nonlinear complex attributes is obtained according to p-stable distribution, and hierarchical sampling method is used to achieve big data classification with nonlinear complex attributes. Cloud computing technology is used to complete sample allocation and clustering of nonlinear complex attribute big data. Experiments show that the clustering effect of complex attribute big data proposed in this paper is good, and the clustering running time is significantly shortened.**

Index Terms—**Big data clustering, balanced allocation, cloud computing, complex attribute, hash function**

I. Introduction

With the application and advancement of cloud computing, artificial intelligence, the Internet of Things, intelligent robots and other new technologies, human civilization has entered the data-intensive DT era from the traditional IT era, and big data has become the most important asset of society. At the third China Electronic Information Expo in 2015, Li Deyi, academician of Chinese Academy of Engineering, delivered a speech with the theme of "Big Data Cognition". In the edge cloud system, computing-intensive tasks are unloaded to the cloud, and the feature learning tasks are implemented at the network edge [1]. In his opinion, big data itself is neither science nor technology, but reflects an objective existence in the Internet era. The big data generated by all walks of life, from TB, PB to EB to ZB, are growing rapidly in three orders of magnitude, which is difficult to recognize with traditional tools. The huge amount of data, low-value density, real-time online, multi-source heterogeneity, has caused great confusion to people's cognition. In order to reduce the computational complexity, the clustering method plays an important role. [2, 3]. In order to make a breakthrough in human cognitive science, it is necessary to make a breakthrough in big data clustering, which is the first step to mine the value of big data assets. Such clustering is interdisciplinary, cross-domain and cross-media, and has become the core competitiveness of many industries. It is very suitable for deriving the clustering trend of big data in visual form [4]. Birds of a feather flock together and people flock together. This is the basic ability of mankind to understand the world and society for thousands of years. It is a universal and fundamental problem that we must face to discover value from big data. Whether it is politics, economy, literature, history, society, culture, mathematics, chemical industry, medicine, agriculture, transportation, geography and other industries, the big data can be found through clustering in macro or micro value. The purpose is to achieve the definition of cluster centroids clustered for each batch [5, 6]. Cluster analysis refers to dividing a data set into multiple meaningful or useful groups (called "clusters"). Cluster analysis can be used for direct knowledge understanding, such as species classification in biology, result grouping in information retrieval, climate model classification in geometeorology, stress type or disease type detection in psychology and medicine, customer classification in business intelligence, etc. It can also serve as the basis for other data analysis and processing techniques, such as data aggregation, data compression, and nearest neighbor discovery. Cluster analysis comes from and serves many research fields, including data mining, statistics, machine learning, pattern recognition, etc. Cluster analysis is a key technology of data mining and a typical unsupervised machine learning method. As a result, the algorithms originally designed with a sequential nature shall be frequently redesigned to realize effective use of computational resources which are distributed [7]. Unlike classification, there is no sample data to learn, so cluster analysis needs to find the similarity between data points from the attribute features of the data set itself, and there is no strict evaluation standard for the quality of clustering results, which is more challenging. Therefore, it has become a hot spot of concern for researchers in these fields. At present, the hot areas of cluster analysis research mainly include scalability of clustering method; Clustering of data with complex shapes (such as non-convex); Clustering of multiple types of data (such as text, graphic image, voice, video); Clustering of high-dimensional data (data with thousands of attributes) and mixed attribute data. Nevertheless, the cost of cloud computation would be extremely high without proper management [8].

Under cloud computing, the big data information

Manuscript received November 02, 2021; revised May 24, 2022. The work was supported by 2017 Zhejiang Provincial Department of education scientific research project "Research on process based practical teaching of higher vocational software technology specialty based on Alibaba cloud (No. y201738080)."

Yanfei Lv is a teacher of Jinhua Polytechnic, Jinhua, 321017, China (e-mail: kmvin_921@163.com).

processing is mainly used to realize data clustering. Simply, it is to adopt the same attribute data feature parameters in big data to achieve comprehensive data analysis. Based on data clustering, create a large database, through the realization of the corresponding pattern recognition and diagnostic analysis services. Target recognition, fault diagnosis, and cloud storage database creation are main research directions of the technology of big data optimization clustering, which is of great application value in this regard. At present, people pay more and more attention to the research on the optimization clustering method of large databases. In FPWhale-MRF, the mapper function makes estimation on cluster centroids which adopt the Fractional Tangential-Spherical Kernel clustering algorithm. This algorithm is based on the integration of the fractional theory with a Tangential-Spherical Kernel clustering approach. [9]. Modern data clustering algorithms mainly include fuzzy C-means clustering method, data clustering method based on network technology, a clustering algorithm based on adaptive beam, and fuzzy K-means clustering algorithm. All these algorithms achieve classification by obtaining the similarity between attribute features of big data information flow. Both fuzzy C-means and fuzzy K-means algorithms achieve clustering optimization through repeated adjustment of clustering results. Thus, it is extremely significant to use the technology of automatic text categorization for the classification and management of this information [10]. In the process of expanding the data scale, its sensitivity to the initial cluster center is also increasing. For data clustering, the fuzzy C-means clustering algorithm has the main disadvantage that it is sensitive to the data in the clustering center of noise and initial value and is prone to fall into local optimization solutions. On the basis of major properties indicated in the methods of partitional clustering of big data, a new categorizing model is designed to ensure scalability in the analysis of mass of data [11].

To solve the above problems effectively, a cloud computation-based clustering method is proposed for big data with nonlinear complex attributes.

II. Cloud Computing Analysis

A. Big data storage framework in cloud computing environment

Cloud computing means the structural model and storage space which are expanded in a dynamical manner by the modern Internet. To achieve effective storage and classification mining of big data under cloud computing, the establishment of plenty of storage mechanism architectures must be realized first. Under cloud computing, large database storage to storage pool implementation through virtualization, cloud computing deployment, by using computer cluster, which mainly includes virtual computer, USB, disk layer structure, in the enterprise operation center can use the terminal used for the application, thus in the distributed computer to calculate. In data mining, the clustering technique is crucial [12]. Figure 1 shows the big data storage architecture under cloud computing. According to Figure 1, if all cloud computing virtual machines have realized indoor allocation, the clustering algorithm can be comprehensively optimized by using the following formula, to achieve physical clustering allocation of big data features under cloud computing through the optimal solution. In case of high dataset volume or dimension, these algorithms cannot meet the requirements for high computational complexity and large memory [13], where the clustering center is represented as V_m and the physical machine is represented as P_m, and the formula is:

$$N = (1/n)(U_t - U_i) + (1/n)(U_t^1 - U_i^1) + (1/n)(U_t^2 - U_i^2) \quad (1)$$

$$X = [x(t_0), x(t_0 + t), ..., x(t_0 + (K+1)t)] \quad (2)$$

Then collect corresponding samples and conduct comprehensive analysis and judgment on them, and then take this sample as the main data. Assuming that the data information flow sample in the large database is set as: $S = X_1, X_2, ..., X_k$, then at time data samples are analyzed in $T_1, T_2, ..., T_k$. It is capable of managing different data sources and formats under several advanced technologies [14]. Now, the big data set X under cloud computing is divided into class c, where 1 is less than c and less than n. Data segmentation is realized as spatial segmentation to obtain the big data storage structure's center vector. In this post, the algorithm of firefly optimization was adopted for finding optimal cluster centers [15]. Since the data blocks uploaded on the client side are all of the same size, they need to be cloud-clustered. After comprehensively analyzing the mechanism of big data storage under cloud computing, the accurate analysis of data clustering can be achieved. Clustering is a simple tool of information mining as well as the best tool for analyzing big data [16]. Extracting features of large database information flow model is shown in Figure 1.

B. Extract the features of information flow model of large database

If the time series under cloud computing is represented by the following formula: $\{x(t_0 + t_i), i = 0,1,...,N\text{-}1)\}$, then X and Y represent the number combination, so the spatial situation of big data with nonlinear complex attributes clustering under cloud computing can be obtained:

$$\begin{aligned} &x(t_0 + (k-1)t) \\ &x(t_0 + (k-1)t + Jt) \\ &... \\ &x(t_0 + (N-1)t) \end{aligned} \quad (3)$$

Among it, $x(t)$ represents the information flow time series of big data with nonlinear complex attributes clustering under cloud computing, J indicates the data reconstruction time window function in the cloud computing background. m indicates the target clustering regulator, and t represents the data processing time. First

of all, the data points density is calculated. Each basic cluster consists of center points with density no less than the given threshold as well as points within density range [17]. The diagonal vector is represented as the distance between the particle in the data clustering center and the target solution, and the training gradient descent can be realized by using error back transmission to achieve feature optimization of big data and input it into the data clustering system, to effectively identify the calculated pattern. Combined with IT convergence, the human-oriented technologies for improving living quality has been constantly developed [18].

III. Big Data with Nonlinear Complex Attributes Clustering Method On the Basis of Cloud Computing

A. Balanced configuration of transmission channel of bid data

In the clustering scheduling of big data, the communication channel needs to be balanced. The combination of linear equalization and fractional interval equalization is adopted to balance the configuration of transmission channel of bid data. A new algorithm of semi-supervised association mining is put forward through adding prior information to the sample data under the premise of avoiding time complexity increase [19]. The output update rule of big data with nonlinear complex attributes clustering is by using multiple regression analysis method:

$$P_j(t+1)=\frac{a_1P_j(t)+a_2P_g(t)}{a_1+a_2} \tag{4}$$

$$mbest(t+1)=\frac{1}{n}\sum_{j=1}^{n}P_j(t) \tag{5}$$

$$X_j(t+1)= P_j(t+1)\pm\beta\times\left|mbest(t+1)-X_j(t)\right|\times\ln(\frac{1}{\mu_j(t+1)}) \tag{6}$$

Where $X_j(t)$ indicates the position of data j in the t generation scheduling task; In the big data clustering scheduling iteration, the ± sign in the formula is adopted to control the data output from 0 to 1. Data clustering is a hopeful analytic technique. It is widely applied to solve the IoT problems and the big-data-based problems [20, 21]. β is called the contraction expansion coefficient of the nonlinear complex attribute communication channel. $\mu_j(t+1)$ is the attenuation vector of the nonlinear complex attribute communication channel varying within the range of [0,1], $mbest(t+1)$ is the best location of the big data clustering center, $P_j(t+1)$ is the best location searched by the $t+1$ generation big data clustering center j, and a_1, a_2 are M -dimensional random vectors. Precomputed matrix trace is taken as the number of clusters for a dataset representing the total number of keywords through using vector representation [22]. The correlation distribution feature quantity $P_g(t)$ of big data in nonlinear complex attribute communication is defined as:

$$P_g(t)=\arg\min\{f(p_j(t))\,|\,j=1,2,...,n\} \tag{7}$$

Where, $f(p_j(t))$ is the fitness value of the optimal position searched by the big data clustering search center j in the T generation. Combining ontology mapping and deep learning methods [23], the transmission channel of bid data is balanced and configured, and the configuration coefficients a_1 and a_2 are determined by the following formula:

$$a_1=c_1r_1 \tag{8}$$

$$a_2=c_2r_2 \tag{9}$$

Where r_1 and r_2 are M -dimensional random vectors; c_1 is the output modulation coefficient of big data in the communication channel, and c_2 is the compensation coefficient of multipath channel.

B. Stratified sampling of big data with complex attributes

During stratified sampling (SS), data objects are first divided into layers which are relatively homogeneous (objects in the same layer are more similar) according to certain criteria, and then a part of samples are selected from each layer to form a representative sample subset. As a common large-scale data analysis technology, SS mainly includes two key steps: stratification and sample allocation, the original large-scale data set is divided into different layers; during sample allocation, the sample subset size should be determined and corresponding samples should be extracted from each layer. The two processes are described as follows.

In order to use hierarchical sampling technology in cluster analysis, large-scale data first needs to be divided into some independent layers, and the data objects in the same layer need to be as similar as possible. As we all know, cluster analysis is a typical unsupervised machine learning method. In other words, the data objects of cluster analysis do not have any label information in advance. For the cluster analysis of large-scale data, the difficulty of using hierarchical sampling lies in how to find an approximate hierarchical variable to stratify the data set. Therefore, it is necessary to use an efficient and simple technology to achieve the purpose. As a randomization algorithm, local sensitive hashing has been extensively applied in various fields because of its high performance in computational efficiency and accuracy. Different from the traditional hash algorithm in computer science, the main purpose of local sensitive hash method is to realize the largest conflict probability of similar objects in the original space. It is

mainly to put similar objects in the same bucket which has high probability and put dissimilar objects in the same bucket which has low probability with a set of hash functions.

Thus, because of its effective neighborhood preserving features and high efficiency, this chapter uses LSH method to layer large-scale data.

The hierarchical scheme of generating different groups by LSH method can be described as below in a formal manner. Given a data set X including N objects, each object is described by d-dimensional feature space R^d, and using a set of hash functions $H=\{h_1,...,h_M\}$, a binary code $y=\{y_1,...,y_M\}$ of M bits can be calculated for each object $x\in X$, that is

$$y=\{h_1(x),...,h_M(x)\} \tag{10}$$

The encoding of bit g is calculated as $y_g=h_g(x)$. Each hash function performs a mapping process $h_g:R^d\to \mathrm{B}$. This binary encoding process can also be considered to map the original data points to the binary value space. For any data $x\in X$, the hash function $h_g\in H$ based on P-STABLE distribution is defined as:

$$h_g(x)=\left\lfloor \frac{w_g{}^T x+b_g}{r_g} \right\rfloor \quad mod2 \tag{11}$$

Where, $w_g\in R^d$ refers to a random vector, where each value is selected from the Gaussian distribution in an independent and random manner. $b_g\in R$ is a real number uniformly selected from the interval $[0,r_g)$, where r_g indicates the window size. It should be noted that the hash value generated by the hash function given in formula (9) is $h_g(x)\in\{0,1\}$. In other words, each object will get an M-bit binary code after the hash process.

Thus, the dataset can get up to $L=2M$ different labels, namely, L layers. Through this transformation, two similar objects (such as x and y) in European space will be assigned to the same layer with high probability. The layered result can become an approximate clustering result. To realize the balance between accuracy and computational efficiency, the number of layers L is set to be the same as the number of classes k in the experiment, namely, $M=\lfloor \log_2(k) \rfloor$.

Next, the data in Table 1 is taken as an example to explain the LSH calculation process in brief. Given that dataset X includes 8 objects, and $H=\{h_1,h_2,h_3\}$ is a set of hash functions. Based on formula (9), after the hash process, each data object will obtain a 3-bit binary code, an example of the LSH scheme is shown in Table 1. Therefore, dataset X will be divided into 4 layers through transformation, the hierarchical results of the dataset are shown in Table 2.

C. Big data with complex attributes sample allocation clustering

In layer sampling, the sample allocation methods are mainly divided into three types: average allocation, proportional allocation and optimal allocation. Since the optimal allocation method considers the sample size and object variance in each layer at the same time, this method will be used in this section to determine the sample size to be sampled in each layer and conduct sampling. After layering the data set through local sensitive hash, the data set X which contains N objects are divided into different layers or subsets, marked as $\{S_1,..,S_L\}$, where $\cup_{l=1}^{L} S_l=X, S_l\cap S_h=\varnothing(1\le l,h\le L,l\ne h)$.

First of all, based on the relevant concepts and theorems of the sampling theory, the number of samples to be sampled n can be calculated as follows:

$$n=\frac{(\sum_{l=1}^{L} N_l\sigma_l)^2}{\sigma^2+\sum_{l=1}^{L} N_l\sigma_l^2} \tag{12}$$

Where, N_l and σ_l represents the number of objects and standard deviation in layer S_l respectively. In fact, when the number of sampled data n is greater than 5.00% of the original data set scale n', the sampling scale N should be further adjusted. By introducing the finite population correction coefficient, the above formula can be further modified as follows:

$$n'=\frac{n}{1+n/N} \tag{13}$$

To facilitate description, in the following discussion, the sample size is still marked as n. The next problem is how to determine the number of samples to be sampled at each layer under the constraint of $\sum_{l=1}^{L} n_l=n$, to minimize the variance of X_s in the sample data set. Formally, this problem can be solved in the following ways:

$$\begin{aligned} &\min Var(\overline{X_s})=\sum_{l=1}^{L}\frac{W_l^2\sigma_l^2}{n_l}, \\ &s.t\sum_{l=1}^{L} n_l=n \end{aligned} \tag{14}$$

Where, $W_l=\frac{N_l}{N}$ represents the proportion of the number of objects in layer S_l to the number of objects in the original data set X; σ_l is the standard deviation of the data in the l-th layer; n_l is the number of samples to

be sampled from the l-th layer. The distribution scheme of samples in each layer can be obtained by Lagrange solution.

$$n_l = \frac{nW_l\sigma_l}{\sum_{l=1}^{L} W_l\sigma_l} \tag{15}$$

The above formula shows that it has large $W_l\sigma_l$. The layer with l value should take more samples. A larger value W_l means that the l-th layer contains a larger proportion of data objects, and more samples should be taken from this layer during the sampling process. If σ_l is large, the data in layer S_l is relatively scattered. To reflect the data variability in layer l, more data objects should be extracted from this layer.

In order to verify the superiority of SS of optimal allocation scheme, the variance of sample data subset under uniform random sampling (URS) and SS is compared. With respect to variance, it is widely known that optimal allocation based SS has better performance than SS based on proportional allocation. In other words, the variance of sampling data allocated according to proportion is greater than that of optimal allocation. Therefore, by comparing the variance of SS and URS under proportional distribution scheme, it is sufficient to reflect the superiority of SS under optimal distribution scheme. Assuming μ and σ^2 represents the mean and variance of the original data set respectively, and X_r represents the sample subset obtained by URS from data set X. Under URS, the sample variance of the subset can be formally expressed as:

$$\begin{aligned} \mathrm{var}(\overline{X_r}) &= \frac{\sigma^2}{n} = \frac{1}{n}\frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l}(x_{i,l}-\mu)^2 \\ &= \frac{1}{n}\sum_{l=1}^{L} W_l\sigma_l^2 + \frac{1}{n}W_l(\mu_l-\mu)^2 \end{aligned} \tag{16}$$

The variance of the sample subset obtained by SS under the proportional distribution scheme can be formally expressed as:

$$\mathrm{var}(\overline{X_r}) = \sum_{l=1}^{L} W_l^2\frac{\sigma_l^2}{n_l} = \sum_{l=1}^{L} W_l^2\frac{\sigma_l^2}{nW_l} = \frac{1}{n}\sum_{l=1}^{L} W_l\sigma_l^2 \tag{17}$$

From the above two formulas, it can be seen that the proportional SS always produces less variance than the URS. Therefore, the sampling data set obtained by SS is more typical than the sampling data set obtained by URS, and realizes the big data clustering of complex attributes.

IV. Experiment

A. Data set analysis

The three mixed attribute data sets in UCI were originally used for classification research, so each data set is followed by a classification label attribute, which facilitates the calculation of classification accuracy. Acute inflammations include pathological indexes and physiological indexes of 120 patients who have acute inflammation. The judgement of whether each patient has cystitis and nephritis lies in one numerical attribute (body temperature) and five classification attributes (different symptoms). The heart disease dataset contains 270 patient information. It extracts 6 numerical attributes and 7 classification attributes from 75 original data attributes to judge whether the patient has heart disease. The credit approval dataset contains 690 user information applying for bank credit cards, including 6 numerical attributes and 9 classification attributes, of which 37 records contain missing data, so the number of instances without missing data is 653. The results in the dataset can be divided into approval and disapproval. These three mixed attribute data sets are widely used in the research of mixed attribute clustering. An introduction of the UCI mixed property dataset is shown in Table 3.

The experiment uses cloud computing distance for clustering, the default parameter p is 2%, and the cluster center is manually selected by decision chart. The experiment is realized by MATLAB 2015 a platform. Because the data set provides real class standards, external validity indicators can be used for comparison, and the four indicators of clustering accuracy (ACC), normalized mutual information (NMI), Rand index (RI) and adjusted hand index (ARI) of clustering results can be calculated respectively. The larger the value of these four indicators, the better the clustering effect.

B. Clustering effect comparison

In order to compare the application effects of these distance measures, this paper uses three commonly used mixed attribute data sets in UCI data set, act inflammations, heart disease and credit approval for DPC cluster analysis. The basic information of the three data sets is shown in Table 1. There are 690 data points in the credit data set, but there are 37 missing fields. Removing the missing data will not affect the comparison of experimental results. Therefore, 653 data points are taken for clustering experimental comparison.

Because there are two decision attributes in the acute dataset, each attribute is a binary classification, so it is treated and calculated as two different datasets, which are recorded as acute1 and acute2. However, the data contents of the two data sets are the same but the class standards are inconsistent, so the results obtained will be biased to some extent. Table 4 below shows the five distance measures on acute dataset and their clustering results.

According to results in the above table, the best clustering result on acute1 is the Gower distance, and the accuracy rate reaches 91.60%, but its performance on acute 2 is relatively poor, only 50.80%. The best performance on acute 2 is K-Prototypes distance, and its accuracy rate reaches 100.00%, but its performance on acute1 is poor, only 40.80%. Among all the five distance measurement methods, the cloud computing distance put forward in this study has

achieved a good balance on the two data sets, and the clustering effect on acute1 and acute2 data sets reaches 84.10% and 75.00% respectively. Ocil distance performs fairly well on acute1, but worst on acute 2.

The heart dataset is divided into two types: with or without disease. Table 5 shows the comparative results of clustering results on the cardiac dataset. The above table shows that K-Prototypes distance can achieve better clustering effect, with a correct rate of 81.11%, followed by cloud computing distance, with a difference of only 0.0037 percentage points. The clustering results of credit dataset are divided into two categories. Table 6 shows the comparative results of clustering results on the credit dataset.

According to the above table, OCIL distance can achieve better clustering results with an accuracy of 84.53%, followed by cloud computing distance with an accuracy of more than 81.00%. Were described from four data sets (Acute data set is divided into 2), the five kinds of distance measurement methods have their own strengths and shortcomings in terms of clustering effect, but on the stability of it, the highest stability of cloud computing distance algorithm is put forward in this study, in the clustering results of all the data sets, although there is no best, but have been second only to the best of good results.

C. Run time comparison

To compare the calculation time complexity of the five distance measures above, the run time of the five distance measures on the Acute, the results of the runtime of the five distance measures on the three datasets are shown in Table 7.

The above table shows that the time complexity of Goodall distance calculation is significantly higher than that of the other four distance measurement methods and increases geometrically with the increase of the number of records. As shown in Figure 2, among the other four distance measurement algorithms, the cloud computing distance algorithm has the shortest run time on the three data sets of acute, credit and credit. Ocil and K-Prototypes distance algorithms have the same time. They are more time-consuming than K-Prototypes distance algorithm in the calculation of dimensional entropy weight. Gower involves square calculation, so with the increase of data points.

Its' time complexity increases the fastest. Cloud computing distance algorithm grows slowly, which shows that cloud computing algorithm is also an ideal distance measurement algorithm in terms of time complexity.

D. Clustering accuracy

Because there are many attributes of big data, it increases the difficulty of clustering large data with nonlinear complex attributes. Because some data attributes are relatively similar, it is very easy to have the problem of unsatisfactory clustering effect. Therefore, the clustering accuracy of clustering method on the basis of cloud computing is verified. Table 8 shows the comparison results of clustering accuracy of different data sets.

According to the comparison results of clustering accuracy shown in Table 8, the proposed clustering method based on cloud computing can cluster different data sets with high clustering accuracy, which can reach 99.8% at most. Therefore, it indicates that this method has strong clustering performance.

V. Conclusion

This paper puts forward a big data with nonlinear complex attributes clustering method on the basis of cloud computing. Build the big data storage framework and extract the features of the information flow model of the big database; According to the hierarchical sampling method, the big data classification of nonlinear complex attributes is realized, and the cloud computing technology is adopted to complete the sample allocation and clustering of big data with nonlinear complex attributes. The following conclusions are drawn through experiments:

(1) The cloud computing distance put forward in this paper achieves a good balance on the two data sets, and the clustering effect on acute1 and acute2 data sets reaches 84.1667% and 75% respectively.

(2) On the acute dataset, the run time of cloud computing algorithm is 0.005s, on the credit dataset, the run time of cloud computing algorithm is 0.028s, and on the credit dataset, the run time of cloud computing algorithm is 0.071s.

(3) On acute dataset and credit dataset, the clustering accuracy based on cloud computing clustering method is 99.8%, and on heart dataset, the clustering accuracy based on cloud computing clustering method is 99.7%.

References

[1] F.Y. Bu, Q.C. Zhang, L.T. Yang, and H. Yu, "An Edge-Cloud-aided High-order Possibilistic c-Means Algorithm for Big Data Clustering," IEEE Transactions on Fuzzy Systems, vol. 12, no.9, pp. 1-16, 2020

[2] K, Omkaresh, S Jena, and V. Ravi Sankar, "MapReduce framework Based Big Data Clustering Using Fractional Integrated Sparse Fuzzy C Means Algorithm," IET Image Processing, vol. 12, no.32, pp. 56-69, 2020

[3] S Singhal, and A Sharma, "Mutative ACO Based Load Balancing in Cloud Computing," Engineering Letters, vol. 29, no.4, pp. 1297-1302, 2021

[4] K.R. Prasad, M Mohammed, L.V. Narasimha Prasad, and D. Kumar Anguraj. "An Efficient Sampling-Based Visualization Technique for Big Data Clustering with Crisp Partitions," Distributed and Parallel Databases, vol.15, no.12, pp. 152-171, 2021

[5] R.M. Alguliyev, R.M. Aliguliyev, and L.V. Sukhostat, "Efficient Algorithm for Big Data Clustering on Single Machine," Journal of Intelligent Technology, vol. 005, no.001, pp. 9-14, 2020

[6] J.L. Wu, N. Endo, and M. Saito, "Cluster Analysis for Investigating Road Recovery in Iwate Prefecture Following the 2011 Tohoku Earthquake," Engineering Letters, vol. 29, no.4, 2021

[7] M. Ianni, E. Masciari, G.M. Mazzeo, M. Mezzanzanica, and C. Zaniolo, "Fast and Effective Big Data Exploration by Clustering," Future Generation Computer Systems, vol. 102, no.02, pp. 84-94, 2020

[8] D.W. Li, S.L. Wang, N. Gao, and Y. Yang. "Cutting the Unnecessary Long Tail: Cost-Effective Big Data Clustering in the Cloud," IEEE Transactions on Cloud Computing, vol.16, no.99, pp. 1-17, 2019

[9] O. Kulkarni, S. Jena, and C.H. Sanjay, "Fractional Fuzzy Clustering and Particle Whale Optimization-Based MapReduce Framework for Big Data Clustering," Journal of Intelligent Systems, vol. 16, no.21, pp. 42-56, 2019

[10] Q.T. Xiao, X. Zhong, and C.H. Zhong, "Application Research of KNN Algorithm Based on Clustering in Big Data Talent Demand Information Classification," International Journal of Pattern Recognition and Artificial Intelligence, vol. 34, no.06, pp. 149-176, 2020

[11] M.A.B. Hajkacem, Chiheb-Eddine Ben N'Cir, and N. Essoussi, "Overview of Scalable Partitional Methods for Big Data Clustering:

Techniques, Toolboxes and Applications," Clustering Methods for Big Data Analytics, pp. 1-23, 2019
[12] Y.L. Zhao, S.K. Tarus, L.T. Yang, J.Y. Sun, YF. Ge, and J.K. Wang, "Privacy-preserving Clustering for Big Data in Cyber-Physical-Social Systems: Survey and Perspectives," Information Sciences, vol. 515, no.76, pp. 132-155, 2020
[13] L.H. Meng, Y.C. Jiao, and Y.T. Gu. "An Easy-to-Implement Framework of Fast Subspace Clustering for Big Data Sets," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 3612-3616, 2020.
[14] S. Ravikumar, and D. Kavitha. "A New Adaptive Hybrid Mutation Black Widow Clustering Based Data Partitioning for Big Data Analysis," Wireless Personal Communications, vol. 43, no.1, pp. 1-27, 2021
[15] G. HimaBindu, C.R. Kumar, C. Hemanand, and N.R. Krishna, "Hybrid Clustering Algorithm to Process Big Data Using Firefly Optimization Mechanism," Materials Today: Proceedings, vol. 53, no.2, pp. 56-87, 2020
[16] J. Caiado, N. Crato, and P. Poncela, "A Fragmented-Periodogram Approach for Clustering Big Data Time Series," Advances in Data Analysis and Classification, vol. 14, no.1, pp. 117-146, 2020
[17] W.J. Lu, "Improved K-Means Clustering Algorithm for Big Data Mining under Hadoop Parallel Framework," Journal of Grid Computing, vol. 18, no.3, pp. 53-65, 2020
[18] H. Jung, and K.Y. Chung, "Social Mining-Based Clustering Process for Big-Data Integration," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no.1, pp. 1-12, 2021
[19] F. Wu, and R. Zhou, "The Application of Parallel Clustering Analysis Based on Big Data Mining in Physical Community Discovery," International Journal of System Assurance Engineering and Management, vol. 54, no.43, pp. 1-9, 2021
[20] A.K. Tripathi, K. Sharma, M. Bala, and A. Kumar, "A Parallel Military Dog based Algorithm for Clustering Big data in Cognitive Industrial Internet of Things," IEEE Transactions on Industrial Informatics, vol. 53, no.19, pp. 1-16, 2020
[21] S.K. Srivastava, and S. Devaiya, "Error of Approximation of Functions, Conjugate to the Functions Belonging to Weighted Lipschitz Class Using Matrix Means," IAENG International Journal of Applied Mathematics, vol. 51, no.4, pp. 837-841, 2021
[22] S.M. Zobaed, E. Haque, S. Kaiser, and R.F. Hussain, "$NoCS_2$: Topic-Based Clustering of Big Data Text Corpus in the Cloud," 2018 21st International Conference of Computer and Information Technology (ICCIT), IEEE, 2019
[23] J. Li, "Contrastive Analysis of English Literature Comparative Literature Based on Bayesian Clustering Approach to Big Data," Cluster Computing, vol. 22, no.3, pp.7031-7037, 2019

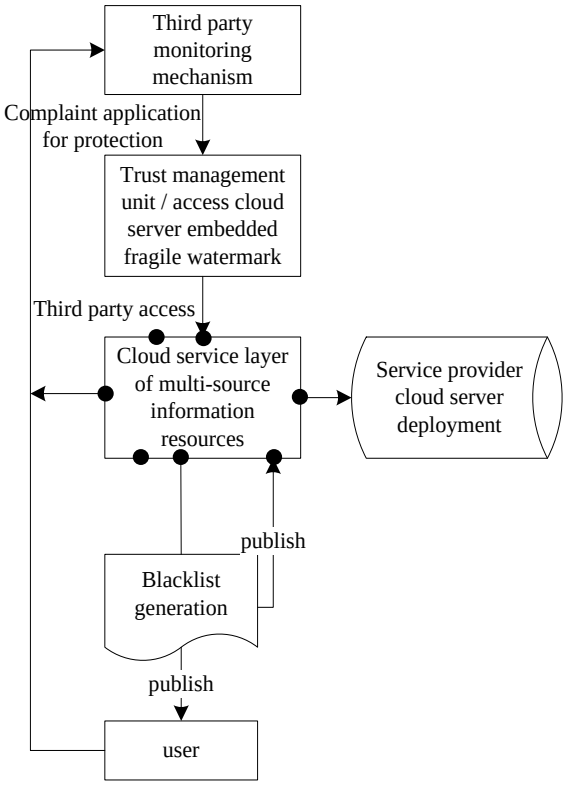

Fig. 1. Big data storage architecture in cloud computing environment.

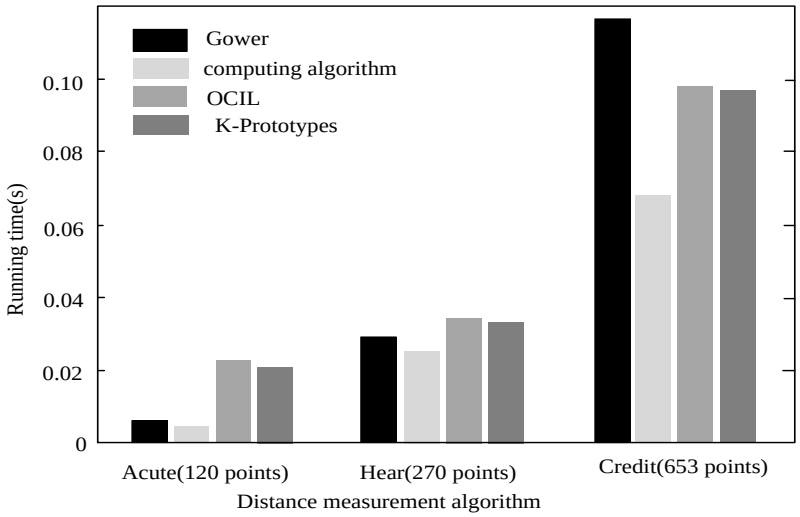

Fig. 2. Comparison of run time.

TABLE 1
EXAMPLE OF LSH SCHEME.

X	H_1	H_2	H_3
X 1	0.0001	0.0001	1.0
X 2	0.0001	1.0	0.0001
X 3	0.0001	0.0001	1.0
X 4	1.0	0.0001	0.0001
X 5	1.0	0.0001	0.0001
X 6	1.0	0.0001	0.0001
X 7	0.0001	1.0	1.0
X 8	0.0001	1.0	1.0

TABLE 2
STRATIFICATION RESULTS OF DATASET X THROUGH H.

Layered label	Object
001	$\{x_1, x_3\}$
010	$\{x_2\}$
011	$\{x_7, x_8\}$
100	$\{x_4, x_5, x_6\}$

TABLE 3
INTRODUCTION TO UCI MIXED ATTRIBUTE DATASET.

Abbreviation	Name	Example	Numeric attribute	Classification properties	Decision attribute	Purpose
Acute	Acute inflammations	120.0	1.0	5.0	2.0	Pathophysiological indexes used to judge acute inflammation
Heart	Heart disease	270.0	6.0	7.0	1.0	Data used to determine whether you have heart disease
Credit	Credit approval	653.0/690.0	6.0	9.0	1.0	It is adopted to judge the customer relationship data of the user applying for credit card and determine whether to grant credit

TABLE 4
COMPARISON OF CLUSTERING RESULTS ON ACUTE DATASET.

Distance algorithm	Clustering results on Acute 1				Clustering results on Acute 2			
	ARI	RI	NMI	ACC	ARI	RI	NMI	ACC
Dgow	0.6919	0.8459	0.6620	91.6	-0.0084	0.4959	0.0001	50.8
Dkpt	0.0256	0.5127	0.0262	40.8	1.0	1.0	1.0	100
Doc	0.4402	0.7199	0.4805	83.3	0.0237	0.5127	0.0147	40.8
Dgd	0.0390	0.5195	0.0358	39.1	0.4856	0.7429	0.3840	85.0
Dudm	0.4629	0.7312	0.4932	84.1	0.2422	0.6218	0.3622	75.0

TABLE 5
COMPARISON OF CLUSTERING RESULTS ON HEART DATASET.

Distance algorithm	ARI	RI	NMI	ACC
Dgow	0.3486	0.6744	0.2661	79.6
Dkpt	0.3843	0.6924	0.3177	81.1
Doc	0.2979	0.6489	0.2279	77.4
Dgd	0.1440	0.5726	0.1961	69.2
Dudm	0.3751	0.6878	0.3107	80.7

TABLE 6
COMPARISON OF CLUSTERING RESULTS ON CREDIT DATA SET.

Distance algorithm	ARI	RI	NMI	ACC
Dgow	0.3463	0.6733	0.2733	79.4
Dkpt	0.0001	0.5036	0.0001	45.5
Doc	0.4776	0.7381	0.3777	84.5
Dgd	0.0052	0.5058	0.0138	55.7
Dudm	0.3875	0.6938	0.2988	81.1

TABLE 7
COMPARISON OF run time OF FIVE DISTANCE MEASUREMENT METHODS ON THREE DATA SETS.

Data set	Number of records	Gower distance	K-prototypes distance	Ocilimproved distance	Good all distance	Cloud computing algorithm
Acute	120	0.00646	0.00466	0.02266	1.9	0.02059
Heart	270	0.02924	0.03261	0.03406	24.3	0.03342
Credit	653	0.11674	0.06805	0.09831	292.4	0.09739

TABLE 8
COMPARISON RESULTS OF CLUSTERING ACCURACY.

Data set	Acute	Heart	Credit
Clustering accuracy/%	99.8	99.7	99.8