Research and Implementation of Hybrid Clustering Algorithm in Big Data Processing

Yan Hou
Qilu Normal University, Jinan, Shandong, 250014, China
122365871@qq.com

Abstract. With the advent of the era of big data, data analysis and processing has become a hot topic of research, and data mining has become a top priority. Based on the discussion of clustering algorithm, this paper proposes a hybrid clustering algorithm. The algorithm solves the dilemma of relying too much on the initial center and falling into local optimum. The experimental results demonstrate the effectiveness and feasibility of the hybrid clustering algorithm.

Keywords: big data; hybrid; data mining; clustering algorithm.

1. The Introduction

With the development of economy, the amount of data is getting larger and larger. At the same time, the era of big data [1] comes. Due to the growth of data, the original data statistics and query cannot meet the current demand [2]. At present, people are more concerned about how to deal with massive data [3] and explore data sets that are valuable to them, so as to obtain the correlation between data according to the characteristics of data sets [4], so as to provide more scientific reference basis for people's decision-making. Therefore, various algorithms of data mining [5] are constantly produced. People begin to analyze the data in different states. Meanwhile, experts and scholars are constantly improving various algorithms. Clustering algorithm [6] is an important analysis method. At the same time, clustering analysis can obtain useful information from data without prior knowledge. Its goal is to divide data sets [7] into many subsets, while data objects can be described by attribute features. Currently, this technology has been applied more widely in many fields, including pattern recognition, image processing, customer relationship management and other fields, bringing huge economic benefits. There are many common clustering algorithms [8], such as hierarchical clustering algorithm, partition-based clustering algorithm, grid-based clustering algorithm, density-based clustering algorithm and model-based clustering algorithm.

As the saying goes, "birds of a feather flock together. A class, in layman's terms, is a collection of similar elements. Cluster analysis originated from taxonomy. In ancient taxonomy, people mainly relied on experience and professional knowledge to achieve classification, and seldom used mathematical tools for quantitative classification. With the development of human science and technology, more and more high to the requirement of classification, so that sometimes only by experience and professional knowledge is difficult to exactly, so people gradually to math tool in reference to the taxonomy, formed the numerical taxonomy, and then to multivariate analysis technology is introduced into the numerical taxonomy to form the cluster analysis. The content of cluster analysis is very rich, including systematic clustering method, ordered sample clustering method, dynamic clustering method, fuzzy clustering method, graph theory clustering method, clustering forecast method and so on.

2. Traditional Clustering Algorithm

Clustering is widely used. In business, clustering can help market analysts distinguish different consumer groups from the consumer database and summarize the consumption patterns or habits of each type of consumers. As a module in data mining, it can be used as a separate tool to discover some deep information distributed in the database and summarize the characteristics of each type, or to focus on a specific class for further analysis. Moreover, cluster analysis can also be used as a preprocessing step of other analysis algorithms in data mining. Clustering analysis algorithm can be

divided for Partitioning the Methods, Hierarchical Methods, density based the Methods, the grid based the Methods, the Model based Methods.

Partitioning methods, given a data set with N tuples or records, will construct K groups, each of which represents a cluster, K<N. And the K groups satisfy the following conditions:

(1) each group shall contain at least one data record;

(2) each data record belongs to and belongs to only one grouping (note: this requirement can be relaxed in some fuzzy clustering algorithms);

For a given K, the algorithm first gives an initial grouping method, and then changes the grouping through repeated iterative methods, so that each improved grouping scheme is better than the previous one. Most partitioning methods are based on distance. Given the number of partitions to build k, the partitioning method first creates an initialization partition. It then uses an iterative reorientation technique to divide objects by moving them from one group to another. A good general preparation for partitioning is to make objects in the same cluster as close or related to each other as possible, and objects in different clusters as far apart or different as possible. There are many other criteria for determining quality. Traditional partitioning methods can be extended to subspace clustering instead of searching the entire data space. This is useful when there are many attributes and data is sparse. In order to achieve global optimization, the partition - based clustering may need to exhaust all possible divisions. In fact, most of the applications adopt popular heuristic methods, such as k-mean and k-center algorithms, to asymptotically improve the clustering quality and approximate the local optimal solution. These heuristic clustering methods are very suitable for discovering spherical clusters in small and medium-sized databases. In order to discover clusters with complex shapes and cluster super-large data sets, the method based on partition needs to be further extended. The algorithms using this basic idea are: k-means algorithm, k-medoids algorithm, CLARANS algorithm;

Hierarchical method (hierarchical methods), this kind of method for a given data set level decomposition, until some condition is met. The concrete can be divided into "bottom-up" and "top-down". For example, in the "bottom-up" scenario, each record is initially grouped into a separate group, and in subsequent iterations, it combines those adjacent to each other into a group until all records are grouped or a condition is met. Hierarchical clustering can be based on distance or density or connectivity. Some extensions of hierarchical clustering methods also consider subspace clustering. The drawback of a hierarchical method is that once a step (merge or split) is complete, it cannot be undone. This strict rule is useful because you don't have to worry about the number of combinations of different options, which will incur less computational overhead. This technology, however, cannot correct bad decisions. Some methods to improve the quality of hierarchical clustering have been proposed. Representative algorithms are: BIRCH algorithm, CURE algorithm, CHAMELEON algorithm, etc.

A fundamental difference between density-based methods and other methods is that it is based on density rather than a variety of distances. In this way, we can overcome the shortcoming that the distance - based algorithm can only find "round" clustering. The idea is that as long as the density of a point in a region exceeds a threshold, it is added to a similar cluster. Representative algorithms include: DBSCAN algorithm, OPTICS algorithm, DENCLUE algorithm, etc.

The first step to solve the graph theory clustering method is to establish the graph corresponding to the problem, the graph node corresponds to the minimum unit of the data to be analyzed, and the graph edge (or arc) corresponds to the similarity measure between the minimum processing unit data. Therefore, there is a metric expression between each minimum processing unit data, which ensures that the local characteristics of the data are easier to process. The graph theory clustering method takes the local connection feature of sample data as the main information source of the clustering.

The clustering algorithm mainly divides the data object into many subsets, and then gets many subsets, each of which is a cluster. The objects in the cluster are as similar as possible, but may be different from other objects in the cluster. The clustering algorithm describes the objects and determines their relationships based on the discovery in the data set, and then divides the objects of the data into meaningful and even useful clusters, finally grouping the data sets implicitly. When cluster analysis is applied to data sets, the whole data structure can be more clearly understood, and

the distribution law of data objects can be observed, so as to understand the development trend of data. The mathematical description of clustering is as follows:

Let's say that $X = \{x_1, x_2, \cdots, x_i, \cdots x_n\}$, each data object is P dimension. At this point, the input is data set X, and X is divided into k subsets. At this point, through clustering algorithm, the results are represented by C. At this point, $c_i(i = 1,2, \cdots, k), k \leq n$ is required to meet the following conditions:

$$c_i \neq \phi, i = 1,2, \cdots, k \tag{1}$$

$$c_i \cap c_j = \phi, i, j = 1,2, \cdots, k \; and \; i \neq j \tag{2}$$

$$\bigcup_{i=1}^{k} c_i = X \tag{3}$$

At this point, here $c_1, c_2, \cdots, c_i, \cdots c_k$ is called the class. It's also called a cluster.

The clustering process generally has several aspects, as shown in figure 1.

Firstly, prepare data, study its structure and types on the basis of mastering relevant original data sets, analyze the data, and finally determine the principles of data selection, mainly including feature standardization and dimension reduction processing;

Secondly, feature selection and extraction: the data is processed centrally first, the most effective features are extracted and then stored in the vector. At last, these features are transformed and new features are obtained.

Thirdly, because different data sets have different characteristics and application backgrounds, different clustering algorithms need to be designed. However, up to now, there is no particularly good clustering algorithm. Therefore, data feature selection should be carried out according to different requirements in order to obtain a better clustering algorithm.

Finally, the evaluation of the results, when the data set used clustering, this process will involve classes. At this point, if the data set of hidden grouping is clustered multiple times, different results will be obtained, so the category at this time may be meaningless. Therefore, it is usually necessary to re-evaluate the results after the clustering is completed, so as to select effective classes. There are mainly three types of evaluation results, such as correlation test, internal evaluation method and external evaluation method.

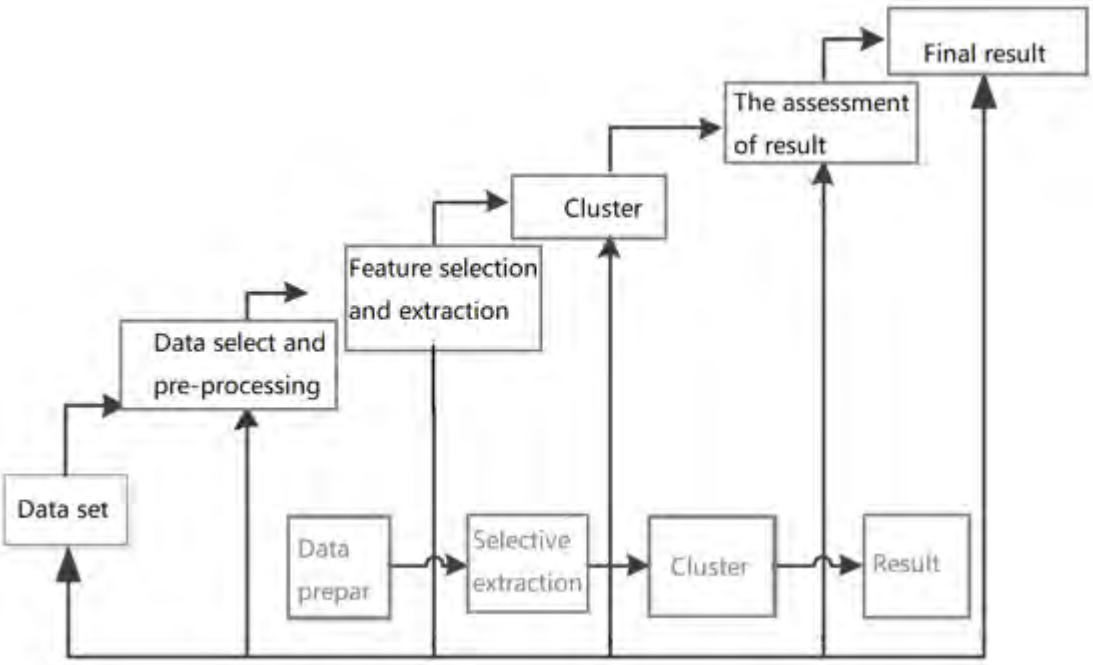

Figure 1. clustering process

However, in most clustering algorithms, data that can only be processed are some inherent numerical attributes or even relatively single classification attributes. However, in various fields of society, the data generated is not a single data type, and usually contains mixed data with multiple numerical attributes. In the process of clustering, only one kind of attribute is considered, and other attributes are ignored, which may lead to the loss of important information. Therefore, it is particularly important to seek mixed clustering algorithm with poor clustering effect.

3. Hybrid Clustering Algorithm

A finite mixture model is a statistical modeling tool that provides an effective mathematical method for simulating complex densities with simple densities. The research on this model can be traced back to a century ago. In 1894, Pearson fitted a set of observation data with the single-variable gaussian mixture model with two mixed components, and estimated the parameter set of the mixed model with the method of moment estimation. In 1977, A.P. Dempster et al. proposed the EM algorithm of maximum likelihood estimation for computing incomplete data, and gave the incomplete data structure of finite mixed model, which solved the difficulty of computing maximum likelihood estimation. Then, the study of finite mixture model entered a new development stage and extended to cluster analysis, speech recognition, neural network and other applications. The core problem of the finite mixture model is:

(1) selection of density of mixed components;

(2) parameter estimation of mixed model.

Gaussian mixture model has been widely used as a finite mixture model due to its simple form and convenient calculation. However, most of the actual data obtained by us are non-linear and non-gaussian, and are limited to the fitting ability of gaussian distribution. As a result, the gaussian mixture model cannot fully, accurately and effectively describe these complex data.

Gaussian model is to use gaussian probability density function (normal distribution curve) to precisely quantize things, and decompose a thing into several models based on gaussian probability density function (normal distribution curve). The principle and process of building gaussian model for image background: the gray histogram of image reflects the occurrence frequency of a certain gray value in the image, or can be considered as the estimation of the gray probability density of image. If the image contains a large difference between the target region and the background region, and there is a certain difference between the background region and the target region in the grayscale, then the gray histogram of the image presents the dual-peak - valley shape, with one peak corresponding to the target and the other one corresponding to the central grayscale of the background. Complex images, especially medical images, are usually multimodal. The problem of image segmentation can be solved by viewing the multi-peak characteristic of histogram as the superposition of multiple gaussian distributions. In the intelligent monitoring system, the detection of moving target is the central content, while in the detection and extraction of moving target, the background target is crucial to the recognition and tracking of the target. And modeling is an important part of background target extraction.

In real life, a large amount of data is generated in many aspects, which is usually mixed attribute data. Therefore, the research on clustering algorithm of mixed attribute data is a hot issue in the field of analysis. Through the k - means combined with algorithm based on the k - modes and formed the k - as algorithm, this algorithm is easy to implement, more can effectively handle the relevant mixed data, but it to rely too much on the initial center, leading to cluster into local optimum easily, and simple using 0 and 1, cannot objectively show thin, result in clustering effect is not ideal. A hybrid algorithm is proposed to solve the existing problems. First, its data set is divided into many kinds, the number of classes is n, making its objective function value small. Select the initial clustering center, any data object, and calculate the average difference degree of the object:

$$d(x_i) = \frac{1}{n}\sum_{j=1}^{n} d(x_i, x_j) \tag{4}$$

Overall differences:

$$G = \frac{1}{n}\sum_{i=1}^{n} d(x_i) \tag{5}$$

Let A class in the clustering process be c_l, and its clustering center is represented by $v_l = \{v_l^r, v_l^c, \}$. At this point, the average value of attribute data is v_l^r, and the value with the greatest number of occurrences is v_l^c, then its measurement formula is:

$$\begin{gathered} d(x_i, c_l) = d(x_i, v_l^r) + d(x_i, c_l) \\ = \sum_{j=1}^{p} \omega \left| x_{ij} - v_{lj}^r \right| + \gamma \sum_{j=p+1}^{m} \sum_s \frac{\delta(x_{ij}, c_{lsj})}{|c_l|} \end{gathered} \tag{6}$$

Among which $\delta(x_{ij}, c_{lsj}) = \begin{cases} 0, x_{ij} = c_{lsj} \\ 1, otherwise \end{cases}$

The steps of the hybrid algorithm are mainly divided into two steps: calculating the distance, which is the distance from the sample to the cluster center, and then putting it into the nearest cluster. Each iteration updates its clustering center, and the next iteration is performed until the convergence of the function is satisfied. However, in the actual situation, the numerical attribute is firstly standardized, and then the information is used to determine the initial cluster center. The specific steps are as follows:

(1) set the number of clustering data sets as K and n, calculate the average difference degree and the overall average difference degree, and sort the average difference degree, as shown in FIG. 2.

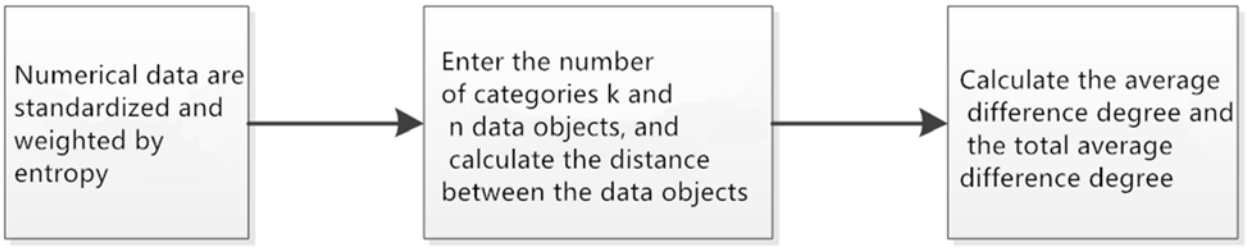

Figure 2. preparation for the basic process

(2) select the largest data as the first initial cluster center, and delete this data at this time;

(3) look for other data with large average difference degree and calculate its distance from the cluster center;

(4) if the overall mean difference degree is greater than the distance of the calculated clustering center, making it a clustering center; otherwise, return (2); Repetition (3) and (4) always make the number of initial clustering centers reach k, and finally output the initial clustering center, as shown in FIG. 3.

(5) calculate and statistic the distance between the data object and each clustering set, and put the data into the concentration closer to it according to the proximity principle;

(6) update the class center, calculate the average value of the numerical attribute, and get the maximum value of occurrence probability in the classification attribute data;

(7) repeat (5) and (6) until stable, and the function value does not change;

(8) output results, and the algorithm flow chart is shown in figure 4.

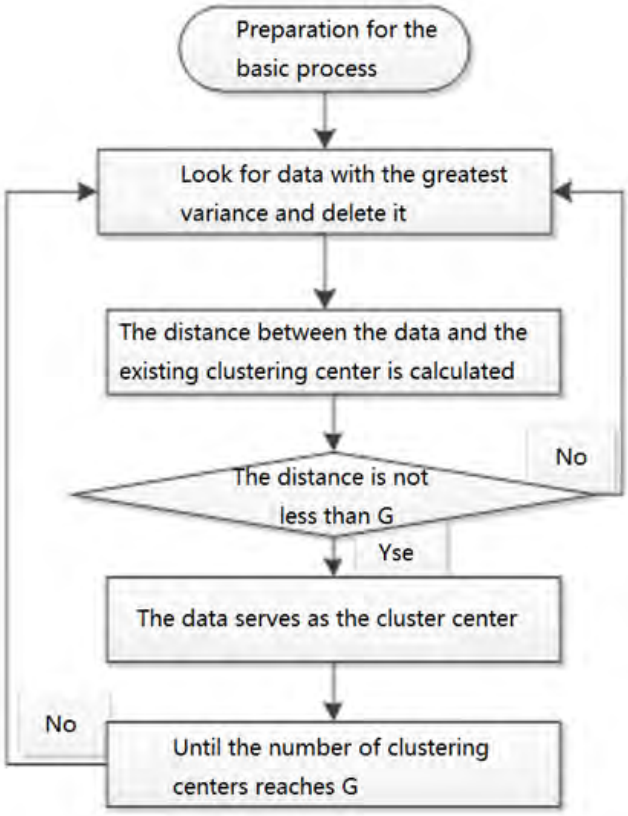

Figure 3. judgment 1

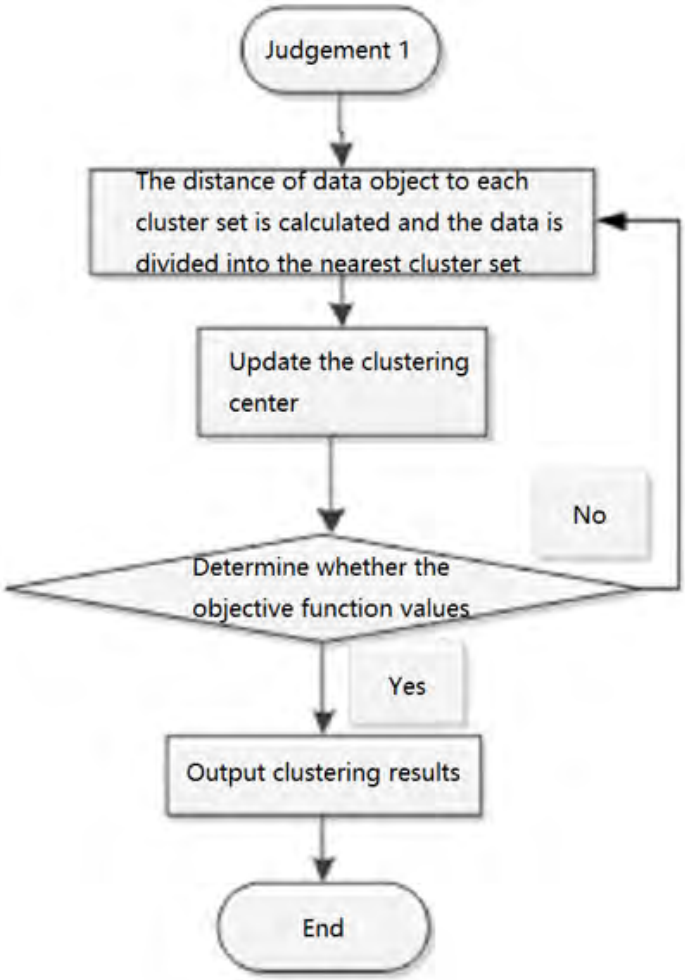

Figure 4. algorithm flow chart

MATLAB to carry out the simulation, its picture is shown in 5. It can be seen from FIG. 5 that as the number of iterations increases, the objective function value will decrease as the number of iterations increases. However, under the same condition, this algorithm is better than k-prototypes algorithm, which proves that the hybrid algorithm has high clustering accuracy.

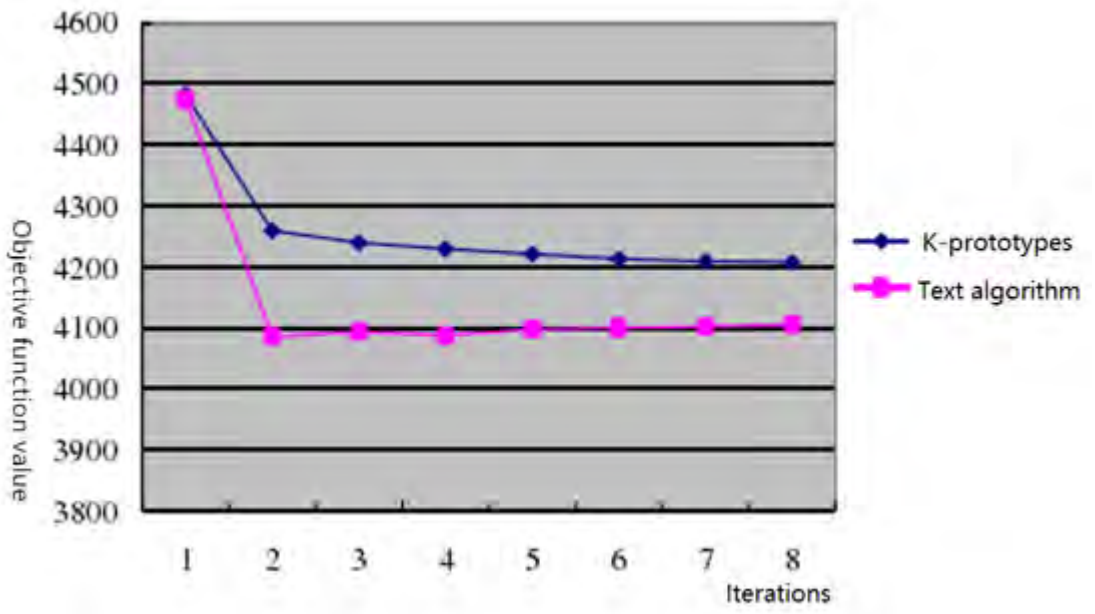

Figure 5. Algorithm Comparison

4. Conclusion

Along with the development of the economy, continuously produce all kinds of data mining algorithm, clustering algorithm is a kind of important analysis method of cluster analysis at the same time can get useful information from without prior knowledge, however, in most of the clustering algorithm, usually can only process data be inherent in certain numerical attributes and even is the classification of the single attribute, this paper proposes a hybrid clustering algorithm. The algorithm solves the dilemma of over-dependence on the initial center and local optimization, and the experimental results also prove the superiority of the algorithm.

References

[1]. Chen Shouwen, Li Mingdong. Implementation of a hybrid mean clustering algorithm [J]. Computer engineering and application,2010,46(18):132-134.

[2]. GREEN R, STAFFELL I, IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, 2014, 61(2): 251-260.

[3]. Tao Zhiyong , Liu Xiaofang, Wang Hezhang. Integration density peak gaussian mixture model clustering algorithm [J/OL]. Computer application: 1-7 [2018-09-29]. HTTP: / / http:// kns. cnki. net/kcms/detail/51.1307.TP.20180725.0820.004.html.

[4]. DIN W I S W, YAHYA S, TAIB M N, et al. MAP: The new clustering algorithm is based on the multitier network topology to engage the lifetime of wireless sensor network[C]//Signal Processing & its Applications (CSPA), IEEE 10th International Colloquium on, IEEE,2014: 173-177.

[5]. Zhu Lizhi, Zhu Wu. An algorithm to achieve the clustering of mixed attribute data flow [J]. Computational technology and automation,2016,35(02):34-37.

[6]. T, JI BAO X, WANG Y, et al. A Fuzzy K - modes - -based Algorithm for Soft Subspace Clustering [C] / / Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on, IEEE, 2011, 2: 1080-1084.

[7]. Chen Xiao and Zhao Jingling. Research and implementation of mixed clustering algorithm in big data processing [J]. Information network security,2015(04):45-49.

[8]. Ga lingxing, Li Zhigang, Zhou Xingshe. Design and implementation of sensor network clustering algorithm for hybrid key management [J]. Computer engineering and application,2007(24):1-3+58.