Research on Clustering Algorithm of Heterogeneous Network Privacy Big Data Set Based on Cloud Computing

Ming-hao Ding(✉)

Department of Computer and Software Technology, Tianjin Electronic Information College, Tianjin 300350, China
dingrui8562@126.com

Abstract. With the rapid development and application of global information technology, big data era has come. China's information security strategy needs to consider the complexity and timeliness of large-scale and heterogeneous network security behavior in big data's time. In order to solve the problem of inaccurate and randomness of single clustering algorithm, a clustering algorithm based on cloud computing for heterogeneous network privacy big data set was proposed. The algorithm utilized the advantages of cloud computing to collect and extract features of big data sets. Then the similarity method was used to carry out the mining process of big data sets, so as to realize the clustering calculation process of big data sets. The algorithm was verified on the UCI dataset. The results showed that the efficiency and accuracy of the cloud computing-based big dataset clustering algorithm were better than the existing ones, indicating that the algorithm design and update strategy were effective.

Keywords: Cloud computing · Heterogeneous network · Big data set · Clustering algorithm · Similarity

1 Introduction

Clustering analysis has been studied for many years and formed a systematic method system [1]. Clustering is an unsupervised machine learning method that takes a group of physical or abstract objects. According to the degree of similarity between them, they are divided into several groups, so that the similarity of data objects in the same group is as large as possible, and the similarity of data objects in different groups is as small as possible [2]. However, the single clustering algorithm has the problems of unstable results and large randomness. Existing research tends to combine the results of clustering large data sets to overcome the shortcomings of clustering.

Research on the clustering of private data sets in heterogeneous networks has appeared in recent years [3], and it has attracted wide attention from all walks of life. However, how to generate the optimal clustering dataset and select the best merging strategy, especially the clustering fusion algorithm for large datasets of classification attributes, is still an unsolved problem. Therefore, it is necessary to conduct research on the generation and mining of cluster members to get the best clustering results.

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 367–376, 2021.
https://doi.org/10.1007/978-3-030-67871-5_33

A cloud computing based heterogeneous network privacy big data set clustering algorithm research is proposed in this paper, and gives the fusion method and strategy of big data set. Firstly, the attributes of each large data set are divided according to the value, and the features are collected and extracted to obtain the initial cluster members. Then, the optimal fusion clustering results are obtained through continuous adjustment and mining. In order to verify the validity of the cloud computing-based heterogeneous network privacy big data set clustering algorithm designed in this paper, the experimental demonstration is carried out. The experimental results show that the cloud computing-based big data set clustering algorithm can improve the data clustering effect and ensure the accuracy of the clustering results, which is extremely effective.

2 Design of Large Data Set Clustering Algorithm

The dispersed large data set matrix is used as an input set of the cloud computing clustering algorithm, and the feature coefficients of each column in the matrix are respectively calculated by the pair of data features. And comparing the matrix characteristics of each big data set within a given threshold, and determining whether the number of points around the point that are greater than the threshold is greater than the data feature. If the characteristic coefficient of a point and any other point is greater than the matrix feature, and the number of features around it and its characteristic coefficient are greater than or equal to the matrix feature, then the point is the core data point, and all the points with the same density as the core point are of one type, and the other type is noise point.

By mining all common feature data in the big data set, and collecting and extracting the characteristics of the common data, once again, using cloud computing technology, mining the clustering features of big data sets to achieve clustering calculation of big data sets, the cloud computing based clustering algorithm flow is shown in Fig. 1.

2.1 Big Data Set Feature Collection and Extraction

Suppose there are n data points $X = \{x_1, x_2, x_3, \cdots x_n\}$, which contain m attributes, the i-th attribute has k_i different values, and the i-th attribute has a weight of ω_i. This paper adopts the simplest and most easy to understand method of generating cluster members [4], that is, by attribute value division, the function expression of the attribute division rule R_i of the i-th big data set is:

$$R_i = \left| C_{i,1}, C_{i,2}, C_{i,3} \cdots C_{i,j} \right| 1 \le i \le m \tag{1}$$

Where, $C_{i,j}$ c represents the j-th data feature of the segmentation result, $\sum_{i=1}^{m} \omega_i = 1$.

Inspired by the literature [5], the data on each attribute is divided into a cluster member, and a unified method is used to divide the different data subsets to obtain the characteristic relationship among the cluster members. Thus, clustering results $R = \{R_1, R_2, R_3, \cdots, R_m\}$ of m cluster members can be obtained, and each cluster member R_i has k_i matrix features.

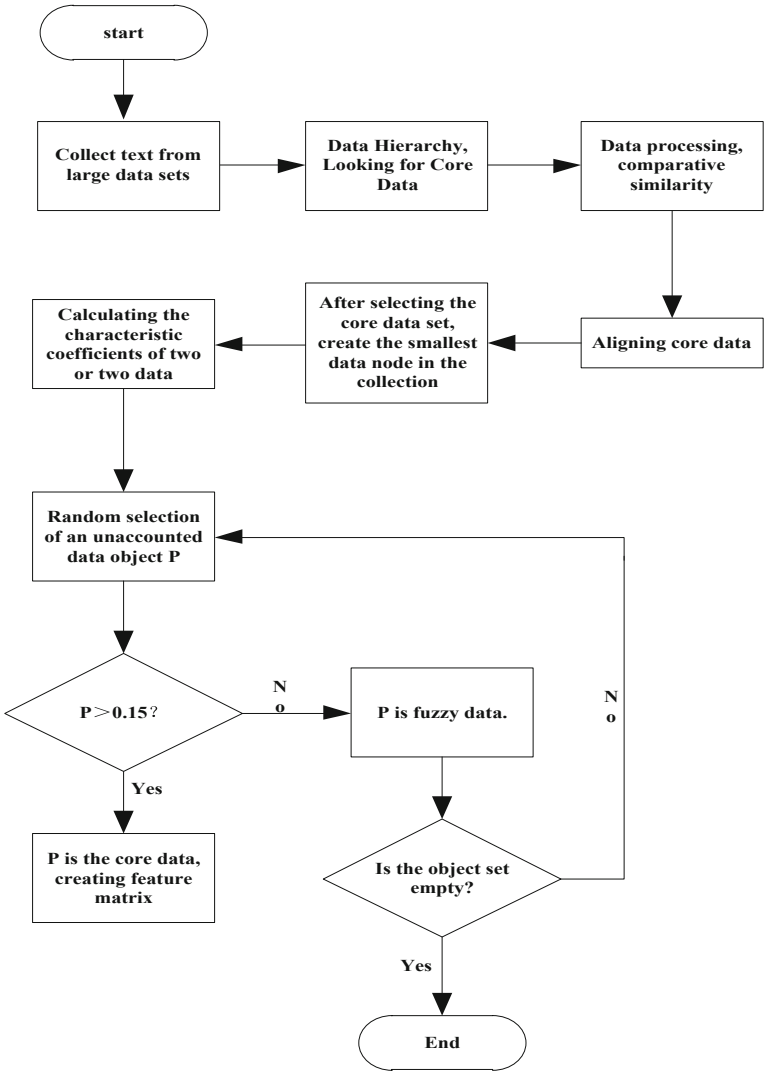


Fig. 1. Cloud computing based clustering algorithm flow chart

The feature collection method for similarity [6] refers to the collection process of metadata in the big data set representing the privacy of heterogeneous networks. By constructing the feature matrix, the clustering combination partition method of multiple large data sets is found, and the similarity between any two data points is used to describe and define the clustering features of large data sets. First of all, we must use cloud computing technology to extract features from the privacy big data sets in heterogeneous networks, and classify them according to their characteristics [7]. Secondly, the content characteristics of the metadata are represented, and the metadata is regarded as a vector space generated by a set of privacy orthogonal terms in a

heterogeneous network. If t_i is treated as a term, $w_i(d)$ is treated as the weight of t_i in the metadata d, and each metadata d can be regarded as a normalized feature vector $V(d) = t_{1i}, w_i(d), t_{2i}, w_i(d), \cdots t_{ni}, w_i(d)$. In general, all data appearing in d is taken as t_i; $w_i(d)$ is generally defined as a function of the frequency $tf_i(d)$ in which t_i appears in d, i.e. $w_i(d) = \vartheta(tf_i(d))$. The frequency function is extracted to obtain the characteristic function of the big data set:

$$\vartheta = \begin{cases} 1, tf_i(R_i d) \geq 1 \\ 0, tf_i(R_i d) = 1 \end{cases} \tag{2}$$

Where, the square root function of ϑ is $\vartheta = \sqrt{tf_i(d)}$; the logarithm function of ϑ is $\vartheta = \log(tf_i(d) + 1)$.

The feature function of the big data set is similarly processed to prepare for the next data mining process.

2.2 Cloud Computing Based Big Data Set Clustering Mining

First, using cloud computing technology, randomly extract metadata features and transform the private data set into structured data that can describe the metadata content. Then use the cluster analysis of the big data set to form a structured metadata tree, and discover the new big data set concept according to the structure, and obtain the corresponding logical relationship. The cloud computing-based big data set mining process is shown in Fig. 2.

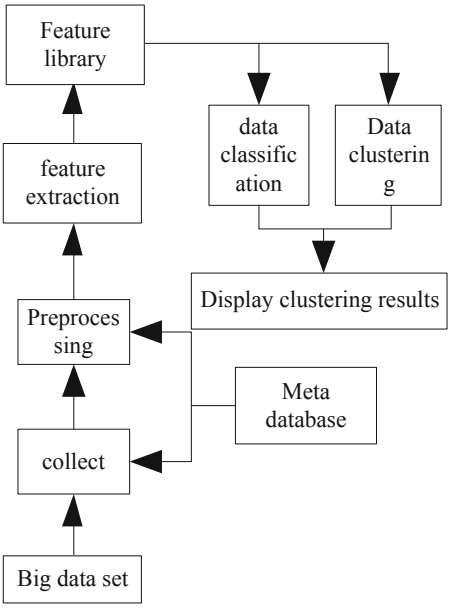


Fig. 2. Big data set mining process diagram

Since the amount of data in a private big data set in a heterogeneous network is very large, the dimension used to represent the metadata feature vector is also very large, and may even reach tens of thousands of dimensions. Therefore, we need to extract the network term with higher weight as the feature item of the metadata to achieve the purpose of dimension reduction of the feature vector. Then, the feature clustering mining process of big data sets is carried out. The big data set cluster mining process is as follows:

(1) Select some of the most representative data features from the original features.
(2) According to the similarity method principle [8], select the most influential feature data set.
(3) Transforming original features into fewer new features by means of mapping or transformation in cloud computing technology [9, 10].
(4) Using the evaluation function method [11], each feature in the feature set is independently evaluated and given an evaluation score, and a predetermined number of best features are selected as feature subsets of the big data set.

Let there be a sample set $X = \{X_1, X_2, X_3, \cdots, X_n\}$ to be classified, and n is the number of elements in the sample, and c represents the number of target clusters. Then there is the following data mining matrix for n elements corresponding to class c:

$$\mu_c = \begin{bmatrix} \mu_1, \mu_2 \cdots \mu_n (n \le \vartheta) \\ \vdots \\ \mu_{c1}, \mu_{c2 \cdots \mu_{cn}} (c \le \vartheta) \end{bmatrix} \tag{3}$$

Where, μ_{cn} represents the mining matrix feature of the n-th element to the c-type data ($1 \le i \le c, 1 \le j \le n$), and meeting $\min J(X, \mu, v) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{cn} d_{cn}^2$, then the problem of clustering this multivariate data set is converted into a simple problem of finding the minimum value of the objective function.

The cloud computing-based big data set clustering mining process is based on the original big data set mining, adding cloud computing technology, adding constraints to the objective function to enforce the cluster search that satisfies the condition, and making the monitoring information constrained clustering search process.

The above assumes an unmarked big data set $X = \{X_1, X_2, X_3, \cdots, X_n\}$, $X_n \in R_n$, divide it into class K, which is $C_1, C_2, C_3, \cdots, C_k$, and the mean of none class is $M_1, M_2, M_3, \cdots, M_k$. Suppose the number of samples in the K-th class is N_K, then $m_K = \frac{1}{N_K} \sum_{i=0}^{k} X_i$, $K = 1 \cdots K$.

According to the European distance and intra-class error squares and criteria, the objective function of cloud-based big data set clustering is $J = \sum_{i=1}^{k} K = 1 \sum_{i=1}^{Nk} |X_i - m_K|^2$. When the algorithm is initialized, the center of each class is randomly selected, so the selection of the initial center determines the quality of the clustering

results. After the introduction of cloud computing technology, a large data set formed by a small number of labeled samples, the large data set contains all K clusters, and each class contains at least one sample to implement a cloud computing-based big data set mining process.

2.3 Implementation of Large Data Set Clustering Algorithm

The cloud computing-based heterogeneous network privacy big data set clustering algorithm is implemented as follows:

The clustering algorithm requires two parameters ε and μ when executed, known as $\mu_c = \begin{bmatrix} \mu_1, \mu_2 \cdots \mu_n (n \le \vartheta) \\ \vdots \\ \mu_{c1}, \mu_{c2 \cdots \mu_{cn}} (c \le \vartheta) \end{bmatrix}$, and ε represents the spatial dimension of the heterogeneous network, up to the dimension [12–14], so no orientation analysis is done.

Search for the number of core data points by checking the ε-domain dimension of the arriving data point in the current time. If the ε field of any data point P contains at least μ data points, create a data matrix with data point P as the core point. Then, by means of breadth search, the data points that can be directly clustered from these core data points are aggregated, and all the obtained density from the data point P is assigned to one class.

If P is the core data point, the cluster data points starting from point P are marked as the current class, and the next step is extended from the center of the matrix. If P is not a core data point, then when the algorithm clusters, the next data point will continue to be processed, in order, until a complete cluster core data point is found. Then select an unprocessed core data point to start expansion, and get the next clustering process, in sequence, until all data points are marked [15, 16].

For data points that are not added to the clustering matrix, they are noise points, and temporarily store them in the invalid area. If the number of data in the invalid area exceeds the maximum range of the preset threshold, the calling algorithm clusters the data in the temporary storage area, and deletes the already clustered data points from the temporary storage area. The dynamic data of the quadratic clustering is recorded as Q, and the clustering calculation process for Q is as follows:

(1) A large data set of mixed attribute features is processed by using different distance calculation methods, and new data point features are calculated by using Eq. (2).
(2) Perform online maintenance on the characteristics of large data sets, and perform mining processing after maintenance.
(3) The clustering algorithm is executed, and if there is data that is not clustered, it is placed in the temporary storage area.
(4) The data feature matrix is again mined and the clustering algorithm is executed until the core data points are found.

The cloud computing technology is used to guide the clustering implementation process of big data sets [17, 18], which solves the problem that the single algorithm clustering quality is not high. First enter the data point $x \in X = \{d_1, d_2, \cdots, d_n\}$ in

memory, d_1 represents the data point in memory. The implementation of the big data set is replaced by a triple, which is equivalent to the center point with the weight to participate in the clustering, the number of data points is the weight, and the clustering result is the mark set, then there is $labels = U^k$. Outputting K sets of disjoint big data matrices $\{X_1\}^{k=1}$ of x, and the objective function is $J \in \sum_{i=1}^{k} J = ik$, then the local optimal clustering process of the big data set is obtained.

So far, the cloud data-based heterogeneous network privacy big data set clustering algorithm design is completed.

3 Simulation Experiment Demonstration and Analysis

In order to ensure the effectiveness of the cloud computing based clustering algorithm for heterogeneous network privacy big data sets, the simulation experiments are carried out.

Set the experimental object to the privacy UCI data set of a heterogeneous network, and perform clustering calculation on it.

In order to ensure the effectiveness of the experiment, the traditional algorithm and cloud based clustering algorithm are compared, and the accuracy of the two algorithms is statistically analyzed. The experimental results are shown in Table 1, Table 2 and Fig. 3.

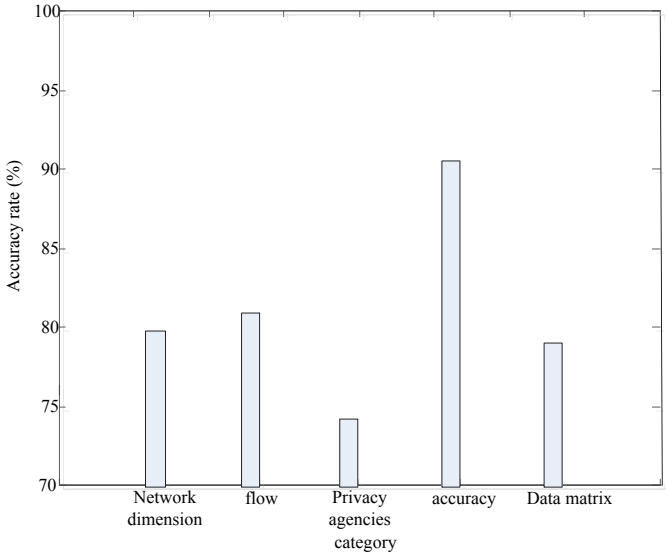
Table 1. Traditional clustering algorithm results

Category	Network dimension	Flow	Privacy agencies	Accuracy	Data matrix
Error rate (%)	0.59	0.85	0.25	0.36	0.48
Cluster velocity measurement (v/ms)	23.6	15.4	53.6	41.2	25.9

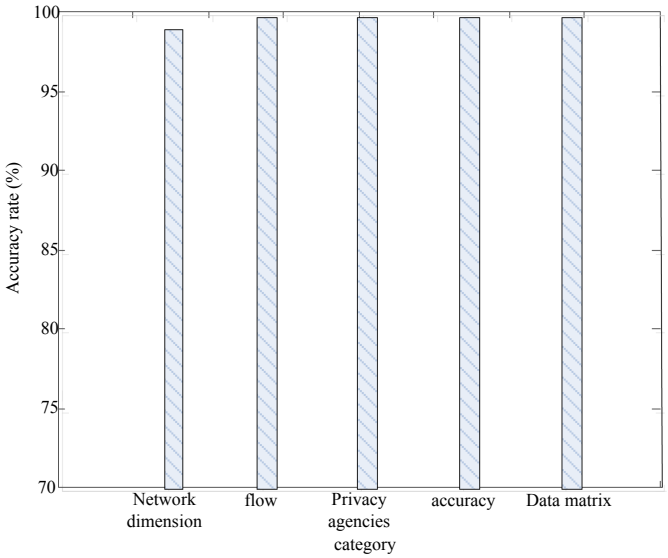
Table 2. Cloud computing based clustering algorithm results

Category	Network dimension	Flow	Privacy agencies	Accuracy	Data matrix
Error rate (%)	0.12	0.05	0.02	0.11	0.03
Cluster velocity measurement (v/ms)	68.3	72.5	95.6	82.6	75.6

According to the data in Tables 1 and 2, the error rate of traditional clustering algorithm is higher than that of this clustering algorithm, and its clustering speed is



(a) Accuracy of traditional clustering methods



(b)The accuracy of clustering method in this paper

Fig. 3. Accuracy Analysis of two clustering algorithms

slower than that of this clustering algorithm. It shows that the heterogeneous network privacy big data clustering algorithm has better effect than the traditional algorithm.

According to Fig. 3, the clustering accuracy of this algorithm can reach 99%, while that of the traditional clustering algorithm is only 91%, which indicates that the clustering accuracy of this algorithm is higher than that of the traditional algorithm.

To sum up, cloud computing based clustering algorithm in the heterogeneous network privacy big data set clustering process, regardless of the clustering speed, or deal with each heterogeneous network privacy structure, traffic and dimension, it is better than the traditional algorithm processing results. It can be seen that the cloud computing based heterogeneous network privacy big data agglomeration algorithm not only improves the clustering accuracy of private data sets in heterogeneous networks, but also improves the stability of the calculation process. The clustering error increases to zero gradually.

4 Conclusion

This paper analyzes and designs the clustering algorithm of heterogeneous network privacy big data sets based on cloud computing, and uses the advantages of cloud computing technology to collect and extract the matrix features of large data sets. Combining the similarity method, mining large data sets and realizing the clustering algorithm design of big data sets. The experimental results show that the cloud computing-based clustering algorithm designed in this paper is extremely efficient. When performing clustering calculation of the privacy big data set in the heterogeneous network, it greatly improves the accuracy of the clustering calculation, and can effectively reduce the clustering error, save the calculation time, and improve the working efficiency of the clustering algorithm. It is hoped that the research in this paper can provide theoretical basis and reference for China's heterogeneous network privacy big data set clustering algorithm.

References

1. Xinchun, Y.C., et al.: Application of multi-relational data clustering algorithm in internet public opinion pre-warning on emergent. Int. Engl. Educ. Res. **1**, 16–19 (2019)
2. Dongmei, C.: Discussions on big data security and privacy protection based on cloud computing. Comput. Knowl. Technol. **15**(15), 101–103 (2018)
3. Ma, R., Angryk, R.A., Riley, P., et al.: Coronal mass ejection data clustering and visualization of decision trees. Astrophys. J. Suppl. Ser. **236**(1), 14–17 (2018)
4. Mingbo, Pan: Research on privacy protection algorithms for network data in large data environment. Microelectron. Comput. **34**(7), 101–104 (2017)
5. Wenzheng, Z., Zaiyun, W., Afang, L.: Relevant analysis of big data security and privacy protection based on cloud computing. Netw. Secur. Technol. Appl. **15**(4), 59–63 (2018)
6. Hodi: Analysis of big data security and privacy protection in cloud computing. Electron. World **25**(16), 98–101 (2017)
7. Yang, H.: Privacy protection of large data security based on cloud computing. Netw. Secur. Technol. Appl. **25**(11), 86–87 (2017)

8. Arman, O., Rahmat, K., Mehrdad, T.H., et al.: Direct probabilistic load flow in radial distribution systems including wind farms: an approach based on data clustering. Energies **11**(2), 310–311 (2018)
9. Jinbo, X., Jun, R., Lei, C., et al.: Enhancing privacy and availability for data clustering in intelligent electrical service of IoT. IEEE Internet Things J. **6**(2), 1530–1540 (2018)
10. Yating, W.: Exploration of big data security privacy and protection based on cloud computing. J. Heihe Univ. **15**(6), 85–87 (2018)
11. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mob. Netw. Appl. **24**(1), 1–4 (2019)
12. Shudong, H., Yazhou, R., Zenglin, X.: Robust multi-view data clustering with multi-view capped-norm K-means. Neurocomputing **311**, 197–208 (2018)
13. Sun, G., Liu, S. (eds.): ADHIP 2017. LNICST, vol. 219. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73317-3
14. Allab, K., Labiod, L., Nadif, M.: A semi-NMF-PCA unified framework for data clustering. IEEE Trans. Knowl. Data Eng. **29**(1), 2–16 (2017)
15. Li, T., Pintado, F.D.L.P., Corchado, J.M., et al.: Multi-source homogeneous data clustering for multi-target detection from cluttered background with misdetection. Appl. Soft Comput. **60**, 436–446 (2017)
16. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
17. Chang, X., Wang, Q., Liu, Y., et al.: Sparse regularization in fuzzy -means for high-dimensional data clustering. Cybern. IEEE Trans. **47**(9), 2616–2627 (2017)
18. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)