Heterogeneous Big Data Intelligent Clustering Algorithm in Complex Attribute Environment

Yue Wang[1] and Jian-li Zhai[2(✉)]

[1] Software College & Nanyang Institute of Technology, Nanyang 473000, China
wangyue66531@163.com
[2] Huali College Guangdong University of Technology, Guangzhou 511325, China
Zhaijianli2033@163.com

Abstract. In order to improve the stability of heterogeneous big data mining operations in complex attribute environment, such as data analysis and cleaning, a heterogeneous big data intelligent clustering algorithm is established. The data cleaning classification method is applied to clean the parameter space in complex attribute environment, and the regular term of sparse subspace clustering is introduced to eliminate the irrelevant and redundant information of heterogeneous big data, and the intelligent clustering index of heterogeneous big data is obtained. By measuring the clustering results, the design of heterogeneous big data intelligent clustering algorithm in complex attribute environment is completed. The experimental results show that the heterogeneous big data intelligent clustering algorithm in complex attribute environment has strong stability in the process of data analysis and cleaning.

Keywords: Complex attribute environment · Heterogeneous big data · Clustering algorithm · Cleaning data

1 Introduction

In recent years, with the increasing utilization of network resources, various industries pay more and more attention to heterogeneous big data mining, especially in complex attribute environment, big data has a lot of characteristic parameters. Has affected the user to the big data utilization degree. For this reason, people need to use the database to carry on the reasonable planning and the effective mining to the heterogeneous big data. Scientific research institutions have proposed some heterogeneous big data mining methods in complex attribute environments, but the mining work in complex attribute environments requires a series of operations such as data analysis, cleaning, conversion, and integration. As a result, the method proposed in the past can not have strong accuracy, stability and practicability at the same time in the mining work [1].

Heterogeneous big data intelligent cluster analysis uses data modeling technology to simulate and analyze the internal structure and distribution of data. From the point of view of data mining, heterogeneous big data intelligent clustering is an unsupervised algorithm. In the absence of prior knowledge, clustering algorithm is used to divide data and

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 357–366, 2021.
https://doi.org/10.1007/978-3-030-67871-5_32

form marker clusters. The research directions of the theory of cluster analysis include the following aspects: First, the ability to process different types of data. Most of the existing algorithms are applied to the analysis of numerical data, but many kinds of data types need to be faced in practical application. Therefore, the limitation of the algorithm in the data processing ability hinders the popularization and application of the algorithm. Second, the ability to identify clusters of arbitrary data shapes. Most of the existing clustering algorithms use standard Euclidean distance to complete similarity measurement tasks, so this algorithm tends to identify spherical clusters. The cluster shape of the actual medium-high dimensional data is mostly non-spherical, so improving the ability of the algorithm to recognize clusters of arbitrary shapes is the key to improving the clustering effect.

In the complex attribute environment of big data mining method, the choice of heterogeneous database is particularly important. Therefore, the RDBMS big data mining method under the complex attribute environment is proposed. By cleaning the parameter space of complex attribute environment and adopting the distributed idea to improve the practicability of the method, the accuracy and stability of mining heterogeneous big data are effectively improved.

2 Design of Heterogeneous Big Data Intelligent Clustering Algorithm Based on Complex Attribute Environment

2.1 Cleaning Parameter Spaces for Complex Attribute Environments

The purpose of cleaning parameter space is to meet the quality requirements of data analysis, so as to fully guarantee the correctness of data analysis [2]. Data cleaning refers to the discovery and correction of corrupt or erroneous records in a recordset, table, or database, and then the replacement, correction or deletion of identified dirty data that is incomplete, incorrect, inaccurate or irrelevant. The process of achieving data consistency. The parameter space cleaning classification methods are as follows:

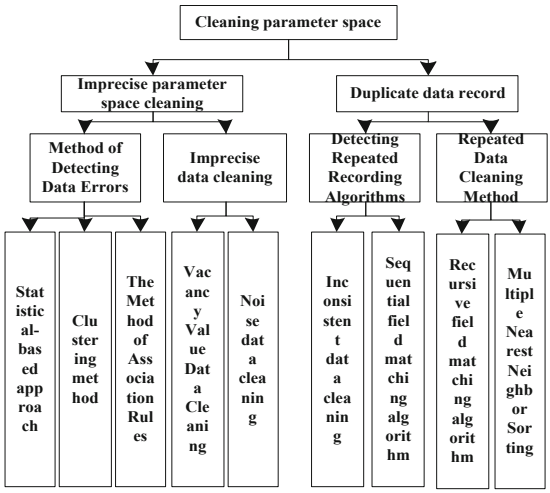


Fig. 1. Classification of parameter space cleaning methods

Figure 1 is a parameter space cleaning method. The first is the cleaning of the imprecise record attributes of the parameter space data sets, and the recognition of the exception attributes in the parameter space data sets [3]. The core idea is to give weight to each attribute first, then count the average value and standard deviation of each attribute field value, and then set a confidence interval for each attribute. Based on whether the attribute value is within the confidence interval to determine whether the attribute is abnormal. The clustering algorithm can judge whether the attribute is abnormal according to the distance between the attribute value and the cluster center, and use pattern recognition knowledge to find the abnormal attribute [4].

The problem of data cleaning is regarded as a statistical inference problem of structured text data in complex attribute environment. It is a classical tool for representation and reasoning of inconsistent knowledge. Before defining Bayesian networks, this paper first gives the corresponding formulas as the theoretical basis. Let Ω be the sample space of experiment E, A is an event of E, Ω is a partition of $p(a) > 0, b_1, b_2, b_3, \ldots\ldots b_n$, $p(b_i) > 0, (i = 1, 2, \ldots, n)$. Then,

$$p(b_i|a) = \frac{p(b_i|a)}{\sum j = 1 p(b_i|a)} \tag{1}$$

$D = T_1, T_2, \ldots, T_n$ represents the input of structured data that contains dirty data. $T_I \in D$ represents one or more tuples with dirty data for the value of the m attribute [5]. Given a candidate replacement set C. Tuple T for possible dirty data in D, it can clean up the database by replacing $T_I \in D$ with a candidate cleanup tuple T with $P_{R(T*|T)}$. Using Bayesian rules in complex attribute environments, it is necessary to take into account in multi-source cleaning that each data source may involve different data fields and different forms of data exist, so the reasons for producing inaccurate data are varied. Inexact data problems in multi-source heterogeneous data environments can be summarized as follows: first, error data: errors in data may be caused by improper data collection or irregular data input, resulting in errors of varying degrees in the data [6]. Second, naming conflict: a naming conflict occurs when the same name is used for a different object or when a different name is used for the same object. Third, data heterogeneity: different representations of the same objects from different sources, such as different component structures, different data types, and different integrity constraints. Fourth, data redundancy: different representations of data from different sources have different version errors. Fifth, in a multi-source heterogeneous environment, even if the same attribute name and data type exist, there may be different value representations or different interpretations across the data source.

2.2 Introducing Regular Terms for Sparse Subspace Clustering in Complex Attribute Environments

When dealing with low-dimensional datasets, traditional clustering algorithms try to find clusters in all dimensions of datasets. But in complex attribute environments, there are usually many independent dimensions. These independent dimensions will hide the existing clusters in the noise data and interfere with the results of the traditional clustering algorithm, while the real correlation data will be distributed on the low-dimensional

structure which can represent the characteristics of the clustering algorithm [7]. In addition, in a very high-dimensional dataset, the distribution of data objects is sparse, and all the data objects are almost equal to each other. This makes a single measure of distance meaningless and can lead to dimensional disaster. Therefore, in this design, the regular term of sparse subspace clustering in complex attribute environment is introduced to eliminate the irrelevant and redundant information in the data set, and clustering is only carried out on the related dimensions. Because the observed data dimension is usually higher than its essential correlation dimension, it is theoretically possible to reduce the dimension of the original space without losing any information. In practice, the dimensionality reduction method is often used to reduce the dimension of high-dimensional data before clustering [8]. There are two commonly used methods of data dimensionality reduction: feature extraction and feature selection.

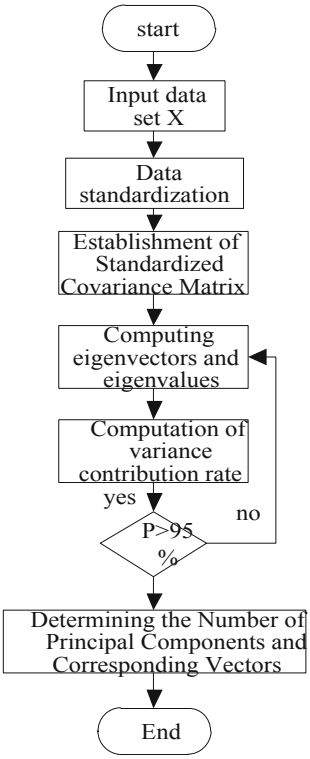


Fig. 2. High-dimensional data dimensionality reduction flowchart

Figure 2 is a concrete operation flow. In the first step, feature extraction is a preprocessing method in projection space, which makes the clustering algorithm only use a small number of newly selected features to cluster. Feature extraction by introducing the regular term of sparse subspace clustering in complex attribute environment, creating linear combination induction data set of data attributes, discovering the potential

structure, generating and selecting new feature vectors, so as to achieve dimension reduction [9]. The second step is to keep the relative distance between the original data objects without deleting any original data attributes when introducing the regular term of sparse subspace clustering in the complex attribute environment, which makes the influence from independent dimensions still exist. Therefore, when there are a large number of independent attributes masking clusters in the dataset, feature extraction will not be able to get the desired effect. In the third step, feature selection is a method to eliminate redundant information by analyzing the entire dataset. It selects the optimal subset from the original data set by searching for various feature subsets and using some criteria to evaluate these subsets. Common search strategies include random search, sampling search and greedy sequential search. Step 4, the evaluation criteria follow two basic models: the wrapper model and the filter model. The fifth step, according to most of the work of supervised learning, finally, select the accuracy measure and classification label to complete the introduction of sparse subspace clustering rules in the complex attribute environment.

2.3 Setting Heterogeneous Big Data Intelligent Clustering Index

Heterogeneous big data clustering validity refers to whether a given fuzzy partition is suitable for all data. The validity index of heterogeneous big data cluster can be used to directly measure the quality of the given clustering results. The good clustering results should be as compact as possible and as far as possible between the clusters [10]. Different literatures put forward different scalar validity measures, but none of them is completely applicable to the evaluation of all clustering results. In this chapter, six validity indicators are selected to evaluate the clustering results.

The first indicator coefficient used to measure the number of "overlaps" between clusters and define it as a formula (2):

$$PC = \frac{1}{N}\sum_{I=1}^{C}\sum_{I=1}^{N} U_{IJ} \tag{2}$$

In formula (2), U_{IJ} indicates the extent to which data point j belongs to category i, C and N indicate the number of clusters and the total number of data samples, respectively.

The second index coefficient, classification entropy (CE): measure the ambiguity of cluster partition and define it as formula (3):

$$CE = \frac{1}{N}\sum_{I=1}^{C}\sum_{I=1}^{N} U_{IJ}\log(U_{IJ}) \tag{3}$$

In formula (3), indicators PC and CE measure whether the clustering results are clear. The higher the value of PC, the more compact the class is, the lower the value of CE, the better the clustering effect [11, 12].

The third indicator coefficient, Partition index (SC): it is the ratio of the sum of the compactness within the cluster and the separation between the clusters. It is the sum of

the individual cluster validity measures normalized by dividing by the fuzzy cardinality of each cluster, which is defined as a formula (4):

$$SC = \sum_{i=1}^{e} \frac{\sum_{i=1}^{n} U_{IJ}}{N_I \sum_{K=1}^{C} |V_K - V_I|} \tag{4}$$

In formula (4), N_I represents sample j of the dataset, V_K and V_I are the i and k cluster centers, respectively. SC can be used to measure the quality of different partitions with the same number of clusters. The lower the value of SC is, the better the clustering results are.

The fourth indicator coefficient, separation index (S): in contrast to partition index SC, the separation index uses the minimum distance separation to achieve the effectiveness of the partition. The smaller the value of S is, the farther the separation between classes is, and the better the clustering results are.

The fifth indicator coefficient, the purpose of XB is to quantify the ratio of total changes within a cluster to the separation of clusters. The size of XB can measure the degree of compactness and separation between clusters. The smaller the corresponding value, the more compact the cluster is and the farther the separation between clusters is, the better the clustering result is.

2.4 Realization of Heterogeneous Big Data Intelligent Clustering Computation

In order to detect and eliminate duplicate records in data sets, it is necessary to solve the problem of how to determine whether the two records are duplicated or not, and to evaluate the similarity of data, that is, the problem of data matching. The simplest attribute set can be obtained by reducing the above attributes. According to the simplest attribute set, the data of the related attributes are extracted, a data table is synthesized, and then the data table is cleaned with similar duplicate data. Therefore, the corresponding records of the records are compared, and the similarity is calculated.

The idea of basic sorting neighborhood law can be summarized in three steps: the first step is to create a sort key: calculate the sort key for each record in the dataset by extracting the relevant field or part of the field. Step 2, sort data: sort the entire data set or part of the data set according to the keys created in the step. Third, data duplication identification: sliding a fixed-size window according to the order of the records, comparing each record with the other records in the window. If the window size window, each new record enters the window compared to the previous record, A record was found to be a "match" for W $\sim$1. In fact, the accuracy of duplicate record detection depends largely on the sort keyword created, which directly affects the matching efficiency and accuracy. If you do not select keywords correctly, you may miss a large number of duplicate records. First, because two duplicate records may be far away from the physical location after sorting, they may never be simultaneously located in the same sliding window and cannot be identified as duplicate records.

Secondly, it is difficult to determine the sliding window size W. If the W is too large, the comparison time will increase, resulting in some comparison unnecessary; if W is too small, some duplicate records cannot be detected. When the size of all the duplicated clusters in the dataset varies greatly, no matter how the size of W is selected, it is not appropriate [13, 14]. In addition, for the whole matching process, the time complexity of the algorithm is O, where n is the total record of the data set, and the flow of heterogeneous big data intelligent clustering algorithm is shown below (Fig. 3).

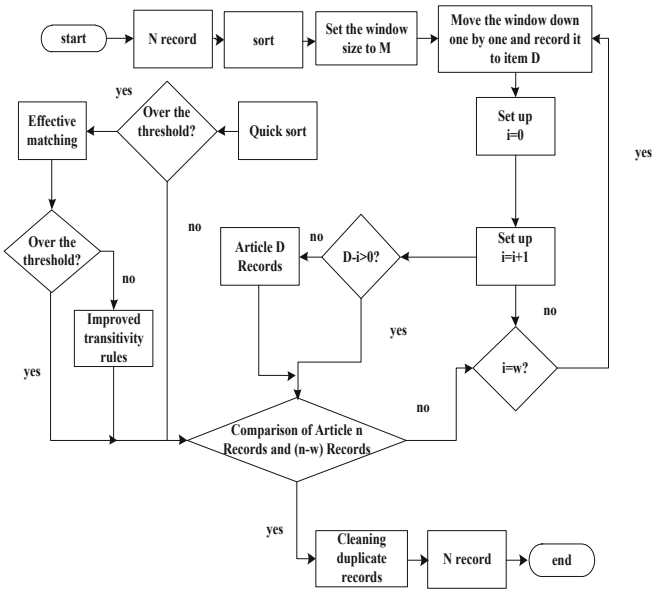


Fig. 3. Heterogeneous big data intelligent clustering proc

Firstly, considering that the window size W is difficult to determine in the SNM algorithm, we analyze the attributes and sort the data sets many times, which makes the repeated records more aggregated, thus entering the same sliding window at the same time. Second, when matching fields, the algorithm assigns a special weight to each attribute, and introduces the concept of effective weight, multiplies the weight by the similarity of the corresponding non-empty attribute, and then combines them to obtain the similarity of the entire record. And it is used to determine whether the two records duplicate the value. Thirdly, in the process of attribute selection, the rationality of the selection is proved by checking the similarity among m specific windows. So far, the design of heterogeneous big data intelligent clustering algorithm in complex attribute environment has been completed.

3 Experimental Conclusion

3.1 Experimental Environment

In order to verify the effectiveness of the algorithm in this paper, simulation experiments are carried out under Matlab 7.0, VS2010 + opencv2.4.13, windows 10, Intel (R) Xeon (R) CPU e5-2603v4 @ 2.20 GHz operating system and 32 GB memory.

Prototype experiment based on hadoop cluster, hive cluster, sqoop cluster and so on. Hadoop is a distributed file system. It can process large-scale data in parallel with hadoop cluster. Hive is mainly responsible for mapping the structured data file into a database table and providing the function of sql query. Then the sql statement is transformed into a MapReduce task and uploaded to the cluster to implement.

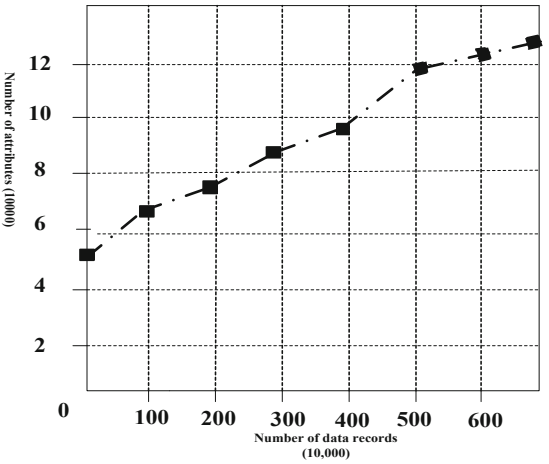
3.2 Data Preparation

Data preparation is the first step of data cleaning. The purpose of this step is to outline the process data and then select the most suitable data sample to model. The main task of this step is to extract data from the database, and check the data set. In order to extract the effective data from the database, determine the operation area, requirement analysis and any changes of the operating conditions, and ensure the efficiency of information extraction in order to extract the effective data from the database through samples and variables.

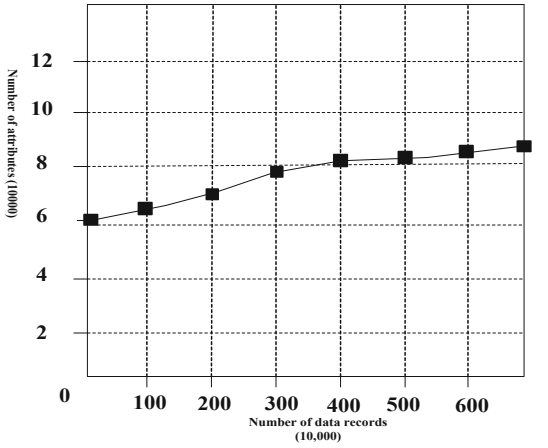
3.3 Experiment and Analysis

Before the contrast test, the data set must be analyzed to know the data characteristics of the data. In this paper, two parameters are selected to describe the data source: the number of attributes in the data record and the noise ratio of the dirty data contained in the data record. Compared with the experimental results, the following is shown:

In Fig. 4, we can see that the number of tuples increases as the number of records entered increases. Figure 4 shows that the proportion of similar duplicates and inconsistencies in the data set is still very high, some have reached more than 25%, the lowest is also more than 10%. It is shown that in the complex attribute environment, not only the data record scale is huge, the number of data attributes also reaches the level of one million, and the data contains the data of different data quality problems of considerable scale, and the number of properties eliminated exceeds the number of verification set. It is shown that the algorithm designed in this paper is relatively stable and the verification experiment is consistent with the real data experiment. It is shown that the experiment design is reasonable and the results are in line with the reality.



(a)Experimental results of traditional algorithms



(b)Experimental results of this algorithm

Fig. 4. Experimental results

4 Conclusion

Due to the large number of heterogeneous big data clustering attributes in traditional heterogeneous big data intelligent clustering algorithms in complex attribute environments, this paper proposes a new heterogeneous big data intelligent clustering algorithm. In a complex attribute environment, by cleaning up the parameter space of the complex attribute environment, introducing sparse subspace clustering rule items, setting heterogeneous big data intelligent clustering indicators, so as to realize heterogeneous big data intelligent clustering calculation. Experimental verification

shows that the intelligent clustering algorithm for heterogeneous big data proposed in this paper has a better effect of heterogeneous big data clustering in a complex attribute environment.

The quality of data analysis in heterogeneous big data intelligent clustering algorithm depends on the quality of data collected from different sources, because data sets in real applications often contain inconsistent data, encrypted data, noise values and data integration caused by a variety of errors. Since the quality of data often fluctuates in the process of data collection, data storage, data fusion and data analysis, data cleaning is not limited to a sub-task of data preprocessing. It runs through every link of data processing. Facing the increasing mass of multi-source heterogeneous data and more complex data structures, in order to improve the efficiency of identifying similar duplicate records and to solve the quality problem of data because of imprecise data, it is necessary to continuously improve in the follow-up research.

References

1. Anonymous: Large data optimal clustering algorithms in cloud computing environment based on PSO. Electron. Des. Eng. **26**(19), 86–89+94 (2018)
2. Qujie: Research on intelligent parallel clustering method for large data in virtual environment. Comput. Measur. Control **25**(6), 257–260 (2017)
3. Yi, M., Ting, X., Shaobin, L.: Research on NoSQL distributed large data mining method in complex attribute environment. Sci. Technol. Eng. **17**(09), 244–248 (2017)
4. Chunhua, H.: Clustering algorithm analysis of multidimensional data de-duplication in large data environment. Comput. Prod. Circ. **32**(11), 151 (2017)
5. Anonymous: Prediction and analysis of energy consumption behavior of integrated energy system users under multi-source heterogeneous large data. Smart Power **46**(10), 92–101 (2018)
6. Li, B.H., Junhua, C., et al.: Distributed clustering algorithms of attribute graph under multi-agent architecture. Comput. Sci. **44**(S1), 407–413 (2017)
7. Linjing, W., Lulu, N., Bin, G., et al.: Sparse fractional feature selection clustering algorithms based on entropy weighting in large data. Comput. Appl. Res. **35**(8), 59–60 + 69 (2018)
8. Houlisa: Clustering algorithm design for eliminating redundant features in large data sets. Mod. Electron. Technol. **41**(14), 56–58+62 (2018)
9. Xiaoyu, C., Xiaojing, L., Haiying, M.: A fast automatic clustering algorithm for large data. Comput. Appl. Res. **34**(9), 2651–2654 (2017)
10. Xiaoyan, T.: A large data text clustering algorithm based on word embedding and density peak strategy. Innov. Appl. Sci. Technol. **6**, 90–90 (2017)
11. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
12. Sun, G., Liu, S. (eds.): ADHIP 2017. LNICST, vol. 219. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73317-3
13. Shuai, L., Weiling, B., Nianyin, Z., et al.: A fast fractal based compression for MRI images. IEEE Access **7**, 62412–62420 (2019)
14. 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)