International Journal of Smart Business and Technology
Vol.9, No.1 (2021), pp.53-62
http://dx.doi.org/10.21742/ijsbt.2021.9.1.05

Research on the Application of Multidimensional Cluster Analysis in Customer Information

Liu Lu

Harbin Institute of Petroleum, Harbin, Heilongjiang 150027, China
77112428@qq.com

Abstract

Cluster analysis is an important technique in data mining. For products to follow customer needs, it is necessary to make a very precise judgment on the distribution of customers in the market and customer groups. When discussing market development trends, it is necessary to classify and aggregate customer groups in the market. Since customer attribute characteristics are often as high as dozens, data mining clustering algorithms are required to process massive customer historical data and classify and aggregate customer groups, so that companies can develop targeted customer products for different types of customers. This article considers the multi-dimensional customer information, combined with the original idea of the K-means algorithm, realizes the classification and aggregation of a large number of customer information through multi-dimensional aggregation, and compares the performance on Hadoop by comparing the data expansion rate and expansion rate. the expansion rate of the K-means algorithm and the K-means under Hadoop the acceleration ratio of means algorithm parallel operation can find that large batches of data (at least tens of millions of data) are more efficient and accurate in a multi-node cluster Hadoop platform. At the same time, the K-means multi-dimensional attribute clustering algorithm is more suitable for the analysis of customer information data with numerous attributes.

Keywords: *Cluster analysis, K-means, Data mining, Hadoop*

1. Introduction

With the continuous development of big data technology, people have collected more and more dimensions of customer behavior data. How to effectively analyze multi-dimensional user behavior data is a problem we need to solve [1]. Customer information is the core and most competitive of an enterprise, and it is often the best to understand customer needs and to develop products for customers, it can take the lead in the market [2]. For products to follow customer needs, it is necessary to make a very precise judgment on the distribution of customers in the market and customer groups. When discussing market development trends, it is necessary to classify and aggregate customer groups in the market. At this time, data mining clustering algorithms are needed to process massive customer historical data. In business intelligence, the clustering analysis algorithm can classify a large amount of customer information in the enterprise. Customers in the same group will have a high degree of similarity in attribute characteristics, which is beneficial to the development of targeted customers for different groups of customers' products. However, because customer attribute characteristics are often as high as dozens, the amount of data is also a difficulty in data clustering. Existing clustering

Article history:
Received (December 12, 2020), Review Result (January 24, 2021), Accepted (March 3, 2021)

Print ISSN: 2288-8969, eISSN: 2207-516X IJSBT

algorithms have encountered bottlenecks in accuracy and time and space complexity when processing multi-dimensional and large-scale data [3]. The research idea for this problem is to combine parallel processing technology and multi-dimensional cluster analysis with the existing more commonly used cluster analysis algorithm K-means, explore more efficient cluster analysis methods, and improve the accuracy of customer information analysis.

Clustering analysis is a relatively basic data processing method in the field of data mining. Data classification through clustering algorithms can divide a data set into several clusters with similar objects within a class but different objects between classes, to find in the data set Potential data patterns and internal connections [4]. Therefore, many experts and scholars at home and abroad have studied various clustering algorithms. K-means algorithm is a classic partition clustering algorithm. Its advantages are simple, easy to implement, and can be used for parallel clustering of large-scale data sets. Generally, when clustering large data sets, the K-means algorithm is much faster than the hierarchical clustering algorithm. Its disadvantage is that the number of clusters k needs to be determined in advance. Because many applications cannot be determined in advance, such as the division of network communities. The k initial clustering centers are randomly selected. Since k initial clustering centers are randomly selected, the algorithm is sensitive to abnormal data.

This paper uses the multi-dimensional characteristics of data, combined with the K-means algorithm and Hadoop's MapReduce algorithm to improve the accuracy of customer analysis.

2. Preprocessing of customer data

2.1. Customer data preprocessing

Customer data preprocessing, this article uses customer information in the real estate industry as an example. The data source is the ERP system and mobile APP. It is subdivided into clue registration, visit and follow-up information, transaction phase-related information, registration information, APP membership promotion, and recommendation information, etc. At this stage, the relevant customer information of all systems needs to be extracted into the corresponding target table, which is mainly divided into customer identity information and customer events. The implementation process of the real estate customer information structure is shown in [Figure 1] below, and the designed customer data structure is shown in [Table 1].

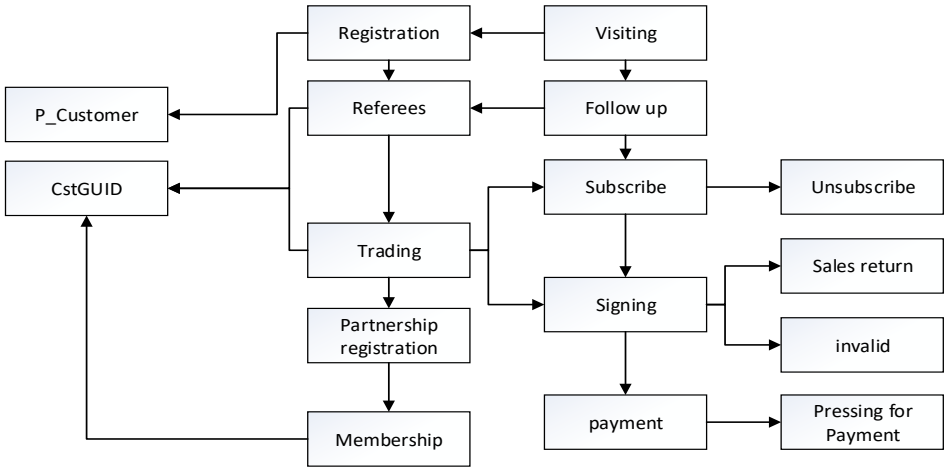

Figure 1. The information structure of real estate customers

Table 1. Structure of customer information

Customer Information	
Basic Information	Decision information
CSTGUID	Marriage
CSTName	BuyersUse
CardType	WorkArea
CardID	HomeArea
Gender	Family
Age	Earning
MobileTel	Work
Email	Hobby
Address	-
WorkAddr	-
Province	-
City	-
EduLevel	

2.2. Customer behavior similarity measurement

The similarity measure is a measure that comprehensively evaluates the similarity between two things. The closer two things are, the greater their similarity measure, and the more distant the two things, the smaller their similarity measure. Assuming that each feature has a different degree of importance, it can be represented by a weight vector ω, which is used to calculate x, y, and the correlation between two user behaviors [5]. The weighted correlation degree is used to calculate the behavior characteristics between two users. The calculation formula for weighted correlation is:

$$m(x;w) = \frac{\sum_i w_i x_i}{w_i}$$

$$cov(x,y,w) = \frac{\sum_i w_i (x_i - m(x;w))(y_i - m(y;w))}{\sum_i w_i}$$

$$cov(x,y;w) = \frac{\sum_i w_i (x_i - m(x;w))(y_i - m(y;w))}{\sqrt{cov(x,x;w)cov(y,y;w)}}$$

3. Design of K-means, a multi-dimensional clustering algorithm based on Hadoop

Given the multidimensional characteristics of customer information attributes and the design of MapReduce algorithm combined with Hadoop, it is multi-dimensionally extended based on the conventional K-means algorithm. The conventional serial K-means algorithm includes the following steps [6].

(1) Select k customer objects from the data objects as the initial cluster centers.

(2) Set the preliminary critical value of the minimum distance, and perform preliminary classification and division by calculating the distance between each data object and the cluster center.

(3) According to the clustering center after division, recalculate the mean value of each cluster. This means value can be changed according to the second step again.

(4) Calculate whether the function convergence is satisfied after each division. If it is satisfied, the algorithm terminates its operation. If the condition is not satisfied, then continue to steps 2 and 3.

It can be seen from the above algorithm steps that the main calculation work of the K-means algorithm is to calculate the distance between each data object and the cluster center according to the set minimum distance, and it has always been possible to divide the data objects into different clusters. Each iteration is performed to initialize the cluster centers based on the previous division so that the data objects can be divided in a more detailed step after each iteration.

There are two common division methods, one is k-means; the other is a k-center point, the latter is more robust than the former, but its complexity is relatively higher, especially for large batches of data. Therefore, from the situation of a large amount of data and many customers information attributes, multi-dimensional K-means clustering is adopted, and each attribute cluster is weighted, and the weight is assigned according to the importance of customer attributes. The iteration of the k-means algorithm is implemented through the Map function and the Reduce function [7].

The most common application of parallel processing technology now is the Hadoop architecture system. It is a cloud computing platform with relatively low cost, less development difficulty, and good performance in the parallel processing of large-scale data. It is characterized by high reliability, relatively low cost, High efficiency [12]. The two core parts of the Hadoop platform framework: HDFS (distributed file system) that provides storage for a large amount of data; MapReduce that provides a computing model for big data. The biggest feature of the Hadoop platform is clustering, which is reflected in its HDFS cluster. There will be one master node (Namenode) in the cluster as the cluster management center, and multiple slave nodes as data nodes. Each node can be an ordinary PC. MapReduce is a programming model for parallel computing of large batches of data. The main idea has two parts, Map (mapping) and Reduce (reduce). The main functions of the Map terminal are as follows.

(1) When the data is Input, the size of the data fragments will be arranged according to the number of sub-nodes. Each data fragment corresponds to a map, and the output result of the Map is temporarily placed in the memory buffer. These data will generate new (key, value) key-value pairs based on the custom Map function. Different types of key-value pairs are also different.

(2) Shuffle is before the Reduce side to ensure that the input is processed and sorted by the Map.

(3) Reduce side: Recursive reduction will be performed on the key-value pairs sent from the Map side. The input parameters are (key, {list value}). After processing through a custom Reduce function, a new (key, value) Key-value pairs.

The default key-value pair (Key, value) of the Map function. To facilitate calculation, customer information data can be led into text form according to attributes. The key here is the displacement of the current text data relative to the starting point, and the value is the corresponding displacement string. After the text is traversed, the distance between the object and each center point is calculated through the value to find the center cluster class with the shortest distance. The designed Map function is as follows.

```
Map((key, value),(key', value'))
{
   Analyze the value at the beginning to get the initial value firstvalue;
   The shortest distance from the center cluster is defined as minvalue, which is the maximum value
   during initialization;
   Dex variable as key;
   K is defined as the number of initial cluster centers;
   For m=0 to k-1
   Do{
      Dis=firstvalue; Define the distance between each object and the m-th cluster center;
      If dis<minvalue
      {
         minValue=dis;
         index=i;
      }
   }
   Key'=index; assign index to key' after each execution of the map function;
   Value'=dis; use dis as the value of value';
   Output (key', value')
}
```

The input source of the Reduce function is the classification merge after the Map, that is (key, V); here the key is the subscript of the merged cluster, and V is the object value of the same cluster, that is, the value' obtained by the Map function; The value of each object of the class is added and divided by the number of objects in the same cluster, which is the value of the new cluster center. The pseudo-code is as follows.

```
Reduce((key,V), (key',value'))
{
   SUM[]; Initialize the array as the cumulative value of each cluster object coordinate.
   NUM=0; initialize variable NUM as the number of objects in the same cluster;
   While(V.hasNext()) //hasNext() is used to judge whether there is the next same cluster object;
   {
      V.next(num); Analyze the displacement and the number of objects in the same cluster from the
      next() function;
      NUM+=num;
   }
   Each value of the array SUM[] is divided by NUM to obtain the new coordinate value of each
   cluster center;
   That is, the key becomes key';
   Value' is the coordinate value corresponding to each object;
   Return (key', value')
}
Repeat the Map function and Reduce until the convergence condition is reached.
```

4. Processing of customer information under Hadoop environment

4.1. Introduction to Hadoop environment, data sources, and index evaluation

This paper explores the use of K-means to achieve cluster analysis of real estate customer information. Based on the amount of data and the subject of inquiry, the deployed Hadoop environment is based on five PCs, one of which is a server virtual machine with 32G memory. The other four are PCs and notebooks, equipped with dual-core 8G memory for the PC and

12G for the notebook. Hadoop is version V2.7.0. The machine is connected and intercommunication through Gigabit Ethernet and LAN established by a switch.

The data comes from a real estate customer, whose needs are based on the existing customer information, visitor registration information, customer purchase information, etc., to dig out the potential customer needs of customers, and adjust the market distribution by analyzing the relationship between different attributes of customers.

Since the attributes of customer information are multi-dimensional, we mainly study some decision-making attributes here. Including the following attributes: gender, age, province, city, industry, education level, marital status, purchase purpose, work area, living area, income level, family status, occupation, hobbies, demand area, intention floor, intention Unit price, source of leads (media advertisement, etc.), etc.

Index performance is often based on changes in data volume and platform performance. Therefore, in the experimental environment, the performance of the processing mechanism is discussed by controlling the changes in data volume and the platform. The expansion rate, acceleration ratio, and data expansion rate are used as evaluation indicators, and the potential Customer information association is also used as an evaluation condition [8].

4.2. Analysis of clustering results

4.2.1. K-means algorithm performance analysis

From the perspective of data magnitude, the running time ratio of tens of millions of data is more efficient than that of millions of data at the same number of nodes [9]. For Hadoop, changes in running time and accuracy caused by changes in the number of nodes can better reflect the advantages of clustered parallel computing [10]. [Figure 2] shows the acceleration ratio of the K-means algorithm in parallel operation on the Hadoop platform. It can be seen from the figure that the acceleration ratio gradually increases with the increase of nodes, and Hadoop parallel operation improves the efficiency of K-means clustering analysis. However, it can also be seen from the figure that the increase in acceleration ratio is the largest when going from 2 nodes to 3 nodes. Another reason for the increase in acceleration ratio is that as the number of nodes increases, the communication overhead between nodes is also gradually increasing. Therefore, when deploying a Hadoop cluster environment, the communication methods and equipment between nodes also need to be considered. At the same time, it can be seen in the figure that under the same Hadoop environment, the acceleration ratio of millions of data is lower than that of tens of millions of data.

Table 2. Algorithm running time

Number of nodes	Data volume	Running time	Accuracy%
2	Million	10min 32s	55
2	Tens of millions	17min 45s	67
3	Million	7min 12s	64
3	Tens of millions	15min 34s	73
4	Million	5min 4s	68
4	Tens of millions	9min 26s	78
5	Million	2min 13s	69
5	Tens of millions	3min 25s	84

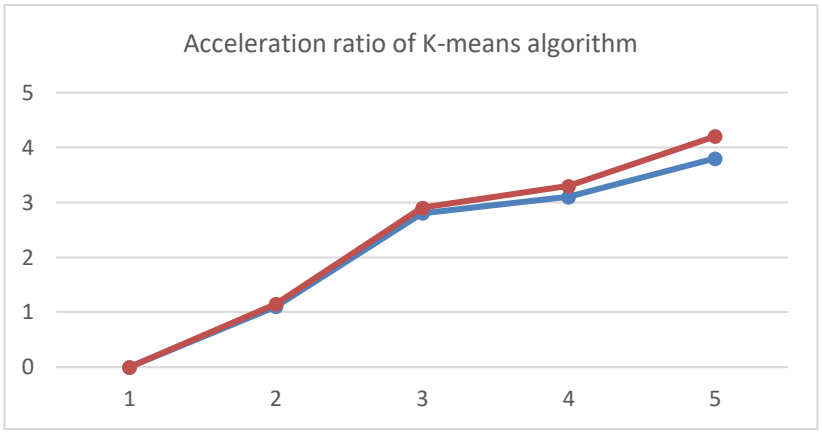

Figure 2. The acceleration ratio of the K-means algorithm on the Hadoop platform

4.2.2. Analysis of customer information mining

First, analyze the three more important attributes of some customer data: gender, income level, and willingness to buy a house. Among them, the income level has three levels: average, medium, and high; the willingness to buy a house has three levels: low, average, and strong. It can be seen that it can be roughly divided into three categories, one is the male population with a strong desire for high income. One is the middle-income female group with an average willingness to buy a house. There is also a group of men with lower incomes and lower willingness to buy houses.

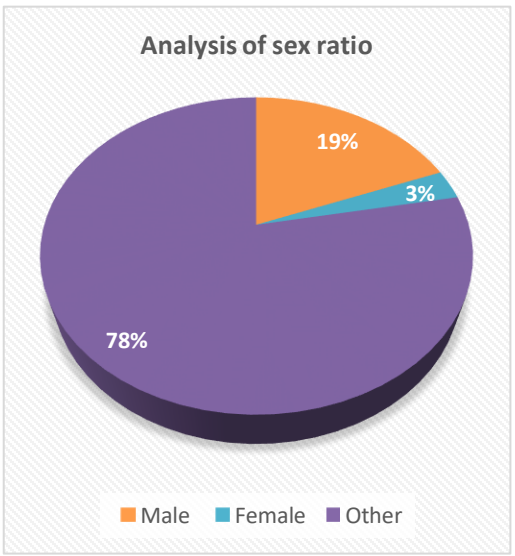

Figure 4. Analysis of sex ratio

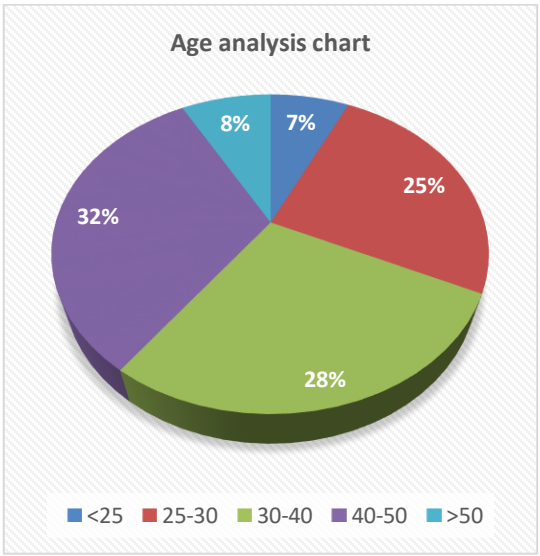

Figure 5. Age analysis chart

Statistics based on the analysis results of customer information. Since the main business of the real estate company is concentrated in Jiangsu, South Jiangsu, and Shanghai, the statistical customer information is mainly concentrated in these places. From [Figure 4] to [Figure 10], we can see the proportions of the main attributes of customers. Combined with [Table 3], we can see that the types of real estate customer groups are greatly affected by age and region. The

customer group is also composed of males between 25 to 40 mostly, and most of the demand is for marriage use. At the same time, income level is also an important attribute that affects the willingness to buy a house. Among them, middle-income people are more willing to buy houses in second-and third-tier cities such as Nanjing, Wuxi, and Suzhou. Therefore, for the real estate market, the construction of residential areas can be increased, and the promotion crowd is mainly 25 to 40 years old.

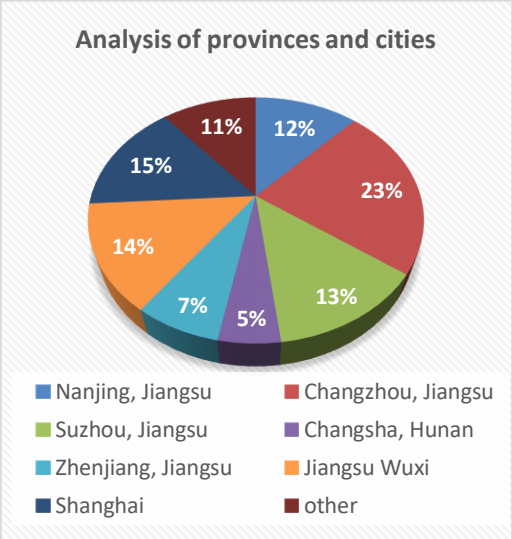

Figure 6. Analysis of provinces and cities

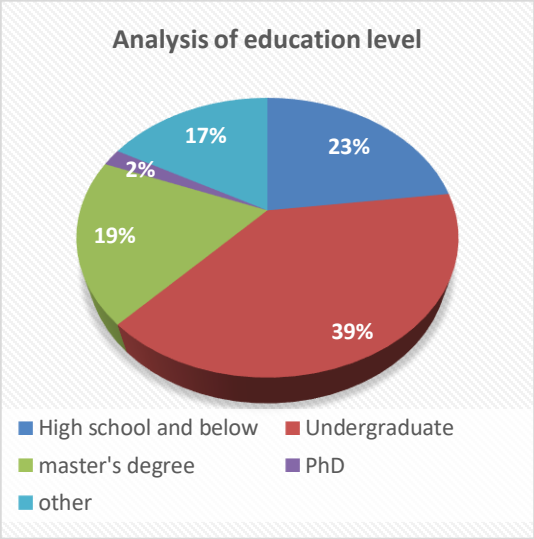

Figure 7. Analysis of education level

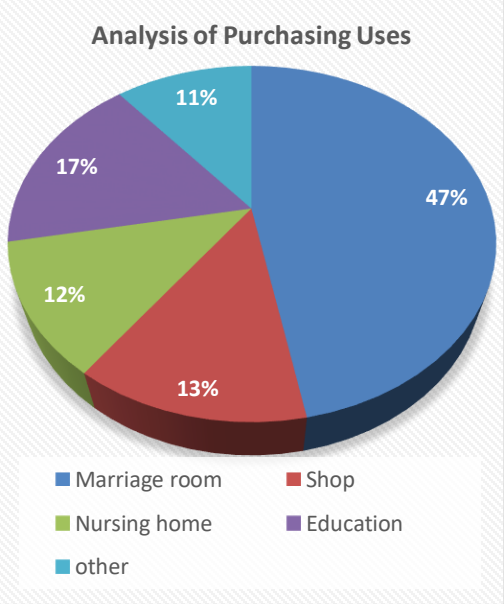

Figure 8. Analysis of purchasing uses

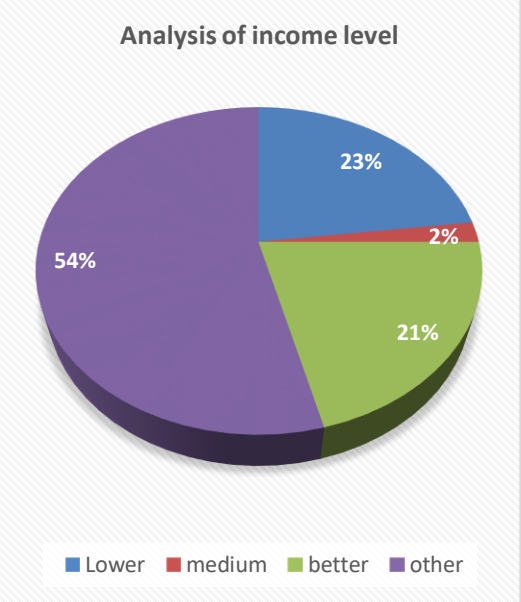

Figure 9. Analysis of income level

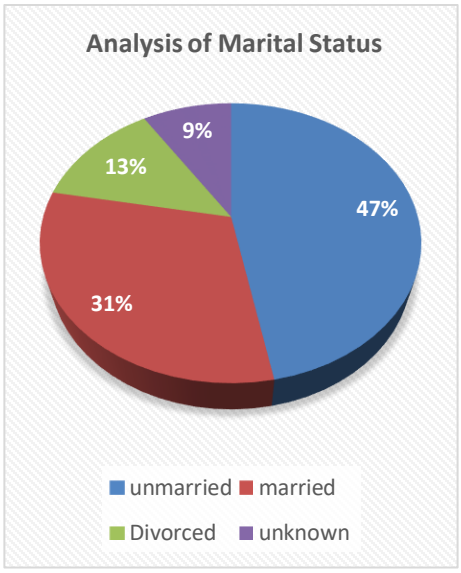

Figure 10. Analysis of marital status

Table 3. Table of cluster analysis results

Clustering Group	Gender	Age	Province City	Education Level	Education Level	Purchasing purpose	Income level	proportion%
1	Male	<25	Nanjing, Jiangsu	Undergraduate	unmarried	Marriage room	general	3.1
2	Male	25-30	Nanjing, Jiangsu	Under-graduate	unmarried	Marriage room	medium	5.3
3	female	30-40	Nanjing, Jiangsu	Under-graduate	unmarried	Marriage room	medium	4.4
4	Male	40-50	Nanjing, Jiangsu	Under-graduate	married	Nursing home	better	4.9
5	Male	>50	Nanjing, Jiangsu	Under-graduate	married	Nursing home	better	1.2
6	Male	<25	Suzhou, Jiangsu	Undergraduate	unmarried	Marriage room	medium	4.2
7	female	25-30	Suzhou, Jiangsu	Undergraduate	unmarried	Marriage room	medium	6
8	Male	30-40	Suzhou, Jiangsu	Undergraduate	married	Nursing home	better	9
9	Male	40-50	Suzhou, Jiangsu	Undergraduate	married	Nursing home	better	7.1
10	Male	>50	Suzhou, Jiangsu	Undergraduate	married	Nursing home	better	1.9
11	Male	<25	Shanghai	Undergraduate	unmarried	Marriage room	medium	1.2
12	female	25-30	Shanghai	Undergraduate	unmarried	Marriage room	medium	3.2
13	Male	30-40	Shanghai	Undergraduate	married	Nursing home	medium	2.6
14	Male	40-50	Shanghai	Undergraduate	married	Nursing home	better	2.3
15	Male	>50	Shanghai	Undergraduate	married	Nursing home	better	1.5
...	...	...	...	...	...	...	...	...

4. Conclusion

Through the research on the Hadoop platform and the K-means clustering algorithm, the cluster analysis of real estate customer information using K-means on the Hadoop platform is realized. By comparing the running time, the expansion rate of the K-means algorithm, and the K-means under Hadoop the acceleration ratio of means algorithm parallel operation can find that large batches of data (at least tens of millions of data) are more efficient and accurate in a multi-node cluster Hadoop platform. At the same time, the K-means multi-dimensional attribute clustering algorithm is more suitable for the analysis of customer information data with numerous attributes.

References

[1] X. Huang and Q. Wu, "Micro-blog commercial word extraction based on improved TF-IDF algorithm," TENCON 2013-2013 IEEE Region 10 Conference (31194), Xi'an, China: IEEE, pp.1-5, **(2013)**

[2] E. Nikolova and V. Jecheva, "Some similarity coefficients and application of data mining techniques to the anomaly-based IDS," Telecommunication Systems, vol.50, no.2, pp.127-135, **(2012)**

[3] N. Xinzheng and S. Kun, "Research on fast parallel clustering and partitioning algorithm for large-scale data," Computer Science, vol.39, no.1, pp.134-137, **(2012)**, DOI: 10.3969/j.issn.1002- 137X.2012.01.030.

[4] K. L. Du and M. N. S. Swamy, "Neural networks and statistical learning," Science Business Media, pp.727-745, **(2014)**

[5] L. Zhang and Q. Guo, "Research on user clustering method based on multi-dimensional behavior analysis," Journal of University of Electronic Science and Technology of China, vol.2, pp.315-320, **(2020)**

[6] J. Liu and H. Guo, "K-means clustering center optimization solution method in cloud computing," Bulletin of Science and Technology, vol.31, no.10, pp.100-102, **(2015)**

[7] X. Jiang, C. Li, W. Xiang, "MapReduce parallelization implementation of K-means clustering algorithm," Journal of Huazhong University of Science and Technology (Natural Science Edition), vol.39, no.1, pp.120-124, **(2011)**

[8] S. Xiang, F. Nie, and C. S. Zhang, "Learning a Mahalanobis distance metric for data clustering a classification," Pattern Recognition, **(2008)**

[9] L. Zeng, "Research on MapReduce parallel computing platform in cloud computing," Harbin Institute of Technology, **(2013)**

[10] B. Yang, D. -Y, Liu, and D. Jin, "Complex network clustering algorithms," Journal of Software, vol.20, no.1, pp.54-66, **(2009)**