Research and application of improved K-means based on MapReduce

Hongqin Wang¹, Hongxia Wang², Li Jiang¹ and Zhengjun Pan¹

¹Department of Software Engineering, South China Institute of Software Engineering, Guangzhou, Guangdong Province, 510990, China
²Department of Automobile Engineering, Henan Institute of Technology, Xinxiang, Henan Province, 450003, China
Hongqin Wang’s E-mail: whq@sise.com.cn

Abstract: With the development of big data, the traditional data mining clustering algorithm K-Means is inefficient and has poor scalability in dealing with massive data. MapReduce on the Hadoop platform was used to realize the parallel processing of the K-Means algorithm, the performance of the algorithm was tested by experiments. The results show that the improved K-Means algorithm has good parallel expansion capability, high efficiency, and great potential when processing big data mining. The algorithm is applied to the big data processing of customer consumption in a restaurant chain, and the effectiveness of the algorithm is verified, which can better serve the decision of restaurant.

1. Introduction
With the development of mobile Internet, the daily data volume of the catering industry is growing exponentially. For massive unstructured data[1], it is inefficient to use traditional data mining clustering algorithm K-means to mine the big data. An improved k-means algorithm with big data advantage is proposed, which provides powerful storage capacity and computing power for big data by using Hadoop platform. Aiming at the limitation of traditional clustering algorithm k-means, Hadoop cluster is built, and MapReduce is used to improve the algorithm. Finally, the parallel algorithm is used to analyze and process the consumption big data of a catering chain store. The improved K-means algorithm is fast and can efficiently process massive data. And it can dig out the high-value customers among the consumers and provide better decision-making for the restaurants.

2. The introduction of clustering algorithm K-means
Clustering refers to the process of separating data items with similar characteristics[3]. In this way, all member objects in the same subset have similar attributes. Different from classification methods, clustering methods emphasize more on automaticity, and gather data into clusters automatically on data sets without category marking. The information in this process is mainly based on the characteristics of data and the interconnection between data.

The K-Means algorithm[4] is described as follows: Firstly, K objects are randomly selected from the database, serving as the center of mass of the initial K clusters. Then, for the remaining data objects in the database, which are calculated their distance from each centroid, and the object is assigned to the cluster with the minimum distance value according to the distance value. Finally, we calculate the centroid of all clusters again. The new centroid is reassigned according to the distance
value.

The error square and criterion function are used to evaluate clustering performance for K-means clustering algorithm. Given the data set $X$, only the description attribute is contained, not the category attribute. Suppose $X$ contains $K$ clustering subsets $X_1, X_2, \ldots, X_K$. The samples in each cluster subset were respectively $n_1, n_2, \ldots, n_K$; The mean representative points (also called cluster centers) of each cluster subset are $m_1, m_2, \ldots, m_K$. $E$ is the sum of squared error of all objects in the database, $p$ is the data object, and $m_i$ is the average value of cluster $X_i$. $E$ can maximizes the independence and compactness of the results. The equation is shown in (1).

$$E = \sum_{i=1}^{k} \sum_{p \in X_i} \| p - m_i \|^2$$

(1)

As the original K-Means algorithm has some problems, such as artificial input of $K$ value, difficulty in determining the initial position of $K$ centers, and isolation point will make the clustering result deviate from the center. We will optimize the K-Means algorithm and realize parallelization based on Hadoop.

3. Algorithm improvement based on Hadoop platform

3.1. Introduction to Hadoop Platform

Data mining is combined with the Hadoop platform[5], adopting the thinking of parallel, data mining platform of Hadoop cluster is built, which integrates data storage, analysis and processing. The Hadoop core module is mainly composed of HDFS and MapReduce. HDFS implements data storage on multiple machines, and MapReduce implements data calculation on multiple machines.

MapReduce is a distributed computing framework that handles large-scale data sets in parallel[6], usually over 1TB of storage. Its main idea is functional programming. Map and Reduce functions are used to complete complex parallel computation. MapReduce distributes large-scale operations on datasets to each node in the cluster for parallel processing of large datasets; each node in the cluster periodically reports completed work and status updates to JobTracker. The working principle is shown in Figure 1.

![MapReduce Working Principle](image)

Figure 1. MapReduce Working Principle

3.2. MapReduce improvement of K-means algorithm

In Hadoop cluster environment[7-8], K-means algorithm is used to parallelize the process. The parallel processing process of K-means algorithm: all data is distributed to different nodes, each node only calculates its own data, each node can read the cluster center generated in the last iteration, and calculate which cluster center the data in this node should belong to, and each node generates a new cluster center according to its own data in the iteration.
The parallel processing of K-means algorithm is an iterative process, which is divided into three stages: Map phase, combine phase and reduce phase.

Map phase: this process mainly reads the data file, reads the cluster center value in the last iteration result, writes the map() function to process, and produces a new cluster center. In the map() function, the global variable centerstring is constructed. Centerstring is used to save the cluster center value of the last iteration and read a data. The function calculates the distance between the primary data and the original cluster center, and assigns the minimum value to centerpoint. The calculated centerpoint is a new cluster center.

Reduce phase: Each reduce receives information about a cluster, including the id of the cluster, the average value of the data points for the cluster, and the number of data points corresponding to the average pointNum. After the Reduce, it outputs the current iteration count, the mean and the number of data points is belong to the centroid.

4. Analysis of improved K-means algorithm

4.1. Experimental environment
When the traditional K-Means algorithm is used for data mining, the amount of data increases to a certain extent, and the computer memory will not be able to support the normal execution of the algorithm. The experimental environment used in this experiment includes: setting 5 virtual machines on the VMware workstation as a Hadoop cluster environment, additionally installing Hadoop, configuring SSH, JDK, Eclipse, Python, and dependent packages such as numpy and pandas.

4.2. Analysis of experimental results
In order to test the performance of the improved K-Means algorithm, 10000, 100000, 1000000, 2000000 and 3000000 records of data were set in the experiment for testing. Each group of experiments was repeated 30 times, and the average value was calculated as the final result. For the accuracy of the experimental results, cluster environment is composed of 5 machines and 3 machines for testing. The test results are shown in Table 1.

| test method/data size | 10000   | 100000  | 1000000 | 2000000 | 3000000 |
|-----------------------|---------|---------|---------|---------|---------|
| original K-means      | 172ms   | 8592ms  | 3476870ms | 7126870ms | 13859330ms |
| optimized K-means     | 3480ms  | 5433ms  | 14853ms  | 25313ms | 37563ms |
| algorithm is used on 3 machines |         |         |         |         |         |
| optimized K-means     | 2123ms  | 5175ms  | 10036ms  | 16678ms | 22673ms |
| algorithm is used on 5 machines |         |         |         |         |         |

Experimental results show that the operation efficiency of the improved K-means algorithm based on MapReduce is much higher than that of the traditional algorithm. The larger the data size, the better the performance.

5. Case analysis

5.1. Data preparation
Taking the customer consumption data of a restaurant chain as the object of study, data mining was carried out, and the consumer group was clustered by using the improved K-means according to the behavioral characteristics of customer consumption, so as to find out high-value customers, which has
practical guiding significance for maintaining customers of the restaurant.

The consumption behavior data of some catering customers are shown in Table 2. CR represents the time interval of the latest consumption, CF represents the consumption frequency, and CM represents the total amount of consumption. Based on these data, customers are divided into different customer groups, and the value of these customer groups is evaluated.

Table 2. Customer consumption behavior characteristics table

| ORDER_ID            | CR  | CF | CM     |
|---------------------|-----|----|--------|
| 2019111200001       | 25  | 10 | 394.00 |
| 2019111200002       | 35  | 3  | 616.30 |
| 2019111200003       | 36  | 7  | 521.00 |
| 2019111200004       | 41  | 5  | 225.00 |
| 2019111200005       | 5   | 18 | 1086.64|
| 2019111200006       | 54  | 2  | 127.80 |
| 2019111200007       | 37  | 5  | 793.8  |
| 2019111200008       | 15  | 2  | 512.20 |
| 2019111200009       | 4   | 21 | 1714.00|
| 2019111200010       | 3   | 11 | 232.81 |
|                     | .....|    | .....   |

5.2. cluster analysis

960 sample data are selected, and the cluster center values are selected respectively. The improved k-means clustering algorithm is used for clustering. The results are shown in Table 3.

Table 3. output results of optimized k-means algorithm

| Grouping category | Customers1 | Customers2 | Customers3 |
|-------------------|------------|------------|------------|
| number of samples | 124        | 622        | 214        |
| proportion of samples | 12.92% | 64.79% | 22.29% |
| Clustering center  | CR -1.62355 | CF -0.11732 | CM 1.30578 |
|                   | CF 1.830445 | -0.02464 | -0.99862 |
|                   | CM 2.18879  | -0.12428 | -0.93114 |

the probability density function diagram for different customer groups is drawn, and the value of different customer groups are compared intuitively as shown in Figure 2, figure 3 and figure 4. Figure 2 is the CR characteristic diagram of three groups of catering customers, figure 3 is the CF characteristic diagram of three groups of catering customers, and figure 4 is the CM characteristic diagram of three groups of catering customers.

Figure 2. characteristic chart of catering customers CR
Through figure 2, figure 3 and Figure 4, the value analysis of customer is carried out:

Characteristics of Customers1: CR interval is relatively small, mainly concentrated in 0-40 days; Consumption times are concentrated in 0~25 times; The consumption amount is 500~2000, This subgroup belongs to the high-consumption and high-value customers.

Characteristics of Customers2: CR interval is at a moderate level, mainly focusing on 20-45 days; Consumption times are concentrated in 2~12 times; The amount of consumption is between 200 and 1000. This subgroup belongs to general customers.

Characteristics of Customers3: CR interval is relatively large, mainly concentrated in 30-65 days; Consumption times are concentrated in 1~6 times; The amount of consumption is between 0 and 200. This group is the lower-value customer group.

Through the analysis of the above examples, valuable data can be quickly analyzed through the improved K-means clustering algorithm.

6. Conclusion
The era of big data has come, in order to improve the speed of data mining in the big data environment.

In this paper, the principle of K-means algorithm is studied based on data mining environment of Hadoop cluster, and the k-means algorithm is optimized by using MapReduce programming idea. And through experiments, it is proved that in the cluster of cloud computing platform, using k-means algorithm to process data mining has better mining efficiency and good expansion performance. Finally, through case analysis, the optimized algorithm is applied to a restaurant chain store customer
consumption big data processing to verify the effectiveness of the algorithm, which can mine out high-value customer groups and better serve for catering decision-making.

**Author's brief introduction**

Hongqin Wang, female, lecturer, the research direction is big data application technology. Hongxia Wang, female, lecturer, the research direction is Automobile big data. Li Jiang, male, lecturer, the research direction is software engineering. Zhengjun Pan, male, associate professor, the research direction is big data application technology and artificial intelligence.

**Acknowledgments**

Fund Project Number: Key Scientific Research Projects of Universities in Guangdong Province (project number: 2018KQNCX393, 2019GKQNCX022). 2017 higher education reform project of Guangdong Province (Guangdong Higher Education letter [2018] 1). Research project of South china institute of Software Engineering.Gu ((project number: ky201816, ky201809, ky201913, JXTD201901, ZYGXK201903).

**References**

[1] Casado R, Younas M. Emerging trends and technologies in big data processing[J]. Concurrency & Computation: Practice & Experience, 2015, 27(8): 2078-2091.

[2] Gordon S, Linoff Michael J, A Berry. Translated by Wenhan Chao, Xiaoming Zhang. (2014). Data Mining Techniques: For Marketing, Scales and Customer Relationship Management, Third Edition. Tsinghua University Press.

[3] Xun Yaling, Zhang Jifu, Qin Xiao, et al. (2017) Data Partitioning in Frequent Item set Mining on Hadoop Clusters[J]. IEEE Transactions on Parallel and Distributed Systems, V28(1): 101-114.

[4] Liangjun Zhang, Zhe Fan. (2016) Hadoop Practice of Big Data Analysis and Mining[M]. Mechanical industry Press.

[5] Benmounah Z, Meshoul S, Batouche M. Scalable. (2017)differential evolutionary clustering algorithm for big data using map reduce paradigm[J], International Journal of Applied Metaheuristic Computing, V8(1): 45-60.

[6] Joseph Issa. (2016) Performance characterization and analysis for Hadoop K-means iteration. Journal of Cloud Computing, 5 (1): 1-15.

[7] Tom White. (2017) Hadoop The Definitive Guide. Tsinghua University Press. 4TH Edition: 141-149.

[8] Lai Yang. (2013) Parallel Approach in Data Mining Based on Hadoop Cloud Platform. Journal of System Simulation. 25(5): 936-944.