Improvement and parallelization of k-means clustering algorithm based on Spark platform

Zhihua Xiang*
Dept. of Information and technology, Guang Dong Polytechnic College, 526020. China
*Corresponding author: Xiangzhihua5240@126.com

Abstract. Clustering algorithm analysis is a very common data mining technology in the current industry research, and has been widely promoted and applied in many social marketing and other related industries. It can not only classify, identify and optimize customers accurately, but also design documents by various data processing methods. K-means clustering algorithm is based on the relevant ideological content of "birds of a feather flock together" to split and synthesize data, so that the similarity of data between different clusters can reach the highest. This paper discusses and improves the algorithm based on Spark.

1. Basic data theory and related technologies of Spark platform

1.1. Spark basic computing framework for big data

Different from other algorithms, Spark adopts distributed computing methods. On the basis of memory computing and through program execution and processing, it greatly reduces the time for disk to store and read data during operation. Therefore, this algorithm can better process, analyze and deeply mine the data based on the iterative learning algorithm of the machine. The calculation model is as follows:

![Spark computing model](image)

RDD and DAG are two very important concepts in the process of data model calculation through Spark platform. In short, RDD can effectively improve the data fusion degree between machines by centralizing and partitioning entities and controlling and processing them. However, RDD realizes free...
caching and conversion in RAM by means of rectangular sharding and effective transformation of operators on the platform. Figure 2 spark ecological model can show that the data will flow directionally with different processing processes, which fundamentally shortens the time of disk operation. Leave more time to the system for deep mining of data [1]. The specific operation flow is shown in figure 3.

![Spark ecosystem](image)

**Fig 2 Spark ecosystem**

**SPARK execution process**

1. Build Spark cluster environment, initialize the running environment.
2. SparkContext requests the resources necessary to run Executor to YARN, etc., and starts CoarseGrainedExecutor Backend.
3. After SparkContext obtains all the resources required for the running environment, Executor applies to SparkContext for Task.
4. SparkContext dispatches tasks to each node in the cluster.
5. Spark builds the DAG diagram and decomposes it into Stage, then manages the Stage, via DAGScheduler and distributes the Task to Executor.
6. After Executor receives the Task, it starts executing and releases the resources occupied after execution. Activate Windows.

**Fig 3 SPARK execution process**

1.2. Basic operation contents of k-means clustering algorithm
Among the many popular data algorithms, K-means is one of the ten classical data mining algorithms. The main content of the algorithm is to divide the target data and guide it to find the cluster point which is closest to its own attribute. In the K-Means algorithm, the aggregation group and the individual are taken as the initial clustering, and the detailed formula is expressed as follows:

\[
E = \sum_{t=1}^{k} \sum_{p \in C_t} \text{dist}(p, c_t)^2
\]  

(1)

\[
m_t(c_t) = \frac{\sum_{j=1}^{n} p_j}{n}
\]  

(2)

It is not enough to reflect it by formula. The detailed algorithm flow is shown in figure 4.
Fig 4 K-Means algorithm process

The implementation process is detailed in figure 5.

Execution process of traditional Kmean algorithm.

1. Set clustering parameter $K$, iteration number $T$, iteration stop threshold $E_d$

2. According to the set parameters, $K$ group data is randomly selected as the initial clustering center.

3. According to formula 4, the sum of Euclidean distance squares from each individual to $K$ clustering centers in $P$ is calculated, and each individual is divided into clusters represented by the nearest center point from the clustering center according to the nearest principle.

4. According to formula 4 the mean value of all individuals in each cluster is calculated, and the clustering center of the cluster is updated to the mean value.
5. To determine whether the iteration stop threshold is less than \( E_d \) or whether the number of iterations is greater than \( T \), if so, the algorithm ends. Output clustering results; otherwise, go to step (3) and continue until the end of the algorithm is reached.

Fig 5 K-Means execution process

The relevant algorithms of K-means are used for targeted data clustering allocation. Finally, after the data clustering is completed, the new clustering operation is carried out on the clustering data in different clusters \([2]\).

2. Improvement of k-means clustering algorithm based on different data sampling

In the process of traditional data operation in the past, k-means clustering algorithm and other relevant contents are usually targeted presets of clustering centers and uniform central sampling of data. In short, the center of this preset point will directly intervene and influence the final effect presented by data in the clustering process. At the same time, in the process of no difference between data, data on the random sampling, the center of the clustering center is selected by a similar rate, but the data in the process of actual operation, its have unequal distribution randomness characteristics, therefore, in the clustering center in the process of selection and the default, according to the distribution of different data to the center of the effective selection.

In this study, under the background of different data sampling, the following improvements are made to the content of k-means clustering algorithm. Firstly, in the process of function definition and data sampling probability, the following correlation operation methods are adopted.

\[
\Phi_x(C) = \sum_{x \in X} d^2(x, C) = \sum_{i=1}^{k} \min_{i \in C} ||x - C_i||^2
\]

\[
p_x = \frac{d^2(x, C)}{\Phi_x(C)}
\]

In this way, the sum of the square between the data points of different attributes and the clustering center is obtained. The calculation model of graph 6SPARK can be seen and minimized by iterative method, and then the effective result of clustering convergence can be achieved.

Fig. 6 SPARK operation calculation model.

3. Improvement of data algorithm on Spark platform

3.1. Improvement of data set sampling stage
In the process of data set sampling, targeted sampling and analysis are carried out according to different characteristics of different data at first, and output is made according to different sampling results, and the output content is taken as the clustering center [4]. In the process of starting the specific algorithm, through the preliminary reading of files in the cluster, an effective RDD1 is formed, and the relevant data content saved in RDD1 is divided into scientific and reasonable partitions with the basic unit size of 64MB. At the same time, a new RDD2 is formed through the targeted operation of different map operators. In addition, iteration and analogy are carried out in the above way to complete the relevant process of sampling. As shown in the figure 7 below:

![Diagram](image)

Fig 7 data set sampling phase diagram

3.2. Improvement of k-means clustering algorithm
Because the k-means clustering algorithm has a relatively complete operation system. Therefore, after the targeted improvement in the sampling stage, the contents that need to be processed in the next step, such as the data set, are greatly reduced in scope and more refined in content. In the application process of k-means clustering algorithm, only the selection of clustering center points for specific operations is needed to complete the whole operation process in the later stage.

4. Content analysis of the results of the research experiment
4.1. Research environment and basic data of the experiment
In the process of this research experiment, the author builds the platform content of the real physical cluster. All of them use the common data and software in the current industry as the main platform system. As can be seen from the fig 8 Spark cluster resource management system YARN configuration information table, through the analysis and operation improvement of different types of data sets, the accuracy and error rate of the data are screened in detail, and the contents of the clustering algorithm and the results of the algorithm are compared and evaluated scientifically. And through the relevant effective way to carry on the targeted detection to its performance. The results show that the algorithm is improved. After that, the accuracy of the algorithm has been effectively improved, and the improved algorithm has better stability and relative accuracy.

![Table](image)

Fig 8 Spark Cluster Resource Management system YARN configuration Information Table
4.2. Basic contents of parallel performance experiment

| clustering algorithm | data set | Average accuracy | Average number of generations | Average time consuming |
|----------------------|---------|------------------|-----------------------------|------------------------|
| K-means              | Wine    | 0.7301           | 7.76                        | 198.33                 |
|                      | Tse     | 0.6333           | 12.03                       | 866.32                 |
|                      | sccts   | 0.4832           | 18.31                       | 2080.89                |

Fig 9. K-means clustering results

Under the condition of certain scale form and relative stability, the parallel and serial time of data execution is compared horizontally, and the most effective effect is judged scientifically. This method is also called acceleration ratio. According to the relevant results of this study, it is shown that because of the different sizes of the data sets, there is no significant effect on the comparison of parallel time. In the process of slicing data into 64MB, the increase of the number of nodes does not significantly improve the efficiency of data operation and execution, or even increase the time it takes, and increase the load in the process of data processing [5]. The detailed data are shown in figure 9 K-means clustering results.

It can be seen that in the process of actual clustering algorithm, we should not only select the size of the data set according to the actual situation, but also select the scientific and reasonable number of nodes according to the number of data sets, and then improve the efficiency of data operation and the stability of data results on the basis of ensuring the rational use of resources. Thus it can be seen that under the environment of Spark platform, the data algorithm can be improved and improved, and realized, which can effectively improve the application range of the future data algorithm in various industries of society.

5. Conclusion

To sum up, the author has made targeted improvement and improvement on the k-means clustering algorithm based on Spark platform. By means of random sampling of the same kind and non-same kind of the original data and horizontal and vertical comparison of the research experimental contents, the accuracy of the research results has been improved to the greatest extent and the error rate has been reduced. Thus, in the operation process of different types of large database scenario, the need to improve on serial data related calculation, can better adapt to the development of different industries in the age of social needs, at the same time, the relevant data such as parallel processing and calculation way also has a more broad space for development.

Acknowledgements

Foundation item: 2018.01-2020.01 Letter No. 2014 from the Department of Higher Education, Ministry of Education (Programming Foundation of 201702071079)

References

[1] Zhefu, wu Tong, zhang Xiao ying. Improvement and parallelization of k-means clustering algorithm based on Spark platform [J]. Internet world, 2016(1):44-50.
[2] Dongfei, Song Hua. Xu Parallelization implementation of improved k-means algorithm based on Spark [J]. Computer system application, 2018.
[3] Peng, liu jiayan, Teng Enjie, ding. Spark based large-scale text k-means parallel clustering algorithm [J]. Chinese journal of information technology, 2017(04):150-158.
[4] Yubo, li Yuwang, Yang Hao, tang. Spark based k-means security interval update and optimization algorithm [J]. Computer technology and development, 2017(8).
[5] qing deng ning Yang. Research on improved parallel k-means algorithm based on Spark framework [J]. Intelligent computer and application, 2018.