Research Article

Lingrui Bu*, Hui Zhang, Haiyan Xing, and Lijun Wu

Research on parallel data processing of data mining platform in the background of cloud computing

https://doi.org/10.1515/jisys-2020-0113
Received Nov 06, 2020; accepted Dec 09, 2020

Abstract: The efficient processing of large-scale data has very important practical value. In this study, a data mining platform based on Hadoop distributed file system was designed, and then K-means algorithm was improved with the idea of max-min distance. On Hadoop distributed file system platform, the parallelization was realized by MapReduce. Finally, the data processing effect of the algorithm was analyzed with Iris data set. The results showed that the parallel algorithm divided more correct samples than the traditional algorithm; in the single-machine environment, the parallel algorithm ran longer; in the face of large data sets, the traditional algorithm had insufficient memory, but the parallel algorithm completed the calculation task; the acceleration ratio of the parallel algorithm was raised with the expansion of cluster size and data set size, showing a good parallel effect. The experimental results verifies the reliability of parallel algorithm in big data processing, which makes some contributions to further improve the efficiency of data mining.

Keywords: Cloud computing, data mining, parallel processing, Hadoop platform, clustering algorithm

1 Introduction

With the rapid development of Internet and Internet of things, the types and quantity of data are increasing, showing an explosive growth trend. In the aspect of data mining, the traditional data storage method has not been able to meet the high concurrent access of massive data. When the mining algorithm is at the level of terabyte and petabyte, there will also be problems of low computing efficiency and poor expansion performance. Therefore, the demand for high-efficient data processing methods is growing [1]. The emergence of cloud computing provides a new possibility for efficient data mining. The function of cloud computing parallel computing can parallelize and transplant mining algorithm to the cloud platform, which has a very good performance in data processing. Lu *et al.* [2] extended the K-means algorithm through tabu search strategy, then realized the parallel processing of the algorithm in spark framework, and verified the advantages of the algorithm in accuracy, scalability and other aspects through calculation experiments. Zhang *et al.* [3] implemented a parallel rendering system on Hadoop framework by combining particle swarm optimization with support vector machine algorithm. Through comparison of different rendering scenes, they found that the system effectively improved the storage and computing efficiency. Chen *et al.* [4] designed a parallel random forest (PRF) algorithm on Apache spark platform. During the training process, the dimension was reduced and the dual parallel method was used. The experimental results showed that the algorithm had significant advantages in aspects such as classification accuracy and scalability. In the present study, a data mining platform was designed on Hadoop, the improved k-means algorithm was parallelized, and the experiment verified that the parallel algorithm was effective in processing big data. The experiment shows

***Corresponding Author: Lingrui Bu:** Shandong Labor Vocational and Technical College, Jinan, Shandong, 250022, China; Email: lrps9uu@126.com
Hui Zhang, Haiyan Xing, Lijun Wu: Shandong Labor Vocational and Technical College, Jinan, Shandong, 250022, China

that the combination of cloud computing with data mining is reliable in solving the data problem, which provides a new idea for solving the problem of low operating efficiency in processing large-scale data and has important values for the development of data mining.

2 Cloud computing and data mining

Cloud refers to a large server cluster, including computing servers, broadband resources, etc. Cloud computing is to distribute computing tasks on the cloud, so that users can access various services on demand [5], which realizes the concentration of computing resources and reduces costs for users [6]. It has unique advantages in data storage, management, sharing [7], etc., with the following characteristics: ① large scale: Google cloud computing, for example, has more than one million servers; ② versatility: cloud Computing can support different applications; ③ virtualization: users can get services at any location and any terminal; ④ high reliability; ⑤ strong scalability; ⑥ low cost.

Data mining refers to the process of finding hidden, unknown and valuable information from a large number of random and noisy data, as shown in Figure 1.

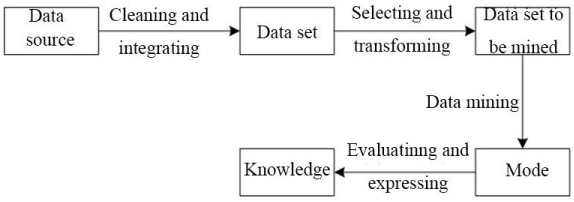

Figure 1: Data mining process

Data mining includes association analysis, classification, prediction, clustering, etc. Its main methods include neural network [8], genetic algorithm [9], decision tree [10], etc. It has very important application values in the fields of finance [11], scientific research [12] and medicine [13]. With the explosive growth of data and information overload, the original technology is increasingly unable to meet the needs of big data processing [14], while the emergence and development of cloud computing provides a new possibility for the efficient mining of big data.

3 Hadoop distributed file system platform

Hadoop (Figure 2) is a distributed computing mechanism [15], which can realize the distributed computing of big data sets and has characteristics of fast processing speed, good fault tolerance and simple use. It realizes the storage of massive data through the distributed file system [16] and then realizes the data computing through MapReduce.

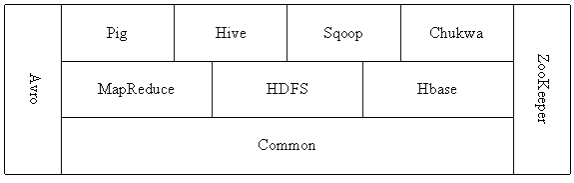

Figure 2: The structure of Hadoop

(1) Hadoop distributed file system: through Hadoop distributed file system, it can realize the distributed storage of massive data on cheap devices. It adopts Master/Slave mode, including multiple namenodes and multiple datenodes: ① namenode: master node, responsible for data distribution and maintenance and storing all information in Fsimage; ② datanode: responsible for executing namenode commands, responding to Hadoop distributed file system read-write requests, and keeping communication with namenode through heartbeat mechanism. To write files to Hadoop distributed file system, clients need to interact with namenode to read and write files on datanode.
(2) MapReduce: it is a software framework of parallel data processing, including Map function, Reduce function and main function: ① Map function: accepting a group of data and transforming it; ② Reduce function: calculating to get a new key/value list; ③ main function: combining job control and file input/output. The calculation model of MapReduce is shown in Figure 3.

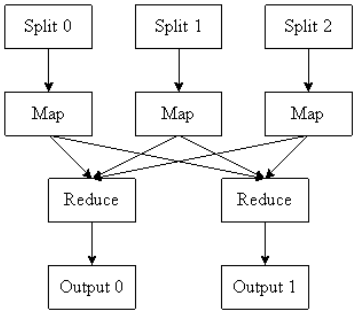

Figure 3: MapReduce calculation model

4 K-means algorithm under Hadoop distributed file system

K-means algorithm is a kind of data mining classification algorithm, which has simple operation and low time complexity. It has been widely used in data mining in industries such as medicine and finance [17], and its flow is shown in Figure 4.

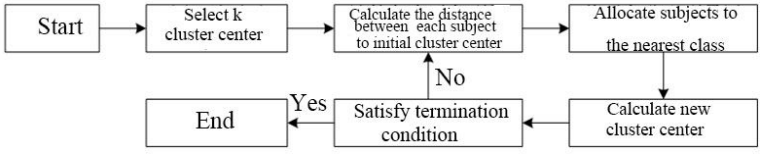

Figure 4: The flow of K-means algorithm

In order to realize the good application of K-means algorithm in big data processing, it was with the idea of max-min distance. The flow of the improved algorithm is as follows:

(1) it is assumed that there are N objects,

$$S_n = \{X_1, X_2, \cdots, X_x\}; \tag{1}$$

(2) an object is randomly selected, for example, X_1 is the first class cluster center, Euclidean distance is used as the measurement index of similarity relation of data object, and the object which has the largest distance with X_1 in S_n is regarded as X_2;

(3) the distances of remaining objects to and X_1 and X_2 are calculated, and the smallest one is supposed as D_{x_i};

(4)

$$\max S_n \{D_{x_i}\} \tag{2}$$

is calculated; if

$$\max S_n \{D_{x_i}\} > m [average (|X_2 - X_1|)], \tag{3}$$

then X_i is supposed as the new clustering center, where $\frac{1}{2} \le m < 1$;

(5) steps (1) – (4) are repeated until there is no new cluster center.

When dealing with big data, firstly, the data can be sampled to reduce the data size. It is assumed that there are totally n samples in the sample set, there are at least $f|C|$ sample points from class cluster C ($0 \le f \le 1$). Suppose the sampling size as s, and it should satisfy

$$s \ge f_n + \frac{n}{|C|} \log\left(\frac{1}{\xi}\right) + \frac{n}{|C|}\sqrt{\left(\log\frac{1}{\xi}\right)^2 + 2f|C| \log\left(\frac{1}{\xi}\right)}. \tag{4}$$

Then, in the sample set, the probability that the number of samples from class cluster C is smaller than $f|C|$ is smaller than ξ, $0 \le \xi \le 1$.

The realization of parallelization of the improved K-means algorithm on Hadoop platform can be divided into three MapReduce processes:

(1) MapReduce1: independent parallel sampling is performed through map function. The data is divided into m parts, and m is the number of tasks of reduce. Then the reduce function clusters the data according to the idea of max-min distance. Several clustering centers and the average distance of each cluster are output.
(2) MapReduce2: the output of reduce in step (1) is processed through map function. Reduce summarizes the output of step (1), merges the adjacent cluster centers, and recalculates the new cluster centers.
(3) MapReduce3: all objects are divided into the nearest clusters by map function, and the new cluster center is calculated by reduce function until the cluster center no longer changes.

The development tool of Hadoop based data mining platform is Eclipse, the development language is Java, and the running environment is Linux x86/x64, which mainly includes three parts:

(1) bottom layer: Hadoop cluster is composed of computers. Linux system is installed and tested by virtual machine. There are two kinds of namenodes in the cluster, which are in running and blocking state respectively. Moreover, only one node provides services, and the other node synchronizes information. When the running node fails unexpectedly, it can carry out timely switching to ensure the availability of the cluster.
(2) service layer: the service engine interface adopts REST interface, which is provided to the user in the form of uniform resource locator (URL). The user can directly interact through hyper text transfer protocol (HTTP) request or embed the interface into the system for secondary development. The files on Hadoop distributed file system are displayed to the user in the form of list, and the user can do operations such as download, delete, etc. Finally, the platform guides the user through the form of web page to complete data mining, and users can get data mining services as long as they choose algorithms and upload data sets. The platform monitors and manages tasks through Apache Ambari tool, which enables users to view the execution of tasks in real time.
(3) user layer: it refers to the developer or the person using the platform service.

5 Experimental results

The experimental cluster used six computers which were equipped with dual core P4 2.8 CPU, 320 G hard disk and 2G memory. Gigabit Ethernet was used. One computer was namenode, and the remaining computers were datanodes. The specific establishment process of the Hadoop platform is as follows.

(1) Ubuntu 14.04 system was installed by means of dual system on every node.
(2) JDK software package was installed, and then whether the software was successfully installed was checked using java-version command.
(3) Key pair generated on namenode through "ssh-keygen-trsa" command. Then public key id_rsa.pub was copied to the authorized-keys file of all nodes to realize secure shell (SSH) configuration.
(4) Hadoop software package was unzipped and installed, and files such as core-site.xml, hdfs-site.xml, and mapred-site.xml were configured.
(5) The Hadoop file package was delivered to other nodes through secure copy (SCP) command.
(6) Namenode was formatted through "Hadoop namenode-format" command, and then all processes were started through "start-all.sh" command.
(7) After the establishment of Hadoop, the MapReduce function was written for the improved K-means algorithm using Java language and operated on the Hadoop platform.

The data set used in this study was Iris data set which is widely used in clustering analysis. The data set includes three categories: Setosa, Versicotor and Viginica. Due to the small capacity of Iris, in order to verify the effect of the parallel algorithm on big data processing, five data sets with different scales were randomly generated by increasing the capacity of original data set through code, as shown in Table 1.

Table 1: Experimental data set

	IrisA	IrisB	IrisC	IrisD	IrisE
Size	10MB	50MB	250MB	1.3GB	6.2GB
Data volume	36000	100000	500000	3000000	11000000
Attribute	50	50	50	50	50
Number of cluster centers	3	3	3	3	3

In order to verify the performance of the improved K-means parallel algorithm, it was compared with the traditional K-means algorithm running in the ordinary single-machine environment. A node in the Hadoop cluster ran the parallel algorithm. The data scale was reduced using equation (1), and then the parallelization was realized according to step (1)-(3) in chapter 4.1. The clustering results of the two algorithms are shown in Table 2.

Table 2: Comparison of clustering results

	Original data	Traditional algorithm	Parallel algorithm
Setosa	70	70 (100%)	70 (100%)
Versicotor	70	64 (91.43%)	69 (98.57%)
Viginica	70	62 (88.57%)	68 (97.14%)

It was seen from Table 2 that the number of samples which were correctly classified by the parallel algorithm was larger than that of the traditional algorithm when Iris of different categories in the data set were divided using the two algorithms; in the clustering of Setosa, the results of the two algorithms were the same; in the clustering of Versicotor, the traditional algorithm correctly divided 64 samples, with a clustering precision

of 91.43%, and the parallel algorithm correctly divided 69 samples, with a clustering precision of 98.57%; in the clustering of Viginica, the traditional algorithm only correctly divided 62 samples, with a clustering precision of 88.57%, and the parallel algorithm correctly divided 68 samples, with a clustering precision of 97.14%. The above results showed that the algorithm designed in this study had better accuracy, better overall clustering effect, and higher clustering quality than the traditional algorithm. The running time comparison of the two algorithms is shown in Table 3.

Table 3: Comparison of single-machine operation time

	IrisA	IrisB	IrisC	IrisD	IrisE
Traditional K-means algorithm	21s	212S	2285s	17548s	Insufficient memory
Parallel algorithm	121s	1569s	2516s	21548s	72158s

It can be seen from Table 3 that the running time of the traditional algorithm was shorter than that of the parallel algorithm in the case of small amount of data. In Hadoop platform, each iteration of the parallel algorithm needs to read, write and transfer data constantly, which consumes some resources. However, in the case of large data (IrisE), the traditional algorithm in single-machine environment would suffer from memory shortage, and the system could not bear the time cost of big data processing, but the Hadoop platform with only one computer node successfully completed the computing task, which showed that the parallel algorithm had significant advantages in big data processing.

The speedup ratio is an index for measuring the parallel performance of the algorithm, and its calculation formula is $S_p = \frac{T_s}{T_m}$, where T_s is the running time of the single processor and T_p is the running time of the multiprocessor. The larger the acceleration ratio S_p, the better the parallel performance. Under different number of computer nodes, the acceleration ratio of the parallel algorithm is shown in Figure 5.

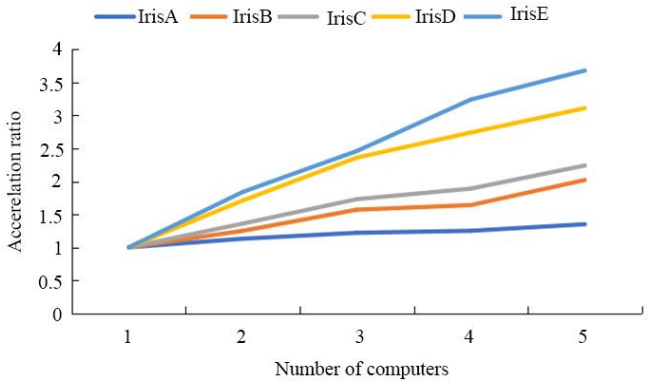

Figure 5: The acceleration ratio of the algorithm

It can be seen from Figure 5 that the acceleration ratio of the algorithm increased with the increase of the number of computers in all data sets, which indicated that the parallel algorithm had higher efficiency in data processing in the cluster environment; the acceleration ratio increased with the expansion of the data set scale, which indicated that the parallel algorithm had better parallel performance in the big data environment.

6 Discussion

Data mining is a process of finding useful information from massive information [18]. With the expansion of information, the implementation of big data mining has become a key and difficult problem [19]. The traditional computer has a large time cost and low processing efficiency in big data processing, and only parallel computing can effectively solve this problem [20]. Therefore, the research on parallel data mining has a very important practical significance.

In this study, a data mining platform was designed based on Hadoop, then the parallelization of algorithm in data mining was studied. The experimental results demonstrated that the improved parallel algorithm had better effect on data set clustering and better division accuracy. In the clustering of Iris data set, the number of samples correctly clustered by the traditional algorithm was significantly smaller than that of the parallel algorithm, indicating the clustering performance of the parallel algorithm. The improved parallel algorithm optimized the principle of initial value in the clustering algorithm, which reduced the independence of the algorithm on the initial clustering center, improved the global convergence ability of the algorithm, and enhanced the clustering effect of the algorithm. In the single-machine environment, when the data amount was small, the operation time of the parallel algorithm was relatively long as it operated on only one computer node, longer than the traditional algorithm, but in the face of large amount of data, the traditional K-means algorithm failed to complete the task of data processing because of insufficient storage, but the parallel algorithm effectively completed the task, which showed the effectiveness of the parallel algorithm in the big data environment; when the scale of data increased, the parallel algorithm still maintained an excellent processing ability. The results of the acceleration ratio of the parallel algorithm suggested that the acceleration ratio showed an increasing tendency with the increase of computer nodes in different data sets; the larger the scale of the data set was, the larger the acceleration ratio was. The above results verified that the parallel algorithm had advantages in big data processing and had good stability and expansibility.

Although some achievements were obtained in parallel data processing, there were still some problems to be solved in the next research:

(1) more mining algorithms need to be implemented on Hadoop platform, such as Apriori, decision tree algorithm, etc.;
(2) Hadoop platform needs to be further optimized to make it play a better data mining effect;
(3) data set in a larger scale needs to be tested;
(4) when the number of data was small, the computing efficiency of the parallel algorithm can be improved.

7 Conclusion

The K-means algorithm in data mining was studied through Hadoop, and it was parallelized and implemented in Hadoop environment. The experiments found that:

(1) the parallel algorithm can complete large-scale data tasks;
(2) the clustering effect of the parallel algorithm was better than that of the traditional algorithm;
(3) the parallel effect of the parallel algorithm strengthened with the increase of computer nodes and data scale.

The experimental results showed that the improved K-means parallel algorithm designed in this study was reliable in dealing with massive data, which makes some contributions to improve the efficiency of data mining and solve the problem of large-scale data processing.

References

[1] K. Siddique, Z. Akhtar, Z. Akhtar, E. J. Yoon, Y. S. Jeong, D. Dasgupta and Y. Kim, Apache Hama: An Emerging Bulk Synchronous Parallel Computing Framework for Big Data Applications, IEEE Access **PP**(2016), 1-1.

[2] Y. Lu, B. Cao, C. Rego, and F. Glover, A Tabu Search based clustering algorithm and its parallel implementation on Spark, Appl Soft Comput **63** (2017), 97-109.

[3] Y. Zhang, Z. Zhu, H. Cui, X. Dong, and H. Chen, Small files storing and computing optimization in Hadoop parallel rendering, Concurr Comp Pract E **29**(2017), 1269-1274.

[4] J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, and K. Li, A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment, IEEE T Parall Distr **28**(2017), 919-933.

[5] Y. Wang, J. Li, and H. H. Wang, Cluster and cloud computing framework for scientific metrology in flow control, Cluster Comput **22**(2019), 1-10.

[6] K. B. Nayar and V. Kumar, Cost benefit analysis of cloud computing in education, Int J Busin Inform Syst **27**(2018), 205-221.

[7] J. Li, Y. Zhang, X. Chen, and Y. Xiang, Secure attribute-based data sharing for resource-limited users in cloud computing, Comput Secur **72**(2018), 1-12.

[8] M. W. Wang, Y. Cui, S. H. Xiao, G. X. Wang, D. Yang, K. Chen, and J. Zhu, Neural network meets DCN, Proc ACM Meas Anal Comput Syst 2018, 2(2):1-25.

[9] Z. Zhang, V. Trevino, S. S. Hoseini, S. Belciug, A. M. Boopathi, P. Zhang, F. Gorunescu, V. Subha, and S. S. Dai, Variable selection in Logistic regression model with genetic algorithm, Ann Transl Med **6**(2018), 45-45.

[10] P. Kai, V. C. M. Leung, L. X. Zheng, S. G. Wang, C. Huang, and T. Lin, Intrusion Detection System Based on Decision Tree over Big Data in Fog Environment, Wirel Commun Mob Com **2018**(2018), 1-10.

[11] R. Geng, I. Bose, and X. Chen, Prediction of financial distress: An empirical study of listed Chinese companies using data mining, Eur J Oper Res **241**(2015), 236-247.

[12] A. Linden, and P. R. Yarnold, Using data mining techniques to characterize participation in observational studies, J Eval Clin Pract **22**(2016), 835-843. .

[13] P. Singh, S. Singh, and G. S. Pandijain, Effective heart disease prediction system using data mining techniques, Int J Nanomed **13**(2018), 121-124.

[14] S. Rallapalli, R. R. Gondkar, and U. P. K. Ketavarapu, Impact of Processing and Analyzing Healthcare Big Data on Cloud Computing Environment by Implementing Hadoop Cluster, Proc Comput Sci **85**(2016), 16-22.

[15] S. Rathee, and A. Kashyap, Adaptive-Miner: an efficient distributed association rule mining algorithm on Spark, J Big Data **5**(2018), 6.

[16] S. Wu, W. Zhu, B. Mao, K. C. Li, PP: Popularity-based Proactive Data Recovery for HDFS RAID systems, Future Gener Comp Sy **86**(2018), 1146-1153.

[17] N. M. Mohammadi, A. Hezarkhani, A. Maghsoudi, Application of K-means and PCA approaches to estimation of gold grade in Khooni district(central Iran), Acta Geochim **37**(2018), 104-114.

[18] X. C. Sheng, X. F. Xue, and Y. P. Cheng, Research on the Parallel Frequent Data Mining Strategy under the Cloud Computing Environment, Appl Mech Mater **719-720**(2015), 924-928.

[19] Y.Liu, W. Cai, and X. Shao, Big data and chemical data mining, Chin Sci Bull **60**(2015), 694-703.

[20] C. F. Tsai, W. C. Lin, and S. W. Ke, Big Data Mining with Parallel Computing: A Comparison of Distributed and MapReduce Methodologies, J Syst Software **122**(2016) 83-92.