Analysis and Research of K-means Algorithm in Soil Fertility Based on Hadoop Platform

Guifen Chen, Yuqin Yang, Hongliang Guo, Xionghui Sun, Hang Chen, Lixia Cai

Analysis and Research of K-means algorithm in soil fertility Based on Hadoop platform

Guifen Chen[1,a], Yuqin Yang[1,b], Hongliang Guo[1], Xionghui Sun[1], Hang Chen[1,2], Lixia Cai[1]
[1]Jilin Agricultural University, Changchun 130118,China
[2] Jilin provincial science and technology information Research Institute, Changchun130033,China
[a]guifchen@163.com,[b]1172126066@qq.com

Abstract: In order to study the K - means algorithm for evaluation of soil fertility, solve the large amount of calculation and high time complexity of the algorithm，this paper proposes the K-means algorithm based on Hadoop platform.First, K-means algorithm is used to cluster for Nongan town soil nutrient data for nine consecutive years;clustering results show that : the accuracy rate increased year by year, and consistent with the actual situation.Then for the K-means clustering algorithm in processing large amounts of data has the disadvantages of high time complexity, This paper uses the K-means algorithm Based on Hadoop platform to realize the clustering analysis of soil fertility of large amounts of data;the results show that: compared with the traditional serial K-means algorithms, improves the operation speed. The above analysis shows that, K- means algorithm is an effective soil fertility evaluation method;Based on Hadoop platform of parallel K-means algorithm has great realistic meaning to analysis of large amount of data of soil fertility factors.

Keywords: K-means algorithm,Hadoop platform,MapReduce model,Soil fertility

1. Introduction

With the wide application of agricultural information technology, 3S technology (GPS, GIS, RS),the Internet of things technology and Expert System (ES) technology are applied extensively in precision agriculture,so that the rapid growth of the agricultural sector data[1].K - means algorithm based on Hadoop platform, can quickly and accurately to the large amount of data of soil fertility carries on the comprehensive evaluation and correct analysis,is of great significance to guide farmers reasonable fertilization. Xiong Chunhong for Jiang Xi tea area soil fertility problems and the status of the heavy metals of fresh tea leaves,she uses mining techniques were analyzed and predicted, creating favorable conditions for soil fertility analysis and comprehensive evaluation[2].Li Lianghou proposed application of clustering anslysis in classifying site type and evaluating soil fertility[3].The research on weighted space fuzzy dynamic clustering algorithm by Chen Gui-fen, proved the effectiveness of soil fertility evaluation[4].The traditional clustering algorithm in processing large-scale data in terms of efficiency of real-time or from the angle of system resources, are not well resolved.In the clustering algorithm, K-means algorithm is the most widely used clustering algorithm based on partition, it has the advantages of rapid and simple; but it has the disadvantage of large amount of calculation and sensitive to the initialization center. In contrast,this paper studies the K-means clustering algorithm based on Hadoop platform of MapReduce parallel programming method, and relevant experiments were carried out for the large-scale soil fertility data .

2.Key technologies and algorithms introduced

2.1. MapReduce model

MapReduce is a distributed parallel computing model proposed by Google Labs, its basic idea[5,6] :(1)The MapReduce database to user program input data set is divided into several small data set,tthen fork copies user processes to other machines in the cluster;(2)A copy of user program has a called master is responsible for scheduling, allocation of jobs to idle worker(Map worker or Reduce worker); (3)After the worker is assigned a

Map job, read data from the input data and extract key value pairs, performs Map computation tasks and generate the intermediate key value pairs,and then cached in memory;(4)The middle of the cache key value to be written to a local disk regularly,and is divided into R districts,each corresponding to a Reduce worker, master is responsible for forwarding the middle of the key positions to Reduce worker;(5)Reduce worker read the intermediate key value pairs and to sort them out,so that the same keys of the key to gather together;(6)After sorting the intermediate key value to be Reduce worker traversal, and then for each unique keys, reduce worker will pass key and values associated with the key to the Reduce function, the output generated by the Reduce function will be added to the output file in the partition;(7)When all the Map and Reduce operations are completed, master wakes genuine user program,MapReduce function call returns user program code.[6] MapReduce operation process shown in Fig1 [7,8]:

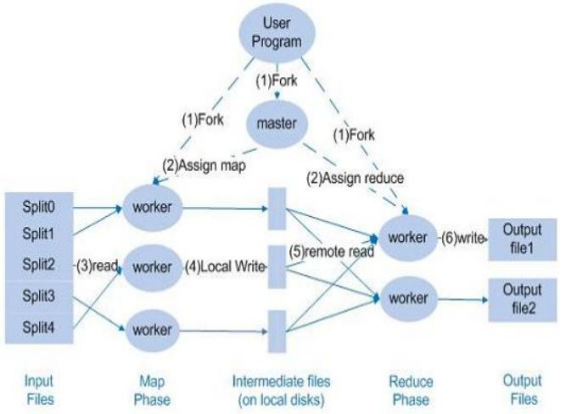

Fig. 1 MapReduce operation

2.2. K-means algorithm

K-means algorithm is a clustering algorithm based on partition method, it is one of the earliest proposed classic clustering algorithm . The main idea of the K-means algorithm is the n object into the k class (cluster) ($k \leq n$).In this experiment we refer to the "National Survey of cultivated land fertility and quality evaluation of technical regulations" in the grading standard,Nong'an basic farmland is divided into six grades[9].First of all, in the whole soil fertility data set we randomly selected k objects,each object represents an initial cluster center or the initial average.For each remaining object, according to its distance from the center of each cluster, assign it to the nearest cluster.And then calculate the average of each cluster,each object in the database is compared with the average value of each cluster,the object is assigned to the most similar clusters.This process is repeated until the cluster objects are"similar", while objects in different clusters are "dissimilar",namely the criterion function converges and make the minimum square error function.

K - means algorithm steps:

Algorithm: K - means. Division of the K- means algorithm based on the average of the objects in the cluster.

Input: the number of clusters k and the database contains n objects.

Output: k clusters, make the least square error criterion.

Methods:

(1) Choose k objects as the initial cluster centers;

(2) Repeat;

(3) According to the average of the objects in the cluster,each object (again) is assigned to the most similar cluster;

(4) Update the average of the cluster, namely calculating the average of objects in each cluster;

(5) To calculate the clustering criterion function;

(6) Until criterion function will not change.[10]

The disadvantage of K-means algorithm:When using K-means clustering algorithm,k values are given in advance.Normally,We don't know a given data set should be divided into many classes is the most suitable, but in this experiment, we are aiming at soil fertility data, according to the soil fertility grading standards and past experience we can easily set the K value, so in this experiment we do not need to consider this problem;The initial cluster centers are randomly selected, select the initial cluster center has a great influence on the clustering results, once the selection is not good initial values, we may not be able to get the clustering results effectively;K-means algorithm needs to be constantly adjusted sample classification and calculate new cluster centers; therefore, the time complexity of the K-means algorithm is relatively large, and increases with the amount of data[11];At the same time, the clustering result of K-means clustering algorithm is easily affected by

noise data.;With the application of information technology in agricultural field, the soil fertility data quantity increases,just K-means algorithm of large amount of calculation and high time complexity aspects,this experiment made improvements, and achieved good results.

3. K-means clustering algorithm to achieve MapReduce

Datasets of MapReduce processing should have such characteristics: it can be broken down into many small data sets, and each of the small data set can be completely processed in paralle[12,13],soil fertility data is well positioned to meet this demand .The basic idea is that K-means clustering algorithm actualizes MapReduce:each iteration start a MapReduce process, MapReduce to complete data records to the cluster center distance calculation and the new clustering center calculation.Fig2 describes the K-means clustering algorithm MapReduce parallel implementation[14].According to the computing needs of MapReduce, preprocessing data records in the form of rows are stored, so that pre-treatment data can be fragmented by row, and no correlation between the pieces of data, fragmentation process is completed by the MapReduce model, without writing code.

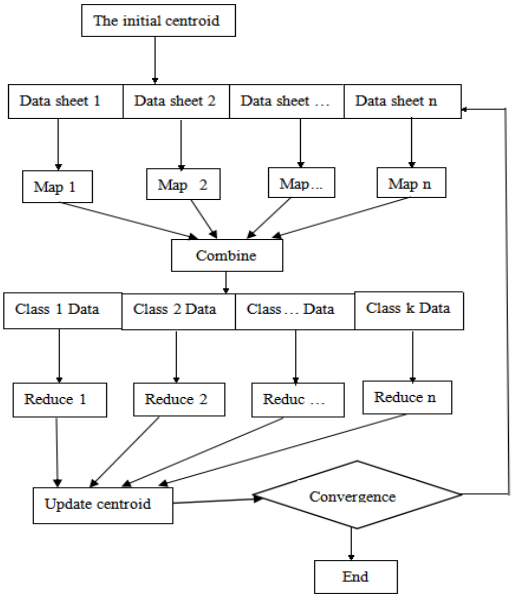

Fig. 2 K - means clustering algorithm MapReduce parallel implementation

3.1. Map function design

The task of the Map function is progressive to read data of soil fertility from the Hadoop Distributed File System (HDFS) file,, calculate the distance from the center to each record and re-mark the new category it belongs to the cluster. The input of Map function is the original data set and the last round of iteration (or initial clustering) clustering center , input data records <key,value>, ie <row number of rows>; output of intermediate results <key, value> that <record category, record> [15,16].

Map function specifically described as follows:

```
Public void map(Object key,Text value ,Context context){
    To calculate the distance the record to each center of mass;
    Compare the above distance;
    The record boils down to the class of the nearest centroid;
    To write <Record Category, Record> into the intermediate file ;
}
```
[7,16]

3.2. Reduce function design

The task of Reduce function is based on the output of the Map function, update the clustering center,for the next round of the Map function. Meanwhile,calculating the value of standard measurement function, for the main function whether iteration is over.Before performing the Reduce function,MapReduce will be merged with the result of output in the middle of the Map function,the intermediate results in multiple <key,value> have the same key value pair combined into a pair. the Reduce function's output <key,value> is <Record Category,

{record collections} > [15,16]. The output <key,value> is<category No., the mean vector + the sum of the squared error of the class>.

Reduce function specifically described as follows:

Public void reduce(Text key,Iterable<Text> values, Context context{

for(The same key for all records) {

Averaging each attribute;

Calculated the record to its centroid distance

The above distance sum ; }

The <Record Category, the mean vector + sum of square error for each class > write the results file;}[16,17]

In the main function called MapReduce process described above, each iteration apply for a new job, until the difference between the sum of the squares of the errors and the last time is less than a given threshold,the iteration ends. Map function last intermediate result is the final results of classification [16,18,19].

4. Experiments and results analysis

4.1. Experimental datasources

In recent years, with the application of information technology in agricultural field, we also get a lot of correlated with soil fertility data.The experimental data is mainly from the national "863" plan- "research and application of corn precise operating system"[20] project demonstration base-NongAn country of JiLin Province for many years conducted aprecision fertilization after soil fertility data applied research.This paper select the representative soil nutrient data to integrated analysis, such as, Alkaline hydrolysis nitrogen, Available potassium and Available phosphorus. From the town of NongAn、WanShun and other towns during 2005 to 2013years.

Tab.1 Part of thesampling data

Town name	Alkaline hydrolysis nitrogen (mg/kg)	Available phosphorus (mg/kg)	Available potassium (mg/kg)	Latitude	Longitude	Elevation
SanGang	139	17	118	44.07632	124.78503	213
SanGang	87	14.5	115	44.07774	124.78433	213
SanGang	80	13.6	126	44.07768	124.78444	217
SanGang	153	13.2	115	44.07841	124.7742	214
SanGang	87	12	120	44.07709	124.77288	220
SanGang	139	13.1	123	44.07123	124.7769	216

4.2. Application of K-means algorithm in the analysis of soil nutrients

K-means clustering algorithm is one of the classic algorithm. First, K-means algorithm is used to cluster for Nongan town soil nutrient data for nine consecutive years(2005-2013),according to the test zone N, P, K data,hierarchical clustering analysis of soil fertility.The experimental results shown in Fig3:

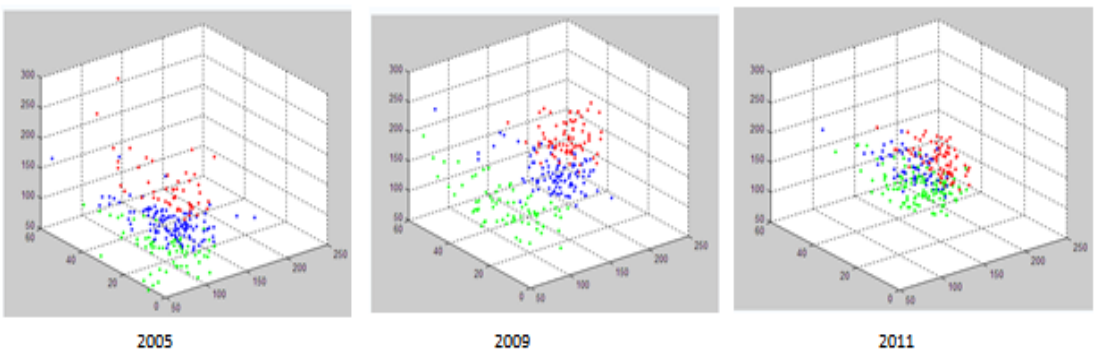

Fig3.Clustering results

The above clustering results showed that:after continuous precision fertilization, three comprehensive similarity of entire plots of soil nutrient data that are Alkaline hydrolysis nitrogen、Available phosphorus and Available potassium increased year by year,soil fertility tends to bebalanced. Experimental results consistent with the actual situation,indicating that K- means algorithm is an effective soil fertility evaluation method. However, with the passage of time, the data of soil nutrients in Nongan county is increasing year by year, environmental factors associated with the soil fertility is considered, the disadvantages of K-means algorithm which are the large amount of calculation and high time complexity becomes more obviously , So we propose the K-means algorithm based on Hadoop platform.

4.3K-means algorithm based on Hadoop platform to achieve

In the Hadoop environment, the data file stored in Hadoop Distributed File System(HDFS) will be automatically divided into a plurality of data block, each size is 64MB. In Map stage , each data block is processed by a Map task, the whole data files can be assigned to different nodes. Assume that each node can perform m tasks, there are n nodes involved in parallel computing, parallel K-means algorithm's time complexity is n*k*t*o/ (m*n), Compared with the time complexity of the serial algorithm n * k * t * o ,parallel K-means algorithm can significantly improve the operation efficiency .

To analyze the advantages of Hadoop platform in large amounts of data processing, in the case of the same hardware configuration environment, the same size of data,We performed serial K-means algorithm and Hadoop platform K-means algorithm computation time comparison. The experimental results are shown in Tab 2, where T1 is the operation time of serial algorithm, T2 is the operation time of the algorithm under the framework of MapReduce.

Tab.2 Comparison of experimental results

Serial	File(KB)	Records(105)	t1（s）	t2(s）
1	1,248	0.32609	4.09	248.81
2	4,991	1.30436	35.13	280.44
3	19,963	5.21744	157.11	461.43
4	39,926	10.43488	246.08	690.97
5	79,851	20.86976	331.45	1217.25
6	109,794	28.69592	979.76	1641.88
7	114,786	30.00028	Out of memory	2507.39

As seen from Tab 2, when the data size is less than 109,794KB, the efficiency of the serial K-means algorithm is higher than K-means algorithm under the framework of single MapReduce, this is because, MapReduce task divides one task into two different stages of Map and Reduce. Data flow implementation process is: (key, value) -- >Map stage -- > (K1, V1),
(K1, V2), (K2, V3) - >sort&shuffle -- > (K1, list (V1, V2), (K2, V3) -- > Reduce stage -- > HDFS, the iterative MapReduce that requires constant read, write and transmit data ;with the increasing of soil fertility data quantity, Running the serial task of the machine,the memory and other resource consumption increases,which will lead to the decline of the performance of the machine; when the data size reach to 114,786KB, it will report that the internal memory is insufficient. While, K-means algorithm under the MapReduce framework is able to cope continues to increase the amount of data, and complete the computing task of large-scale data successfully.

5 Results and Discussion

Research on the NongAn town 2005 to 2013 soil nutrient data clustering showed that when dealing with large amounts of data, Based on Hadoop platform of parallel K-means algorithm have good operational results and practical significance .

(1)K-means algorithm is used to cluster for Nongan town soil nutrient data for nine consecutive years(2005-2013).Experiments show that after continuous precise fertilization ,comprehensive similarity of entire plots of soil nutrient data increased every year ,the soil fertility tends to equilibrium .Experimental results consistent with the actual situation, indicating that K- means algorithm is an effective soil fertility evaluation method.

(2)K-means algorithm based on Hadoop platform, When dealing with small amount of data, Serial K-means algorithm efficiency is better than under the MapReduce framework parallel K-means algorithm. Because when a small amount of data , in the MapReduce framework of K-means algorithm for each iteration, restarting a new JobTracker, the startup and interactive process will consume some resources.

(3)K-means algorithm based on Hadoop platform, when dealing with large amount of data , Hadoop platform can be well done the calculate of the large amounts of the soil fertility information data , which fully reflects the capabilities and advantages of the Hadoop’s processing big data .Running the serial task of the machine,the memory and other resource consumption increases and causes the machine performance degradation,which lead to report insufficient memory and this can not meet the needs of data growth.

(4)With the increase of the amount of data, the Value that the difference of running time under the MapReduce framework divided by the total time tend smaller ,which reflects the stability and reliability of Hadoop platform.

Acknowledgements

This work was supported by the national “ 863 ” project (2006AA10A309), National Spark Plan (2008GA661003) and Shi Hang of Jilin province projects (2011- Z20).

References

1. Turner BL, Meyer WB. Land use and land cover in global environmental change: considerations for study [J]. Int.SoiSci.,J.,1991,130:669-680.

2. Chunhong Xiong,Assesment and Prediction on Heavy Metals Status for Soils and Fresh Tea Leaves in Jiangxi Major Tea Regions Based on GIS Data Mining Technology[D]. Nanchang：Nanchang University,2011.

3.Lianghou Li,Jiyue Li.Application of Clustering Analysis in Classifying Site Type and Evaluating Soil Fertility.2010 Third International Conference on Education Technology and Training (ETT 2010) ,2010:468-471.

4.Guifen Chen,Liying Cao, Guowei Wang.Application of Weighted Spatially Fuzzy Dynamic Clustering Algorithm in Evaluation of Soil Fertility[J]. Scientia Agricultura Sinica, 2009, 42(10): 3559-3563.

5.Yanjiang Qian. Research and implementation of large-scale dataclustering techniques [D]. ChengDu: University of Electronic Science and Technology,2009.

6.Longfei Li.Research and implementation of Hadoop +Mahout intelligent terminal based cloud application recommendation engine.[D]. Chengdu: University of Electronic Science and Technology, 2013.

7.Jianjiang Li,Jian Cui,Ran Wang.Review of MapReduce parallel programming model[J].Journal of Electronic, 2011, 39(11)：2635-2641.

8.Peng Liu.Open the shortcut leading to the actual Hadoop cloud computing [M] .Beijing:Electronic Industry Press, 2011: 60-74.

9. Guifen Chen. Research and Application of Spatial Data Mining Technology for Precision Agriculture [D]. ChangChun: Jilin University,2009

10.Wubin Pan.Research and Application of Mining parallel K-means based on meteorological data cloud.[D], Nanjing: Nanjing Information Engineering University, 2013.

11.M.C.Naldi,R.J.G.B.Campello.Evolutionary K-means for distributed data sets[J]. Neurocomputing. 2014, 127(15) 30-42.

12.Prng Liu.Open the shortcut leading to the actual Hadoop cloud computing[M].Beijing: ElectronicIndustry Press,2011:60-74.

13.Lizhe Wang,Jie Tao,Rajiv Ranjan,Holger Marten, Achim Streit,Jingying Chen,Dan Chen.G-Hadoop: MapReduce across distributed data centers for data-intensive computing[J]. Future Generation Computer Systems. 2013,29(3):739–750.

14.Ting Zhou,Junying Zhang,Cheng Luo.Realization of K-means clustering algorithm based on Hadoop[J]. Computer Technology and Development , 2013.7.23（7）18-21

15.Srirama S N,Jakovits P,Vainikko E.Adapting scientific computing problems to clouds using MapReduce [J].Future Generations Computer Systems , 2012 , 28(1): 184-192.

16.Guilan Xie , Shengxian Luo.Research on Application of Hadoopbased on MapReduce model[J].Software world, 2010,29（8）：4-7.

17.Weizhong Zhao,Huifang Ma,Yanxiang Fu,Zhongzhi Shi.Parallelk-meansclustering algorithm designed Hadoop cloud-base computing platform[J].Computer Science 2011. 10. 38（10）166-167.

18.Yuanfang Li,Shikun Deng,Yupiao Wen.PageRank algorithm block matrix under Hadoop-MapReduce [J].Computer Technology and Developme,2011, 21(8):6-9.

19. Xiaoping Jiang,Chenghua Li,Wen Xiang,Xinfang Zhang,Haitao Yan. K-means clustering algorithm MapReduce parallel realization [J].Huazhong University of Science and Technology (Natural Science) ,2011.6.39 (SupplementI）120-124.

20.Guifen Chen,Li Ma,Hang Chen.Research status and development trend of precision fertilization technology[J]. JilinAgricultural University,2013,35(3):253-259.