A Parallel Clustering Method Study Based on MapReduce

Sun Zhanquan
Shandong Provincial Key Laboratory of Computer Network
Shandong Computer Science Center
Jinan, Shandong, 250014, China
sun30@indiana.edu

Abstract—**Clustering is considered as one of the most important tasks in data mining. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. It has been widely applied to many kinds of areas. Many clustering methods have been studied, such as k-means, Fisher clustering method, Kohonen neural network and so on. In many kinds of areas, the scale of data set becomes larger and larger. Classical clustering methods are out of reach in practice in face of big data. The study of clustering methods based on large scale data is considered as an important task. MapReduce is taken as the most efficient model to deal with data intensive problems. In this paper, parallel clustering method based on MapReduce is studied. The research mainly contributes the following aspects. Firstly, it determines the initial center objectively. Secondly, information loss is taken as the distance metric between two samples. The efficiency of the method is illustrated with a practical DNA clustering problem.**

Keywords- Clustering; Information bottleneck theory; MapReduce; Multidimensional Scaling; Twister

I. INTRODUCTION

With the development of electronic and computer technology, the quantity of electronic data is in exponential growth [1]. Data deluge has become a salient problem to be solved. Scientists are overwhelmed with the increasing amount of data processing needs arising from the storm of data that is flowing through virtually every science field, such as bioinformatics [2-3], biomedical [4-5], Cheminformatics [6], web [7] and so on. Then how to take full use of these large scale data to support decision is a big problem encountered by scientists. Data mining is the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. It has been studied by many scholars in all kinds of application area for many years and many data mining methods have been developed and applied to practice. But most classical data mining methods out of reach in practice in face of big data. Computation and data intensive scientific data analyses are increasingly prevalent in recent years. Efficient parallel/concurrent algorithms and implementation techniques are the key to meeting the scalability and performance requirements entailed in such large scale data mining analyses. Many parallel algorithms are implemented using different parallelization techniques such as threads, MPI, MapReduce, and mash-up or workflow technologies yielding different performance and usability characteristics [8]. MPI model is efficient in computation intensive problems, especially in simulation. But it is not easy to be used in practical. MapReduce is a cloud technology developed from the data analysis model of the information retrieval field. Several MapReduce architectures have been developed now. The most famous one is developed by Google, but the source code is not open. Hadoop is the most popular open source MapReduce software. The MapReduce architecture in Hadoop doesn't support iterative Map and Reduce tasks, which is required in many data mining algorithms. Professor Fox developed an iterative MapReduce architecture software Twister. It supports not only non-iterative MapReduce applications but also an iterative MapReduce programming model. The manner of Twister MapReduce is "configure once, and run many time" [9-10]. It can be applied on cloud platform. It will be the popular MapReduce architecture in cloud computing and can be used in data intensive problems.

Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. Many classical clustering methods have been studied and widely applied to many kinds of field, such as k-means, Fisher clustering method, Kohonen neural network and so on[11-12]. K-means is a popular clustering method. But it is difficult to determine the initial centroids which have great effect on the clustering result. On the other hand, the distance measure of k-means can't measure arbitrary correlation between samples. Information bottleneck (IB) theory is proposed by Tishby [13]. It is a data compression method based on Shannon's rate distortion theory. The clustering method based on IB theory was widely studied in recent years. It measures the distance between samples with the quantity of information loss caused by merging. It has been applied to the clustering of image, texture, and galaxy successfully [14-15] and got good results. But when the scale of data set becomes larger and larger, classical clustering method will not work to deal with large scale data set. How to develop clustering methods based on MapReduce to process large scale data is an important issue. It is the development trends of big data science.

In this paper, a novel clustering method based on MapReduce is proposed. It combines IB theory with centroid based clustering method. Firstly, IB theory based hierarchy clustering is used to determine the center of each Map computational node. All sub-controids are combined into one centroid with the IB theory in Reduce computation node. For measuring the complicated correlation between samples, information loss is used to measure the distance. The clustering method is an iterative model. The clustering method is programmed with iterative MapReduce model Twister. For visualizing the clustering results, interpolation MDS is used to reduce the samples into 3 dimensions[16]. The reduced clustering results are show in 3D coordination with Pviz software developed by Indiana University. Bioinformatics is an important application field of large scale data analysis. Lots of bioinformatics data will be generated all over the world every day. DNA clustering is taken as an example to illustrate the efficiency of the proposed clustering method.

II. IB PRINCIPLE

In this paper, IB clustering method is used to determine the initial clustering centroids. It will be realized in MapReduce model. The clustering method is introduced as follows.

The IB clustering method states that among all the possible clustering of a given object set when the number of clusters is fixed, the desired clustering is the one that minimizes the loss of mutual information between the objects and the features extracted from them. Let $p(x,y)$ be a joint distribution of the "object" space X and the "feature" space Y. According to the IB principle we seek a clustering $\hat{X}$ such that the information loss $I(X;\hat{X}) = I(X;Y) - I(\hat{X};Y)$ is minimized. $I(X;\hat{X})$ is the mutual information between X and $\hat{X}$

$$I(X;\hat{X}) = \sum_{x,\hat{x}} p(x)p(\hat{x}\mid x)\log\frac{p(\hat{x}\mid x)}{p(\hat{x})} \quad (1)$$

Given a random variable X and a distortion $d(x_1,x_2)$ measure, we want to represent the symbols of X with no more than R bits. The rate-distortion function is given

$$D(R) = \min_{p(\hat{x}|x)|I(X;\hat{X})\le R} Ed(x,\hat{x}) \quad (2)$$

where $Ed(x,\hat{x}) = \sum_{x,\hat{x}} p(x)p(\hat{x}\mid x)d(x,\hat{x})$.

The loss of the mutual information between X and Y caused by the clustering $\hat{X}$ can be viewed as the average of this distortion measure

$$\begin{aligned} d(x,\hat{x}) &= I(X;Y) - I(\hat{X};Y) \\ &= \sum_{x,\hat{x},y} p(x,\hat{x},y)\log\frac{p(y\mid x)}{p(x)} - \sum_{x,\hat{x},y} p(x,\hat{x},y)\log\frac{p(y\mid \hat{x})}{p(y)} \\ &= ED(p(x,\hat{x})\,||\,p(y\mid\hat{x})) \end{aligned} \quad (3)$$

where $D(f\,||\,g) = E_f \log(f/g)$ is the KL divergence. Wecan obtain the rate distortion function

$$D(R) = \min_{p(\hat{x}|x)|I(X;\hat{X})\le R} (I(X;Y) - I(\hat{X};Y)) \quad (4)$$

which is exactly the minimization criterion proposed by the IB principle, i.e. finding a clustering that minimize the loss of mutual information between the objects and the features.

Let c_1 and c_2 be two clusters of symbols, the information loss due to the merging is

$$d(c_1,c_2) = I(c_1;Y) + I(c_2;Y) - I(c_1,c_2;Y) \quad (5)$$

Standard information theory operation reveals

$$d(c_1,c_2) = \sum_{y,i=1,2} p(c_i)p(y\mid c_i)\log\frac{p(y\mid c_i)}{p(y\mid c_1\cup c_2)} \quad (6)$$

where $p(c_i) = |c_i|/|X|$, $|c_i|$ denotes the cardinality of c_i, $|X|$ denotes the cardinality of object space X, $p(c_1\cup c_2) = |c_1\cup c_2|/|X|$.

It assumes that the two clusters are independent when the probability distribution is combined. The combined probability of the two clusters is

$$p(y\mid c_1\cup c_2) = \sum_{i=1,2}\frac{|c_i|}{|c_1\cup c_2|}p(y\mid c_i) \quad (7)$$

The minimization problem can be approximated by a greedy algorithm based on a bottom-up merging procedure. The algorithm starts with the trivial clustering where each cluster consists of a single data vector. In order to minimize the overall information loss caused by the clustering, classes are merged in every step, such that the information loss caused by merging them is the smallest. The method is suitable to both sample clustering and feature clustering.

III. CLUSTERING BASED ON MAPREDUCE

The parallel clustering method can be divided into two parts. The first part is to determine the initial center point. The second part is to obtain the global center point through iteration and get the final clustering results.

A. Initial centroid calculation

Given data set D with n samples, it is divided into m partitions $D^1, D^2, \cdots, D^m$ with $n_1, n_2, \cdots, n_m$ samples separately. Operate clustering on each partition $D^i = \{D_1^i, D_2^i, \cdots, D_{n_i}^i\}, i = 1, \cdots, m$ with the clustering method introduced in section 2. We can obtain the sub-centroids $C^i = \{C_1^i, C_2^i, \cdots, C_{n_i}^i\}, i = 1, \cdots, m$. All sub-centroids are collected together to generate new data set $C = \{C^1, C^2, \cdots, C^m\}$. The new dataset are clustered with information bottleneck theory. Then we can obtain the initial global center C^0. In the calculation equation (7), the number of each clustering is considered. The centroid vector should include the numbers of samples that generate the centroid. The realization of the calculation process based on Twister is shown in figure 1. Firstly, partitioned datasets are distributed to each computational node. In each Map computational node, IB is operated on each dataset to obtain the sub-centroid. All sub-centroids are collected in Reduce node to generate new dataset. The new dataset is analyzed with IB to generate the initial centroid of the global dataset.

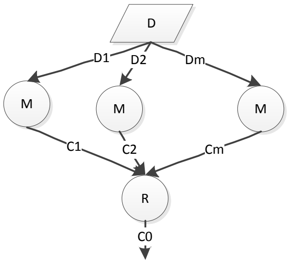

Figure1.Initial centroid calculation process

B. Centroid based iterative clustering

After getting the initial center C^0, it is used to calculate the final centroid. The process is as follows. For each sample in each sub dataset $\boldsymbol{x} \in D^i$, calculate the distance between the sample and all the samples in center data set C^0. In the calculation, information loss (6) is taken as the distance measure. Set k empty dataset $P^1, P^2, \cdots, P^k$. The sample will be added to dataset P^i if the distance between x and the center c_i^0 is minimum. Recalculate the center of C^i with the dataset P^i according to (7). After obtaining the new sub-centroids $C^1, C^2, \cdots, C^m$, the new centroid C^0 is calculated according to the following equation.

$$c_i^0 = \sum_{j=1}^{m} \frac{\left|c_i^j\right|}{\left|c_i^1 \cup c_i^2 \cdots c_i^m\right|} c_i^j \qquad (8)$$

Through calculation the difference between the old C^{old} and the new generated C^{new} to determine whether the iteration will stop. The iteration process based on Twister is shown as in figure 2. The samples are partitioned and deployed in each computational node in the first step. The initial C^0 got from the first step is mapped to each computational node. In each Map node, sub-centroids are recalculated. All sub-centroids are collected in Reduce node and the global centroid C^0 is regenerated according to (8). The new centroids are feedback to main computational node and the difference between the old C^0 and the new generated C^0 is calculated. The iteration will stop when the difference is less than the prescribed threshold value. The difference between two iterations is measured with Kull-back divergence.

$$\delta = \sum_{i=1}^{l} x_i^{new} \, log \frac{x_i^{new}}{x_i^{old}} + \sum_{i=1}^{l} x_i^{old} \, log \frac{x_i^{old}}{x_i^{new}} \qquad (9)$$

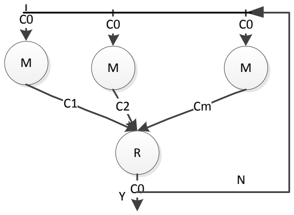

Figure 2.Iteration process to calculate the final centroid

IV. EXAMPLES

A. Data source

The initial data set was received from Dr. Mina Rho in Indiana University. They are some 16S rRNA data. We select 100043 DNA data to clustering analysis. In the data file, each DNA record is expressed with a G, A, C, and T strings. The example is analyzed in India cluster node of FutureGrid. Each node installs Ubuntu Linux OS. The processor is 3GHz Intel Xeon with 10GB RAM.

B. Preprocess

Calculate the probabilities of $\{A, C, T, G\}$ and $[AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT]$ of each string. The sample strings are transformed into 16 dimensions vector. The samples described with probabilities are taken as the input of the clustering.

C. DNA clustering

The original data set are partitioned into 100 partitions and deployed to 8 computational nodes. Each data set partition is clustered with IB introduced in section 2. Then 100 sub-centroids are obtained. One Reduce computation point is used to combine all the sub-controids into one centroid with IB theory. They are taken as the initial centroid of the centroid clustering method. The initial centroids are mapped to each computation node. The centroid of the partition is recalculated according to the section 4.2 iteratively. The clustering will stop when the stop rule is met. The computation times based on different cluster number are listed in table1.

TABLE I. COMPUTATIONAL TIME BASED ON DIFFERENT CLUSTERING

Centroid number	Partition number	Computational nodes	Computation time
3	100	8	14465.368
5	100	8	14512.32
10	100	8	15132.52

D. Clustering result visualization

In this example, 4000 samples are selected to be pre-mapped into 3 dimension space. Firstly, distance matrix is calculated and Euclidean distance is taken as the measure. Then the mapped vectors are calculated according to the distance matrix with MDS method. The left samples are mapped into low dimension with interpolation MDS method. The number of nearest neighbor is set $\mathrm{k} = 10$. After dimension reduction, the clustering results are shown as in figure 3,4, and 5 respectively.

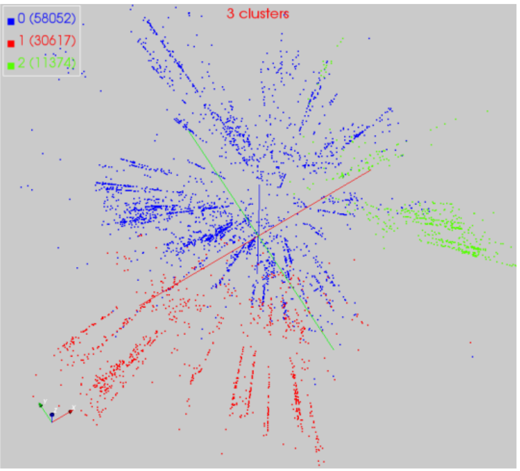

Figure3.Clustering results of 3 clusters

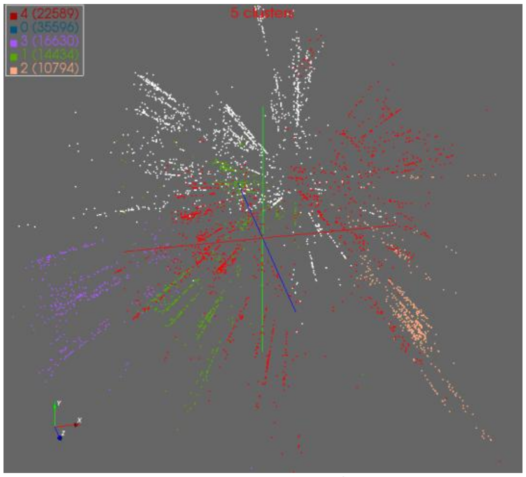

Figure 4.Clustering results of 5 clusters

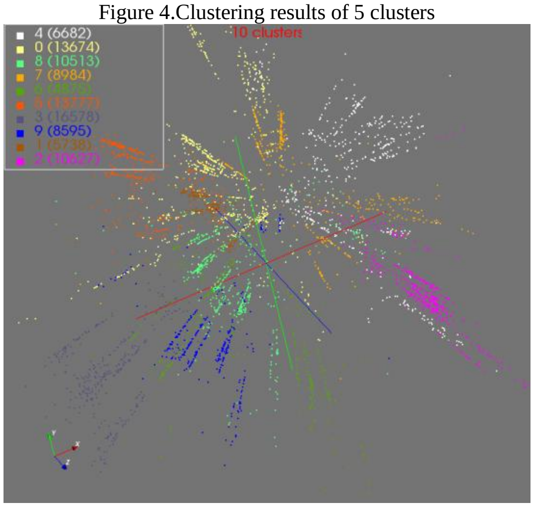

Figure 5.Clustering results of 10 clusters

V. CONCLUSIONS

Clustering of large scale data is an important task in many application fields.Clustering methods study based on MapReduce is an important task of data mining. The proposed clustering method based on MapReduce is an efficient method for large scale data analysis. It provides an objective method to determine the initial clustering centroid. The information loss is used to measure the distance between samples. It can measure any complicated correlation between samples. The clustering results are visualized with interpolation MDS method. Through the DNA clustering example analysis results, we can see that the clustering results is satisfactory. The information loss can measure the complicated correlations between DNA sequences. The clustering result is not affected by the initial cluster centroid. It provides a novel means to solve large scale clustering problems.

ACKNOWLEDGMENT

This work is partially supported by national youth science foundation(No. 61004115), Provincial Outstanding Research Award Fund for young scientist (No. BS2009DX016) and National Science and technology support program(2012AH09B03).

REFERENCES

[1] J R Swedlow, G Zanetti, C Best. Channeling the data deluge. Nature Methods, 2011, 8: 463-465.

[2] G C Fox, X H Qiu et al. Biomedical case studies in data intensive computing.Lecture Notes in Computer Science, 2009, 5931: 2-18

[3] X H Qiu, J Ekanayake, G C Fox et al. Computational Methods for Large Scale DNA Data Analysis. Microsoft eScience workshop, 2009

[4] J A Blake, C J Bult. Beyond the data deluge: Data integration and bio-ontologies. Journal of Biomedical Informatics, 2006, 39(3), 314-320.

[5] J Qiu. Scalable Programming and Algorithms for Data Intensive Life Science. A Journal of Integrative Biology, 2010, 15(4): 1-3

[6] R Guha, K Gilbert, G C Fox, et al. Advances in Cheminformatics Methodologies and Infrastructure to Support the Data Mining of Large, Heterogeneous Chemical Datasets. Current Computer-Aided Drug Design, 2010, 6: 50-67.

[7] C C Chang, B He, Z Zhang. Mining semantics for large scale integration on the web: evidences, insights, and challenges. SIGKDD Explorations, 2004: 6(2):67-76.

[8] G C Fox, S H Bae, et al. Parallel Data Mining from Multicore to Cloudy Grids. High Performance Computing and Grids workshop, IOS Press, 2008:311-340

[9] B J Zhang, Y Ruan et al. Applying Twister to Scientific Applications. Proceedings of CloudCom,IEEE CS Press, 2010: 25-32

[10] J Ekanayake, H Li, et al. Twister: A Runtime for iterative MapReduce. The First International Workshop on MapReduce and its Applications of ACM HPDC, ACM press, 2010: 810-818

[11] R Maitra, AD Peterson, AP Ghosh. A systematic evaluation of different methods for initializing the K-means clustering algorithm. Knowledge creation diffusion utilization,2010: 1-11

[12] H Simon. Self-organizing maps. Neural networks - A comprehensive foundation (2nd ed.). Prentice-Hall, 1999.

[13] NTishby, C Fernando, WBialek. The information bottleneck method.The 37th Annual Allerton Conference on Communication, Control and Computing, 1999, pp. 1-11.

[14] JColdberger, SGordon, HGreenspan. Unsupervised image-set clustering using an information theoretic framework. IEEE transactions on image processing, 2006, 15(2): 449-457.

[15] NSlonim, TSomerville, NTishb. Objective classification of galaxy spectra using the information bottleneck method. Monthly Notices of The Royal Astronomical, 2001, 323: 270-284.

[16] S H Bae, J Qiu, G Fox. Adaptive Interpolation of Multidimensional Scaling. International Conference on Computational Science, 2012:393-402