MapReduce Parallel Implementation of Improved K-means Clustering Algorithm on Spark Platform

Huang Suyu, Tan Lingli

School of Computer Science, Wuhan Donghu University, Wuhan, Hubei

Keywords: Spark platform; K-means clustering; MapReduce; Parallelization

Abstract: Cloud Computing is the development of Distributed Computing, Parallel Computing and Grid Computing. Cloud computing is a new distributed parallel computing environment or mode. The emergence of cloud computing makes the networking and service of data mining technology become a new trend. Clustering is different from classification. In the classification model, there are sample data whose class labels are known. The purpose of classification is to extract classification rules from the training sample set for class identification of objects whose class labels are unknown. In clustering, it is necessary to divide all data objects into clusters according to some measure without knowing the information about the classes of the target data in advance. Therefore, cluster analysis is also called unsupervised learning.

1. Introduction

According to different parallel strategies, K-means clustering algorithm can be divided into two categories: data parallel and control parallel. Data parallelism first divides the entire data set into several small subsets, and then performs the same operations or instructions for each subset. Control parallelism is the simultaneous execution of different operations or instructions.

From the point of view of data mining, data parallelism is superior to control parallelism in many aspects. First, the execution process of the algorithm based on data parallel strategy is the same as that of the corresponding serial algorithm. Therefore, it is easy to implement parallel versions of some serial algorithms and simplify the development process. Second, data parallelism has higher platform independence. Because the control flow of data parallel algorithm is still parallel, there is no need to design a special control flow for each parallel platform. Thirdly, the parallel algorithm based on data has better scalability and is suitable for processing large-scale data sets.

In view of the massive data scale in the current practical application, many parallel clustering algorithms have been proposed by scholars at home and abroad. Dhillon proposes a parallel K-means algorithm, which uses Message Passing Interface (MPI) to communicate between processors. In the iteration, each processor only processes the part of data allocated to it. After processing, communication between processors is needed to update the coordinates of the center points, so that the next iteration can be carried out. Du et al. proposed a parallel hierarchical clustering algorithm. The running platform of the algorithm is a cluster of distributed memory architectures with large network bandwidth and small transmission delay. Experiments show that the proposed algorithm has good scalability. Feng et al first theoretically analyzed the related problems of parallel clustering algorithm using data parallel strategy in cluster system, including the selection of acceleration ratio and communication strategy, and then proposed a parallel hierarchical clustering algorithm. Experiments prove the theoretical analysis in this paper. Olman et al. proposed a parallel clustering algorithm for bioinformatics data. In this algorithm, the minimum spanning tree is used to solve the clustering problem.

Boutsinas et al. put forward a framework of clustering algorithm, which can expand the computing scale of clustering algorithm and solve the clustering problem of large-scale data. Xu proposes a parallel DBSCAN algorithm, which uses a distributed spatial index structure called dR* tree in a fully shared framework network composed of multiple computers. Experiments show that the algorithm has an approximate linear acceleration ratio. Judd et al. proposed three pruning techniques and integrated them into the parallel clustering tool P-CLUSTER. The performance of

  DOI: 10.25236/csam.2019.069

P-CLUSTER was significantly improved by using computational pruning techniques. Other parallel clustering algorithms include Rasmussen 1989, Olson 1995, Foti 2000, Dash 2007, etc.

Although many parallel clustering algorithms have been proposed by scholars at home and abroad, users need to devote a lot of energy to data partitioning, task scheduling, load balancing, processor communication and so on. For the general users, these technologies are not easy to master, and people urgently need a platform which is easy to learn and use for the design of parallel programs, so that people can devote more energy to the design of clustering algorithm itself.

In recent years, cloud computing as a new business computing model has received extensive attention. Hadoop is a cloud computing platform that makes it easier to develop and process large-scale data in parallel. Its main characteristics are: capacity expansion, low cost and high efficiency. Reliability, etc. Hadoop is an implementation of MapReduce computing model. With the help of MapReduce, it is easy to write parallel programs, run them on the computer cluster, and complete the calculation of massive data.

According to K-means algorithm and theory of parallel data clustering based on cloud computing platform, this paper uses square error function as clustering criterion to make the algorithm simple, fast and effective in dealing with large data sets.

2. Clustering Algorithm

Clustering is a process of dividing a data set into subsets and making the data objects in the same set have high similarity, while the data objects in different sets are not similar. The similarity or dissimilarity measure is based on the value of describing attributes of data objects, which is usually described by the distance between clusters. The basic guiding principle of clustering analysis is to maximize the similarity of objects in classes and minimize the similarity of objects between classes.

The purpose of clustering algorithm is to obtain the most essential "class" properties that can reflect these sample points in N-dimensional space. This step does not involve domain experts, it does not consider any domain knowledge except the set of knowledge, does not consider the specific meaning of feature variables in its domain, only considers it as a one-dimensional feature space.

The selection of clustering algorithm depends on the type of data, the purpose and application of clustering. Generally, clustering algorithms can be divided into the following categories:

(1) partitioning method: Given the number of partitions to be constructed k, the partitioning method first creates an initial partition. Then an iterative relocation technique is used to improve the partition by moving objects between partitions. At present, K-means algorithm and K-medoids algorithm are two popular heuristic partitioning methods.

(2) Hierarchical method: decompose a given set of data objects hierarchically. BIRCH, CUREIN and CHAMELEON are typical hierarchical clustering algorithms.

(3) Density-based method: Distance-based clustering can only find spherical clusters, but it is difficult to find clusters with arbitrary shapes. For this reason, density-based clustering is proposed, which can be used to filter noise data and find clusters with arbitrary shapes. DBSCAN, OPTICS and CLIQUE are three representative methods.

(4) Model-based method: Assuming a model for each cluster to find the best fit of data to a given model, the model-based algorithm can locate clustering by constructing a density function reflecting the spatial distribution of data points, or automatically determine the number of clustering based on standard statistics.

(5) Grid-based method: The grid-based method quantifies the object space into a finite number of cells to form a grid structure on which all clustering operations are performed.

3. Evidence Analysis Process

Document [2] improves K-means and proposes a distributed clustering algorithm DK-means, which can be used as a parallel implementation of K-means clustering algorithm. In this algorithm, there are p sites in the distributed system, from which Sm is the main site and the other P-1 sites are

the slave sites. Firstly, K cluster centers are randomly generated at the main site. $\{c_1,\ldots,c_k\}$ As the global initial cluster centers, they are broadcasted to all slave stations. According to these centers, each station confirms the cluster of local data objects, and gets the local cluster centers by formula 1. At the same time, the total number of local cluster centers and corresponding cluster data objects is obtained from the site. $\{(c_{i1},n_{i1}),\ldots,(c_{ik},n_{ik})\}$ $(1\le i\le p)$ It is transmitted to the main site, which calculates the global cluster center based on the cluster information.

$$c_j=\frac{n_{1j}\times c_{1j}+n_{2j}\times c_{2j}+\cdots+n_{pj}\times c_{pj}}{n_{1j}+n_{2j}+\cdots+n_{pj}},(1\le j\le k) \tag{1}$$

Iterate this process until the global discriminant function E value is stable, that is, the global cluster center is stable. In the clustering process, DK-means does not need to transmit a large number of data objects between sites, but only needs to transmit the cluster center and the total number of data objects in the cluster. The traffic of DK-means is very small, so the efficiency of DK-means is very high.

Theorem The clustering result of DK-means algorithm is the same as that of K-means algorithm for centralized clustering of distributed data.

It is proved that DK-means algorithm is implemented in distributed environment. Each site is divided into K clusters and the center points are respectively K clusters.$\{c_{i1},\ldots,c_{ik}\}$, among,$1\le i\le p$, $c_{ij}=\frac{1}{n_{ij}}\sum_{y\in W_{ij}}y,1\le j\le k,$ n_{ij} Cluster w_{ij} Total number of data objects in. Then global cluster centers

$$\begin{aligned}c_{ij}&=\frac{n_{1j}\times c_{1j}+n_{2j}\times c_{2j}+\cdots+n_{pj}\times c_{pj}}{n_{1j}+n_{2j}+\cdots+n_{pj}} \\ &=\frac{n_{1j}\times\frac{1}{n_{1j}}\sum_{y\in W_{1j}}y+n_{2j}\times\frac{1}{n_{2j}}\sum_{y\in W_{2j}}y+\cdots+n_{pj}\times\frac{1}{n_{pj}}\sum_{y\in W_{pj}}y}{n_{1j}+n_{2j}+\cdots+n_{pj}} \\ &=\frac{\sum_{y\in W_{1j}}y+\sum_{y\in W_{2j}}y+\cdots+\sum_{y\in W_{pj}}y}{n_{1j}+n_{2j}+\cdots+n_{pj}}\end{aligned} \tag{2}$$

K-means is used to cluster distributed data centrally to get k clusters, and then cluster center points. c_s $(1\le s\le k)$:

$$c_s=\frac{1}{n_s}\sum_{y\in W_s}y=\frac{1}{n_{1s}+n_{2s}+\cdots+n_{ps}}\left(\sum_{y\in W_{1s}}y+\sum_{y\in W_{2s}}y+\cdots+\sum_{y\in W_{ps}}y\right) \tag{3}$$

so $c_j=c_s$,

The theorem is proved.

Card completed.

4. Complexity Analysis of Algorithms

For any parallel and distributed clustering algorithm, there are two aspects of complexity, namely time complexity and communication complexity Tcomm. In the calculation process, the main calculation step is to calculate the distance of the corresponding central vector of each data point; in the communication process, data, central vector and other related information need to be transmitted from one site to other sites. Firstly, the complexity of Distributed Clustering Algorithm in one repetitive step is analyzed. Let Tdata be the actual travel time of a data item; and Tstart the time required to establish a connection. Because of the parallel execution, it only needs to transmit data once, so the complexity of no step is as follows:

Ttime=Tstart+KTdata

Similarly, the complexity of calculating distance is as follows:

Ttime=KnTdist

In formula: Tdist is the time to calculate the distance of a single data point; n = N/P. Now let T be the number of cycles required by K-means algorithm, then the complexity of the whole algorithm is as follows

Ttime=T｛Tstart+KTdata｝

Tcomm=TKnTdist.

Because of the developed network, the time of resume connection can be neglected. Therefore, the complexity expression of the algorithm in this paper can be written in the following form:

Ttime=TKTdata,Tcomm=TKnTdist.

5. Summary

The K-means algorithm does not depend on the order. Given an initial class distribution, the generated data classification is the same regardless of the order of the sample points. Based on large-scale data operation, it is obvious that K-means on a single computer can not satisfy the data clustering processing, and the continuous iteration will test the operation time. In this paper, K-means is parallelized, which makes the operation time greatly reduced. Here, K-means is discussed.

Acknowledgements

Project of Hubei Natural Science Foundation in 2018. (Project Number:2018CFC876)

References

[1] Kloudas K, Mamede M, Rodrigues R. Pixida: optimizing data parallel jobs in wide-area data analytics [J]. Proceedings of the Vldb Endowment, 2015, 9(2):72-83.

[2] Boubela R N, Kalcher K, Huf W, et al. Big Data Approaches for the Analysis of Large-Scale fMRI Data Using Apache Spark and GPU Processing: A Demonstration on Resting-State fMRI Data from the Human Connectome Project.[J]. Frontiers in Neuroscience, 2016, 9(62):492.

[3] Hayden D S, Chien S, Thompson D R. Using Clustering and Metric Learning to Improve Science Return of Remote Sensed Imagery[J]. Acm Transactions on Intelligent Systems & Technology, 2012, 3(3):1-19.

[4] Seshadri K, Shalinie S M, Kollengode C. Design and evaluation of a parallel algorithm for inferring topic hierarchies[J]. Information Processing & Management, 2015, 51(5):662-676.

[5] Li K, Wei A, Zhuo T, et al. Hadoop Recognition of Biomedical Named Entity Using Conditional Random Fields[J]. IEEE Transactions on Parallel & Distributed Systems, 2015, 26(11):3040-3051.