K-means parallel multi-relational clustering algorithm for spatial data

Huang Suyu, Wang Song

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

Keywords: Clustering algorithm; parallel; relational; spatial data; cluster

Abstract: This paper is a study of parallel clustering algorithm K-means. Firstly, the design idea of K-means clustering algorithm on a single computer is introduced. Secondly, the design idea of K-means clustering algorithm in cluster environment is elaborated in detail. When K-means clustering algorithm is faced with massive data, the complexity of time and space has become the bottleneck of K-means clustering algorithm. On the basis of fully studying the traditional K-Means clustering algorithm, this paper presents the design idea of parallel K-Means clustering algorithm, and gives the estimation formula of its acceleration ratio. The correctness and validity of the algorithm are proved by experiments.

1. Introduction

Data Mining, also known as Knowledge Discovery in Database (KDD), is a process of extracting unknown and potentially valuable information or patterns from a large number of incomplete, noisy, fuzzy and random data. With the rapid development of computer technology and the popularization of network, people have more opportunities to use convenient methods to exchange information with the outside world. However, the influx of data increases the difficulty of obtaining useful information. How to obtain valuable information from a large number of data has brought difficulties to the implementation of data mining system. Because of the high complexity of processing these data, the computing power of the system is difficult to meet the requirements. At this time, the limited computing resources provided by traditional single server often can not meet the requirements, and large-scale parallel computing needs to be realized by means of distributed computing technology. Clustering is an important technology in data mining and an effective means to analyze data and find useful information from it. Based on the idea of clustering, it grouped data objects into several kinds or clusters, which made the objects in the same cluster have high similarity, but the objects in different clusters differ greatly. Through clustering, people can identify dense and sparse regions, and find interesting relationships between global distribution patterns and data attributes. K-means is a basic partition method in clustering analysis. Square error and criterion function are often used as clustering criteria. So we use the distributed clustering method based on HADOOP to improve the efficiency of clustering.

2. Clustering Algorithms

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.

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.

The purpose of clustering algorithm is to obtain the most essential "class" properties that can

  DOI: 10.25236/csam.2019.068

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. Hadoop Platform

Hadoop is a distributed infrastructure developed by the Apache Foundation. Users can develop distributed programs without knowing the details of the distributed infrastructure. Make full use of the power of the cluster for high-speed computing and storage. Hadoop implements a distributed file system (HDFS). HDFS has high fault tolerance and is designed to be deployed on low-cost hardware. And it provides high throughput to access application data, which is suitable for applications with large data sets. HDFS relaxes POSIX requirements to allow streaming access to data in the file system.

Hadoop has many elements. At its bottom is the Hadoop Distributed File System (HDFS), which stores files on all storage nodes in the Hadoop cluster. The upper layer MapReduce engine of HDFS consists of JobTrackers and TaskTrackers.

MapReduce is an efficient distributed programming model and an implementation method for processing and generating large-scale data sets, MapReduce computing.

The workflow of each stage of the model is as follows:

(1) Input: An application based on MapReduce framework of Hadoop platform usually needs a pair of Map and Reduce functions provided by implementing appropriate interfaces or abstract classes, which should also specify the location of input and output and some other running parameters. At this stage, the large data files in the input directory will be divided into several independent data blocks.

(2) Map: MapReduce framework regards application input as a set of < Key, value > key-value pairs. At Map stage, the framework calls user-defined Map functions to process each < Key, value > key-value pair and generates a new set of intermediate < Key, value > key-value pairs. The types of these two groups of key-value pairs may be different.

(3) Shuffle: In order to ensure that the input of Reduce is the ordered output of Map, in the Shuffle phase, the framework obtains < Key, value > key pairs for each Reduce through HTTP, and the MapRedus framework groups the input of Reduce phase according to the Key value, because the output of different Maps may have the same Key.

(4) Reduce: In this stage, the intermediate data is traversed to execute user-defined Reduce function input parameter of < Key, {list of value}> for each unique Key, and the output is a new <

Key, value > key-value pair.

(5) Output: This stage completes a typical MapReduce process by writing the results of Reduce output to the location specified in the output directory.

4. Analysis of K-means Clustering Algorithms

MacQue, a K-means algorithm proposed in 1967, is a classical clustering algorithm widely used in scientific research and industrial applications. The core idea of K-means algorithm is to divide n data objects into K clusters so as to minimize the sum of squares from data points in each cluster to the cluster center.

Input: Number of clustering k, data set containing N data objects.

Output: K clusters.

(1) Select K objects arbitrarily from N data objects as the initial clustering center.

(2) Calculate the distance from each object to each cluster center, and assign the object to the nearest cluster.

(3) After all objects are allocated, the centers of K clusters are recalculated.

(4) Compared with the K clustering centers calculated previously, if the clustering centers change, turn (2) or turn (5).

(5) Output clustering results.

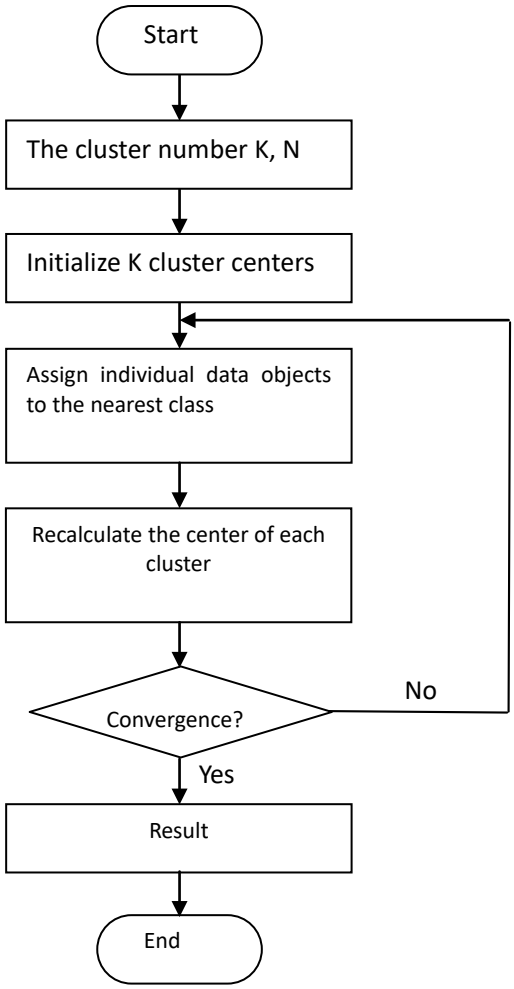

Fig.1 The workflow of the K-means algorithm

The workflow of the K-means algorithm is shown in Figure 1.

Firstly, K objects are randomly selected from N data objects as initial clustering centers, while the remaining objects are assigned to the most similar clusters (represented by clustering centers) according to their similarities (distances) with these clustering centers. Then the cluster centers of each new cluster (the mean values of all objects in the cluster) are calculated. Repeat this process

until the standard measure function begins to converge. Generally, Euclidean distance is calculated. The concrete calculation formula (formula 1) is as follows:

$$d_{(i,j)} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2 + \cdots + (x_{in} - x_{jn})^2} \quad (1)$$

The advantage of K-means algorithm is that it can deal with large data sets. K-means algorithm is relatively scalable and efficient, because its computational complexity is O (n K t), where n is the number of objects, K is the number of clusters, t is the number of iterations, usually T < n, K < n, so its complexity is usually expressed in O (n).

Next, a group of two-dimensional data is taken as an example to illustrate the clustering process of K-means.

Table 1 Two-dimensional data

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x	1	2	2	3	9	10	10	11	15	16	16
y	2	2	5	3	14	13	15	16	6	5	8

The spatial distribution is shown in Figure 2.

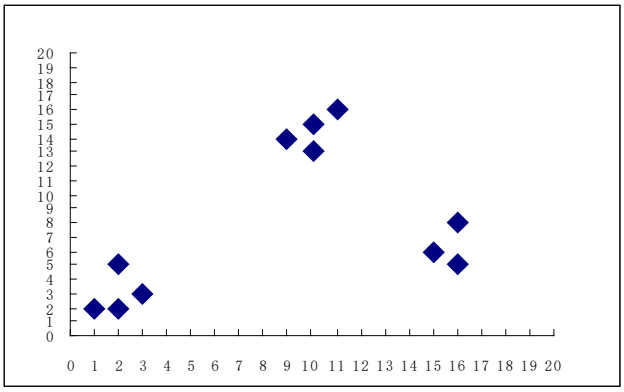

Figure 2 Data distribution

Enter k = 3, KL = x1, K2 = x2, K3 = x3.

At the beginning of the algorithm, the first three data are selected as the initial clustering centers, and the clustering after one iteration is shown in Figure 3.

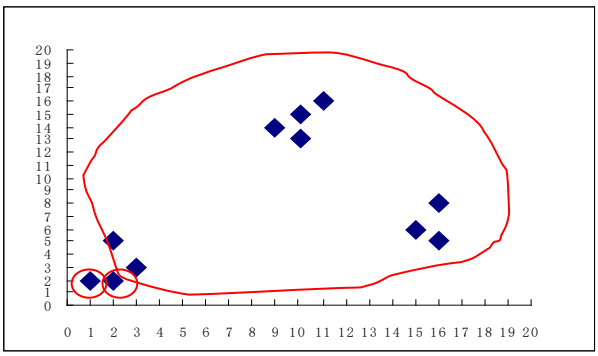

Figure 3 First iteration

After repeated iterations, the final optimal clustering results are shown in Figure 4.

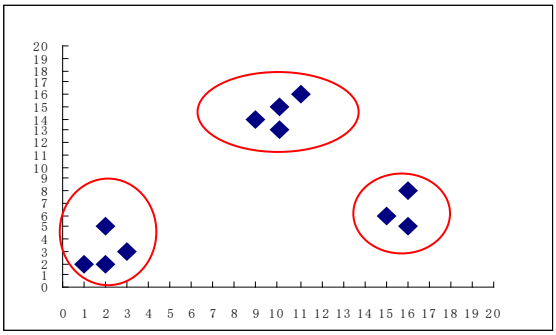

Figure 4 Clustering results

In addition, K-means algorithm does not depend on order. Given an initial class distribution, regardless of the order of sample points, the generated data classification is the same.

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.

5. Summary

The data in this paper run on a single computer. We use Weka as the experimental platform. The full name of Weka is Waikato Environment for Knowledge Analysis. It is a free, non-commercial (corresponding to SPSS's commercial data mining product, Clementine), open source machine learning and data based on JAVA environment. Data Mining software. The experimental data are the same as the theoretical data. Parallel platform data we use virtual machine technology, virtual two computers equipped with Red Hat Enterprise Linux 4.

Acknowledgements

The Youth Natural Science Fund Project of 2017 Wuhan Donghu University. (Project Number: 2017dhzk006).

References

[1] Schadt E E, Edwards S W, Guhathakurta D, et al. A comprehensive transcript index of the human genome generated using microarrays and computational approaches [J]. Genome Biology, 2004, 5(10):R73.

[2] Halkidi M, Vazirgiannis M, Batistakis Y. Quality Scheme Assessment in the Clustering Process[C]// European Conference on Principles of Data Mining & Knowledge Discovery. 2000.

[3] StefanSteiniger, RobertWeibel. Relations among Map Objects in Cartographic Generalization[J]. American Cartographer, 2007, 34(3):175-197.

[4] Abadpour A. A sequential Bayesian alternative to the classical parallel fuzzy clustering model[J]. Information Sciences, 2015, 318(C):28-47.

[5] Albano M, Chessa S, Nidito F, et al. Dealing with Nonuniformity in Data Centric Storage for Wireless Sensor Networks[J]. IEEE Transactions on Parallel & Distributed Systems, 2011, 22(8):1398-1406.

[6] Halkidi M, Vazirgiannis M, Batistakis Y. Quality Scheme Assessment in the Clustering Process[C]// European Conference on Principles of Data Mining & Knowledge Discovery. 2000.

[7] StefanSteiniger, RobertWeibel. Relations among Map Objects in Cartographic Generalization [J]. American Cartographer, 2007, 34(3):175-197.