Research on Clustering Algorithm Based on Big Data

Background

Chunqiong WU¹,²
¹Yango University, FuZhou, FuJian, 350015
²Engineering Research Center of Business Intelligent in Bin Data for Fujian Province, FuZhou, FuJian, 350015

Abstract. In recent years, machine learning has developed rapidly and clustering is an important research direction in the field of machine learning. After continuous and in-depth research, the clustering algorithm for small-scale data sets has made great progress and many effective clustering algorithms have been studied. However, these algorithms are difficult to obtain satisfactory results when dealing with large-scale data sets. The reason is mainly due to the high computational complexity and the weak ability to process high-dimensional data. It is more and more important to process these massive data in time. Therefore, the research on clustering algorithm for large-scale data sets has become one of the important tasks in the field of machine learning. This paper summarizes the meaning of big data and clustering algorithm, sorts out the clustering algorithm suitable for big data and gives suggestions on the research of clustering algorithm under big data. It is hoped that some research achievements can be achieved in the following studies.

1. Introduction

Clustering is a kind of unsupervised machine learning technology, which is used to mine the intrinsic similarity of data and divide the data set into several subsets. Each data subset is a cluster, the samples within the cluster are similar to each other, and the samples between different clusters are not similar. In general, the similarity of samples is characterized by Euclidean distance, Markov distance, Manhattan distance, Pearson distance, Chebyshev distance, cosine similarity, Jaccard similarity and probability density. Clustering techniques have been widely used in real life, such as customer grouping in commercial activities, gene sequence classification in bioinformatics, spam identification in the Internet, and analysis of industry electricity usage behavior in the electricity market.

With the advent of the era of big data, data collection and storage has become relatively easy and convenient. Large-scale data sets of GB-level and even TB-level storage are emerging one after another. The size of data sets of big data is growing at an unimaginable speed, which brings great challenges to data processing. Therefore, clustering research for large data sets is constantly emerging. So far, clustering algorithms for different types of small and medium-sized data sets have made a historic breakthrough in clustering accuracy. However, these algorithms still have many problems when dealing with large data sets. The main defects are high computational complexity and long computing time, which is unacceptable.

2. Overview of big data

Nowadays, big data is becoming more and more popular. There is no unified standard for the definition of big data. In essence, big data contains a variety of meanings and large capacity is its most
basic meaning. It also has some characteristics different from "very large data" and "massive data". So far, after consulting literature, many literatures have defined big data and this paper believes that three definitions are more important.

Architectural definition: According to the National Institute of Standards and Technology of the United States, "Big data refers to the capacity of data, the speed of data acquisition and the presentation of data, which limit the ability to analyze and process data using traditional relational methods and require the use of horizontal extension mechanism to improve the processing efficiency". Moreover, to further subdivide big data, it can be divided into big data science and big data frameworks. Big data science is a research covering big data acquisition, adjustment and evaluation technology; The big data framework is a software library and algorithm for distributed processing and analysis of big data among the clusters of computing units.

Attributive definition: The IDC of the International Data Center was the first to study big data and had a certain impact on big data. In 2011, IDC defined big data for the first time in its report: “Big data technology describes a new era of technologies and systems designed to extract data from large-scale, diverse data through high-speed capture, discovery and analysis techniques. This definition embodies the four main characteristics of big data, namely volume, diversity, velocity and value. It is called the definition of "4Vs" and this definition is widely used.

Comparative definition: In 2011, McKinsey's research report defined big data: “Data sets that go beyond the ability of typical database software tools to capture, store, manage and analysis data.” This definition does not describe the characteristics of big data, etc., and does not explicitly describe the characteristics of big data. This definition contains an evolutionary view of what kind of data set can be considered big data.

This article considers that all big data definitions are acceptable and each definition reflects a specific aspect of big data. Through in-depth research and analysis, this paper believes that big data has the following characteristics. First, the most critical factor in distinguishing between big data and traditional data is the capacity of the data set. For example, Facebook reported that 2.7 billion users logged in and commented every day in 2012. Second, there are three forms of big data: unstructured, semi-structured, and structured. Traditional data is often structured and easy to label and store. Most of the data generated by Facebook, Twitter and other users is now unstructured and there are certain difficulties in storage. Third, the speed at which big data is generated makes high demands on the rate of analysis and processing of data sets. For applications with high real-time requirements, such as fraud detection and RFID data management, big data enters the enterprise as a stream and needs to process the data as quickly as possible. Finally, the use of a large number of data mining methods for large data sets can extract important values from massive data of low value density. Figures 1 and 2 show the hierarchical architecture of big data and the key steps in big data processing.

![Figure 1. Layered architecture of big data system](image)
3. Research on Clustering Algorithm of Big Data Set

There are two types of execution of the algorithm: serial processing and parallel processing. Serial refers to the program being executed sequentially on a single actuator. Parallel refers to the simultaneous execution of a program on multiple processors. This paper believes that parallelism is more suitable for the processing of big data for the characteristics of big data sets, serial and parallel. Therefore, this paper mainly summarizes the clustering algorithms suitable for parallel and gives suggestions for the current research situation of the algorithm. The essence of parallel algorithm is to divide and conquer big data. Under the distributed computing framework, the training of clustering algorithm is completed by different computing paradigms. This paper mainly analyzes the clustering algorithm based on MapReduce computing framework, the clustering algorithm based on Spark computing framework and the clustering algorithm based on Storm computing framework, as follows.

3.1 Clustering algorithm based on MapReduce computing framework

MapReduce is a parallel computing model proposed by Google. It is oriented to large-scale data sets. The core of the computing model is to divide the data and schedule the computing tasks, namely Hadoop distributed file system (Hadoop distributed file system, HDFS) automatically divides a large data to be processed into multiple data blocks, each data block corresponds to one computing task and automatically schedules the computing nodes (Map nodes or Reduce nodes) to process the corresponding data blocks and simultaneously. The execution status is monitored and is responsible for the synchronization control performed by the Map node. For the same type of clustering algorithm, the main difference between different parallelization strategies in the MapReduce framework is reflected
in the design of the Map function and the Reduce function. The following is a brief introduction to the MapReduce-based k-means algorithm.

The k-means algorithm is a typical representative of clustering algorithms. In 2009, ZHAO and other first base MapReduce framework designs realized parallel k-means clustering algorithm PK-Means (parallel k-means). The simulation results show that PK-Means has the ability to deal with 8 GB large-scale data clustering problems on a cluster with 4 compute nodes. In 2011, Jiang Xiaoping designed a parallel k-means clustering algorithm based on 10 computational nodes based on the MapReduce framework. In 2014, CUI et al. designed a k-means clustering algorithm based on MapReduce framework to eliminate iterative correlation. On a cluster system with 16 compute nodes, it was verified that the algorithm has 4,296,075,259 9-dimensional samples. It can be seen that the MapReduce-based k-means algorithm has achieved good results in processing large-scale data sets.

3.2 Clustering algorithm based on Spark computing framework

Apache Spark is an open source cluster computing framework based on memory computing. It has the advantages of MapReduce, but it is better than MapReduce. The intermediate output can be stored in memory, so there is no need to read and write HDFS. Therefore, Spark is more suitable for data mining and machine learning algorithms that require iteration. Although Spark was launched late, it has been widely recognized as a big data computing engine because of its unique advantages in dealing with big data computing problems. For the research progress of big data calculation and analysis under the Spark framework, please refer to the literature. The representative work based on the Spark framework clustering algorithm is given below: The fuzzy c-means algorithm based on Spark.

The earliest implementation of the algorithm is Liang Peng. In 2015, Liang Peng and others continued to research and finally realized the scalable nucleation fuzzy c-means algorithm scalable KFCM under the Spark framework. The idea of kernel function is introduced into the algorithm. After mapping the samples of the original space to the high-dimensional space and then dividing the data, the clustering ability of the parallel fuzzy c-means algorithm is significantly improved. In 2016, Wang Guilan sent an article on parallel fuzzy c-means algorithm based on custom cache sensitive matrix operation in Spark framework. The algorithm designed in this paper makes the data movement between computing nodes be reduced, making the clustering algorithm the linear relationship between performance and data volume is guaranteed. BHARILL et al. also designed an iteratively optimized fuzzy c-means algorithm SRSIO-FCM based on scalable random sampling under the Spark framework.

3.3 Clustering algorithm based on Storm computing framework

Apache Storm is a high-fault, distributed, high-real-time big data processing system with a high degree of focus on streaming data processing. Therefore, it is also called Hadoop in real-time processing. The input stream of the Storm cluster is managed by the spout component, which sends the data to the bolt. If Bolt does not save the data to memory, it will pass the data to other bolts. The Storm cluster converts the data sent by spout between multiple bolts. A Storm cluster consists of two types of nodes: one master node and multiple working nodes. The master node is responsible for distributing tasks, and the worker nodes are responsible for performing tasks.

4. Suggestions for improvement of clustering algorithm

Based on the above analysis, cluster-based distributed computing has become the main method for dealing with big data clustering problems. This is the embodiment of the strategy of “divide and conquer” big data. Although the above algorithms have certain processing power in the processing of big data sets, these algorithms also have obvious defects. These algorithms need to be improved. These defects are mainly manifested in: (1) The data is not properly divided; (2) too much dependence on the parallelism of the algorithm; (3) insufficient automation; (4) insufficient processing of the data
in time, unable to obtain important value from the data in time. In response to these problems, this paper proposes the following four suggestions:

(a) The division of data requires reasonable planning to improve the feasibility of the algorithm. Through statistical analysis of the data, the regularity of the data is obtained, so that the data can be more scientifically and reasonably divided.

(b) Data parallelism replaces algorithm parallelism. Although MapReduce and Spark computing frameworks have good big data processing capabilities, but each has its own shortcomings, MapReduce handles the performance of iterative calculations poorly. Spark's memory is calculated to the limit of computer memory capacity. Therefore, you can explore the strategy of using data parallelism instead of algorithm parallelism.

(c) Automation of clustering algorithm training. Minimizing human intervention in the clustering process is one of the keys to dealing with big data clustering problems. Because in the context of big data, for the massive data generated in practical applications, in many cases we do not know what kind of information is contained in the data, which corresponds to the diversity and value of the 4V characteristics of big data.

(d) The ability to capture the value of data in real time, especially for data that requires high real-time performance. In the final analysis, it is caused by the above three points. After solving the above three points, the real-time performance of the data will be greatly improved.

5. Conclusion
After continuous research, some research results have been achieved in big data, such as big data search, big data storage, big data mining, etc., but still cannot meet the needs of current big data. Researching real-time, highly robust new high-efficiency clustering algorithms for big data has become a key task to be solved in the deep exploration of the hidden value of big data.

In the field of data mining, the final result of many clustering algorithms is sensitive to the correct setting of parameters, which leads to these algorithms far from being called mature and practical intelligent machine learning algorithms. In the big data environment, it is necessary to study and design a more efficient intelligent automatic clustering algorithm. Therefore, the clustering algorithm for big data needs constant research to meet the needs of current big data.

Acknowledgements
Fund Project: Engineering Research Center of Business Intelligent in Bin Data for Fujian Province

References
[1] Luo Enzhen, Wang Guojun, Li Chaoliang. Research on clustering algorithm for multidimensional data deduplication in big data environment[J]. Small Computer Systems, 2016, 37 (3) : 438-442.
[2] Lin Qingxin. Exploring the application of K-means clustering algorithm under big data in network security detection[J]. Network Security Technology and Application, 2017, (3): 92-93.
[3] JI Lianghao. Research on clustering algorithm based on density biased sampling [J] Journal of Chongqing Uni-versity of Posts and Telecommunications Natural Science Edition, 2007, 19( 6) : 729-732.
[4] Li Xuelong, Gong Haigang. Review of Big Data Systems[J]. Chinese Science: Information Science, 2015, 45(1) : 1–44
[5] Zhou Runwu, et al. Parallel optimal sampling clustering K-means algorithm for big data processing [J]. Computer Applications, 2016, 36( 2) : 311 - 315.
[6] Qu Jie. Research on Big Data Intelligent Parallel Clustering Method in Virtual Environment[J]. Journal of Computer Measurement and Control, 2017, 25 (6) : 257-260.
[7] ZHAO Weizhong, MA Huifang, HE Qing. Parallel k-means clustering based on MapReduce [C] // Proceedings of the 1st International Conference on Cloud Computing. Beijing: Springer-Verlag, 2009: 674-679.

[8] IANG Xiaoping, LI Chenghua, XIANG Wen, et al. Parallel implementing k-means clustering algorithm using Map Reduce programming mode[J]. Journal of Huazhong University of Science and Technology Natural Science Edition. 2011, 39( s1) : 120-124.

[9] MENG Xiangrui, BRANDLEY J, YAVUZ B, et al. MLlib: machine learning in apache Spark[J]. Journal of Machine Learning Research, 2016, 17( 1) : 1235-1241.

[10] LIANG Peng. Research on spark oriented fuzzy c-means clustering algorithm [D]. Harbin: Harbin Institute of Technology, 2015.

[11] WANG Guilan, ZHOU Guoliang, SA Churila, et al. Parallel fuzzy c-means clustering algorithm in Spark[J]. Journal of Computer Applications, 2016, 36 ( 2 ) : 342-347.