Research on AI and big data application framework: Mahout

Ren chen

Renchen School of Economics and Bussiness Adminstration , Hefei University

renchenwork@163.com

Abstract

Neural network, from a single input layer to the middle layer and then to the output layer, each layer compares the predicted value with the expected value. Through continuous data training, we can finally get a model that meets the expectations. Big data is a tool to provide "materials" for the algorithm. The data on the actual network is miscellaneous, and not all things are useful information. The data required for training is only a small part of the vast data. At this time, it is necessary to filter, classify and integrate the data, The final data obtained through big data is the data suitable for machine learning or neural network training. The unique data characteristics of big data and the superior performance of neural network determine that the two complement each other. On the one hand, big data realizes data value mining through the excellent analysis ability of neural network; On the other hand, neural network makes full use of massive heterogeneous data for training and learning. This paper combines big data and neural network, and expounds the use of big data AI framework Mahout framework [1].

Keywords*: big data analysis; machine learning; K-Means; Hadoop; Mahout;*

1. Introduction

The rapid development of information technology has given birth to another disruptive technological change in the IT industry after the Internet, Internet of things and cloud computing: *big data*. Since its birth in 1997, the term "big data" has been the focus of attention from all walks of life. Its huge scientific research value and public service value need to be recognized, developed and utilized. The research on neural network algorithm based on big data is of great significance.

However, the current machine learning samples are mostly small samples, and the prediction model is not ideal to a great extent. In the process of learning optimization model, machine learning based on large samples requires a lot of time and hardware cost, resulting in low research efficiency. For the current machine learning problem, this paper quotes the Apache mahout framework and studies how this framework combines machine learning with big data to optimize the model learning process.

2. Background

2.1 Hadoop

2.1.1 The core of Hadoop framework

The core of Hadoop framework[2] is as follows:

Hadoop common[3]: provides infrastructure for other Hadoop modules.

HDFS[3]: a highly reliable and high-throughput distributed file system. The framework is shown as Figure 1.

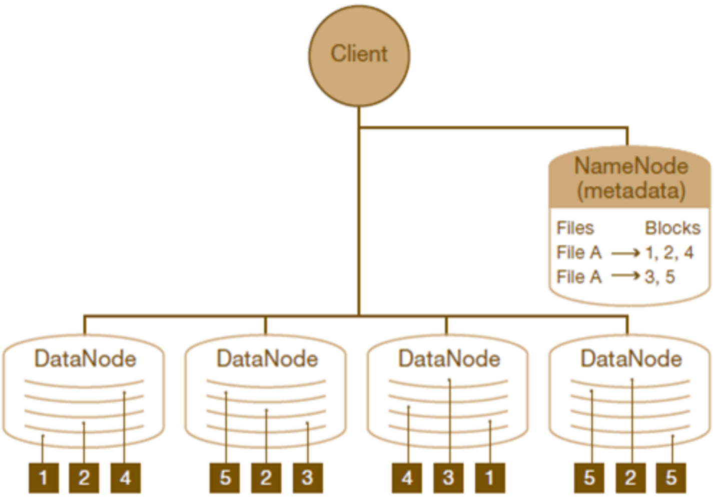


Figure 1. HDFS framework

MapReduce[4]: a distributed offline and computing framework - distributed computing framework.

Yarn[5]: a new MapReduce framework, task scheduling and resource management distributed resource management framework.

The characteristics of Hadoop framework is shown as Figure 2.

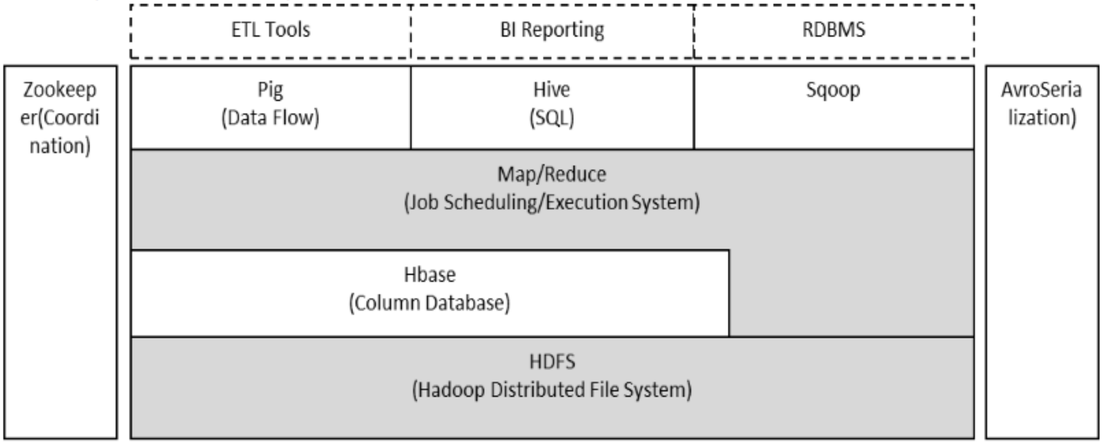


Figure 2. Hadoop framework

2.1.2 Hadoop five nodes:

NameNode (management node):

NameNode manages the command space of the file system. *NameNode* records the location information of the data node where each block in each file is located, but it does not persist and store this information, because this information will be reconstructed from the data node when the system starts.

DataNode:

They are the working nodes of the file system. They store and retrieve data according to the scheduling of the client or the *NameNode*, and regularly send the list of blocks they store to the *NameNode*.

Secondary-NameNode:

The *Secondary-Namenode* is a secondary daemon used to monitor HDFS status. Like *NameNode*, each cluster has a *Secondary-NameNode* and is deployed on a separate server. Unlike *NameNode*, *Secondary-NameNode* does not accept or record any realtime data changes. However, it communicates with *NameNode* to save snapshots of *HDFS* metadata regularly.

ResourceManager(RM):

RM is resource management. In yarn, the resource manager is responsible for the unified management and allocation of all resources in the cluster. It receives resource reporting information from each node (*NodeManager*) and allocates these information to each application (actually application manager) according to certain policies.

NodeManager(NM):

Nm is the agent of *ResourceManager* on the slave machine. It is responsible for container management, monitoring their resource usage, and providing resource usage reports to the scheduler of *ResourceManager*.

2.2.Mahout

Mahout is an open source machine learning software package under Apache. The currently implemented machine learning algorithm mainly includes three parts: collaborative filtering or recommendation engine, clustering and classification. *Mahout* aims to establish an extensible machine learning software package from the beginning of design to deal with the problem of big data machine learning. When the amount of data being studied is too large to run on one machine, *Mahout* can be used to analyze the data in the Hadoop cluster. Some parts of mahout are implemented directly on Hadoop, which makes it capable of big data processing, which is also mahout's biggest advantage. *Mahout* only provides the library of machine learning, not the graphical user interface, and mahout does not include all machine learning algorithm implementations, which can be regarded as a disadvantage. However, mahout is not "another machine learning software", but to become an "extensible machine learning software for processing big data", Therefore, more and more machine learning algorithms will be implemented on *Mahout*. The following Figure 3 is the framework diagram of the combination of Mahout and Hadoop.

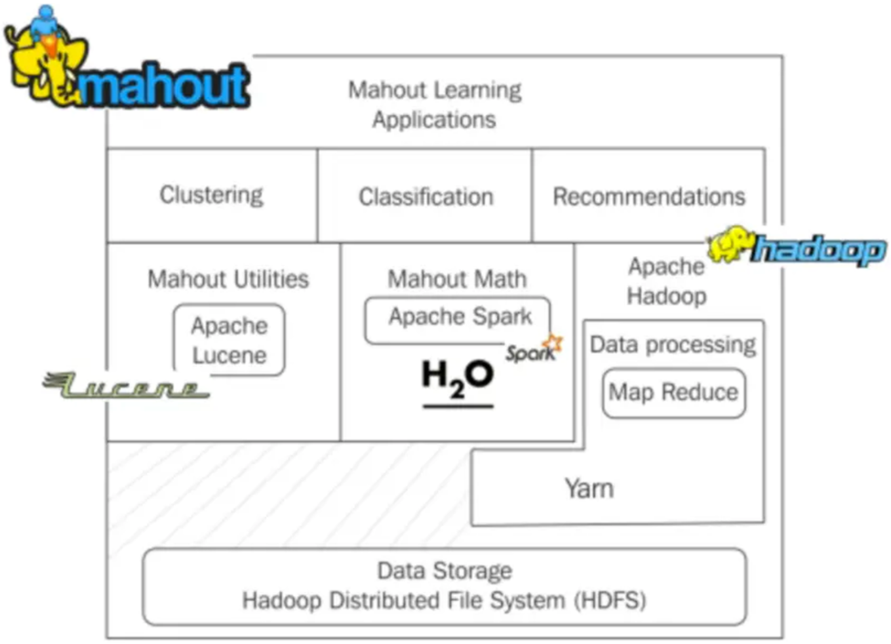


Figure 3. the framework of Mahout and Hadoop

2.3.K-Means

In the *K-Means* algorithm, *K-Means* that the cluster is *K* clusters, and *Means* means that the mean value of the data in each cluster is taken as the center of the cluster, that is, the centroid of each class is used to describe the cluster. The principle of *K-means* algorithm is relatively simple, but it has defects, that is, it may converge to the local optimal solution (the local optimal effect is not as good as the global optimal effect), and the convergence speed is relatively slow on large-scale data sets. In other words, *K-means* algorithm is a locally optimal iterative algorithm affected by the initial value.

The specific process of *K-means* algorithm is to calculate the center point closest to all the center points for each data point, and then classify this point as the cluster represented by this center point. After an iteration, recalculate the center point for each cluster class, and then find the nearest center point for each point. In this way, the cluster class does not change until the first and second iterations. The simple process of K-Means is shown as Figure 4.

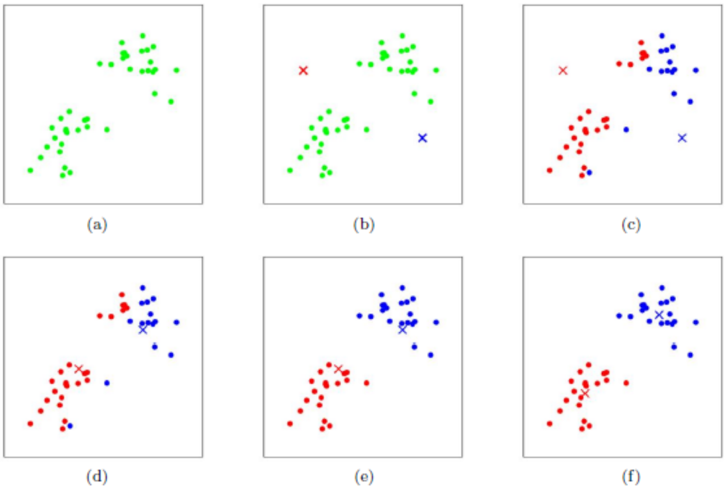


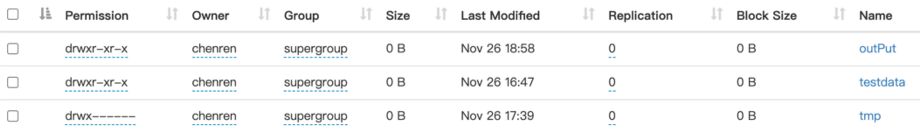
Figure 4. the process of the algorithm of K-Means

3. Dataset

The following dataset was donated by *Tom Brijs* and contains the (anonymized) retail market basket data from an anonymous Belgian retail store. More details can be found in supporting information. The data used in the experiment are available at https://github.com/RCwukaka/hadoop-retails-dataset/blob/master/retail.dat

4. Methods and Results

The detail of System deployment process is shown at https://github.com/RCwukaka/hadoop-retails-dataset/blob/master/README.md. After deploy, Hadoop file distribution is shown at Figure 5



□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	chenren	supergroup	0 B	Nov 26 18:58	0	0 B	outPut
□	drwxr-xr-x	chenren	supergroup	0 B	Nov 26 16:47	0	0 B	testdata
□	drwx------	chenren	supergroup	0 B	Nov 26 17:39	0	0 B	tmp

Figure 5. Hadoop file distribution

In the mahout-0.9 package, there is an integrated *K-Means* algorithm package. This research use integrated algorithm analyse the dataset.

By K-Means algorithm analysed, the results are saved in retail_kmeans.txt. the results are available at github. In this result, we can get that *Mahout* and *Hadoop* can analyse big data.

5. Conclusion

Mahout is an algorithm library that integrates many algorithms and provides the implementation of some scalable classical algorithms in the field of machine learning. Mahout can be effectively extended to Hadoop clusters.

6. Limitation and Future Research

In machine learning, Java is also a good choice because Java is easy to debug. Mallet, deep learning4, Weka and MOA are the most used java libraries for machine learning.

Although the syntax of Python is simpler, it is a fully mature general-purpose programming language. For this reason, a large number of machine learning and artificial intelligence are implemented in Python, which has a code base and a huge ecosystem. Of course, pytorch have distribution-framework for neural network. The package of torch.distributed provide this function.

So, in the future, we will focus on how to combine Python and Hadoop for big data DP analysis.

7. Reference

[1] Sean Owen etc., Mahout in Action, Manning Publications, 2011

[2] Garry Turkington. Hadoop 基础教程[M]. 张治起译. 人民邮电出版社 第 1 版, 2014.

[3] 蔡斌, 陈湘萍. Hadoop 技术内幕：深入解析 Hadoop Common 和 HDFS 架构设计与实现原理[M]. 机械工业出版社, 2013.

[4] 董西成. Hadoop 技术内幕：深入解析 MapReduce 架构设计与实现原理[M]. 机械工业出版社, 2013.

[5] 董西成. Hadoop 技术内幕：深入解析 YARN 架构设计与实现原理[M]. 机械工业出版社, 2013.