Decision Tree Algorithm for Big Data Analysis

Yan Hou
Qilu Normal University, Jinan, Shandong, 250014, China
122365871@qq.com

Abstract. With the development of the times, the amount of data processing is increasing, which brings new research challenges to data storage and processing data, but for traditional decision tree algorithms, it is far from meeting current data processing. Requirements, therefore, put forward new requirements for the decision tree algorithm. This paper expounds the decision tree algorithm, introduces the algorithm platform briefly, and finally forecasts the development direction of the decision tree algorithm.

Keywords: big data; analysis; decision tree; algorithm; platform.

1. Basic Concept of Decision Tree

Decision tree [1] is usually a predictive model of tree structure in machine learning [2]. Typically, its internal nodes test a property [3], but leaf nodes represent the final category. This model can solve many basic problems [4], such as optimization problem, multi-stage decision problem and so on. It also restores the decision-making process. In addition, it can also disassemble complex processes into many simple decisions [5], and then clearly explain the whole decision-making process.

Decision tree is a tree structure used to classify instances. A decision tree consists of nodes and directed edges. There are two types of nodes: internal nodes and leaf nodes. Where an internal node represents a test condition for a feature or attribute (used to separate records with different characteristics), a leaf node represents a classification. Once we have constructed a decision tree model, it will be very easy to classify on its basis. To do this, start with the root node, test a feature of the local instance, and assign the instance to its children (that is, select the appropriate branch) according to the test structure. When this branch may reach a leaf node or another internal node, the new test condition is used to recursively execute until it reaches a leaf node. When we reach the leaf node, we get the final classification result.

In the field related to machine learning and data mining [6], machine learning algorithm [7] is usually difficult to understand the calculation process, which basically fails to meet such demands. But the model of decision tree [8] is the appropriate scheme. Decision tree algorithm is one of the machine algorithms which is often adopted. The technology of decision tree is developing and many algorithms are derived.

The Classification task is to determine which predefined target class an object belongs to. Classification is not only a universal problem, but also the basis of other more complex decision-making problems. It is also the largest algorithm family in machine learning and data mining technology. Many of the algorithms we introduced earlier (such as SVM, naive Bayes, etc.) can be used to solve classification problems. As we begin this article, let's briefly review what classification is. Suppose we now have a feature set, as shown in the table below, that collects the symptoms of several patients and the corresponding symptoms. Symptoms include the degree of headache, cough, body temperature and sore throat, and the combination of these features corresponds to the classification of a disease.

The essence of the classification problem is that given such a data set, it requires us to train (or establish) a model. When a new set of eigen vectors emerges, we are asked to predict (or judge) which category an object with such a set of eigen vectors should belong to. In terms of we now given example, suppose you are a doctor, now admitted a new patient, and then through visits you know his symptoms (including the degree of the degree of headache, cough, body temperature and throat is sore), then you will be according to you have established a good model to judge whether the patients have to common cold or influenza. The number of categories of classification problems can be two or more. Binary classification problem is the simplest classification problem, and multiple classification problem model can be built on the basis of the binary classification model. The iris data set that we have been using in previous articles is a typical multiclassification problem, and the ultimate goal of the problem is to determine which of a given flower should fall into the category of Sentosa, versicolor, and virginica.

2. Big Data Challenges

With the development of technology, information can be digitized into data and processed and analyzed by computer. In some fields, such as the Internet, finance and medicine, many data sets can generate many records every day. In addition, products such as smart devices allocate multiple sensors, which in turn generate large amounts of data. And the development of network makes data storage more convenient. Many industries conduct more and more analysis and processing of information in order to obtain useful information from big data, so as to mine the useful information and adopt it. Therefore, the new development of big data analysis algorithm becomes the current focus. Figure 1 shows the system framework.

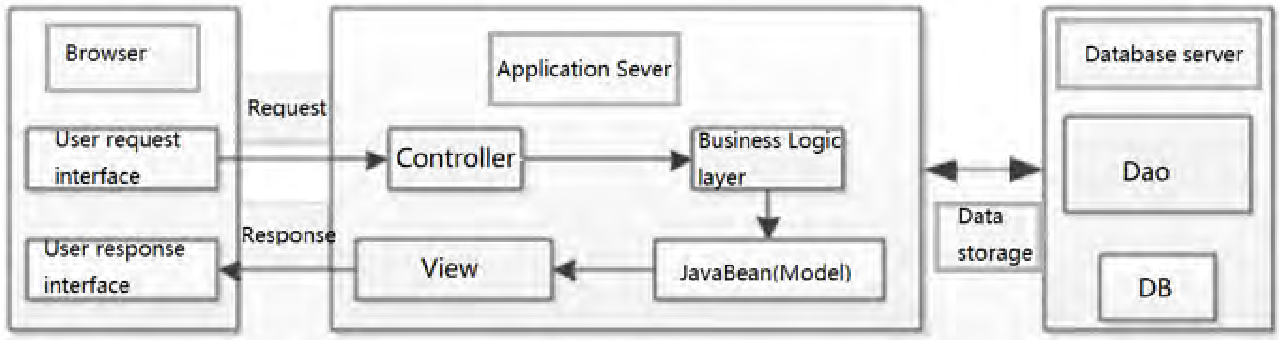

Figure 1. System framework

For the algorithm of decision tree, the features of big data acquisition contain a lot of quantity, there is a lot of redundancy, and even some low-quality or irrelevant feature values. The processing of them not only consumes a lot of computing resources, but also results in the oversize tree structure. It also affects the accuracy of its predictions. Therefore, it is an important means to improve the efficiency and accuracy of machine learning to screen the features of data sets and find out the features that meet the requirements.

The high sample capacity of big data itself makes the decision tree algorithm unable to be built in memory, in addition, it will consume a lot of operation time. At the same time, the problem is how to deal with the data efficiently. It is reasonable to adopt cluster storage and bandwidth resources and reduce the computing load by task parallelism. Then how to carry out parallel computation of decision tree algorithm in machine learning can be used as a feasible implementation method of decision tree algorithm.

The mining of big data is a process of discovering valuable and potentially useful information and knowledge hidden in massive, incomplete, noisy, vague and random large databases, which is also a decision support process. It is mainly based on artificial intelligence, machine learning, pattern learning, statistics, etc. The mining of big data is a process of discovering valuable and potentially useful information and knowledge hidden in massive, incomplete, noisy, vague and random large databases, which is also a decision support process. It is mainly based on artificial intelligence, machine learning, pattern learning, statistics, etc.

Classification is to find out the common features of a group of data objects in the database and classify them into different classes according to the classification pattern. The purpose of classification is to map the data items in the database into a given category through the classification model. It can be applied to the application classification and trend prediction. For example, Taobao stores divide the purchase situation of users into different categories in a period of time, and recommend related commodities to users according to the situation, thus increasing the sales volume of stores. Many algorithms can be used for classification, such as decision tree, KNN, Bayes, etc.

The regression analysis reflects the property of the attribute value of the data in the database, and the dependence between the attribute values is found by the function expressing the relation of the data map. It can be applied to the study of the prediction and correlation of data sequences. In marketing, regression analysis can be applied to all aspects. For example, through the regression analysis of the sales in this quarter, the sales trend in the next quarter is predicted and targeted marketing changes are made. Common Regression algorithm include: Least squares method (Ordinary further Square), Logistic Regression, Logistic Regression), step by step Regression (Stepwise Regression), Multivariate Adaptive Regression Splines as well as the Locally Estimated Scatterplot Smoothing.

Clustering is similar to classification, but different from the purpose of classification, it is to classify a group of data into several categories according to the similarity and difference of data. Data belonging to the same category are very similar, but data similarity between different categories is very small, data correlation across classes is very low. Common clustering algorithms include k-means algorithm and Expectation Maximization algorithm (EM).

Association rules are hidden associations or interrelationships between data items, that is, you can derive the occurrence of other data items based on the occurrence of one data item. The mining process of association rules mainly includes two stages: the first stage is to find all the high frequency project team from the massive raw data; The second extreme is to generate association rules from these high frequency project teams. Association rule mining technology has been widely used in financial industry enterprises to predict the needs of customers. Banks improve their marketing by bundling information that customers may be interested in for users to understand and obtain. Common algorithms include Apriorism algorithm and Eclat algorithm.

As an advanced artificial intelligence technology, neural network is very suitable for dealing with non-linear problems and processing problems characterized by fuzzy, incomplete and imprecise knowledge or data due to its self-processing, distributed storage and high fault tolerance. Typical neural network models are mainly divided into three categories: the first type is a feedforward neural network model used for classification prediction and pattern recognition, which is mainly represented by functional network and perceptron. The second type is the feedback neural network model for associative memory and optimization algorithm, represented by the discrete model and continuous model of Hopfield. The third category is the self-organizing mapping method for clustering, represented by the ART model. Although the neural network has many models and algorithms, there are no uniform rules for which models and algorithms are used in data mining in specific fields, and it is difficult for people to understand the learning and decision-making process of the network.

3. Algorithm Optimization of Decision Tree

The optimization of feature algorithm is mainly classified in the original feature set to form a new subset, and finally processed and analyzed through the algorithm. This method is easy to extend and apply in practice. This selection algorithm can be divided into two categories, mainly filter and wrapper. The filter is mainly an algorithm to measure the useful information within the collection. It is a preprocessing process for subsequent classification, and the information is evaluated and screened through the correlation coefficient, sample distance and other relevant indicators.

Because of the relatively large size of the data set, when the memory calculation, not all the data can be processed at once. Data often needs to be placed on disk temporarily. Because the decision tree algorithm usually needs to read and write the data, so its data volume is very large, which makes its reading and writing speed very slow. This requires optimization of the decision tree construction algorithm to reduce data read and write operations. This has become an important optimization direction of decision tree algorithm. SLIQ algorithm is one of the optimization algorithms, which mainly uses the form of breadth priority and pre-ordering to reduce the number of reads and writes, so as to improve the efficiency.

For distributed algorithms, previously C4.5 was mainly extended through the framework of ccNUMA, which enhanced the processor's ability to read data and thus improve its running speed. It was an early distributed algorithm. Google then came up with PLANET, a scalable distributed framework. The controller is its core, regulating the construction of the whole tree. It is usually used to train the tree model of big data. The controller can also be applied to compute clusters, and it can be distributed through MapReduce. In addition, the decision tree model of integrated learning can also be solved by distributed methods.

Decision tree algorithm is a method to approximate discrete function values. It is a typical classification method, which firstly processes the data, generates readable rules and decision trees by the induction algorithm, and then analyzes the new data by the decision. Essentially, decision tree is the process of classifying data through a series of rules. The decision tree method first appeared in the 1960s to the end of 1970s. ID3 algorithm is proposed by J Ross Quinlan, which aims to reduce the depth of the tree. But the number of leaves was neglected. The C4.5 algorithm is improved on the basis of ID3 algorithm, and a great improvement is made on the missing value processing, pruning technology and derived rules of the predicted variables, which are suitable for both classification problems and regression problems.

Decision tree algorithm constructs decision tree to discover the classification rules contained in data. How to construct decision tree with high precision and small scale is the core content of decision tree algorithm.

Decision tree construction can be done in two steps. The first step, generation of decision tree: the process of generating decision tree by training sample set. In general, the training sample data set is a data set that has a history and a certain degree of comprehensiveness according to actual needs and is used for data analysis and processing. The second step, the decision tree cutting technology: the decision tree pruning is decision tree of on one phase generated test, calibration and repair process, main is to use the new sample data in a data set (referred to as test data sets) check occurring in the process of decision tree to generate preliminary rules, will that affect the prediction accuracy of equilibrium branch off.

For the optimization algorithm, if it is to be the source of flow algorithm, VFDT needs to use the inequality form of Hoeffding to save and count the corresponding information of any leaf stage, and replace the inter-decision part with the node, and finally a decision tree can be formed. When new data is input, the data is usually classified at the node, and the statistical information is updated timely. Since VFDT only processes data once, the time cost is reduced, but its disadvantages are also very obvious. The main problem is that it cannot process continuous values, leading to concept drift and low accuracy. Along with the progress of The Times, the VFDT algorithm is also constantly changing. The new VFDTc algorithm can process the numerical data more effectively and improve the accuracy of prediction.

4. Service Platform

Apache software foundation is a non-profit organization, mainly used to support open source software projects. Many machine learning platforms in different companies or communities in the initial development stage will eventually become its projects, such as Spark, Storm and so on. As a relatively early infrastructure, Hadoop can be developed and distributed at the upper level by users, which USES a large number of clusters for computing and storage. In particular, earlier versions were unable to implement interfaces in other languages.

The open source clustered computer system is available with Apache Spark, which centers on the MLlib machine learning library, is a commonly used machine learning algorithm, and also has the functions of data calculation, testing and generation. At the same time, under the developer's research, this kind of decision tree is introduced into Spark, which makes the algorithm develop continuously.

Apache Storm can simplify the mechanism, and the Storm is free. Storm also offers online machine learning. Because of the multi-stream development of data, and the continuous development of decision tree algorithm, Storm is a good choice because of its real-time character.

5. Future Development

First, enhance the quality of the data attributes. In the study of big data, some attributes of many information and data may be missing, and the absence of attributes has a profound impact on the study of decision tree algorithm, which is most likely to lead to inaccurate or wrong research results. How to deal with the loss of these attributes will become a research topic in machine learning and the development direction of this algorithm.

Second, change and control the sample proportion of different categories. When data analysis is processed, the number of samples may be too different due to different categories, which may cause some samples to be ignored by the system when using decision tree algorithm for data classification. Therefore, how to control the classification of small amounts of data is also the direction of model development.

Thirdly, the study of the new decision-making model. Because the law of data can change with time, new data cannot match the original parameters well, and its model also changes with time, so it is difficult to describe the data better. Therefore, how to update and make decisions based on the changing trend of data becomes the trend of future algorithm model development.

Finally, information is processed and used effectively. In some scenarios, large amounts of data are generated in a short time, such as transaction records on e-commerce websites, media data of social networks, and People's Daily call records. A large amount of data cannot be stored completely, but only observed for a limited time, and the processing capacity is lower than the data rate. How to process and use information efficiently is also the research direction of the algorithm.

References

[1]. Pang Zhenda. Research on application of data collection and data mining in mobile communication network [D]. Beijing Jiaotong university,2018.

[2]. Zhang Fayang. Research on classification mining algorithm of stream data based on STORM [D]. Nanjing university of posts and telecommunications,2016.

[3]. Hou Liduo. Application study of decision tree algorithm in engineering quality supervision decision support system [D]. Guizhou university,2016.

[4]. Li Mingyue. Application of decision tree algorithm in bank telemarketing [D]. Huazhong university of science and technology,2016.

[5]. Dai Yanli. Analysis of decision tree algorithm in data mining and its application [J]. Science and technology communication,2015,7(23):33-34.

[6]. Long Zhiyong. Optimization and application of decision tree algorithm based on parallelization [D]. Zhejiang university,2015.

[7]. Du Liying. Analysis of decision tree algorithm based on data mining [J]. Journal of Jilin institute of architectural engineering,2014,31(05):48-50.

[8]. Li Wei. Application and parallel study of decision tree algorithm [D]. University of electronic science and technology,2014.