Research on Classification Algorithm in Data Mining

Liu Caili

Xi' An International University , Institute Of Technology, Department Of Computer Science And Technology, Shanxi Xi 'An 710077, China

Keywords: Classification Algorithm, Data Mining

Abstract. With the continuous development of information technology and computer industry, data processing has become a top priority. We want to do a good job of data processing, it is necessary to apply to the data classification algorithm, which as a key technology in data mining can be a good job to complete the data processing. In this paper, by comparing several different data classification algorithms, to find their similarities and differences to further promote the data classification algorithm to lay the foundation

Introduction

In the context of the current era, many industries have introduced a large data mining concept, which not only to the computer industry has brought development opportunities, but also brought challenges. Because you want to do a good job of large data mining related work, we must master the data classification algorithm, and data classification algorithm can be regarded as a data mining in a difficult. With the deepening of data analysis, people have developed a variety of classification algorithms to continuously reduce its difficulty. Usually based on the data classifier as the benchmark, the corresponding data classification, including the decision tree, Bayes class, based on the association rules and the use of database technology category, this article will be a brief description of them.

The Concept of Data Mining

Data mining is an interdisciplinary field that is influenced by multiple disciplines, including database systems, statistics, and machine learning, visualization, and information science. Data mining is essentially a decision support process. The main technical means are statistical methods, including mathematical statistics, multivariate statistical methods, and econometrics and time series analysis methods. In addition, operational research, artificial neural network and expert system technology development, but also for data mining provides a new way of thinking. Its main feature is a high degree of automatic analysis of the original data, inductive reasoning, from which to explore the potential model to predict the behavior of customers to help decision makers make the right decisions. DM technology is based on statistics and artificial intelligence. Artificial intelligence is a way to solve the real world problems by means of automata as a means of simulating human macroscopic explicit thinking behavior. Artificial intelligence goals are very high, in addition to the need for complex algorithms, but also need a specific system. But DM only uses artificial intelligence in some of the already mature algorithms and techniques.

The Introduction of Classification Algorithm

In the knowledge discovery, classification technology come out on top, is the first choice of technology KDD. Classification is essentially the task of data analysis, a large number of disorganized data in accordance with certain rules. In data mining, there are many classification algorithms, the specific choice of which method, according to the actual situation. Commonly used classification algorithms are: decision tree, classification based on association rules, Bayesian belief

  DOI: 10.35532/JCES.V1.012

network, backward propagation, support vector machine and so on. There are two processes of classification: the first process is to form a classifier; the second process is to classify the data through the classifier.

Build the Model. The formation of a classifier as the first stage of classification, to create a classification model, used to describe the pre-defined sample data set, which is a learning process, this process is called the learning stage. The classification model is formed by first extracting the sample dataset in the database, forming the model according to the sample dataset, and the sample data set is also called the training tuple, which is composed of the database tuples and the associated class labels. In all categories that have been defined in advance, if each tuple is one of all categories, then the class label attribute is determined and is discrete and disordered. The training tuples are composed of individual tuples in the training set are tuples randomly extracted in the data set, also known as samples, instances and data points. In this phase, the class labels for each training tuple are told. This stage of the formation model is carried out under supervision, so called supervision and learning, so supervision is the important work of this stage. In the cluster, the training group's class label is unknown, that is to say in advance do not know the collection you want to learn, this is the biggest difference.

This stage of data classification in a data set using a classification model is also often considered to be a learning map, also called a function. For example, in bank loans, a large number of lenders in the extraction of part of the data as a sample data set to form a classification model. According to the model to determine whether the application for lenders risk, reduce the risk of banks.

Use the Model to Classify. The second stage of data classification is modeled using models. In this phase, it is important to assess whether the predictor of the classifier is of the utmost importance. As in the study period, some of the noise data in the training tuple, the vacant data resulting in the resulting classifier abnormalities, the emergence of excessive fitting phenomenon. Therefore, in order to solve this problem, a test set independent of the training tuple is generated to test whether the model is abnormal or not. The test set consists of test tuples and class labels. So you want to achieve a more optimistic and satisfactory assessment, usually using the sample set to measure the accuracy of the classifier. If the accuracy of the classification model in the data sample set meets the requirements, then a large number of data can be classified by the model formed in the previous stage. For example, we can summarize the classification rules by analyzing the previous loan application data and then approve or reject the new loan applicant according to the rules.

The Classification Algorithm Overview in Data Mining

Classification is an important topic in data mining. The purpose of classification is to learn a classification function or classification model (also often referred to as a classifier) that can map data items in a database to a given category. Classification can be used to extract models that describe important data classes or to predict future data trends.

The purpose of the classification is to analyze the input data and find an accurate description or model for each class by showing the characteristics of the data in the training set. This description is often expressed as a predicate. The resulting class description is used to classify future test data. Although the class tags of these future test data are unknown, we can still predict the classes to which these new data belong. Attention is predicted, but not sure. We can also have a better understanding of each of the classes in the data. That is, we have gained knowledge of this class. There are three classifiers to evaluate or compare scale:

Predictive accuracy: Predictive accuracy is one of the most widely used comparative scales, especially for predictive classification tasks. The currently accepted method is the 10-level hierarchical cross validation method.

Computational complexity: computational complexity depends on the specific implementation details and the hardware environment. In data mining, the complexity of space and time will be a very important part because the operation object is a huge database.

The simplicity of the model description: the more concise the model description is more popular for descriptive classification tasks; for example, the classifier construction using rules is more useful.

Most of the classification algorithms are memory-resident algorithms, and recently there have been some scalable classification techniques that can handle large amounts of data that reside on disk. There are many classification techniques, such as decision tree, Bayesian network, neural network, genetic algorithm, k a nearest classification and so on. The focus of this paper is to discuss the algorithm in decision tree in detail.

The Decision Tree Classification Algorithm

Traditional Algorithms. C4.5 algorithm as a traditional data classification algorithm has a very obvious advantages, such as the rules easy to understand, easy to get started with the actual operation. But with the popularity of computers, the scale of data becomes more and more huge, its complexity is growing. C4.5 has been unable to meet the new era of data classification processing work. And because of the rules of the decision tree classification algorithm, it is decided that the data should be scanned and sorted repeatedly in the process of data classification. Especially in the construction of trees, this shortcoming is more obvious. This will not only affect the speed of data analysis, but also a waste of more system resources. For large data mining, C4.5 is more incompetent, because C4.5 algorithm is very limited scope, can only handle less than the amount of system memory data, memory cannot keep too large data sets, C4.5 even There will be unable to run the situation.

Derivation Algorithm. (1) SLIQ algorithm and SPRINT algorithm are improved by C4.5 algorithm, on the basis of some technical improvements, such as enhanced data sorting technology, and adopted a breadth of priority processing strategy. This allows the SLIQ algorithm to record the number of data processing well and has considerable scalability to provide the basic conditions for handling large data. However, the SLIQ algorithm has some drawbacks, because it is based on the C4.5 algorithm, so when data processing, still need to keep the data set in memory, which led to SLIQ algorithm can handle the size of the data set The limit. That is, if the length of the data record exceeds the predetermined length of the sort, the SLIQ algorithm is difficult to complete the data processing and sorting work. (2) SPRINT algorithm is to solve the SLIQ algorithm in the size of the data set by the memory limit of the problem developed. The SPRINT algorithm redefines the data analysis structure of the decision tree algorithm and changes the traditional algorithm to keep the data set in memory. It is worth mentioning that it does not like the SLIQ algorithm as the list of data stored in memory, but it is integrated into the data list of each data set, which not only avoid duplication of data when the data caused by the slow scan, and Frees up memory pressure. Especially in the large data mining, because the data base is too large, in each data set of the property list to find the required data can greatly save the analysis time, the data classification work has become more convenient. However, the SPRIT algorithm has some shortcomings. For a list of data that does not have a separable attribute, the result may not be very accurate because it can only be analyzed within the data set, resulting in its scalability being limited.

Other Classification Algorithm

Bayesian Classification Algorithm. Bayes classification algorithm is the use of probability and statistics developed by an algorithm, in the current data classification is widely used. But its disadvantages are also obvious, because the Bayes classification algorithm needs to make some assumptions about the characteristics of the data before the analysis, and this assumption often lacks the theoretical support of the actual data, so it is difficult to be accurate and effective in the data analysis process The On top of this, the TAN algorithm has been developed to improve the accuracy of the hypothesis proposition of the Bayes classification algorithm, that is, to reduce the independent assumptions between arbitrary attributes of NB.

CBA Classification Data Algorithm. The classification algorithm based on association rules is CBA classification data algorithm. This algorithm generally requires the use of data construction classifier, in the process of data analysis, first search for the right category for the category of

association rules, which is called CAR; and then selects the appropriate data set from CAR. CBA algorithm is mainly used in the Apriori algorithm technology and it can make the underlying data association rules to the surface, to facilitate the sorting. But because of its data classification is prone to omission, so often set the minimum support for the 0 to reduce the missing data, which resulted in the optimization of the algorithm cannot fully play, reducing operational efficiency.

MIND and GAC-RDB Algorithm Classification Algorithm. In the context of large data mining, the future data classification algorithm should be based on the database technology-based classification algorithm. Although a long time ago there have been some specialized research database staff found and put forward based on the database technology classification algorithm, but has not been practical use. Because in the data mining and data analysis, it is difficult to integrate it with the database system, for now, MIND and GAC-RDB algorithm can better solve this problem.

MIND algorithm and decision tree algorithm are somewhat similar, are through the construction of data classifier for data analysis. But the MIND algorithm uses UDF methods and SQL statements to be associated with the database system implementation. In the data analysis, UDF method can greatly shorten the time for the analysis of the data characteristics of each node, so as to provide a theoretical basis for the integration of the database. SQL statement is through the analysis of the properties of the data set in order to select the most appropriate split attribute, and then sort the data, thus saving the data classification time. But MIND algorithm cannot directly in the database system to achieve the query function, more importantly, the algorithm's maintenance costs are too high, is not conducive to universal.

GAR-RDB algorithm in the MIND algorithm based on the more improved can make full use of the database system for aggregation operations, that is, to achieve the integration of the database system. The algorithm has the advantages of accurate classification, rapid analysis, faster implementation, and scalability is also excellent. More importantly, it can take full advantage of the database provided by the query function, thus avoiding the phenomenon of repeated scanning data sets shorten the analysis time, saving system resources.

Conclusion

Large data mining is the trend of the development of the times, so the importance of data classification algorithm will also be with the show. By analyzing several different algorithms, we can compare the accuracy of data analysis, scalability and results to select the most suitable data classification algorithm. They have varying degrees of their own advantages and disadvantages, so continue to study in depth to develop a better classification algorithm.

References

[1] Qian Shuangyan. On data mining in the data classification algorithm[J]. Electronic production. 2014 (13)

[2] Liu Hongyan, Chen Jian, Chen Guoqing.Data mining algorithm in data mining [J]. Journal of Tsinghua University (Science and Technology) 2002 (06)

[3] Jiang Wei. Data mining technology classification algorithm analysis[J]. Computer Knowledge and Technology. 2009 (01)

[4] CHEN Ping, QIAO Xiuquan, LIU Zhen, TIAN Xiao-ping.Design and performance analysis of parallel algorithm for decision tree in data mining grid [J]. Journal of Beijing University of Posts and Telecom 2009 (S1)

[5] Dong He, Rong Guangyi.Comparative Analysis of Data Classification Algorithm in Data Mining [J]. Journal of Jilin Normal University (Natural Science Edition). 2008 (04)