Data Mining Technology And The Research And Analysis Of The Algorithm

Qiyu Han[1,a], Panqing Wang[1,b], Shuo Wang[1,c]

[1]Department of Information Engineering ，Mechanical Engineering College, Shi Jiazhuang 050003，Hebei，China

[a]505151776@qq.com,[b]13081101815@163.com,[c]1031799669@qq.com

Keywords: data mining,computing technique, the comparison of different algorithm and methods.

Abstract. At the age of "Big Data"， the information has the characteristics of large volume, high complexity, fast growth and so on,as a kind of data processing, data mining is the key technology of the information of the era of "Big data" .The technology of data mining involves a lot of relevant research methods.This paper expounds the concept of data mining technology and data mining process and method of data mining technology and algorithm are compared,at the age of "Big Data".

1. Introduction

At present, the latest research on data mining techniques at home and abroad is mainly embodied in further research on the method of discovering knowledge, and in recent times it is mainly about relevant improvement of methods of Bayes and Boosting, and the close connection of knowledge discovery of database and database and the application of method of regression forecasting in knowledge discovery of database. It shows, in the big data time, how to be based on different situations to choose related algorithm and techniques of data mining appears very important.

Data mining is one step of Knowledge-Discovery in Databases. The present comparatively recognized definition is put forward by W.J.Frawley，G.Piatetsky.Shapiro, etc.: data mining is to extract the knowledge people are interested from data of large-scale database. That knowledge is connotative , potential and useful information unknown in advance, and the knowledge extracted is showed by concept, rules, regulations, models and other forms. From the definition , we can see data mining defines mining object into database. In addition, there is a more generalized version: data mining generally refers to the process of automatically search hidden association rule learning information from plenty of data. Data mining usually has relation with computer science, and realize the above-mentioned target through statistic, online analytical handling , IR, machine learning, expert system(law depending on the past experience) and model cognition, etc.. This shows that when we define the mining object of data mining, we can’t only be limited to database, but from various file system and other data set organized in any forms.

2. The process, techniques and algorithms of data mining

2.1 The process of data mining.

The process of data mining, as shown in figure 1, includes analysis object, data preparation, data mining, outcome evaluation and result application these five stages.

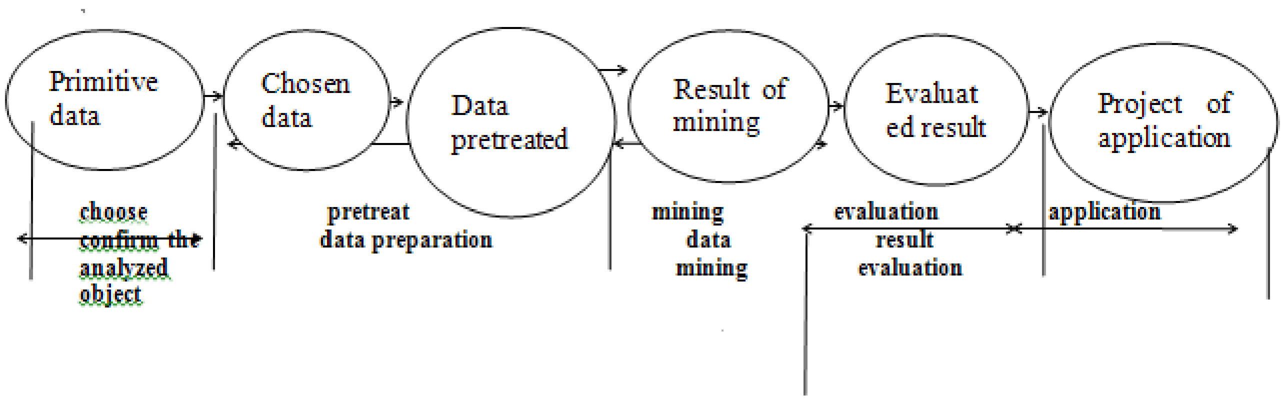

Figure 1. The flow chart of data mining

2.2 The techniques of data mining.

As for the development of present data mining, there are mainly eight techniques of data mining: neural network, genetic algorithm, decision-making tree, rough set , method of covering the positive example and rejecting the counter-example, statistical analysis, fuzzy sets and mining object, etc..

2.3 Data mining algorithm.

Aimed at the population and application of present data mining techniques in various fields, the international authoritative academic organization The IEEE International Conference on Data Mining (ICDM) voted ten classic algorithms in the field of data mining: C4.5, K-Means, SVM, Apriori, EM, PageRank, AdaBoost, KNN, Naive Bayes and CART.

3. The comparative study of the major data mining algorithms

3.1 Apriori Algorithm.

One of the most popular data mining methods, first of all, we need to find all the frequent set in Apriori, these frequencies need to meet certain conditions set, it appears as frequent and predefined minimum support at least. Then generate strong association rules set by the frequency, these rules must meet the minimum support and confidence. Then use the L1 = find frequent frequency set v1-itemsets (D) found to produce the desired rule, generates all rules contain only items of the collection, only one of wherein each of the right side of a rule, here is the definition of rules. Once these rules are generated, then only those users more than the minimum rules given credibility was only to stay. In this algorithm to be able to generate all the frequent sets, we use the recursive method.

Apriori algorithm is a method of using a candidate set to find frequent item sets. It is a full use of anti-monotonic level search algorithm. If an item set is not frequent, it is a superset of any non-frequent.

But for Apriori algorithm is the existence of two inadequate; that may produce a large number of candidate sets, and you may need to repeat the scan database, these two shortcomings could lead to much lower efficiency of Apriori Algorithm, at the same time also led to excessive candidate set algorithm program will generate too much data redundancy, it makes the algorithm to reduce the operating efficiency program.

3.2 C4.5 algorithm.

Classification technology is a tool for data mining algorithms are often used. C4.5 algorithm is an extension of ID3, it is capable of generating a decision tree classifier represented, and it may be easier to understand the form of a rule set to represent classifier.

C4.5 in the algorithm not only inherits the ID3 algorithm itself has the advantage, and in the following aspects of the ID3 algorithm corresponding improvement, it makes the algorithm more perfect;

a) With information gain rate select Properties, to overcome the bias select multiple values of attributes with attribute information gain when choosing inadequate.
b) Prune the tree construction process.

c) Able to complete the process of continuous discrete attributes.
d) Not able to complete the data processing.

Therefore C4.5 algorithm has a corresponding advantage, classification rule is to produce easy-to-understand and has high accuracy.

C4.5 algorithm but there are insufficient regard; during construction of the tree, The need for many existing data sets of sequential scanning and sorting, resulted in greatly reduced the efficiency of the algorithm C4.5. Moreover, C4.5 algorithm requires data set has great limitations; C4.5 algorithm is suitable for only those data sets can reside in memory, if the training set is large enough, and cannot be accommodated in the memory, C4.5 algorithm program will not run.

3.3 EM algorithm.

EM algorithm is one of the most common data mining algorithm, mainly used in the statistical calculation. Maximum expected (EM) algorithm is to find the probability that the model parameters maximum likelihood estimation or maximum a posteriori estimation algorithm, wherein the probability model relies on unobservable hidden variables (Latent Variable). EM is often used in computer vision and machine learning data clustering (Data Clustering) and other related fields.

EM algorithm is to provide a flexible finite mixture distribution, Calculation Methods of mathematical modeling and cluster-based data sets. Common hybrid clustering model can be used for continuous data and predict potential density function. These hybrid models likelihood expectation maximization algorithms by maximum to fit .EM is also an iterative algorithm .Commonly used in the probabilistic model parameters contain hidden variables (latent variable) maximum likelihood estimation or maximum a posteriori estimate.

The main purpose of the EM algorithm is to provide a simple iterative algorithm, and calculate density function is verified.

EM algorithm biggest advantage is relatively simple calculation method, meanwhile, the algorithm has high stability. But as an iterative algorithm will be relatively obvious disadvantage, the algorithm is to find the optimal time is easy to fall into local optimum, it would be impossible to find the global optimum.

There are currently 10 kinds of mainstream classical algorithm, frequency of use is the most widely used of these three algorithms, the following three algorithms will be compared (Table I).

Table I. The comparison of algorithm

Comparison algorithm name	Applications	Aims	Processing Unit	Prototype	Shortcoming
C4.5	Machine learning and data mining classification	Supervised learning	Tuple	Decision Tree	Requirements for data collection has great limitations
Apriori	Association rule	Search the database to obtain candidate set of support item sets	Frequent item sets	Recursive algorithm	a) The set of candidate large amount, reduce the efficiency b) excessive data redundancy
EM	Data clustering machine learning and computer vision	Looking for maximum likelihood estimation of the parameters or maximum a posteriori estimation algorithm	Hidden variable	Iterative Algorithms	General disadvantage iterative algorithm, the algorithm is easy to fall into local optimum

4. A comparative study of data mining technology is mainly

4.1 Artificial neural networks;

Artificial neural network is a system of human brain structure and function simulation, in the parallel processing of information distribution and its main features, it can be achieved on an analog image thinking. In a neural network, storage and processing of information are one, that is reflected in the information stored in the distribution of interconnected neurons. This distributed storage, not only does not make the information when a part of the destruction of damaged, to resume as soon as possible, enhanced fault tolerance of the network, but also makes network on with noise or defective input has a strong ability to adapt, Lenovo Enhanced holographic memory and network capacity. Artificial neural networks can perform a variety of data mining tasks classification, clustering, feature mining.

4.2 Decision Tree.

Decision tree is an artificial intelligence method for building a classification model, it is a sample-based inductive learning a common way. Inside the tree node properties were tested and led by the node based on the attribute value determination branch, in conclusion the decision tree leaf node. In the decision tree method, decision is set by the data sets generated classification rules.

Rules can be extracted from tree to get into simple rules and get two steps to streamline the rule properties;

A. For generating a good tree, can be obtained directly from the rules. Each path from the root to leaf could be a rule. Rules adopted if ... then expressed in the form.
B. Streamlining of Rule Properties. From simple rules obtained in the previous step, it may contain many unrelated attributes.

4.3 Prediction.

Data applied to a given input prediction predict a continuous (or ordered) value. Data is currently the most widely used method for forecasting a return. Regression analysis is one extremely versatile data mining method that is used to describe and analyze the correlation between variables inherent laws inherent law prompted variables, suggesting variables. Regression analysis can be used to model contact one or more independent or predictor variables and a (continuous value) dependent or response variables. We have many variables or between approximately linear correlation and linear regression analysis method is simple, complete theory, so the linear regression model is often preferred as a data mining.

In response to these three mainstream data mining technology, we will be more than three technologies related comparison (Table II)

Table 2. The comparison of data mining technology

Technical Comparison	Field	Task	Model	Treatment	Characteristic
Neural Networks	Neural Networks	They can complete a variety of data mining tasks classification, clustering, feature mining	Human brain architecture	Interconnected neurons reflect the distribution of information on the storage	In the parallel processing of information distribution and its main features
Decision Tree	Machine learning method (inductive learning method)	Internal node attributes tested and judged by the node leads to branch based on property values	Classification Model	Internal node attribute test, and based on the property value judgment led by the branch node and leaf node conclusion	You can get into simple rules and access rules to streamline property
Prediction	Regression	Dependent or	Linear	Used to	It based on the law

	analysis (most widely used)	response variable modeling links between one or more independent or predictor variables and a (continuous value).	regression model (most commonly used)	describe and analyze the correlation between variables, suggesting that the internal laws of variables, suggesting that the internal laws of variables.	of development of things inside, so this method is more accurate

5. Summary

This article mainly analyzes and researches several main algorithms and techniques of data mining at the time of big data, and through comparing the features of several algorithms and techniques, realizes further comprehension of techniques of data mining. But there are also some sufficiency in this article. This article only illustrates and compares several mainstream algorithms and techniques, but doesn't compare them in experiment. And, in the learning and working later, I will tend to wield these algorithms and techniques to collect data through experiment and make comparison.

References

[1].Jianwei Dai，Zhaolin Wu，Mingdong Zhu，Jianhua Gong，et al.Data engineering theory and technology.National defence industry press，2010，p.180-220.

[2].Yongqing Wang，et al.Principle and method of artificial intelligence.Xi 'an jiaotong university press，1998，p.412-421.

[3].Michael Negnevitsky，et al.Artificial Intelligence A Guide To Intelligent Systems.China Machine Press，2011，p165-216，p365-421

[4].Kai Chang：Data mining based on neural network classification algorithm comparison and analysis.A master's degree，Anhui University，China，2014.p13.