Research on Data Mining and Statistical Analysis

Xiaoyao Lu[1, a]

[1] School of Statistics and Mathematics of Zhejiang Gongshang University, Hangzhou, Zhejiang, China, 300018

[a]email,

Keywords: Data Mining, Statistical Analysis

Abstract. Since data mining methods and technology, the most basic and most important method is statistical methods and statistical theory also spawned a number of new data mining methods, the study of data mining and statistical methods can be applied not only to practitioners use data mining to provide advice and guidance, but also to analyze data using statistical characteristics of data mining, and promoting scientific and technological development and the creation of social wealth to lay the theoretical foundation for researchers and practitioners.

Introduction

With the development of science and technology, the use of database technology to store data management, the use of machine learning methods to analyze data to dig out a lot of hidden knowledge behind the data, combined with this idea now form the very popular attention Popular areas of research: database knowledge discovery, which is the KDD data mining technology in one of the most critical aspects.

In 1995, in Montreal, Canada convened the first session of the "Knowledge Discovery and Data Mining" International Conference, data mining is the word quickly spread. Data mining is from a lot of, incomplete, noisy, fuzzy, random data to extract implicit in them, it is not known in advance, but is potentially useful information and knowledge. Data mining is an interdisciplinary and it brings together databases, artificial intelligence, statistics, visualization, parallel computing, different disciplines and fields in recent years by the widespread concern. Data mining and statistics are closely related. Statistical data mining appears to provide a new application area, but also to the statistical theory of a challenge, it will undoubtedly promote the development of statistics.

The Main Task of Data Mining

Data mining is the target from a large number of data, we found that the relationship of the law after its hiding or between data, thereby serving the decision. The main tasks of data mining are:

Data Summary. The purpose is to summarize the data of the data concentrated to give a comprehensive description of its population. By summarizing the data, the raw data to achieve an overall grasp. The easiest method is to use statistical data summarized in the conventional method to calculate the sum of the individual items in the database, the mean and variance and so on.

Classification. The main function is to use a classification function or model that can be assigned based on the attribute data of the data into different groups. So that we can use this model to analyze existing data and predict new data will belong to which group. For example, we can bank outlets into good, fair, poor three types, based on this classification and analysis of various properties of three kinds of bank outlets, such as the position, profitability, etc., and determine their classification and mutual critical relationship.

Correlation Analysis. Data in the database are generally relationship exists, there is some regularity between the value that is to say two or more variables. Relational Model using a typical case is "diapers and beer" story. In the United States, some young father after work often go to the supermarket to buy baby diapers, supermarkets and thus found a rule, buying baby diapers young fathers, 30% to 40% of people at the same time to buy some beer. Then adjust the supermarket shelf display, the diapers and beer together, the results of a significant increase in sales.

Clusters. When the data to be analyzed lack of descriptive information, or can not be organized into any classification mode, you can use cluster analysis. Cluster analysis is in accordance with a similar level metrics, the user data is divided into a series of meaningful subset. Statistical methods Cluster analysis is a means to achieve clustering, which is mainly based on geometric clustering distance. Artificial Intelligence Clustering is based on the concept described.

Main Data Mining Method

The modern world is a data-driven world, the data source is infinite and enormous data set is stored in a central data warehouse. In recent years, in order to concentrate extract valuable information from these massive data found a new method from the raw data of the explosive growth in knowledge, data mining and its applications are widely studying.

At home and abroad, sophisticated data mining methods have a lot of research, based on their knowledge or discovery or mining association model types into knowledge, knowledge classification, knowledge clustering, classification or regression prediction knowledge; international academic exchanges will last two years on exhibit many new mining methods: mining community, mutual nearest neighbor query method based on cloud theory, evidence theory.

Correlation Analysis Method. Association analysis method is by finding association rules in the data warehouse valuable property or project to tap an association between the type of knowledge or information. Its mission is to reduce the number of difficult disorder to follow the data, making it a small amount, to facilitate observation of static data or information to understand. Find related business knowledge in the field of application of the most classic example is market basket analysis.

Classification Analysis Method. Data Mining and Knowledge model, classification is more important as a model, classification analysis method is a very necessary and efficient analysis methods, data mining classification tools, the function of a more comprehensive decision tree classification, main idea is a sign of the test data and correctly classified. Learning to learn recursion tree belongs to the general use of top-down manner, a property value from the beginning of the root Jian namely, to determine how down branch, and then each internal node in the decision tree to continue comparing property values, according to the judgment of the property value, step by step down the branch structure, and eventually the leaf nodes of the tree conclusion.

Cluster Analysis Method. Cluster analysis is a statistical analysis method, also known as cluster analysis, classification main sample or index between. In the business world in general is based on a different calculation method of cluster analysis, clustering method is divided into block-based method based on hierarchical clustering, density-based clustering method, grid-based clustering method based on model clustering methods; the specific algorithm is divided into k-means algorithm, K-medoids algorithm, Clara algorithm, Clarans algorithm. k-means algorithm to accept input k; and after n data objects into different clusters, cluster number of k and achieve the following objectives: the poly-category objects inside high similarity; and the polyethylene category outside similarity smaller objects. It means clustering similarity can use the internal properties of the object category poly metrics calculated from the mean value has become the center of the polyethylene category. Other clustering algorithms compare the above Table 2-3.

Prediction Analysis Method. Prediction method is a continuous mining method of knowledge is an important value data or information to predict. Traditional application methods are: time series analysis, linear and nonlinear regression analysis, gray system model analysis, Markov analysis method; now the forefront of the application of neural network algorithm is mainly and support vector machine algorithm that can be used to extract be able to describe collections or important data to predict future trends in the data model, mainly used in business sales, market share prediction of possession.

Application of Data Mining

How to use data mining methods and mining historical bridge sub objective data so that it can create the conditions for greater access and business services and technological innovation, is the main problem of data analysis, managers, policy makers and research scholars concern. The following analysis of the existing data mining methods in business services, applied research hotspot medical science, life science.

Application of Data Mining in Commodity Retail. In the retail industry, the most humble of data mining method is Correlation Analysis, as early as 1994, by the United States, IBM's Rakesh Agrawal Al Madan research center sales transaction database sales of products between each found association rules, and guide decision-making. Its application conditions must be based on retail merchandise sales process accumulated large amounts of data. Planning within the retail display shelves of goods, with concurrency, application and determine the amount of the commodity purchase digital network platform for retail data mining system and all involve the use of data mining association rules.

Application of the Data Mining in Insurance Industry, Financial Industry and Communications Industry. From the perspective of business applications, data mining support decision-making process can be described in three steps: First, data collection, and then use data mining techniques to extract useful knowledge, and finally to extract relevant knowledge to assist decision-makers to make decisions . This part of the field of insurance, financial services, communications data mining techniques to analyze, data mining pointed out that the conditions of use and range of application-related technology and methods.

Application of the Data Mining in Biopharmaceutical and Genetic Aspects. Biology itself is a rich and complex field of study, the amount of data is much greater need of auxiliary data mining technology and guidance. Especially in the DNA sequence similarity search and compare aspects, features and analysis of genomic gene sequence simultaneously aspects of biological data visualization and visual data mining, protein structure prediction and so high in research, reflecting the growth potential of data mining and data mining research open up a vast space.

Application of Data Mining Methods. From existing applied research shows that shopping basket analysis association rule mining is home of the association rules found Boolean association rules. In fact, in other types of application of association rules is quite rich, in addition to Boolean association rules, as well as quantitative association rules. The key difference between the two main attributes corresponding to the rule is to deal with different data types. If there are items under consideration and otherwise belong to Boolean association rules mainly deal with discrete data, but also can handle the kind of data that reflect the relationship between variables; if for numeric fields mining process then it is quantitative association rules, the It reflects the association between quantitative or property, describe the association rules in accordance with the quantization interval division value of open, often also involves the dynamic properties of discrete values.

Data Mining Process and the Statistical Theory

Statistical methods to provide scientific theory, principles and methods for the study of the precise number of phenomena, that is, provide the tools and means. Data mining process consists of five modules, namely, the problem statement and clarify the assumptions, data collection, data preprocessing, model evaluation, interpretation model and draw conclusions; data mining tasks mainly summarized as: data preparation, data reduction, data for learning. Analysis of the data for statistical purposes in the method of data mining processes and tasks have played a significant role; simultaneous data mining also contributed to the development of statistical methods, because with the increasingly expanding data sources and data structures complex, requiring the simultaneous development of statistical data mining to solve the many problems faced. Before application of data mining sticks, first of all to practical problems or to provide a clear description of the phenomenon. After describing the problem, the model will usually unknown relevance to specify a set of variables, and then specify the related form of a general as an initial hypothesis, and current issues

under study illustrates a given hypothesis. If the model before converting from the reality that there is no scientific and reasonable assumptions, then people perceptions of the real world it is impossible to rise to the theory of the stage; and because the realistic problems or complex phenomenon, involving broad face, a variety of factors the object of study are linked, and these factors are primary and secondary points, so the hypothesis would need to analyze what is the main factor, to simplify the problem for data description, and then grasp the essence of the problem. Therefore, the problem statement and assumptions set forth in this first step of data mining conditions common to assume that the principal component analysis, sampling, select the feature properties and other statistical methods.

Generating and collecting data in two ways, one is an expert (modeler) control data generated design experiments; the other is the expert does not affect the data generated by observation, data generated in this way is random. In most data mining are used observation. In data generation and collection process to apply statistical methods in statistical probability and random numbers, we need to apply random events, random numbers to determine the credibility of the data generated. The method of data collection is a widely used method through the random number generating experimental data set and applied probability and size to determine credibility. After data collection is complete, the sampling distribution is totally unknown, or that its distribution is part or the data collection process is not explicitly given. Modeling and interpretation of the results of the final very important point is to understand the data collection is how it affects their theoretical distribution, for which you need to know a priori knowledge of these probability distributions. Similarly, to determine the model used to evaluate, test model, the model is applied to the data from the same samples of unknown distribution is also important. If different distributions, the model evaluation can not be successfully used in the final results of the application. In short, the process of data collection application of random events, random testing, experimental design, random probability, probability distribution, prior knowledge, with the distribution of the sample and other statistical methods.

The experts do not impact the process data of observation, the data generated randomly to existing databases, data warehouses and data marts. Data preprocessing data mining often includes at least two common tasks: First, monitoring and removal of outliers, whose function is to prepare data; Another task is scaling, coding, and select the feature, which aims to achieve data dimension reduction. . So one of the tasks of data mining is to detect, including changes in the deviation detection and outlier detection, designed to identify the most important changes in the data set, look for outliers to determine if the mutation of things happen. This step is mainly applied to important statistical methods in data preparation analysis. Data mining is the task of the two data dimension reduction, mainly used in statistical analysis of principal component analysis, factor analysis, regression analysis. Data mining model is a "large-scale" structures or relationships for most cases and summary, and the pattern is a partial structure to meet in a few cases or small data space area, namely in data mining models only It is a local model. Select and implement appropriate data mining technology is the main task of the model evaluation phase, in fact, to achieve this task is based on several models from which to choose the best model is the additional tasks. The basic principles of the theory of learning from data and uncover mainly dependent on the prior probability, posterior probability, probability theory and Bayesian formula and other multivariate statistical analysis, a Bayesian network statistical areas, statistical learning mature, from a data successfully learn and apply these technologies to give an objective and credible assessment model, in order to find an appropriate model.

Analysis of Statistical Methods in Data Mining Tasks

Statistical Methods in Data Preparation. Required to prepare data mining can be divided into two categories, structured and unstructured data. Structured data is a two-dimensional table structure that can achieve data logical expression, that line data. The traditional way is stored in the form of two-dimensional table is stored in the data where it is generally considered to be traditional data; however, the two-dimensional logic in the database table is not convenient for data storage, is called

unstructured data, common unstructured data include, but are not limited to, all text format, reports, HTML, images and audio / video. The data is not strictly between structured data and unstructured data, can easily expand its fields if needed, that an unspecified number of data fields, defined as semi-structured data, the Exchange store is a semi-structured the data represents. Semi-structured and unstructured data together are called non-traditional data, also known as multimedia data. For these traditional and non-traditional data, before applying data mining techniques, it must be given clear and specific expression, that is, from the abstract to the general problem of the Standard Model.

Statistical Methods in Data Reduction. For small and medium-sized data sets, data preprocessing step is sufficient; but for large data sets, data Statute is an essential intermediate step after the statute in order for data mining. Data statute includes two aspects, one is to achieve a dimension of the Statute, the second is the numerical value of the Statute. In general, the data coding or transform dimension reduction, so that you can get the original data "compression" means reduction or obtained in the form of the original data. And typically effective data reduction method of wavelet transform and principal component analysis of the current study is the widely popular. Statistical methods Karhunen-Loeve (K-L) or Hotelling transform method is the most popular method for large data sets to achieve Statute. K-L transform matrix transformation can have a variety (second order matrix, covariance matrix, within the overall class scatter matrix, etc., when the K-L transform matrix of covariance matrix, equivalent to principal component analysis.

Statistical Methods in Data Learning. Many of the latest development model data according to the method are from learning ability of biological systems, in particular human ability to learn to get inspiration. In fact, the study of biological systems based on data-driven approach to the environment is unknown, statistical properties. Prediction learning process can be summarized into two stages: the first stage is input to learn that the focus on learning estimation system from known samples or unknown relevance. The second stage is the predicted output is applied to estimate the correlation in the future when new input, new output system to predict. Reasoning step two stages corresponding to two types of classic: inductive and deductive.

Inductive learning is defined as a process that input and output measurements or observations with limited system to estimate the unknown input process-related or system configuration output. According to inductive learning theory, all data must go through the learning process of the organization, each pair of input and output with a simple instance of the term "sample" to represent.

Conclusion

The study of data mining is in the ascendant, and its prospects have been confirmed in the international arena. Data mining is an interdisciplinary, data mining techniques related to: databases, artificial intelligence, statistics, visualization, parallel computing, different disciplines and fields in recent years by the widespread concern. Data mining and statistics are closely related, so how statistical data mining services, which is a problem in "data mining" the rapid development of today, statisticians must be answered. Statistical data mining challenge to bring will undoubtedly promote the development of statistics.

References

[1] Jia Xinzhang, Li Jingyuan. Statistics and Decision, Vol. 6 (2004) No 53, p.25-26

[2] Peng Sue, Wang Yunhui, Wang Qunyong. Geomatics World, Vol. 12 (2005) No 27, p.74-76

[3] Jing Jianfen, Pattern Recognition and Artificial Intelligence, Vol. 30 (2004) No 19, p.144-145

[4] Wang Kuailiang. Statistics and Decision, Vol. 29 (2008) No 27, p.21-23

[5] Zhang Gongxu, Sun Jing. Geomatics World, Vol. 8 (2003) No 27, p.57-60