Research on the Application of Statistical Methods in Data Mining

Chen Xu
(Hongshan College of Nanjing University of Finance and Economics, Nanjing, Jiangsu, 210003)

Abstract: China's science and technology are constantly developing, and digital mining technology is also constantly improving. In data mining, the most basic method we apply is statistical method. With the development of statistical technology, many new data mining technologies are emerging. In order to provide data mining workers with more valuable information and research experience, and further promote the development of data mining technology, we need to continue to study the application of statistical technology in data mining. This paper introduces the significance and current situation of data research, and focuses on several typical statistical methods and technologies involved in data mining and their practical applications.

Keywords: data mining; Research status; Statistical techniques; practical application

1. The concept of data mining

At present, there is no uniform definition of data mining technology, and people mainly focus on its background and views. At present, the most comprehensive and highly recognized data mining technology is defined as: extracting hidden, regular and other potential information from a large amount of information, and creating and processing these information. As an important cross discipline, data mining can integrate database, artificial intelligence, machine learning, statistics and many other sciences together to achieve innovation and development of technology and theory. Among them, database, artificial intelligence and statistics are the three pillar theories of data mining. The purpose of data mining is to discover all kinds of hidden knowledge and information to reveal the inherent laws of things. Data mining is applied to a variety of methods, including statistical knowledge, genetic algorithm, rough set method, decision-making method, fuzzy logic method, etc. It can also apply proximity visualization technology, pattern recognition technology, etc., which can make data mining more scientific and orderly on the support of all the above technologies. Data mining generally goes through the following five steps: first, elaboration and hypothesis; Second, data collection and sorting; Third, data preprocessing; Fourthly, model evaluation; Fifth, the conclusion analysis. The five processes of data mining are a process that requires continuous iteration. In the iterative process, the keyword "discovery" can be used to define the progress of each iteration. The discovery process can be obtained automatically or manually.

2. The relationship between data mining and statistics

Generally speaking, the most important role of statistics is to study the basic principles and methods of statistics. Specifically, statistics is the process of collecting, sorting, classifying, analyzing and using digital data, while data is the induction and summary of all kinds of data, which can be regarded as a cognitive and reasoning method. Statistics is the use of statistics, probability theory and other professional knowledge to make statistics and analysis of things with different attributes, so as to find the relationship and development between them. In this process, the most important method is statistical analysis. Before the emergence of data mining, people have been used to statistical analysis technology and often use statistics to analyze the relationship between data. However, we cannot simply regard data mining as an expansion and replacement tool of statistics, but we must fully understand the differences between the two methods and analyze their application characteristics according to their differences. Most statistical analysis technologies are developed based on mathematical principles and technologies, so the general prediction is more accurate and the effect is better. Data mining technology is a new technology that combines statistics technology.

Statistics and data mining have the same goal, that is, to continuously mine the structure of data. Because the goals of statistics and data mining are the same, many scholars and experts regard data mining as a branch of statistics. However, this understanding is not accurate, because data mining has a great impact on ideas, tools, methods, etc., especially in computer science. For example, through the learning of database technology and artificial intelligence, we can pay attention to more similarities between statistics and data mining, but the differences between them are still very large. Data mining refers to the process of constantly mining a large amount of data information, which can fully mine the data relationship in the data model, and has a high degree of attention to the observed database processing.

3. Application research of data mining method

3.1 Research on the application of data mining in financial industry and communication industry

In recent years, data mining technology has been widely used in finance, insurance, communications and other fields, including medical insurance, life insurance, automobile insurance, etc. The application of data mining in the financial industry includes credit card issuance, customer credit rating, customer relationship management, customer information classification, etc. All these need to be predicted through big data. In the communication industry, data mining has a wider application range. It can conduct accurate marketing by analyzing the

user's behavior, thus greatly reducing the operating costs of operators, expanding the scale of enterprises, and improving the competitiveness of enterprises in the communication market.

3.2 Research on the application of data mining in commodity sales

In the retail industry, the practical applications of data mining include: commodity shelf placement rules, sales data statistics and analysis, purchase, sales and inventory management statistics, and consumption behavior analysis on the digital network platform. Through the data analysis of data association, these application scenarios with big data association can construct the data model of associated marketing, which is also called the "shopping basket analysis" model, It can give full consideration to the interests of retailers and sellers, and on this basis, through the complementarity, relevance, substitutability and other data characteristics of marketing elements such as commodities and brands, and then through specific all association relations to achieve internal in-depth guidance and decision-making. The association rules involved in the association relations can be obtained by data mining, which needs to be realized through machine learning algorithms, At present, the more mature algorithms are: rule mining algorithm with constraints, mining algorithm based on interest, granular computing mining algorithm, etc. The use conditions of association relationship must have a large amount of big data, and the data must be independent of each other, and the data dimensions must be sufficient.

3.3 Application of Data Mining in Scientific Research such as Biopharmacy and Genetic Engineering

Biomedicine and genetic engineering are very complex systems engineering. In the process of pharmaceutical production, there are millions of combinations between molecules, and gene decoding in genetic engineering has a large amount of data. Data mining technology can improve data processing speed. The amount of biological data is very large and complex, so the processing of data needs more advanced technical means to assist and guide, such as the research on DNA sequence retrieval and comparison, genome characteristics and sequence analysis, protein structure prediction, biological data visualization, etc., which shows the great potential of data mining.

4. Statistical theory analysis in data mining process and task

4.1 Analysis of statistical methods in data mining

Before using data mining technology, the first problem to be solved is to clearly describe the problem. After the problem is described, the modeler will define a group of related variables, and then determine the relevant relationship according to these variables, which is the so-called "initial assumption". If there is no link of scientific and reasonable assumptions between the reality and the model, the perceptual knowledge of the real world will be difficult to translate into theoretical research. Moreover, the real problems are very complex, involving a wide range of fields, and the relationship between different factors and phenomena will be primary and secondary. Therefore, we must separate these factors and problems, so as to make our research object simple, We can use data mining technology to solve problems. Therefore, problem statement of research phenomena is the most important precondition of data mining technology, which can include principal component analysis, sampling analysis, selective feature attribute analysis, etc.

4.2 Statistical Methods in Data Collection

Generally, there are two common methods for data collection of research objects: one is that model builders use scientific experiments to collect data; Secondly, detailed observation does not affect the data generation scenario, so as to collect data. Most data mining technologies are conducted through the latter. For the generation and collection of data, the probability and random number methods in statistics are often used. It is necessary to closely combine random events, random numbers, etc. to identify the generated data and verify its reliability and validity. The most common collection method is the random number method, which obtains a group of test data through observation, and then judges the reliability of the data according to its probability until the collection is completed. Because the distribution of samples is unknown, a comprehensive analysis of the model and results must be carried out to understand the impact of data collection on the theoretical distribution and the probability distribution of such data. Similarly, the evaluation model, test model and the distribution of relevant data in the model are also important. If the distribution forms of these models are different, the evaluation model cannot be applied to the final results. In a word, statistical methods such as random event, random probability, random experiment, probability distribution and prior knowledge are systematically used in data collection of data mining.

4.3 Statistical Methods in Data Preprocessing

In the application of observation method that does not affect the data generation process, data can be randomly generated from existing databases and preprocessed. Generally, data preprocessing includes the following two tasks: first, monitoring and eliminating outliers, which is a kind of data preparation; the second is scaling, encoding and selection. The goal of this processing program is to reduce the selected data. This is because data detection is an important link in data pre-processing. The focus of data detection is data changes, deviations, outliers, etc. Its purpose is to find the most important changes in the data set, find outliers, and determine whether any transactions have

changed. This step requires the data preparation method in statistics. Another task of data preprocessing is to reduce the real data to different degrees. The specific tools involved in this process include factor analysis, regression analysis, principal component analysis, etc.

4.4 Statistical Methods in Model Evaluation

Data mining generally requires the help of a data model. The data model is a "super large" data structure, which can show the huge association between data. For patterns in data mining, it can reflect the "local characteristics" of the data structure, and then more accurately describe and summarize a few key areas of data. In practical applications, patterns in data mining are usually presented with a local model. The most important task in the model evaluation stage is to select and implement appropriate data mining technologies, which usually requires the integration of multiple local models to achieve, and the selection of models is an additional task. This process needs to follow the discipline knowledge of multivariate statistical analysis, such as prior probability and posterior probability, and apply these technologies to objectively evaluate the model, and finally identify the most appropriate model.

4.5 Statistical methods in interpreting models and drawing conclusions

Data mining technology provides decision basis for decision-makers to make strategies through modeling. Staff need to effectively explain and explain the model, so as to ensure that the model can be more suitable for real use scenarios. At present, the main development trend of data mining is to apply higher dimensional models to obtain data results with higher accuracy, and then verify the reliability and validity of the results with scientific application tools and means to explain the progress. When explaining the results, the staff need to use the statistical tables and variables in statistics to clearly present the results. If the analysis is conducted solely from the data mining process, statistical methods and theories can really play a very important role in data mining. All links need statistical methods and theories to support, from the preparation stage to the processing process, to the final result verification and conclusion generation, It can be seen that there is an inevitable correlation between data mining technology and statistical analysis methods. Only through a high degree of integration of statistical theory and computer technology can scientific processing methods be raised to a higher level, thus providing a new development direction for the combination of statistics and information technology.

5. Analysis of Statistical Methods in Data Mining Tasks

5.1 Statistical Methods in Data Preparation

The standard data mining mode is based on typical cases and describes them in detail. Generally, the characteristics are required to meet the measurement conditions of most cases, which also caters to the sampling theory in statistics. The data in data mining has structural characteristics. Generally, it is represented by tables or single relationships, which is called relational data structures. The columns in the tables can reflect the characteristics of the data, while a row in the tables can reflect the characteristics of the target entity. The row can be regarded as a sample, while the standardized model of data mining can be regarded as a collection of samples generated by sampling in essence, The characteristics in the relation table must be presented through most sample measurements.

5.2 Statistical Methods in Data Reduction

In small and medium-sized data sets, data pre-processing is very necessary, but in large-scale data sets, data classification has become an indispensable link, and a large amount of data needs to be compressed. Generally speaking, large-scale data reduction includes two steps, the first is to achieve dimensional reduction, and the second is to achieve numerical reduction. Usually, data coding or data transformation can complete data dimension reduction. At present, wavelet transform and principal component analysis are widely used data dimension reduction methods, which are also typical and effective reduction methods.

5.3 Statistical Methods in Data Learning

The learning ability of biological system can provide reference for data development model, especially for human thinking model and learning ability. Biosystem learning is to conduct mathematical statistics on the unknown environment in a data-driven way. For example, the predictive learning process well reflects this process. The predictive learning process includes two stages. The first stage is the input learning stage, which is to estimate and analyze the known samples, and then predict their relationships. The second stage is the input stage, which is to predict through estimation, Analyze what will happen to the output next time a new input condition is encountered. This classical reasoning process is also known as induction and deduction.

6. Conclusion

Data mining is to process and visualize large-scale data, and its ultimate goal is to discover the knowledge generation process, and then predict the future trend. Data mining technology can use different analysis methods and analysis tools to build mathematical models for massive data, establish relationships between data attributes, and make scientific decisions using models and associations. In the process of data mining, the most common method is statistical sub learning, and the related statistical methods in data mining are to enrich and

improve the theory of data mining, and provide a lot of experience and reliable data statistical means for data mining. Of course, with the development of the Internet, data mining is increasingly combined with machine learning and deep learning, but statistical methods are still the most critical and interpretable theoretical methods.

References:

[1] Li Junjie. Deep Mining of Ship Traffic Data in Modern Statistical Theory [J]. Ship Science and Technology, 2019, 41 (24): 31-33

[2] Pang Jianping. Analysis on the Application of Probability Theory and Mathematical Statistics in Data Mining [J]. Technology and Market, 2018, 25 (11): 101-102

[3] Wang Jun, Zhao Yingjun. Statistical Thinking Based on Data Mining and Training of Programming Talents [J]. Computer Age, 2018 (09): 96-98

[4] Hu Huimin Research on large-scale network anomaly detection method based on data mining [D]. Xi'an University of Electronic Science and Technology, 2018

[5] Mou Hongmin. Discussion on the Application of Statistical Methods in Data Mining [J]. China Civil Business, 2017 (12): 261