Application Analysis of Probability Theory and Mathematical Statistics in Data Mining

Fengmin Liu

Jilin Vocational and Technical Institute Communications, Changchun, Jilin 130012,China

Keywords: Data mining; Probability theory; Mathematical statistics; Statistics; Application

Abstract: Data mining refers to the process of applying data analysis and data discovery algorithms to obtain potentially available patterns or guiding rules from a database. With the development of computer and network technology, there are a large number of large and wide-ranging data. It is impossible to analyze such data by the traditional statistical method of simple summary and analysis according to the specified mode. Probability theory and mathematical statistics are technologies used in data statistics, but they also play a very important role in data mining. Data mining and statistics should learn from and permeate each other, divide their work and work together to contribute to the mining of valuable knowledge hidden behind complex phenomena. Therefore, data mining is a cross-disciplinary subject. It promotes people's application of data from low-level simple query to mining knowledge from data to provide decision support. This paper briefly analyzes the application of probability theory and mathematical statistics in data mining.

1. Introduction

Data mining is a process of extracting unknown and valuable data information from a large number of noisy, fuzzy, incomplete and random data [1]. The steady progress of computer hardware technology has provided mankind with a large number of data collection equipment and storage media. The maturity and popularization of database technology have enabled the amount of data accumulated by human beings to grow exponentially. Its purpose is to explore the visual presentation of large-scale non-numerical information resources, such as numerous files or line-by-line program codes [2] in a software system, and to help people understand and analyze data by using techniques and methods in graphics and images [3]. The traditional models of uncertain knowledge include neural network, rough set theory and fuzzy logic. With the help of computer, machine learning, artificial intelligence and other theories and technologies, it has irreplaceable advantages. The emergence of data mining is exactly the new development direction of probability statistics to adapt to this change. In today's increasingly information-based society, a large number of data are accumulated and stored in large database systems. It can not only query the past data, but also find out the potential connection between the past data and carry out higher-level analysis, so as to better make ideal decisions and predict the future development trend.

2. The Concept of Data Mining

Data mining is a process of revealing meaningful new relationships, trends and patterns through careful analysis of a large amount of data. Through intuitive communication of key aspects and features of information, it further realizes in-depth insight into sparse and complicated spatial data, and visualization technology plays an important role in non-spatial fields. When using a database management system to store massive data, it is necessary to analyze the data and mine the knowledge contained behind the massive data by means of machine learning. The combination of the two contributes to the generation of data mining. Therefore, in this sense, data mining is the process of mining useful knowledge from databases, data warehouses and other data storage methods. This description emphasizes the diversity of data mining in the form of source data [4]. The methods and techniques of data mining can be divided into five categories: inductive learning methods, biomimetic techniques, formula discovery, statistical data mining techniques, and fuzzy

  DOI: 10.25236/iwmecs.2020.039

mathematics methods. These methods can be used to detect abnormal data. In data mining, some quantitative data and qualitative data, qualitative data and quantitative data are often mixed together, and some data are missing, which requires the combination of relevant data processing technology and data mining technology in statistics. Because it is a data mining function, it can obtain data distribution independently, observe the characteristics of each cluster, analyze specific data at the same time, and provide preprocessing steps for other algorithms.

3. The Relationship between Statistics and Data Mining

Statistics has an orthodox theoretical basis, but now there is a new discipline with a new owner, and claims to solve problems that statisticians used to think were in their fields, which is bound to attract attention. In statistics, the representative knowledge process extracted from data often emphasizes the representation of mathematical models. The model occupies a central position. It can not only summarize the relationship between the analyzed variables, but also make a summary description of the data, which has universal applicability. In statistical analysis methods, there are mainly variance analysis, correlation analysis, principal component analysis and regression analysis [5]. Statistics, on the other hand, refers to the process of statistics and analysis of various attribute relationships using professional statistics, probability theory principles, etc. Through analysis, the association and development rules between attributes are successfully found. Statistics is a science that studies statistical principles and methods. Specifically, it is the principle and method of studying how to collect, sort out and analyze the digital data reflecting the overall information of things, and based on this, infer the overall characteristics. At the same time, judging from the massive data to be processed by data mining and the complexity of the data, the traditional statistical methods of inference and inspection based on general assumptions have shown great limitations. In the mining phase, users can use visualization tools to perform simple operations on various algorithms. Visualization technology is still needed to show and explain the discovered knowledge in the stage of representing results.

4. Application of Statistical Methods in Data Mining

4.1. Cluster analysis

Cluster analysis in multivariate statistical analysis provides a variety of clustering methods. Problems that are not known in advance should be divided into several categories, and if the specific classification of observed individuals is not known, the observed data should be analyzed and processed. Data mining has many new features. First of all, data mining is faced with a huge amount of data, which is also the reason for data mining. Secondly, the data may be incomplete, noisy and random. Depending on the type of data mined or the given data mining application, the data mining system may also integrate technologies in the fields of spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, etc. Statistics regards each sample as a point in the m-dimensional space, and defines the geometric distance in the space. the points that are close to each other are classified into one class, and the points that are far away from each other should belong to different classes.

Statistics is a common method of concept clustering. We discuss the characteristics of concept clustering algorithm on the side of Cobweb algorithm. Cobweb algorithm creates hierarchical clustering in the form of a classification tree. Each node in the classification tree corresponds to a concept (class) [6]. It also contains a probability description of the class.

Probability description includes the probability of concept and conditional probability in the form of $Q(C_i = M_{ij} \mid K_l)$.

Where $C_i = M_{ij}$ is an attribute-value pair and K_l is a concept class. Cobweb algorithm uses classification metrics to guide tree creation. classification utility is defined as follows:

$$\frac{\sum_{l=1}^{n} Q(K_l)[\sum_i \sum_j Q(C_i = M_{ij} \mid K_l)^2 - \sum_i \sum_j Q(C_i = M_{ij})^2]}{n} \quad (1)$$

Where: n is the number of concepts or classes; Probability $Q(C_i = M_{ij} \mid K_l)$ indicates the similarity within the class. The larger the value, the more predictable the attribute-value pair is to be a member of the class. The probability $Q(C_i = M_{ij})$ indicates that the class is different, and the greater the value, the smaller the sharing of class attribute-value pairs.

The methods of defining distance usually include absolute distance, Euclidean distance, Chebyshev distance, etc. It should be said that all algorithms serve a certain mining system. The research of data mining system is to establish a scientific system structure, which is conducive to the reuse and embedding of mining algorithms and the organic combination of algorithms and other modules of the system [7]. It allows certain correlation between attributes within certain classes and certain inheritance between classes, i.e. some classes share certain model parameters in the class hierarchy.

4.2. Probabilistic analysis network

In the process of data processing, its learning speed is one hundred times faster than that of the BP algorithm for the same problem, and its accuracy is also relatively high, which fully shows that the probabilistic analysis network is always faster than the weight-threshold network in some performance [8]. Using mathematical models, we can reveal the internal structure of things, analyze the relationship between variables, and make statistical inference and prediction. If a certain method can bring the desired results, its rigour can be temporarily ignored. Data mining is a pattern establishment and pattern recognition based on existing data. It is to mine information and discover knowledge without clear assumptions. Therefore, it is naturally hoped that fewer comprehensive variables will be used to replace more original variables, and at the same time, it is required that these comprehensive variables can reflect as much information as possible of the original variables and are not related to each other. Simulate the function of human neurons, adjust and calculate the data through input layer, hidden layer and output layer, and finally obtain the results for classification and regression. Faced with today's complex data, it has not only met certain criteria such as independent co-distribution, but may only meet a small proportion of statistical models. It emphasizes the processing of massive databases and reveals the patterns existing in the data and the relationships among the data. It can be seen as an automatic exploratory analysis of a large number of complex data sets by computers.

4.3. Principal component analysis

Principal component analysis (PCA) is an analysis method that focuses information scattered on a group of variables on some comprehensive indexes (principal components) in order to use principal components to describe the internal structure of data sets and achieve the purpose of data interpretation. Analyzing the interrelation between variables and using the learning and statistical inference functions prompted by Bayesian theorem can realize data mining tasks such as prediction, classification, clustering, causal analysis, etc. It quantitatively studies the laws of certain specific phenomena (such as social and economic development). However, with the continuous expansion of the scope of statistical research and the effective application of statistical methods in the social and natural fields. In general, the mathematical treatment is to make the original P indexes into a linear combination as a new comprehensive index. If the first selected linear combination, that is, the first comprehensive index, is taken as F1, we hope to reflect as much information as possible on the original index [9]. It does not require the discovery of knowledge that is universally applicable, nor does it require the discovery of new natural science theorems and pure mathematical formulas, nor is it the proof of any machine theorem. In fact, the position of computer knowledge in data mining is just like that of mathematics in physics and other disciplines. It is only an essential tool.

We cannot regard it as a branch of mathematics because physics and other disciplines apply mathematics extensively. The statistical simulation tool is used to analyze the data, and the variation result of the average convergence complement length of the data is obtained, and then the variation rule of the data is analyzed.

4.4. The application of Bayesian network in data mining

Bayesian network originates from the research in the field of artificial intelligence. It is a combination of graph theory and probability theory. It can help people to use probability statistics to reason about uncertainty and analyze data. Statistics require high data quality; However, data mining is a process of directly extracting potentially useful information or knowledge from a large number of incomplete, noisy, fuzzy and random practical application data. This is the essence. According to the authoritative definition of statistics in Encyclopedia Britannica, all methodological science that studies how to collect, analyze, express and explain data is statistics. Especially with the application and research of Bayesian network in machine learning, probability theory, mathematical statistics and data mining are closely linked. Bayesian network discrimination makes up for its deficiency. Stepwise discriminant method is to select variables with strong discriminant ability through variable discriminant ability test to screen discriminant variables and establish discriminant functions. This method is relatively more effective. Knowledge analysis and integration of useful data with probability statistics can provide experience for future decision-making. We are constantly exposed to random events, probabilities, random variables and their distribution, and we need to understand the deeper meaning and achieve mastery through a comprehensive study. In practice, we found that a method can bring the desired results, but due to the rigor of statistics, it cannot be used. Obviously, traditional statistics lack the necessary spirit of adventure.

4.5. Regression analysis

Statistics also has corresponding methods from different angles to realize correlation analysis. Variance analysis is divided into single-factor and multi-factor variance analysis, which is a statistical method for analyzing experimental (or observational) data. A new data block structure is formed, and then better data blocks are formed through heredity, regeneration and mixing until the optimization of the data structure is completed, and then the optimal solution of the data is obtained. People replace the reorganization operation by extracting information from the preferred solution set, and then use the distribution probability of this information to generate new solutions, thus realizing chain learning of the algorithm. The basic idea of this method is that probability distribution determines each clustering state, and each data in the model is generated by multiple probabilities in the distribution state. The basic idea is that each cluster is determined by a probability distribution, and each data can be regarded as a mixed distribution of multiple probability distributions. Convert the data into an analysis model. This analysis model is established for mining algorithms. The key to the success of data mining is to establish an analysis model that is truly suitable for mining algorithms. For example, when analyzing sales data in a retail industry, what is the data itself? Before all the data is known, users are not interested and only some original data is enough. Of course, this difference is not absolute.

5. Conclusion

The method of model query and optimization that is being developed in data mining is an interesting research area at present. In view of the maturity of statistical theory and the universality of statistical application, statistical data mining technology has become the most mature data mining technology at present. However, the clustering method in data mining cannot be simply equated with the clustering method in statistics, which adds a large number of database technologies and machine learning methods and pays more attention to the efficiency of the algorithm. With the continuous expansion of data sources and the increasing complexity of data structures, relying solely on data mining technology has gradually revealed its inability to meet its needs. However,

the synchronous development of statistics is continuously enriching and perfecting data mining technology. Statistics can play an important role in data mining: excavation science, and they are closely linked. Statistics should cooperate with data mining and jointly develop forward in a perfect way.

Acknowledgements

The Social Science Research during the period of 13th five-year plan of the Education Department of Jilin Province, Topic: The Research and Practice of the Higher Vocational College General Education's Cooperativity Effect in the Ideological and Political Education, Project Number: JJKH20180634SZ.

References

[1] Dong Liqing. On the Application of Mathematical Statistics in Total Quality Management. Time Report: Academic Edition, no. 3, pp. 342-342, 2015.

[2] Li Jingyuan, Gao Fabao. Exploring the distribution law of examination scores of “Probability Theory and Mathematical Statistics” in college public courses—taking Yangzhou University as an example. Statistics and Applications, vol. 006, no. 003, pp. 333 350, 2017.

[3] Duan Shijuan. On the application of probability statistics in investment decision-making. Time Report: Academic Edition, no. 3, pp. 341-341, 2015.

[4] Rong Ni. Cognitive application and teaching methods of mathematical statistics. Chinese extra-curricular education (Elementary Education Edition), no. 011, pp. 121,123, 2017.

[5] Xu Xiangdong. Strategy analysis of teaching management using probabilistic statistics. Statistics and Management, no. 8, pp. 191-192, 2016.

[6] Zhai Xue. Research and analysis of probability theory and mathematical statistics based on big data. Science and Technology Economics Guide, v no. 025, pp. P.141-, 118, 2016.

[7] Guo Xuetao. Using probability theory and mathematical statistics theory to optimize the product quality supervision and spot check system. China Inspection and Testing, no. 03, pp. 44 + 64-66, 2017.

[8] Huang Jiazeng. Application analysis of modeling thought in the teaching of "Probability Theory and Mathematical Statistics". Market Forum, no. 007, pp. 82-84, 2018.

[9] Luo Lin, Wang Yazi, Liang Yanyan. Discussion on the teaching of probability theory and mathematical statistics. Journal of Zhoukou Normal University, vol. 033, no. 005, pp. 57-59, 2016.