Applied Analysis of Probability Theory and Mathematical Statistics in Data Mining

Jidong Zhao[1(✉)], Xiaoxuan Gong[2], and Cheng Zhenhua[3]

[1] Basic Teaching Department, ShanDong JiaoTong University, Weihai 264209, Shandong, China
zhaojidong0913@163.com
[2] Yunnan College of Business Management, Yunnan 650106, China
[3] Sports Department, Modern College of Northwest University, Xi'an 710130, Shaanxi, China

Abstract. With the improvement of the national scientific and technological level, China has entered the era of big data. If we want to achieve better development in the era of big data and make rapid progress in the domestic socio-economic level, we need to find the universal laws of data in massive, complex and low correlation data, which requires the use of data mining and probability theory and mathematical statistics. Probability theory is a branch of mathematics that studies probability, random variables and random functions. It has been applied in many different fields such as engineering, computer science, economics and so on. Data mining is the summary and analysis of a large number of data, while probability theory and mathematical statistics are the more detailed analysis of data based on data mining. The two complement each other and achieve each other. Probability distribution is a function that assigns probability to the result or event according to the occurrence or non-occurrence of the result or event in the experiment or observation. Apply probability theory and mathematical statistics to data mining to improve the accuracy and efficiency of data mining. Therefore, in order to improve the quality of data mining, it is necessary to apply statistical methods effectively. On this basis, the specific application of statistics in data mining is analyzed through the connection of probability theory, mathematical statistics and data mining. Combined with specific algorithms, the application of statistics in data mining is discussed. This paper studies and analyzes the characteristics of data mining, probability theory and mathematical statistics, and discusses the specific application of statistics in data mining.

Keywords: data mining · probability theory · mathematical statistics · application analysis

1 Introduction

The so-called thought refers to the result of objective existence reflected in human consciousness through thinking activities, which is the basis of all human behaviors. People are great because of thought, and people are noble because of thought. Mathematical thinking refers to the result of thinking activities that the spatial form and quantitative relationship of the real world are reflected in people's consciousness. Mathematical

Y. Zhang and N. Shah (Eds.): BigIoT-EDU 2023, LNICST 580, pp. 90–100, 2024.
https://doi.org/10.1007/978-3-031-63130-6_12

thought is the essential understanding of mathematical facts and theories after generalization. The basic mathematical thoughts are the foundational, summative and most extensive mathematical thoughts that are embodied or should be embodied in basic mathematics [1]. They contain the essence of traditional mathematical thoughts and the basic characteristics of modern mathematical thoughts, and are developing historically. Through the cultivation of mathematical thinking, the ability of mathematics will be greatly improved. To master mathematical ideas is to master the essence of mathematics.

There are both differences and connections between mathematical thought and mathematical method. The main difference is that mathematical thought is the soul of mathematics and plays a guiding role in the practice and exploration of mathematics. In other words, mathematical thought is the guiding principle of mathematical work. Any research on mathematical work is guided by mathematical thought, while using mathematical methods to solve daily mathematical problems, Mathematical methods, like tools used in daily life, generally focus on the practical work of mathematics. They are the manifestation and realization means of mathematical ideas, and the concrete manifestation of the application of theory to practice; Compared with mathematical methods, mathematical ideas are more abstract, universal and general, which shows that mathematical ideas are implicit. That is to say, mathematical ideas are often hidden in daily mathematical practice, but mathematical methods are often concrete and operable [2]. The mathematical methods used in the solution of a mathematical problem or the exploration of mathematical theory are both concrete and available.

At the same time, mathematical thinking and mathematical methods are closely related, interdependent, and inter-dependent. Mathematical thinking guides mathematical workers to flexibly use mathematical methods to solve mathematical problems and explore mathematical knowledge, and mathematical methods are more to verify mathematical thinking and enrich mathematical thinking in mathematical practice. Mathematical thinking and mathematical methods are inseparable in the daily mathematical exploration work, and the boundary between them is not clear and fuzzy. Therefore, in the actual mathematical learning and research, we always refer to these two concepts as mathematical thinking methods. There are many commonly used mathematical thinking methods, such as reduction thinking method, limit thinking method, number and shape combination thinking method, classification discussion thinking method, function and equation thinking method, analogy thinking method, statistical thinking method, construction thinking method, and so on [3].

Although probability and mathematical statistics are used for data statistics, they are also widely used for data mining. Data mining is to discover potential patterns and rules in the database and apply them to better understand and use data. Data mining is a means to obtain useful information from databases or other databases through computer technology and modern communication technology. By collecting, sorting and processing a large amount of data, we can mine the knowledge and rules hidden in the data for users to achieve higher work efficiency. Data mining is a new science. It is a data classification with large, incomplete and discrete attributes. It can extract data and make it more reasonable; New, unified and valuable information, which provides data support for public decision-making, is an effective means to comprehensively analyze and solve various practical problems. In today's increasingly mature social and economic

development, all industries need better data analysis and utilization capabilities to make timely and accurate judgments, so as to promote the development of the industry. Data mining mainly uses computer technology to analyze complex data through advanced algorithms, explore the internal relationship between data, and find some rules in the data; Provide users with decision-making basis [4].

In the era of big data, probability theory, mathematical statistics and data mining all play an important role in data management. In data management, probability theory and mathematical statistics are the most commonly used methods of induction and summary, which can make messy data orderly and facilitate technicians to predict future data. Data mining is a new technology, which generally uses computers with high computing power to process data. Through data mining, massive and complex data can be sorted out to make these data present regularity and uniformity, so as to facilitate the interpretation and analysis of these data by technicians. Both data processing methods can find the laws existing in huge data and predict and analyze the future in advance.

2 Related Work

2.1 Bottleneck of Data Analysis

Mathematical statistics, namely statistical theory, is a methodological science that studies how to collect data, sort out data, and analyze and infer data. Its purpose is to explore the internal quantitative regularity of the data of things, and to achieve a scientific understanding of objective things. Mathematical statistics plays an important role in all walks of life in reality. The solutions to a series of problems such as seed selection, fertilizer selection and farming conditions in agricultural activities are all related to mathematical statistical methods. Generally, the investigation results are obtained through proper design and statistical analysis of field experiments. The statistical methods of mathematical statistics, such as experimental design, regression design, variance analysis and multivariate analysis, play a wide role in the trial production of new products, improvement of old products, reform of technological processes, use of raw materials and search for appropriate formulations in industrial production [5]. The significance of mathematical statistics in social and economic fields can not be ignored, for example, in the investigation and prediction of social population, and in the analysis of ability in psychology, mathematical statistics knowledge should be used. In a word, in modern society, people's life, economic development, social policies and so on can not be separated from mathematical statistics to pave the way.

More and more applications are constantly producing a large amount of data: from medical image processing, gene sequencing to weather observation data, satellite data: from financial industry, insurance industry, communications industry to pharmaceutical industry, e-commerce, traditional retail industry; From the system log, microblog, SNS user relationship network, website click stream… we are producing a huge amount of data that could not be imagined in the past at any time.

Let's take a look at the following set of data: Tencent currently has more than 783.9 million QQ active accounts (the maximum number of simultaneous online accounts reached 167.3 million), 592.8 million QQ space active users and 469 million microblog users; The registered users of Sina Weibo have also exceeded 400 million, and 100

million new microblogs will be generated every day; China Mobile's 700 million users will contribute 360 billion minutes of voice calls, 69.6 billion MB of network data traffic and 62.4 billion SMS messages every month on average in 2012. These users' personal information, communication records, network traffic records, SMS records, and business customization records constitute almost the largest user behavior information database in the world; As the world's largest bank by market value, ICBC provides services to 4.11 million corporate customers and 282 million individual customers through its 16648 outlets. Every deposit and withdrawal data, every online banking login or transfer data, will be accurately recorded by ICBC. According to IBM's report, 2.5 EB (1EB = 1024 PB) of data was generated every day in 2012, and this number is doubling every 40 months—whether you accept it or not, the era of big data has arrived [6].

The data generated by all walks of life is not only huge in quantity but also diversified in type. In addition to traditional structured data, we are exposed to more and more types of unstructured data: system logs, user complaint letters, SNS user friends, microblogs, pictures, audio; Video, sensor data… Not all data can be applied to the relational database model we are familiar with in the past. New technologies are needed to meet the needs of big data.

In recent years, column-stores have received widespread attention. In short, the column number database is a database that stores the data table columns separately. The attribute values belonging to the same column are continuously stored on the disk, while the traditional database stores data item by item. Because all attribute values of a column belong to the same data type, a columnar database is easier to compress than a traditional database. Unlike a traditional database, which cannot omit null values in a record, a columnar database does not need to store null values. These characteristics make a columnar database occupy less disk space. The columnar database is more suitable for OLAP (online analytical processing) than the traditional database, because OLAP generally only needs to read a part of the columns in the table. If the columnar database is used, only the data of the corresponding columns need to be read, but if the traditional database is used, the entire table needs to be read [7]. The ParAccel Analytic Database database adopts column-oriented storage technology and MPP technology, which combines the characteristics of fast query speed and compression of columnar database with the scalability of large-scale parallel processing technology.

2.2 Relationship Between Statistics and Data Mining

Statistics is a discipline that studies the basic principles and methods of statistics, mainly including the relevant knowledge of mathematics and probability, which is easy to collect, organize and process. In these three aspects, data processing is the most important and commonly used link in statistical work (as shown in Table 1). The data analysis mainly adopts analysis of variance, correlation analysis and regression analysis. Because of many defects in traditional data processing methods, they have been widely used in modern enterprise operations. With the rapid development of Internet and computer technology, people's demand for information is also increasing, so data mining has become a necessary means. Data mining refers to the use of knowledge discovery capabilities in databases or data warehouses to deal with practical problems, so as to

further improve the company's operation and management capabilities. Data mining is the in-depth analysis, summary and mining of a large number of data.

Table 1. Data processing and analysis

$k = \{0,1\}$	$n \geq 0$ $0 \leq p \leq 1$	$\lambda \geq 0$
p > 0	$k \in \{0, \ldots, n\}$	$k \in \{0,1, 2, \ldots\}$
$\frac{q-p}{\sqrt{pq}}$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\lambda^{-1/2}$
$\frac{6p^2-6p+1}{p(1-p)}$	$\frac{1-6p(1-p)}{np(1-p)}$	λ^{-1}
p	np	λ
pq	np(1–p)	λ
$-q\ln(q) - p\ln(p)$	$\frac{1}{2}\ln(2\pi nep(1-p)) + O(\frac{1}{n})$	

(1) Statistics, like data mining, extract the structural features of the data from a large number of data, and analyze the internal connections and attributes between the data.
(2) Data mining technology is a major direction of the current development of statistical technology, which puts forward a new idea for statistical analysis and data processing.
(3) As an important link in statistical analysis, data mining has been widely used in database, intelligent processing, machine learning and other aspects.
(4) Statistics and probability theory is a relatively mature and widely used technology in the field of data mining. Its application will promote the in-depth and effective development of data mining.

Both data mining and statistics are used to sort out, summarize and analyze huge data, find common rules in the data, and predict and make decisions in the future. As for the relationship between the two, data mining is an important branch of statistics. As a new statistical method, data mining provides statistics with a new idea of data statistics, that is, using advanced computer technology, database technology, etc. to carry out statistical analysis on huge data. Probability theory and mathematical statistics, as the important content of traditional statistics, have been developed for many years. Their accurate and efficient processing methods will also promote the level of data mining work to be improved again [8]. The two statistical methods should learn from each other and make progress together to promote the rapid development of statistics.

3 Application of Statistics in Data Mining

(1) Probabilistic analysis network

ANN is a widely used data mining algorithm that is composed of a series of networks called data nodes. The processing method of input, output, numerical value and other nonlinear data is adopted to adjust the positioning of each node, so as to realize the data analysis. Data mining requires a large amount of qualitative data, quantitative data and other data, which are data lost, so it is necessary to combine data mining and data mining technology to mine the rules behind the data. Based on this, the application of probabilistic statistical network in data mining is deeply discussed. Probabilistic analysis network is a data analysis model based on mathematical statistics. Its basic principle is to allocate the association between the sample points, so as to obtain the overall trend change. Probabilistic analysis networks are a general framework for statistical inference in a variety of complex systems. Probabilistic networks are used for data mining, including data pattern recognition, optimization problems in non-linear regression, and other data utilization and processing. Probabilistic analysis method is a new method that can accurately judge the different data distributions in the existing situation, so as to draw uncertain conclusions. This method also uses the traditional weight-threshold neural network (BP) as a new weight-threshold network, with a learning speed of more than 100 times faster than BP, and with higher accuracy. At the same time, the association of the original data is probability to achieve the purpose of prediction. This fully proves that in a sense, in a sense, the computational rate is faster than the weight-threshold network. Meanwhile, the model can be predicted based on the input information, so the model has high accuracy and high reliability. In practice, because the characteristics of the network are the specific structure of its nodes and the high randomness, it needs to be emphasized in practice. The Markov Chain Table (Markov) is an algorithm that can synthesize and quantify a large number of nonlinear data comprehensively, and deeply discusses the transmission probability and convergence between different states. This method focuses on solving the optimal parameters of the network model [9]. By studying the relationship between various data, and analyzing the transfer matrix in the uncertain data network, we conclude that the average convergence and complement, and the evolution trend of the data is studied.

(2) Application of Bayesian networks in data mining

A series of work on data mining, data cleaning, transformation and visualization was conducted, and the results of data mining were verified and analyzed. The intrinsic correlation and logicality between the data are studied by using decision tree, neural network and Bayeses methods, and they are also described intuitively by using data mining technology.

Bayesian neural networks are widely used in many fields, such as machine learning, Bayesian, machine learning, etc. Due to the accelerating speed and speed of the computer operation, the operation amount of the Bayesian network is getting bigger and larger, and it is more and more difficult to meet the demand of massive and high precision. Using Bayesian principle and probability allocation (physical or Bayesian) method, logical connection can be easily solved between data; such as prediction, analyzing data analysis, mining clustering, etc. In bioinformatics theory, a large number of complex gene expression data are often achieved by computer. Probabilistic evolution

methods have good robustness and good convergence, so they have attracted wide attention from domestic and foreign scholars. I often encounter some problems in practice. For example, whether the sample data has a certain reliability and uncertainty, or there is some unknown information, and then using the Bayesian method for abnormal analysis, usually use a close way to analyze the internal connection of each data; to realize the processing of data.

The work steps of data mining are divided into the following steps: First, the staff should clean up the massive data, eliminate some irrelevant data, and retain valuable data for analysis. Secondly, it is necessary to analyze, summarize and summarize the retained data, and transform the massive data through statistical means to reveal the universal laws of data. Finally, the transformed data results should be represented by statistical charts or other forms of expression to make their images easy to understand and understand. It is worth noting that the staff should finally analyze the data results to judge the accuracy of the results. In the whole data mining work, we need to use neural network technology, Bayesian algorithm and other technologies. Through these statistical technologies, we can help staff quickly obtain the universal rules between data. Bayesian network plays an important role in the whole development of statistics [10]. With the passage of time and the progress of theoretical and technical level, Bayesian network has gradually become a calculation method of uncertain knowledge and data reasoning, which is inextricably linked with probability theory and mathematical statistics. People need to use Bayesian network in the process of studying machine learning. By using Bayesian network, staff can summarize and analyze massive data, get the general rules between massive data, and promote the rapid progress of machine learning research, as shown in Fig. 2. In addition, the data disorder diagram in Bayesian network solves the problem of complex and obscure data statistics results. By combining the data statistics results with the probability diagram model, the staff can easily obtain the rule information of the data results. It is worth noting that in the process of using Bayesian networks, approximate methods are generally used to analyze data and obtain the general rules of data (Fig. 1).

(3) Data mining based on the probabilistic evolution algorithm

Genetic Analysis (GA) is a commonly used algorithm in data mining, which is also commonly used when processing some data. The groups were then divided into offspring of different sizes. With variation as assistance, data reconstruction and other optimization, so that GA can establish modules in a large number of nonlinear data, data processing, after processing, after screening, recombination, gene composition; before completing the data structure optimization, get the best data, get the best data, get the best results. Using this method, the redundancy of multiple data groups can be reduced, and the data processing accuracy can be improved. However, in practice, because of the data recombination, it often leads to the data fragmentation, so that the routine algorithm can not effectively process the data. In the case of data change, in order to retain the original information, it must be redesigned to avoid the original knowledge loss, resulting in a reduced operational efficiency of the system. This type of data destruction is called a link locking problem (as shown in Fig. 2). The paper makes the traditional changes to the random distribution to make it closer to the real situation. Based on this idea, this paper gives a new solution idea, and uses it in data mining, so as to obtain a number of

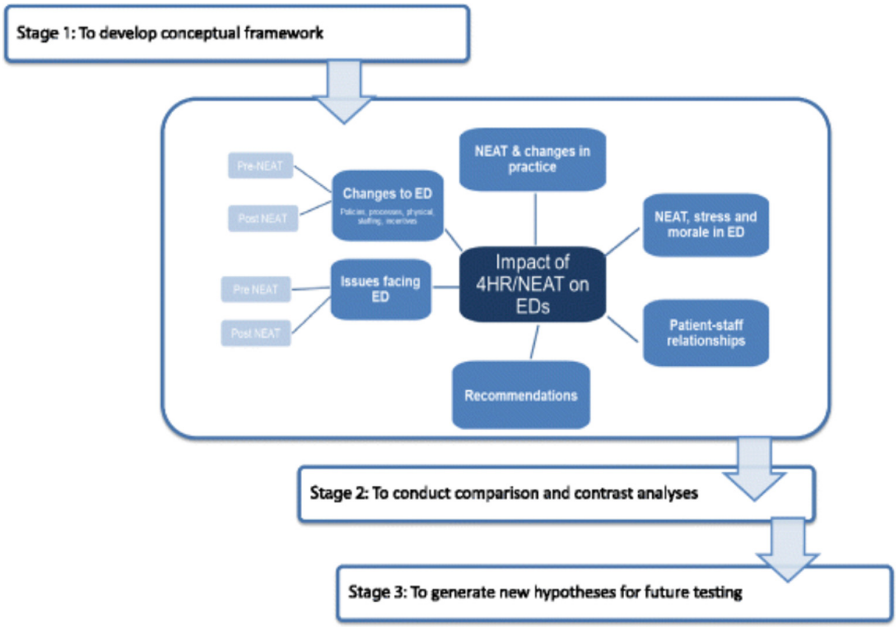

Fig. 1. Application of Bayesian network in data mining

large problems. The proposed algorithm can construct the corresponding classification mode according to the different decision tables. The method is based on the basic ideas and basic principles of PMEA, and it is used for stochastic evolution, by expanding the method through its analysis to reduce the operational complexity and make it more efficient. The experimental results show that this algorithm is very effective. Probabilistic evolution method is widely used in various industries because of its good convergence performance and good robustness. It can quickly find the hidden laws and laws, and analyze some inaccurate or uncertain information; so its application prospect is very broad. However, in PMEA, import the predicted information into PMEA to solve the difficulties encountered in the learning process; and enhance the efficiency and accuracy of data mining. Therefore, we need to construct a precise, reasonable mathematical model. Which are the most critical questions in the evolution of possibilities. Traditional stochastic evolution algorithms do not solve these problems very well, and therefore do not yet serve as a general, universal collection of data available for multiple complex systems. This paper conducts an in-depth study on this issue. On this basis, combining the stochastic evolution methods, such as: population size, selection mechanism, ratio, and other problems, the current probabilistic evolution methods are the main [4].

Compared with other statistical methods, the accuracy of statistical results obtained by genetic algorithm will be higher. However, because genetic algorithm is a interlinked process, once there is a mistake in a calculation link, it will lead to a deviation in the final statistical results. In the actual operation process of the staff, the data block is often lost due to data reorganization, and the local data block is abnormal, resulting in serious deviation of the data result, which can not effectively summarize a large amount of

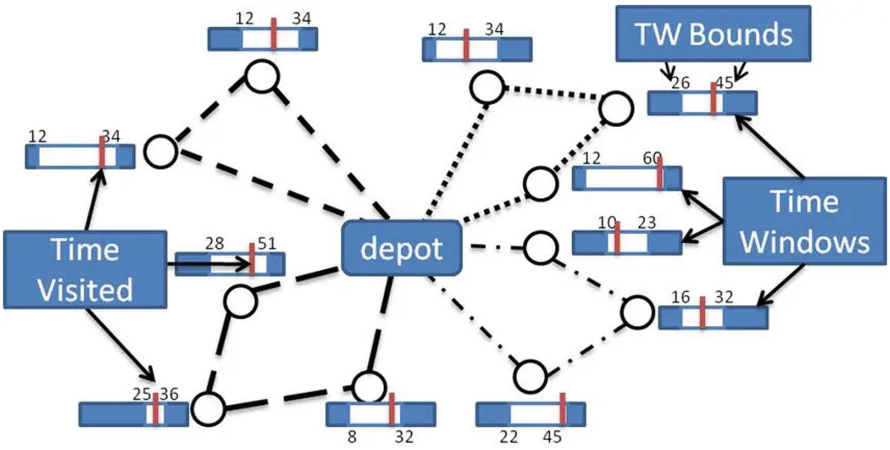

Fig. 2. Link lock

data. Therefore, the staff should always pay attention to the calculation process during the implementation of genetic algorithm, and select excellent data to replace the data reorganization during the data reorganization to ensure the smooth completion of genetic algorithm. The introduction of genetic algorithms in data mining can help staff solve a large number of problems, effectively deal with massive and complex data, and obtain the general rules of data. In the actual operation process, staff can also use compressed genetic algorithm to assist in solving, which can reduce the difficulty of calculation to a certain extent and improve the speed of genetic calculation. Nevertheless, there are still many problems in the process of probabilistic evolutionary computation, such as the selection requirements and selection proportion of data at the initial stage of genetic algorithm, which are important issues that need to be solved urgently.

4 Provide New Research Directions

The development of mathematical statistics and probability theory often comes from the processing of actual data and the full use of human needs to explore the changes between massive data. Therefore, extracting useful knowledge from a large amount of data and turning it into available information has become an important concern in the current statistical community. As an emerging technology, data mining, its application scope is expanding, especially in the operation of enterprises. Therefore, based on the real situation, through the analysis of a large number of data, it is found that the current statistical method in China is not perfect, that is, how to determine the data allocation mode and sampling size. The above problem is an important problem facing the current statistical community. This is because when doing data processing, the classification according to the “sample” and “overall” will constantly change to meet the needs of data processing. Due to the large amount of data, it is difficult to accurately define the sample and the whole, and the data change is also diversified; is the gradual characteristics of the data consistent with the advance prediction. At the same time, because the number

of samples is too small or too large, it will lead to a large number of computational costs. Moreover, there are irregularities such as random error, noise and outliers in real data processing. Therefore, how to find low probability events with typical significance in massive data is a very important problem. Because of the real data test, we assume that there are enough small probability events, because there are very few samples, but when the number of data reaches some extent, these data will change. This complex problem cannot be well solved by traditional statistical analysis methods. In addition, due to the increasing amount of data, the traditional mathematical statistical analysis method can no longer meet the needs of large capacity, and the traditional mathematical statistical analysis method also has many deficiencies, such as the need for manual conduct. However, if the above problems are not solved well, it is difficult to deal with a large amount of data effectively. In this way, statistical methods can be better adapted to changes in the data, thus improving the efficiency of data processing.

If Chinese enterprises want to improve the level of economic development in the era of big data, they must make good use of existing database technology. Although the complicated data information is obscure and difficult to understand, if you can understand the knowledge content hidden behind the data information, it will help enterprises make correct choices in the future, avoid possible risks, and promote the overall improvement of the national economic level. In the face of massive data information, data mining technology is born. In the early days, people developed data mining technology and focused on analyzing data through artificial intelligence. With the continuous exploration of data mining technology, people gradually shifted the focus of data mining technology from artificial intelligence to probabilistic statistical methods. Only by combining probability and statistics with data mining technology can we accurately control the future trend in the era of big data. First of all, data mining and statistical methods should be organically combined to promote the progress of statistics, so that statistics can still play the role of induction and data collation in massive data statistics. Secondly, the data mining work should be monitored in real time, and adjusted quickly when the data mining work is abnormal to ensure the accuracy of the data results. Finally, we should establish a visual data mining prototype system to promote the development of statistics.

5 Conclusion

In short, data mining, as an emerging technology, is increasingly expanding, especially in the field of economic management. Data mining refers to extracting useful knowledge and rules from massive amounts of data and performing a series of operations to draw the expected conclusions. The application of mathematical statistics and probability theory in data mining plays a great role in promoting the development of data mining, especially in data processing and data analysis. By learning statistics and probability theory, we can have a relatively complete understanding of the characteristics and change laws of the data.

References

1. Discussion of three paradoxes in the teaching of probability theory and mathematical statistics. Creative Educ. Stud. 09(3), 617–620 (2021)

2. Paliy, I.: Probability Theory and Mathematical Statistics (2021)
3. Gao, J., Zhi, L., Sun, J., et al.: Research on the Application of Blended Teaching in Probability Theory and Mathematical Statistics Based on MOOC + SPOC + Flipped Classroom, vol. 6. Science Publishing Group (2021)
4. Sheynin, O.: Theory of probability and statistics as exemplified in short dictums (2021)
5. Hu, Y.: Fire risk assessment of urban utility tunnels based on improved cloud model and evidence theory. Appl. Sci. **13** (2023)
6. Jonis, M., Monnier, M.C., Schmid, O.: Analysis of regulatory framework and standards applied to organic wine-making in Europe (2022)
7. Lu, L., Zhou, J.: Research on mining of applied mathematics educational resources based on edge computing and data stream classification. Mob. Inf. Syst. **2021**(7), 1–8 (2021)
8. Schmidt, J.H., Mccann, C.: ESG challenges in the construction of UK balanced portfolios for private investors: an analysis of the availability and performance of ESG funds across various asset classes. J. Appl. Financ. Banking **12** (2022)
9. Ummarino, G.A.: Mathematical and physical properties of three-band s+– eliashberg theory for iron pnictides (2023)
10. Xuan, D., Di, Z., Yikai, C., et al.: Characteristic mode analysis: application to electromagnetic radiation, scattering, and coupling problems. Chin. J. Electron. **33**, 1–14 (2023)