Fast Mining Method of User Behavior Data based on Bayesian Network

Yuting Jiang
Information Engineering Institute, Jiangsu Maritime Institute, Nanjing, 211100, China

Abstract: The conventional user behavior data mining method can achieve the data mining of user behavior. However, due to the large volume of data filtering, there is a shortcoming of user behavior mining response time. Therefore, a fast mining method based on Bayesian network for user behavior data is proposed. The Bayesian network is introduced, the overall framework of the mining model is built and a fast mining algorithm for user behavior data is designed to complete the establishment of a fast mining model of user behavior; Based on the user behavior data selection mechanism, relying on the mining information input sequence and the construction of the output protocol, the user behavior can be quickly mined. The experimental data show that the proposed user behavior data mining method has a higher average response time of 47.23% compared with the conventional user behavior data mining method, and has high data mining accuracy, which is suitable for the rapid mining of user behavior data.

Keywords: Bayesian network; User behavior; Data mining; Mining method

1. Introduction

The conventional user behavior data mining method can achieve the data mining of user behavior. Due to the large amount of data screening and limited by mining ensemble, there is a shortcoming of user behavior mining response time, which is not suitable for rapid mining of user behavior data [1]. Therefore, a fast mining method based on Bayesian network for user behavior data is proposed. Constructing a set of variables, introducing a Bayesian network, determining the hardware framework of the mining model, constructing the overall framework of the mining model in the data mining process, relying on user behavior data interval calculation, and quickly mining data interval calculations to construct a fast mining algorithm for user behavior data and establish a quick mining model of user behavior; Based on the user behavior data selection mechanism, the mining information input sequence is constructed, the mining data output protocol is built, and the proposed Bayesian network-based user behavior data mining method is completed. In order to ensure the effectiveness of the designed user behavior data mining method, the user behavior test environment is simulated, and two different user behavior data mining methods are used to simulate the mining response time and mining accuracy. The test results show that the proposed user behavior data mining method is highly effective.

2. The Construction of User Behavior Fast Mining Model

2.1. The importing of Bayesian network

Bayesian network is a probabilistic network based on Bayesian's "Review on Probability Problem Solving" published by Biometrika, and mathematicians represented by Robbins constantly optimize this probabilistic network. Bayesian network[2]. Bayesian network is a mathematical method for behavioral probability statistics and data analysis mining.

If a variable is A={A1, A2,...,An}, where the data of the variable set A satisfies a certain causal relationship D, and the local variable conforms to the relevant probability distribution R, then a Bayesian network can be constructed. The schematic diagram of its Bayesian network is shown in Figure 1[3].

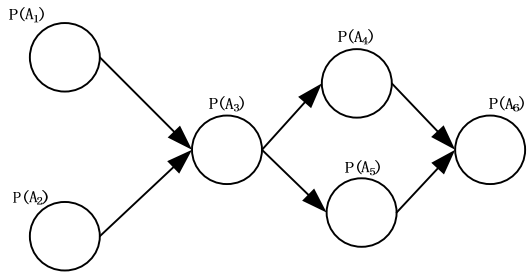

Figure 1. Bayesian network diagram

Fig.1 shows that based on the user behaviors A1 and A2, the user is inferred to operate A3 next according to the causal relationship D and the correlation probability distribution R. At the same time, according to operation A3, the derivative calculation is performed, A4 and A5 are

estimated, and the conclusion A6 is obtained. Its derivative calculation satisfies the following formula [4]:

$$h_f = s\frac{v^2}{2g} \quad (1)$$

Where hf represents user behavior data and v represents user behavior willingness, g represents the basic quantitative information of the user, and the basic quantitative information of the user is obtained according to the basic information of the user login, and can be represented by formula 2. s represents the Bayesian coefficient of value. The basic quantitative information formula of its users is given below [5]:

$$Z = \lim_{0\to\infty} A_n \bigcap_{i=1}^{n} X_i lk f\,! \quad (2)$$

In which Xi represents the user's gender, f represents the user's age, l represents the user's behavioral bias, and k represents the user's willingness to browse.

It can be obtained from Equation 1. The value of the Bayesian value coefficient will directly affect the accuracy of the user behavior estimation results A4 and A5. At the same time, obtaining the response time of accurate Bayesian value coefficient will directly affect the speed of user behavior mining. The selection of the Bayesian coefficient of value and the causal relationship D, the correlation probability distribution R, are satisfied with the relationship of Table 1 [6].

Table 1. Bayesian Value Coefficient Selection

Causal relationship D	Probability distribution R	The Bayesian coefficient of value S
[0.00,0.25]	[0.00,0.50]	[0.00,0.41]
[0.25,0.50]	[0.00,0.50]	[0.41,0.62]
[0.50,0.75]	[0.50,1.00]	[0.62,0.83]
[0.75,1.00]	[0.50,1.00]	[0.83,1.00]

In order to quickly obtain the Bayesian coefficient of value, the causal relationship D and the range of values of the associated probability distribution R are determined firstly, the range of Bayesian value coefficients is determined next according to Table 1, and then the specific Bayesian value coefficient value is determined according to the Robbins function, which greatly saves the time for the Robbins calculation to obtain the Bayesian coefficient. Its Robbins function calculation program is shown below [7].

By determining the causal relationship D, the correlation probability distribution R, and using the Robbins function calculation program, the value of the value of the leaves is determined relatively quickly. Through the calculation of the user's basic quantitative information, the derivative calculation is imported into the Bayesian network to realize the introduction of the Bayesian network.

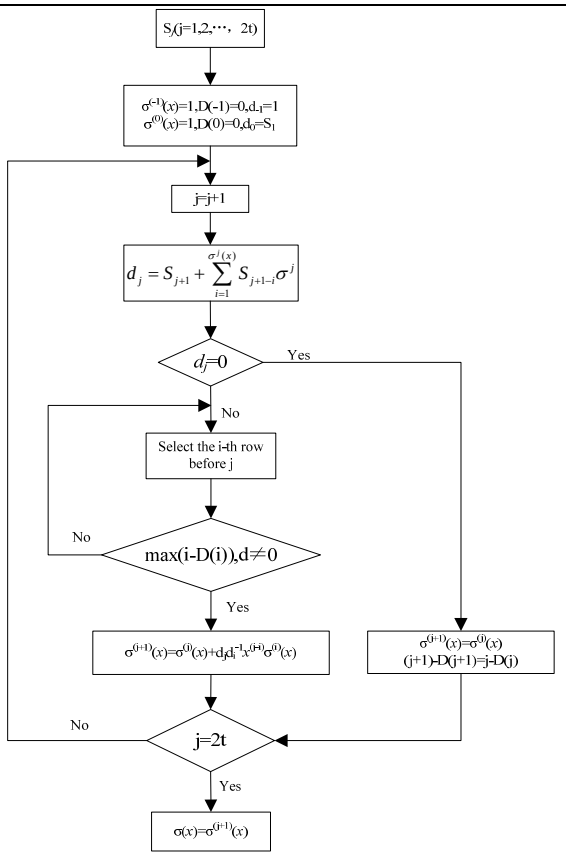

Figure 2. Robbins function calculation program

2.2. Mining model overall frame design

Bayesian network data mining framework, mainly including the ability to safely run Bayesian network user behavior data rapid mining model, central processing unit (CPU), random access memory (RAM), hard disk, display interface card (graphics card, video card, Graphics card) and related equipment such as input and output, the minimum requirements for the hardware composition of its Bayesian network data mining framework are shown in Table 2.

Table 2. Bayesian Network Data Mining Framework Hardware Composition Minimum Requirements

Name	Type	Parameter
CPU	Intel i9 7900X	CPU frequency: 3.3GHz Maximum Turbo Frequency: 4.3GHz Number of cores: ten cores Number of threads: twenty threads
Memory bank	Cras II	Memory capacity: 16GB Memory type: DDR4 Memory frequency: 3200MHz Number of pins: 288pin Slot type: U-DIMM
Hard disk	HGST 7K1000 1TB 7200	Hard disk capacity: 1000GB Single disc capacity: 500GB Cache: 32MB

		Speed: 7200rpm
Graphics card	iGame GTX 1080Ti Vulcan X OC	Core frequency: 1480/1733MHz Memory frequency: 11000MHz Power interface: 8pin+8pin Memory capacity: 11GB Memory bandwidth: 352bit

Relying on the minimum requirements of Bayesian network data mining framework hardware composition, the overall framework of mining model is constructed, and the user behavior data rapid mining algorithm is equipped to achieve the construction of user behavior fast mining model.

2.3. Establishment of a fast mining algorithm for user behavior data

To ensure the rationality of fast mining, a fast mining algorithm for user behavior data is established. Using the user behavior data rapid mining algorithm, the mining calculation parameters are obtained, and the user behavior data is determined to quickly mine the calculation interval to achieve the rapid mining of user behavior data. The user behavior data rapid mining algorithm mainly includes user behavior data interval calculation and fast mining data interval calculation. The user behavior data interval calculation algorithm uses △Pp(k), △Pi(k), △Pd(k) calculation method to obtain the data interval parameter △P(k), and the algorithm path diagram is shown in Figure 3.

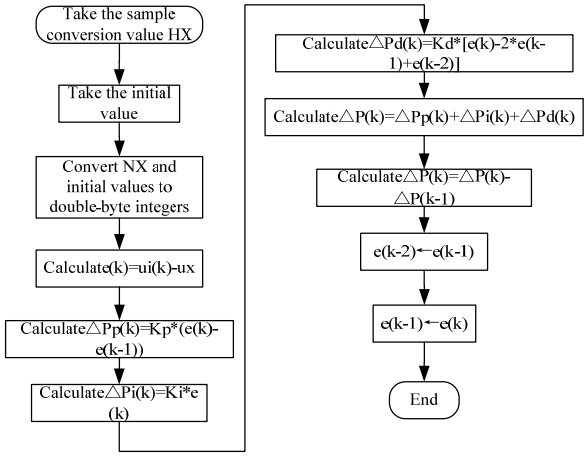

Figure 3. User behavior data interval calculation path diagram

Where △Pp(k) represents user behavior data, △Pi(k) represents the user behavior prediction function, and △Pd(k) represents the behavior prediction coefficient. The path diagram of the fast mining data interval calculation algorithm is shown in Figure 4.

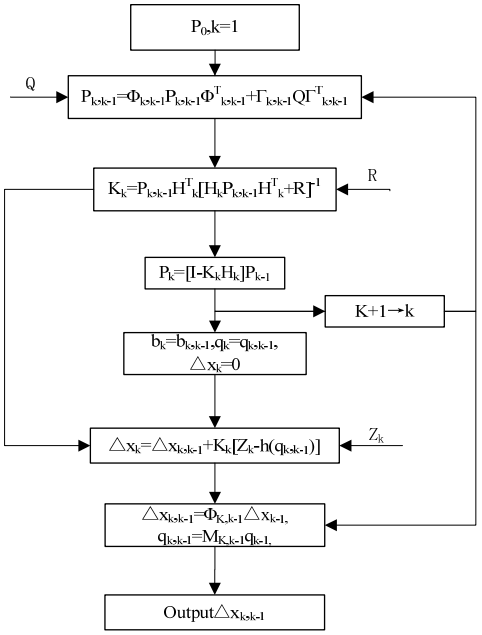

Figure 4. Fast mining data interval calculation algorithm path diagram

In the formula, k, k-1 represents the k, k-1 time user mining data, x represents the data amount, q represents the mining correlation parameter, Φ represents the mining data function, and P represents the fast mining data interval. Through the calculation of user behavior data interval and the rapid mining of data interval calculation, the establishment of fast mining algorithm for user behavior data is completed.

Utilizing the introduction of Bayesian network and the overall framework design of mining model, relying on the establishment of rapid mining algorithm for user behavior data, the construction of user behavior fast mining model is achieved.

3. Achievement of User Behavior Rapid Mining

3.1. The construction of user behavior data selection mechanism

Effective user behavior data selection is of great significance for rapid data mining. The quality of data selection will directly affect the accuracy of the fast mining method. The data selection process is to obtain the data parameters satisfying the data mining conditions through the database. The main processes include data extraction, data filtering, and data loading [8].

The data extraction is to construct a HMPS table for the user behavior database, which specifies the data related requirements. The selection principle is, the selected user behavior data must be representative; the selected user behavior data source information must be accurate [9];

the selected user behavior data related records must be complete; when selecting user behavior data, a certain point selection should not be repeated; the randomness and representativeness of the selection must be guaranteed; the selected user behavior data is inserted into the HMPS table [10].

The data is filtered by extracting the data of the HMPS table, selecting a certain representative parameter for marking, mathematically processing the marking parameters. The data extraction process is the preprocessing of the data screening process. If the data filter process is put directly into the user behavior database, the filtering speed recognition speed will be slower. Through the data extraction process, mathematical screening is carried out, and the data structure is ensured through two layers of data selection structure [11]. The loading of data means that some selected data is loaded into the mining model to perform data mining.

3.2. The construction of mining information input sequence

In order to improve data mining capabilities, the HSFD data transmission sequence method is introduced to transmit user behavior data mining. The HSFD data transmission sequence has the characteristics of small influence by environmental factors, stable operation of the equipment, and high compatibility with the carrier data. The HSFD data transmission sequence is to transmit data according to the set band mode by constructing a data transmission sequence set [12].

The data set includes a data record point R, a data node N, a node attribute M, a node content E, a mapping relationship O, a mapping condition P, and an output D. The R/N stage is edited using the Fr data model, the N/M stage is edited using the Fe data model, the M/O stage is edited using the Fo data model, and the O/D stage is edited using the Fp data model, as shown in the data transmission sequence diagram 5[13].

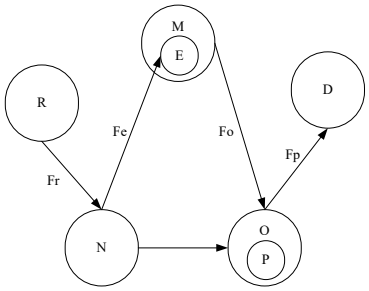

Figure 5. Schematic diagram of the data transmission sequence

3.3. The construction of a mining data output protocol

The mining data output protocol is a working protocol for outputting mining results, mainly including connection services and display services. The connection service uses the ESP protocol for data transmission, and the ESP protocol, also known as the transmission control protocol, controls the transmission data [14]. The display service uses the SHP protocol for data transmission. The SHP protocol is also called the user data packet protocol. The ESP protocol and the SHP protocol are applied to different stages of data transmission, but the basic rules of the protocol are the same. It can be divided into two phases, mainly including the addressing phase and the establishment of the connection phase.

The addressing phase is when the operator triggers a command and the mobile terminal issues a request to transmit data, indicating that the application data is interconnected by multiplexing. It is transmitted according to the data transmission sequence [15].

Establishing a connection phase is to define a sequence of messages according to the transmitted data transmission sequence; A timer is set for each message and a maximum delay is set for the timer to implement data transmission to the message. For those messages that have not received the acknowledgment message beyond the maximum delay, they are considered to have been lost, and the connection is released and needs to be retransmitted. The data transmission diagram is shown in Figure 6.

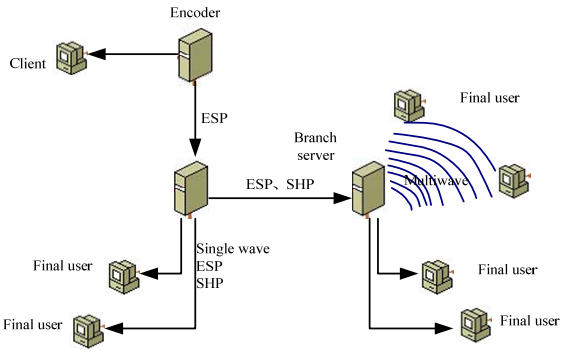

Figure 6. Schematic diagram of data transmission

Based on the user behavior fast mining model construction, relying on the user behavior data selection mechanism and the mining information input sequence to determine the user behavior data filtering and input mining. Based on the construction of mining data output protocol, the display of user behavior data mining results is performed, and the user behavior data based on Bayesian network is quickly mined.

4. Simulation Experiment

In order to ensure the effectiveness of the Bayesian network-based user behavior data rapid mining method proposed in this paper, simulation experiments are performed. During the test, different user behaviors were used as test objects, and the simulation test of mining response time and mining accuracy was carried out. Simulations of different data complexity of user behavior and user behavior are performed. In order to ensure the validity of the experiment, the conventional user behavior data mining method was used as the comparison object, and the simulation test results were compared twice, and the test data was presented in the same data chart.

4.1. Data preparation

In order to ensure accuracy of the simulation experiment, experimental parameters are set. In the simulation test process, different user behaviors are used as test objects. Two different user behavior data mining methods are used to simulate the mining response time and mining accuracy, and the simulation test results are analyzed. Since the analysis results obtained in different methods are different from the analysis methods, it is necessary to ensure the consistency of the test environment parameters during the test. The test data setting results in this paper are shown in Table 3.

Table 3. The Test Data Setting

Programs	Range/parameters of execution	Remarks
Simulation users	Male:Female=1:1 Age 20:30:40=1:1.5:1	Using two methods to perform experiment
Behavior data type	Simple behavior & complex behavior [0,0.5&0.5,1.0]	Taking classified experiment
Simulation system	WUD-21E	

4.2. Mining response time simulation test

During the experiment, two different user behavior data mining methods were used to perform data mining in different user behavior data difficulty simulation environments, and the changes of mining response time were analyzed. Due to the use of two different user behavior data mining methods, the analysis results cannot be directly compared. For this purpose, third-party analysis and recording software is used to record and analyze the test process and results, and the results are displayed in the comparison curve of the test. In the simulation test result curve, the third-party analysis and recording software function is used to eliminate the uncertainty caused by the simulation laboratory personnel operation and the simulation computer equipment factors. Simple behavior mining response time simulation experiments is performed only for different user behaviors, different user behavior data mining methods and different user behavior data difficulty. The comparison curve of the test results in the simple behavior data environment is shown in Figure 7.

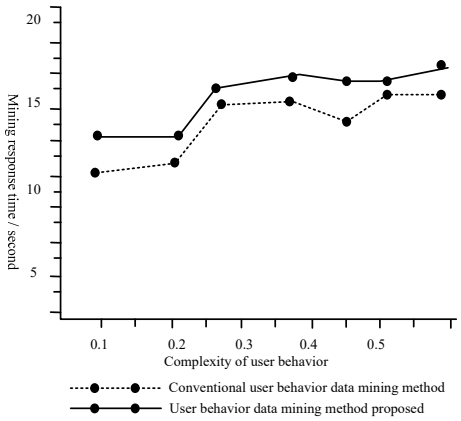

Figure 7. Comparison of test results in a simple behavioral data environment

According to the test curve results, the third-party analysis and recording software is used to perform arithmetic mean processing on the proposed user behavior data mining method and the mining response time and mining accuracy of the conventional user behavior data mining method. The average response time of the proposed user behavior data mining method is 14.25 seconds, and the average response time of the conventional user behavior data mining method is 18.56 seconds. The proposed user behavior data mining method improves the mining response time by 7.06% compared with the conventional user behavior data mining method. In the simple behavior data environment, there is little difference between the two user behavior data mining methods.

At the same time, only different user behaviors, different user behavior data mining methods, different user behavior data difficulty, and mining behavior simulation test of complex behaviors are performed. The comparison curve of the experimental results of the complex behavior data environment is shown in Figure 8.

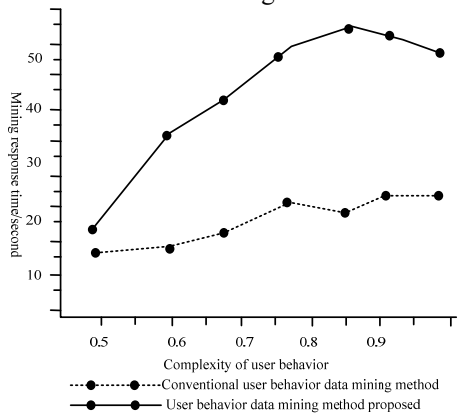

Figure 8. Complex behavior data environment test comparison results curve

According to the test curve results, the third-party analysis and recording software is used to perform arithmetic mean processing on the proposed user behavior data mining method and the mining response time and mining accuracy of the conventional user behavior data mining method. The average response time of the proposed user behavior data mining method is 15.89 seconds, and the average response time of the conventional user behavior data mining method is 51.69 seconds. The proposed user behavior data mining method improves the mining response time by 69.26% compared with the conventional user behavior data mining method. In the context of complex behavioral data, the proposed user behavior data mining method is much shorter than the conventional user behavior data mining method.

4.3. User behavior mining accuracy simulation experiment

During the test, two different user behavior data mining methods were used to work in the simulation environment to analyze the changes in mining accuracy. Due to the use of two different user behavior data mining methods, the analysis results cannot be directly compared. To this end, third-party analysis and recording software was used to record and analyze the test process and results, and the results were displayed in the comparison test curve of this test. In the simulation test result curve, the third-party analysis and recording software function is used to eliminate the uncertainty caused by the simulation laboratory personnel operation and the simulation computer equipment factors. The mining accuracy simulation test is performed only for different user behaviors and different user behavior data mining methods. The comparison curve of the test results is shown in Figure 9.

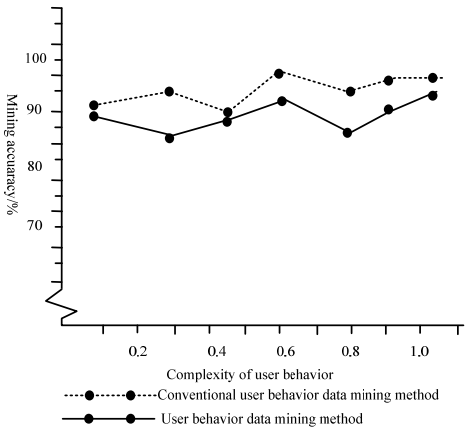

Figure 9. Accuracy test comparison curve

According to the results of the test curve, the third-party analysis and recording software is used to perform arithmetic mean processing on the proposed user behavior data mining method and the mining accuracy of the conventional user behavior data mining method. The average accuracy rate of the proposed user behavior data mining method is 98.56%, and the average accuracy rate of the conventional user behavior data mining method is 85.43%. The proposed user behavior data mining method improves the mining accuracy rate by 13.13% compared with the conventional user behavior data mining method. In the different behavior data environment, the proposed user behavior data mining method has higher accuracy than the conventional user behavior data mining method. According to the above test data, the arithmetic mean value is used to calculate that the proposed user behavior data mining method is 47.23% higher than the conventional user behavior data mining method. At the same time, it has high data mining accuracy and is suitable for rapid mining of user behavior data.

5. Conclusion

This paper proposes a fast mining method for user behavior data based on Bayesian network. Based on the construction of user behavior fast mining model, the user behavior data selection mechanism is determined, the mining information input sequence is constructed, and the data output protocol is mined to achieve the research of this paper. The experimental data shows that the method designed in this paper is extremely effective. It is hoped that the research in this paper can provide a theoretical basis for user behavior data mining methods.

References

[1] Wang Fei, Yue Kun, Sun Zhengbao, et al. Evaluation data analysis and dynamic behavior modeling based on Bayesian network[J]. Journal of Computer Research and Development, 2017, 54(7): 1488-1499.

[2] Yang Yuewen, Hou Huijuan, Yang Wei, et al. Mining and Forecasting Method of Transmission Parameters of Transmission Lines Based on Bayesian Model[J]. Power System Technology, 2017, 41(11): 3648-3654.

[3] Chen Wei, Zhu Biao, Zhang Hongxin. BN-Mapping: Visual Analysis of Geospatial Data Based on Bayesian Network[J]. Chinese Journal of Computers, 2016, 39(7): 1281-1293.

[4] Liu Guangyi, Li Ou, Song Tao, et al. Efficient data transmission method for wireless sensor networks based on Bayesian network[J]. Journal of Electronics & Information Technology, 2016, 38(6): 1362-1367.

[5] Shi Chunyan, Yan Yuting, Wang Lei. Multi-modal health data fusion method based on Bayesian network for body area network[J]. Journal of Transduction Technology, 2017, 30(10): 1602-1607.

[6] Wu Jingwei, Yang Min. Satisfaction Analysis Model of Urban Conventional Bus Service Based on Bayesian Network[J]. Journal of Southeast University(Natural Science Edition, 2017, 47(5): 1042-1047.

[7] Zhang Ming, Wang Fuzhang, Cheng Chao. Fault clustering of urban rail transit equipment and Bayesian network early warning[J]. Computer Engineering and Applications, 2016, 52(11): 259-264.

[8] Mi Hongying, Wei Cong, Li Hongrong, et al. Data mining of the rule of differentiation and treatment of coronary heart disease

with angina pectoris and qi stagnation syndrome based on the theory of sinus theory[J]. Journal of Traditional Chinese Medicine, 2018(5):381- 385.

[9] Ren Jianye, Xu Ming, Lu Jiahui. Research progress on clinical drug law and syndrome type of traditional Chinese medicine based on data mining[J]. Chinese Journal of Traditional Chinese Medicine, 2017(10): 4579-4582.

[10] Liu Fei, Zhang Shaowu, Gao Hongyan. Generic algorithm construction algorithm based on partial mutual information and Bayesian scoring function[J]. Journal of Northwestern Polytechnical University, 2017, 35(5): 876-883.

[11] Di Ruohai, Gao Xiaoguang, Guo Zhigao. Small data set BN modeling method and its application in threat assessment[J]. Chinese Journal of Electronics, 2016, 44(6): 1504-1511.

[12] Fan Tiebing, Ning Qiuping, Yang Zhixu, et al. Explicit analysis of tacit knowledge of traditional Chinese medicine based on data mining technology[J]. Chinese Journal of Experimental Traditional Medical Formulae, 2017(10):221-226.

[13] Zhang Zhongyuan, Lin Zhifeng, Liu Dong, et al. Transformer fault diagnosis based on normal cloud model and improved Bayesian classifier[J]. Journal of Electric Measurement & Instrumentation, 2017, 54(4): 50-56.

[14] Du Chao, Wang Zhihai, Jiang Jingjing, et al. Bayesian Classification Algorithm for Data Stream Based on Exposure Mode[J]. Journal of Software, 2017, 28(11): 2891-2904.

[15] Qian Chao, Dai Liang, Lin Shan, et al. Traffic State Prediction Based on Random Repeat Mountain Climbing Method[J]. Journal of Transportation Systems Engineering and Information, 2016, 16(1): 162-168.