Analysis and Prediction Method of Student Behavior Mining Based on Campus Big Data

Liyan Tu(✉)

Inner Mongolia University for the Nationalities, Tongliao 028000, China
tlyimun@163.com

Abstract. How to effectively mine students' behavior data is an important content to improve the level of student information management. The platform of student behavior analysis and prediction based on campus big data is established, and the value of big data produced by students' campus behavior is analyzed. The behavior data of students' consumption laws, living habits and learning conditions are collected, modeled, analyzed and excavated around the large data environment, and the student behavior is predicted and warned by the stratified model of students' behavior characteristics. The experimental results verify the effectiveness of the methods used, and the behavior characteristics can be analyzed according to the behavior characteristics of the students, and the students' behavior will be guided to the overall health direction in a timely manner.

Keywords: Big data · Student behavior · Prediction model · Data mining

With the continuous development of information technology, cloud computing and data mining technology have been widely applied. The digital campus and the campus management system service platform are increasing, and the data accumulated in the campus information environment have also increased greatly. The data of the students' behavior (learning behavior, life behavior and heart behavior) in the corresponding business system has formed a relatively complete big data on campus. Environment, traditional campus management concepts and data analysis methods have been unable to meet the growing demand for data processing. How to manage and share the campus data efficiently, and optimize the student management by using the large data mining method, and provide a clearer and detailed data service for the students' campus life according to the analysis results are the hot spots of the current student management. It can be seen that making full use of students' school behavior data to build digital campus and intelligent campus makes the level of campus information upgrade, which is the problem facing the construction of campus service system.

1 The Connotation of Data Mining

Data mining refers to the extraction of information hidden in the data which have potential value in the massive and messy data. Through analysis, it can provide people with decision-making process. The main implementation process of data mining is data acquisition, data preprocessing, feature extraction, feature selection, data mining,

G. Gui and L. Yun (Eds.): ADHIP 2019, LNICST 302, pp. 363–371, 2019.
https://doi.org/10.1007/978-3-030-36405-2_36

model evaluation. Estimate. It is a process of continuous optimization. Data preprocessing, data mining and model evaluation are important components of data mining, as shown in Fig. 1.

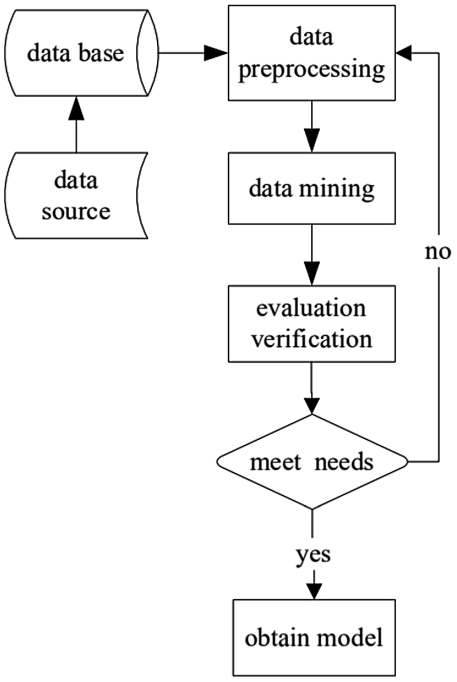

Fig. 1. Process of data mining

1. Data preprocessing; data preprocessing is the precondition of data mining, it is the preprocessing of data information by computer technology. The main function is to clean and screen the invalid and invalid data, and then integrate and transform the data, and lay the foundation for the establishment of the model.
2. Data mining: different mining algorithms are different in data extraction and processing, and the results are different. The most suitable and effective mining algorithms can be selected according to different data characteristics and business requirements.
3. Model assessment: we need to evaluate the model to detect whether the results obtained through data mining meet the expected requirements. If the mining results do not meet the requirements, it is necessary to re select data or uses other mining algorithms.

2 Platform Structures of Student Behavior Analysis and Prediction Based on Big Data

2.1 The Process of System Work

As shown in Fig. 2, based on the multi source data, such as student consumption, academic achievement, attendance management, book borrowing and so on, the students' behavior is analyzed and the rules and habits of students' life are predicted. First, the data is preprocessed, multi source data is fused, and the data is stored in the distributed system HDFS to ensure the consistency of data in the relational database, so that the data can be easily converted. At the same time, Scala module is applied to cluster analysis and association rule mining to complete student behavior classification, student behavior prediction and early warning.

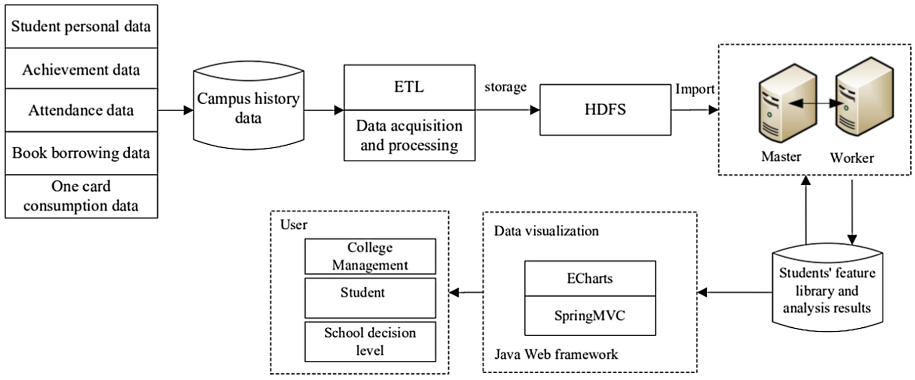

Fig. 2. System overall business process

2.2 Data Acquisition and Preprocessing

Students' data are characterized by wide source and huge quantity, which makes data quality not very high, so they need to be pre processed before data analysis. The process of preprocessing is to clean up some unrelated data in the data which are less associated with the mining targets, thus improving the quality of the data and providing more valuable data for the subsequent data processing.

1. Data cleaning: in order to ensure the integrity of data, data is cleaned by merging data and unified type, eliminating redundant and missing data.
2. Data conversion: a large portion of data collected is user's historical data. On this basis, the data are transformed according to the mining objectives of this paper, and statistics, clustering and classification methods are used to compress and standardize the data.
3. Data integration: through the analysis of the original data, it is found that because the data are from different systems, there are many duplicated data attributes, and this kind of data is integrated in the preprocessing stage, thus reducing the effect of the data dimension.

3 Analysis and Feature Extraction of Students' Behavior

Before data mining, we need to preprocess the collected data. Data preprocessing is a key step in the whole process, which generally includes data acquisition, cleaning, analysis and feature extraction. Data collection is to obtain student behavior data, data cleaning is to remove the abnormal values and missing values in the student's behavior data, and to avoid the impact of these data problems on data mining. Data analysis is to better understand the data and have a more comprehensive understanding of the data in advance, and the final feature is proposed. In order to reduce the complexity of the data mining, the original data is converted to the data which can be used directly by data mining, which can effectively improve the results of the final data mining. By analyzing the usability of data and evaluating students' behavior in school, we build a student behavior characteristic database.

3.1 Consumption Indicators

By analyzing the students' consumption behavior in the school, the student consumption records, such as term consumption amount, monthly average consumption amount, single maximum consumption and consumption times, are used as evaluation indicators to find out the students' consumption pattern and consumption level. The evaluation index of the law of student consumption is shown in Table 1.

Table 1. Evaluation index of students' consumption

Index name	Range of value
Term consumption	1–10000
Monthly average consumption	1–2000
Single maximum consumption	1–50
Frequency of consumption	1–500
Monthly consumption peak	1–500
Consumption level	High/secondary
Consumption habits	Regularly/normal

3.2 Indicators of Students' Living Habits

In order to evaluate the students' habits and habits effectively, the students' time of practice, physical exercise, time and place of activities are used as evaluation indicators to analyze the data collected, so as to understand the regular habits of the students. The evaluation indicators of students' living habits are shown in Table 2.

Table 2. Evaluation index of students' living

Index name	Range of value
Regular diet	0–30
Early rise index	0–30
Physical exercise index	0–30
Internet time	0–240
Monthly consumption peak	1–500
Consumption level	High/secondary
Consumption habits	Regularly/normal
Place of regular activity	Library/dormitory
Habits and customs	Regularly/healthy

3.3 Student Learning Index

In order to analyze the students' degree of effort and academic achievement, the data are analyzed with the evaluation index, such as attendance rate, book reading, learning length and learning habits, so as to understand the students' normal learning situation, as shown in Table 3.

Table 3. Average value of each student's index

Index name	Range of value
Attendance rate in class	0–1
Book reading quantity	0–80
Number of access to Libraries	1–1000
Long learning time	1–240
Average achievement	1–100
Number of failures	Many/less
Learning habit index	Excellent/inferior

3.4 Analysis of Behavior Results

The data from the management system of the digital campus of Jilin University is used as the data source, which includes the students' Campus consumption records, library loan records, class attendance records, library entrance records, students' records, students' records, physical exercise records, and campus network from April 2016 to April 2018, which are 20000 undergraduate students of the school. The access records of the collaterals, etc. After the data is integrated, the data is converted into the HDFS module by using the Sqoop tool, and the data is pre processed on the basis of the Spark platform. Then, the statistical analysis is carried out, and the student behavior feature library is established, and the indexes in the feature library are used for testing. Taking the students' effort analysis as an example, the students are divided into 9 categories according to the students' effort and achievement index on the Spark platform, and the average value of each student's indexes is shown as shown in Table 4.

Table 4. Clustering results of learning effort

Category	Student ratio	Attendance index	Achievement index	Book reading	Number of access to libraries	Long learning time	Course passing rate
1	6.89	0.58	51.98	25	35	88	0.68
2	18.01	0.87	84.92	49	61	216	0.97
3	11.22	0.76	82.75	35	55	175	0.87
4	3.05	0.33	45.59	21	21	98	0.50
5	5.08	0.46	63.88	18	28	112	0.65
6	21.15	0.87	79.02	52	58	231	0.98
7	11.24	0.76	66.12	36	73	165	0.86
8	11.06	0.62	75.18	28	34	198	0.75
9	12.30	0.63	67.22	24	32	129	0.62

It can be seen that the students who work hard and have excellent results account for 18.01% of the total number of students. Most of the students can study hard, although the distribution of results is different. Only a small number of students do not work hard enough, and their grades are very poor. These students account for only 3.05%. 5.08% of the students are qualified, but the degree of effort is not enough. If we urge them, we will make further progress. According to the evaluation criteria and the clustering results of students, the results are compared with the real situation. The results show that the cluster analysis method is reasonable and effective.

4 Student Behavior Prediction and Early Warning Experiment

4.1 A Stratified Model of Students' Behavior Characteristics

Students' behavior characteristics are extracted from the student behavior database, and the students' age, gender and their colleges are used to build a set of students' behavioral characteristics.

$$S = \{C_1, C_2, C_3, \ldots, C_n\} \tag{1}$$

In order to distinguish the contribution degree of different characteristics to the model, we assign different weights to different student characteristics in the set and satisfy the following conditions:

$$\sum_{i=1}^{n} w_i = 1 \tag{2}$$

In order to establish a hierarchical model of students' behavior characteristics:

$$S = W_i C_i \tag{3}$$

Based on student behavior strati fication model, students' behaviors are predicted and predicted. The overall framework of the model is shown in Fig. 3.

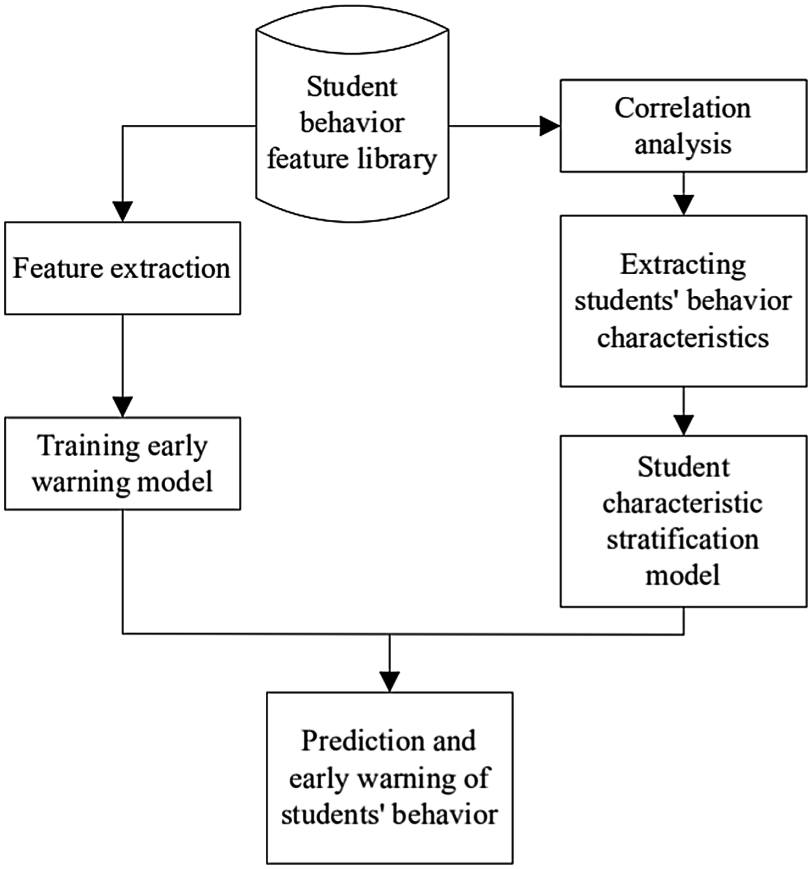

Fig. 3. Student behavior stratification prediction early warning model

4.2 Student Behavior Prediction Experiment

The cross validation method is used to analyze the correctness of the model of student behavior stratification prediction, and the relationship between the prediction results of students' behavior characteristics and the real value is reflected by the average relative error and the standard error. Based on the correctness of the methods used by different student scale comparison and analysis, and compared with the traditional forecasting methods, the scale of the test data is 100500100020005000 and 10000 students

respectively. The average relative error of prediction is shown in Fig. 4. The average relative error of student behavior characteristics is shown in Fig. 5.

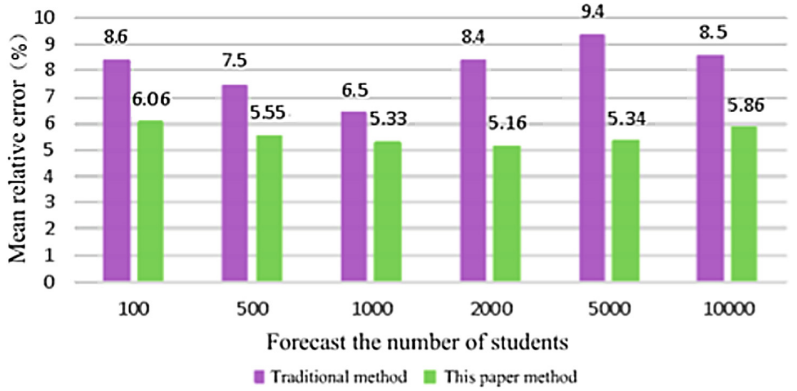

Fig. 4. Average relative errors of prediction

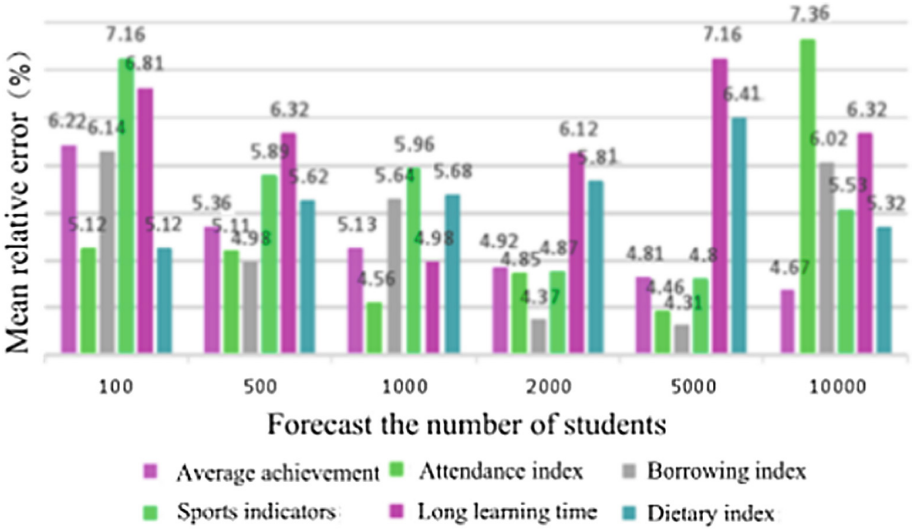

Fig. 5. Average relative errors of student behavior

It can be seen that compared with the traditional prediction method, the average relative error of the student behavior measurement model based on feature stratification is small, it is kept at about 5%, the prediction accuracy is high, and it has a good prediction effect. With the change of the relative error of the prediction of the number of students, the method is more expansible, and it will not reduce the accuracy of prediction by predicting the changes in the number of students. The relative error

distribution on the behavior characteristics of each student is more uniform, indicating that the average relative error of the students' behavior characteristics in each dimension is relatively small, which is suitable for the prediction of muti-dimensional students' behavior.

5 Conclusions

The accumulation of campus student behavior data provides a data basis for student behavior analysis and prediction. The student behavior analysis and prediction platform based on the large data of the campus is set up. The data mining technology is used to preprocess, statistics and analyze the original data, so as to establish a feature library that can describe the individual behavior of the students, and predict the students' behavior based on the stratified model of student behavior characteristics. The experimental results show that the methods used can effectively mine students' consumption rules, living habits and learning conditions. The prediction results are in good agreement with the actual situation. Compared with the traditional forecasting methods, the accuracy is high and the error is small. It is beneficial to the school to master the students' learning and life situation, and the education can be carried out pertinent.

References

1. Lambiotte, R., Kosinski, M.: Tracking the digital footprints of personality. Proc. IEEE **102** (12), 1934–1939 (2014)
2. Sun, A., Ji, T., Wang, J., et al.: Wearable mobile internet devices involved in big data solution for education. Int. J. Embed. Syst. **8**(4), 293 (2016)
3. Hasbun, T., Araya, A., Villalon, J.: Extracurricular activities as dropout prediction factors in higher education using decision trees. In: 2016 IEEE 16 International Conference on Advanced Learning Technologies (ICALT), pp. 242–244 (2016)
4. Hammoud, S.: MapReduce network enabled algorithms for classification based on association rules. Brunel University School of Engineering and Design Ph.D. theses (2011)
5. Maillo, J., Triguero, I., et al.: kNN-IS: an iterative spark-based design of the k-nearest neighbors classifier for big data. Knowl. Based Syst. **117**, 3–15 (2017)
6. Arias, J., Gamez, J.A., Puerta, J.M.: Learning distributed discrete Bayesian network classifiers under Map Reduce with Apache spark. Knowl. Based Syst. **117**, 16–26 (2017)