A BP Neural Network-Based Early Warning Model for Student Performance in the Context of Big Data

Chengxiang Shi and Yun Tan

Department of Mathematics and Information Engineering, Chongqing University of Education, Chongqing, China

Correspondence should be addressed to Chengxiang Shi; shicx@cque.edu.cn

Received 19 July 2022; Revised 24 August 2022; Accepted 3 September 2022; Published 28 September 2022

Academic Editor: Sweta Bhattacharya

Copyright © 2022 Chengxiang Shi and Yun Tan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nowadays, educational data mining technology has received more and more attention from scholars in China, and the application of correlation between student behavior data and student achievement to teaching management has become a hot research topic. Starting from the study of the potential association between book borrowing and student achievement in the big data environment, the paper analyzes the correlation between book borrowing and student achievement based on the Apriori algorithm and concludes that there is a strong correlation rule between book borrowing and student achievement. Based on BP neural network prediction algorithm, the paper constructs an early warning model for student performance by predicting book borrowing through course performance. The absolute value of the error between the predicted value of book borrowing and the real value of borrowing is used as a basis to make early warning for students’ performance, so as to realize the monitoring of students’ learning situation, thereby providing theoretical suggestions for teachers’ teaching and promoting the school’s management of students.

1. Introduction

Under the background of big data, people will generate a large amount of behavioral data every day. With the development of database technology, many potential correlations can always be mined from people’s behavioral data. The development of a country’s education is related to the country’s future; therefore, how to link data mining technology with education has become an important hot spot at present. This paper aims to analyze students’ book borrowing behavior through big data mining technology, realize early warning of students’ academic performance, so as to optimize student management, improve teachers’ work efficiency, and better serve students.

Traditional student learning behavior analysis is mostly carried out through questionnaire surveys, manual evaluation, and other methods. This not only is time cost high and errors are large but also there are many non-objective factors that make the results often not objective enough. With the development of computer networks and the Internet, scholars at home and abroad have begun to use educational data mining technology to analyze students’ behaviors; for example, foreign scholar Professor Andrew Kyngdon [1] established an automatic scoring model based on a new neural network model in his research. Compared with traditional research methods, this research result is more scientific and improves the accuracy of the results. In the research on student behavior based on big data, there are also many studies on the correlation between library borrowing data and students’ learning behavior; for example, Chinese scholars Yang Xinya [2], Wu Xudong [3], and others conducted research on the correlation between book borrowing and students’ grades through SPSS and neural network algorithms and found that there was a greater correlation between student borrowing and student performance. Although studies have shown that there is a large correlation between student borrowing behavior and student performance, there are relatively few applied studies on this association.

Machine learning has been a popular choice in analyzing students’ performances. As an example, the study in [4, 5] used two datasets for the classification and analysis of
students’ performance using five machine learning algorithms. As part of the study, eighteen experiments were conducted which helped in predicting the performance of the students. The study in [6] presented a prediction model to predict students’ performance in secondary education. As part of the study, five classification algorithms were used, namely, Logistic Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), XGBoost, and Naïve Bayes, considering dataset from two Portuguese school reports and surveys. The imbalanced dataset was pre-processed using K-Means SMOT (Synthetic Minority Oversampling Technique) before the actual classification was performed. The study also used an interpretable LIME (Local Interpretable Model-Agnostic Explanation) model for all the classifiers used in the study.

Therefore, the paper collected desensitized book borrowing information for five professional students from a school and a college in the past four years from 2016/01 to 2020/09 through communication with the school library management center and the academic affairs office, and the performance information of all students in the 2016 grade. Firstly, the collected data is processed, and then the Apriori algorithm is used to study the association rules between book borrowing and students’ grades, and it is found that there is a strong correlation between book borrowing and students’ grades. Finally, on the basis of the mining results of association rules, the BP neural network algorithm [7] is used to construct a student performance early warning model to monitor the performance of students in various majors. By comparing the early warning status of student performance of different majors, it provides a valuable reference for the talent training direction of the school.

The unique contribution of the paper includes:

(i) Implementation of Apriori algorithm to study the association between book borrowing and students’ grades

(ii) Implementation of BP neural network algorithm to develop early warning model pertaining to students’ performance in order to monitor the performance of students in various majors

2. Data Sources and Data Processing

2.1. Data Sources. With the support of the school library management center and the Academic Affairs Office, the paper collected desensitized library borrowing information of students in a college of a university and the grade information of all students in the class of 2016 in a subordinate secondary college in the last four years during 2016/01-2020/09. Among them, library borrowing data contains book barcode, processing time, request number, and other field information. And student performance data information contains course name, course nature, course grade, and other field information.

2.2. Data Pre-Processing. In order to facilitate later studies, the raw data need to be cleaned to retain the data of research value. The paper mainly uses SPSS statistical analysis software to conduct preliminary data processing. The specific treatment method is as follows: Because the system may malfunction during the students’ borrowing process, resulting in errors such as missing information and garbled code in the borrowed book information entry. Therefore, before analyzing the data, the paper uses SPSS to filter out the problematic data using book call number and book name as keywords. For example, for the column of book call number, clear data types obviously have abnormal data. Considering that there are students who change departments midway, take a break from school, serve in the military, and retake courses after being discharged from the military resulting in grade data that is not meaningful to study, this paper uses students’ names as keywords to count the total number of courses taken by students in their four years of study, and if a student does not have the same number of courses as other students in the same major, the student is considered to be in the above situation and then clear the information of this part of students. The book borrowing information through data cleaning retains the book borrowing information data of students in a certain college of a university for the past four years from 2016 to September 2020, with a total of 18091 items, and the grade data retains the grade information data of all students in the 2016 class of the college, with a total of 33758 items. For the following research, the paper uses SPSS to classify the data of book borrowing after preliminary processing, and summarizes the book borrowing information of students of the same major in a table, and then uses the COUNTIF function to check the book borrowing information. Statistically, the COUNTIF function helps to count data that meets a particular criterion. This helps to calculate the total amount of books borrowed by each student in the past four years. For the preliminary processed score data, the paper uses SPSS to classify the data and summarize the scores of students of the same major in a table. Finally, the book borrowing data and grade data are integrated according to different majors and different students. The integrated data is as follows in Table 1 (due to too much data, the text only shows the data of some students in a certain major):

2.3. Data Conversion. The data type of the Apriori algorithm requires category data, so the book loan data and the grade data need to be converted before implementing the algorithm.

2.3.1. Conversion of Book Loan Data. The library’s book circulation information necessarily reflects some behavioral characteristics of students. The paper uses the indicator of the total number of books borrowed in the book lending information to reflect students’ enjoyment of reading. It is understood that in order to promote reading, the university requires students to read at least 30 books during their four years of college for completing the reading block for credit. According to this requirement, the paper classifies students’ preferences for reading into three categories, and the classification rules and data conversion results are as follows in Table 2 (due to too much data, only some students’ data are shown in the article):
As Table 3 shows, for example, if Zhang * Yao, a student of Computer Science and Technology Major (School Enterprise Class), borrows books for a total of 60 times, then his corresponding reading level is A.

2.3.2. Conversion of Grade Data. The paper uses the professional composite score to describe the good or bad professional performance of students, where the professional composite score is calculated as:

\[
\text{professional composite score} = \frac{\sum \text{Score in each major course} \times \text{Corresponding professional course credits}}{\sum \text{Credits in each major course}}.
\]

The composite score is considered as combining the items representing a variable to create a score or data point pertaining to the same variable. The score forms reliable and authentic measures of latent and theoretical constructs. In academics, the composite score is also considered the sum of passing scores.

The overall professional grades of each student were calculated, and the students’ overall professional grades of the five majors were clustered into three categories by SPSS, and the clustering results are shown in Table 4:

As shown in the above table, the paper classifies student performance into three categories: excellent, good, and poor, where category 1 indicates that student performance is poor, category 2 indicates that student performance is good, and category 3 indicates that student performance is excellent. The achievement data were transformed according to the clustering results, and the transformed data (partial data) are shown in Table 5:

As the above table shows, for example, Ai * Ling of Internet of Things Engineering (School Enterprise Class) has a comprehensive professional score of 91.42, which corresponds to a performance level of Category 3.

3. Study of the Correlation between Book Borrowing and Student Achievement

3.1. Principle of Association Rule Apriori Algorithm. Association rules are a type of data mining function used to mine hidden connections between data from a dataset, denoted as \( X \rightarrow Y \), where the set of items \( X \) is called the...
3.2. Model of Correlation between Book Borrowing and Student Achievement

3.2.1. Model Assumption. There must be some connection between students’ use of library resources and student achievement. For the association between library borrowing information and student achievement, the paper makes the following hypothesis: there is some connection between student achievement and the total number of books borrowed, and students with better achievement tend to enjoy reading.

3.2.2. Model Building and Solving. The paper implements Apriori algorithm by Python, and the minimum support and minimum confidence in the algorithm are set to 10% and 60%, respectively, and the results obtained are as follows in Figure 1:

From the analysis results, it can be obtained that the lift of both obtained association rules is greater than 1, indicating that the obtained association rules are meaningful. With a minimum support of 10%, the following association exists between students’ enjoyment of reading and student achievement: 60.28% of the students who liked reading more had a good overall score and 60.64% of the students who had a poor overall score did not like reading.

4. Research on Performance Alert Model Based on BP Neural Network

4.1. Theory Related to BP Neural Network Algorithm. Neural networks are divided into biological neural networks and artificial neural networks. Artificial neural networks have the ability of parallel processing of information, self-learning ability, and inference ability, and are widely used in multiple fields such as pattern recognition intelligent robots, automatic control, and prediction estimation. BP (back propagation) neural network is the most traditional neural network, which was proposed in 1986 by a research group led by Rumelhart and McCelland. The BP neural network model topology includes an input layer, a hidden layer, an output layer, and at least one hidden layer. Between the input and hidden layers are the weights of the network, indicating the strength of the connection between the two neurons. Any neuron in the hidden layer or output layer integrates the information from the neuron in the previous layer, simulating the principle in biology that a neuron must be stimulated to be triggered, and outputs the integrated information as the neuron in that layer. The algorithm steps are as follows:

Step 1: A sample is taken from the training sample set and its input information is fed into the network.
Step 2: The output of each layer node is calculated from the network forward.
Step 3: Calculate the error between the actual output and the desired output of the network.
Step 4: Starting from the output layer and working backwards to the first hidden layer, the individual connection weights of the entire network are adjusted in the direction of error reduction according to certain principles.

Table 3: Table dividing the degree of students’ preference for reading (partial data).

| Name        | Specialty                                      | Total number of borrowings | Borrowing level |
|-------------|------------------------------------------------|---------------------------|-----------------|
| Zhang•Yao   | Computer Science and Technology Major (School Enterprise Class) | 60                        | A               |
| Feng•Ya     | Computer Science and Technology Major (General Admission Class) | 61                        | A               |
| Tan•qin     | Mathematics and Applied Mathematics Major       | 38                        | B               |
| Ran•Le      | Mathematics and Applied Mathematics Major       | 31                        | B               |
| Zhang•Wei   | Mathematics and Applied Mathematics Major       | 24                        | C               |
| Hao•Fei     | Internet of Things Engineering Major (School Enterprise Class) | 18                        | C               |

Table 4: Clustering center for student achievement by category.

| Clustering center | Achievement level |
|-------------------|-------------------|
|                   | 1            | 2            | 3            |
|                   | 70.03        | 76.65        | 84.15        |
Step 5: The above steps are repeated for each sample in the training sample set until the required error is achieved for the entire network training sample set.

The BP neural network independently trains the input and output data according to the error back propagation algorithm, finds a reasonable connection between the input and output data, and realizes the prediction of the result. Compared with prediction methods such as least squares regression and grey model, it has strong nonlinear modeling ability, so the prediction of nonlinear data is more accurate. Therefore, this paper chooses to use BP neural network algorithm to predict students’ grades.

Table 5: Transformed data.

| Name       | Specialty                                                        | Professional composite score | Achievement level |
|------------|------------------------------------------------------------------|------------------------------|------------------|
| Ai*Ling    | Internet of Things Engineering Major (School Enterprise Class)   | 91.42                        | 3                |
| Pai*       | Mathematics and Applied Mathematics Major                        | 73.09                        | 1                |
| Bu*jin     | Internet of Things Engineering Major (General Admission Class)   | 75.90                        | 2                |
| Cao*       | Computer Science and Technology Major (School Enterprise Class)   | 77.20                        | 2                |
| Cao*Wei    | Computer Science and Technology Major (General Admission Class)   | 89.45                        | 3                |
| Chen*Li    | Mathematics and Applied Mathematics Major                        | 76.47                        | 2                |
| Chen*Ha    | Internet of Things Engineering Major (General Admission Class)   | 89.52                        | 3                |
| Du*you     | Mathematics and Applied Mathematics Major                        | 77.60                        | 2                |

| Candidate 1-item set | Item set | Support |
|----------------------|----------|---------|
| 1                    | 18.73%   |
| 2                    | 57.97%   |
| 3                    | 23.31%   |
| A                    | 14.14%   |
| B                    | 42.63%   |
| C                    | 43.23%   |

| Frequent 1-item set  | Item set | Support |
|----------------------|----------|---------|
| 1                    | 18.73%   |
| 2                    | 57.97%   |
| 3                    | 23.31%   |
| A                    | 14.14%   |
| B                    | 42.63%   |
| C                    | 43.23%   |

| Candidate 2-item set | Item set | Support |
|----------------------|----------|---------|
| 1, C                 | 11.35%   |
| 2, C                 | 24.10%   |
| 3, C                 | 7.77%    |
| 1, B                 | 5.58%    |
| 2, B                 | 25.70%   |
| 3, B                 | 11.35%   |
| 1, A                 | 1.79%    |
| 2, A                 | 8.17%    |
| 3, A                 | 4.18%    |

| Frequent 2-item set  | Item set | Support |
|----------------------|----------|---------|
| 1, C                 | 11.35%   |
| 2, C                 | 24.10%   |
| 2, B                 | 25.70%   |
| 3, B                 | 11.35%   |

Rules for correlating reading levels with student achievement

| Rule   | Support | Confidence | Lift  |
|--------|---------|------------|-------|
| B—>2   | 25.70%  | 60.28%     | 1.04  |
| 1—>C   | 11.35%  | 60.64%     | 1.40  |

Figure 1: Results of association rule analysis.

4.2 Performance Alert Model Modeling. From the correlation analysis between student grade data and total student book circulation data in the previous chapter, it is known that the following correlations exist between grades and total book circulation: Students who enjoy reading more tend to have good overall scores, and those with poor overall scores tend to dislike reading. Based on the above study, this chapter uses BP neural networks to establish an early warning model of student performance. The model takes students’ major course grades as input and the total number of books borrowed as output and is trained by BP neural network to predict the total number of books borrowed by students.
Set the threshold $K$ for making the achievement warning, where $K$ is calculated by the formula:

$$K = |\text{Projected total number of books borrowed} - \text{True total number of books borrowed}|. \quad (3)$$

If the error between the predicted result and the true value is greater than a certain threshold $K$, a grade warning is made for the student, indicating that there is a gap between the actual learning situation of the student and the result of the test score feedback, or that there is a suspicion of falsification of the student’s book borrowing, which requires the teacher of the class to pay more attention to this part of the students in class and the counselor to strengthen the guidance of this part of the students’ learning and life. The flow of the performance alert model is shown in Figure 2:

4.3. Implementation of Performance Alert Model

4.3.1. Discussion of the Achievement Warning Model Threshold $K$. Since the four majors, namely, Mathematics and Applied Mathematics Major, Internet of Things Engineering Major (General Admission Class), Computer Science and Technology Major (School Enterprise Class), and Computer Science and Technology Major (General Admission Class), include further 21 major courses, the paper uses the grade data of the students in the above four majors and the data of the total number of books borrowed to determine the threshold $K$. This helped in triggering the warning through model training. Students’ grade data and students’ total book borrowing data are imported into MATLAB, with the grade data package as input and the total book borrowing data as target, into the BP neural network toolbox, wherein the model is tuned to test data, validation data, and training data ratios by the distribution of errors between the predicted and true values of book borrowing. Set the hidden layer of BP neural network as 10 layers, 25% of the data as test data, 15% of the data as validation data, and 60% of the data as training data, and the least squares optimization algorithm was selected to train the data to get the most desirable results, the results are as follows in Figure 3:

The error plot shows that the prediction accuracy of the model test set and validation values is high, indicating that the use of the model can achieve the purpose of predicting the total number of books borrowed through the professional course grades. The poor prediction accuracy of the training set indicates that there is a discrepancy between some students’ learning performance and their actual learning situation.

After investigation, it is learned that approximately 20% of the students in the college will receive an academic warning from the school’s academic affairs office in their senior year, on the eve of graduation. Combined with the error histogram in Figure 4, more than 78% of the students’ book borrowing total volume predicted value and the true value of the error absolute value is within 25. Thus, in the paper, the threshold $K$ for the achievement warning model to trigger achievement warning is set to 25. If $K > 25$, achievement warning is made, and if $K$ is within the interval of $[0, 25]$, achievement warning is not made.

4.3.2. Model Implementation and Results Analysis. The professional course grades and the total number of books borrowed of the students of the college are imported into MATLAB, the grade data package is used as input, the total number of books borrowed is used as target, the hidden layer of BP neural network is defined as 10 layers, 25% of the data is used as test data, 15% of the data is used as validation data, 60% of the data is used as training data, and the least squares
optimization algorithm is selected to train the data. After running the data using MATLAB, the predicted and true values of the total number of books borrowed are obtained as follows in Figure 5:

A summary of the percentage of students who made performance warnings for each major is obtained as follows in Table 6:

Comparing the percentage of students who made achievement alerts in the five majors, the following results can be obtained: The most students in the Internet of Things Engineering Major (General Admission Class) had the highest number of performance warnings, followed by the following in descending order: Mathematics and Applied Mathematics Major, Computer Science and Technology major (School Enterprise Class), Computer Science and Technology Major (General Admission Class), and Internet of Things Engineering Major (School Enterprise Class). Through understanding with the professional counselors, comparing the proportion of the number of academic warnings made by the Academic Affairs Office of the students of
each major before graduation and the proportion of the number of early warnings analyzed by the results of the model, it is found that the difference between the two is within 5%. This shows that the model has a high early warning on student performance and has a certain reference value.

Comparing the proportion of students with grade warnings in the general admission class and the school enterprise class in Table 7, it can be found that the performance warning of the school enterprise class is significantly better than that of the general admission class.
Table 6: Summary of the percentage of students making performance warnings.

| Major                                           | Percentage of students making alerts |
|-------------------------------------------------|--------------------------------------|
| Mathematics and Applied Mathematics Major       | 33.06%                               |
| Computer Science and Technology Major (General Admission Class) | 13.22%                               |
| Computer Science and Technology Major (School Enterprise Class) | 22.22%                               |
| Internet of Things Engineering Major (General Admission Class) | 35.60%                               |
| Internet of Things Engineering Major (School Enterprise Class) | 10.45%                               |

Table 7: Proportion of early warning students for different school-running nature.

| The nature of running a school                  | Percentage of students making alerts |
|------------------------------------------------|--------------------------------------|
| General admission class                        | 27.27%                               |
| School enterprise class                        | 16.51%                               |

This result is consistent with last year’s results of the school’s national-level subject assessment.

In previous research, although some scholars have applied the BP neural network algorithm to the student performance early warning model, the research object is often a student of a certain major, and the error of the model is about 10%. This paper studies the early warning situation of students’ performance in different school-running nature and majors, and the number and type of research objects are more than in previous studies. From the two dimensions of the nature of running a school and the major, by comparing with the actual situation, it is verified that the error of the model is within 5%, which improves the accuracy of the model.

5. Conclusion

This article mainly uses the four-year university grade data and book borrowing data of a certain school as the research object. First, using the Python programming language, the Apriori algorithm is used to mine the association rules of the book borrowing and student score data. It is found that there is a certain positive correlation between student grades and the total amount of books borrowed. Students who like to read will not have bad grades, and students with better grades also prefer to read. Then, based on the conclusion that there is a strong association rule between student performance and the total amount of book borrowing, using the BP neural network algorithm and taking student performance as input and the total amount of student book borrowing as the target, after neural network training, predict the total number of student books borrowed and use the absolute value of the difference between the predicted result and the true value to reasonably set the threshold $K$, thereby constructing the early warning model of student performance. This model realizes the monitoring of student performance and is widely used in the management of students’ learning behavior. Using this model to give early warnings to students of five different majors in a certain university, compare the warnings of students in different majors, and learn that in university study, more attention should be paid to cultivating students’ practical ability, combining theoretical knowledge with practice, and encouraging students to go out of campus and go to internships in enterprises corresponding to their majors.

However, this article still has the following shortcomings: One is that the data samples in this article are not large enough, which leads to the unsatisfactory promotion value of the association rules. In the follow-up research, it is necessary to increase the data samples to improve the promotion of the association rules. The second is the insufficient use of book borrowing information. It only uses the information of the total amount of borrowed books of students and does not use the data of the book category. However, by reading the literature, it can be found that the category information of borrowed books is very important for studying student performance. The significance of the research requires further exploration and analysis. The third is that this article does not make better improvements to the algorithms involved, and further study and research are needed for three issues in the future.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was partly financially supported through grants from the Chongqing Higher Education Teaching Reform Research Key Project (No. 222166), the Chongqing Science and Technology Bureau Technology Innovation and Application Development Key Project (No. cstc2020jscx-dxwtBX0044), the Chongqing Science and Technology Bureau Technology Innovation and Application Development General Project (No. cstc2020jscx-mxsmX0152), and the Scientific Research Project of Chongqing University of Education (No. KY202107B).

References

[1] S. Dan, L. Dongbo, and F. Xia, "Research on course grade prediction and course early warning based on multi-source data
analysis,” Research on Higher Engineering Education, vol. 1, pp. 189–194, 2020.

[2] Y. Xinya, Y. Hui, and Z. Zuoling, "A preliminary study on the relationship between students’ grades and borrowing behavior in higher education: the case of Chongqing University," Digital Library Forum, vol. 9, no. 9, pp. 23–26, 2013.

[3] X. Wu, C. Zhengjun, F. Luyuan, and L. Yingxi, Application of BP neural network algorithm in the study of correlation between book borrowing and student achievement, vol. 13, Computer Knowledge and Technology, 2017.

[4] G. T. Reddy, M. P. K. Reddy, K. Lakshmannana et al., "Analysis of dimensionality reduction techniques on big data," IEEE Access, vol. 8, pp. 54776–54788, 2020.

[5] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student performance prediction and classification using machine learning algorithms," in Proceedings of the 2019 8th International Conference on Educational and Information Technology, pp. 7–11, Cambridge, UK, March 2019.

[6] K. M. Hasib, F. Rahman, R. Hasnat, and M. G. R. Alam, "A machine learning and explainable AI approach for predicting secondary school student performance," in In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0399–0405, Las Vegas, NV, USA, January 2022.

[7] T. R. Gadekallu, D. S. Rajput, M. Reddy et al., "A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU," Journal of Real-Time Image Processing, vol. 18, no. 4, pp. 1383–1396, 2021.