Hindawi
Complexity
Volume 2021, Article ID 5577167, 10 pages
https://doi.org/10.1155/2021/5577167

Research Article

Online English Teaching Course Score Analysis Based on Decision Tree Mining Algorithm

Xiaojun Jiang

School of Translation Studies, Xi'an Fanyi University, Xi'an, Shaanxi 710105, China

Correspondence should be addressed to Xiaojun Jiang; seanrain@xafy.edu.cn

Received 25 February 2021; Revised 16 March 2021; Accepted 24 March 2021; Published 2 April 2021

Academic Editor: Wei Wang

With the advent of the Big Data era, information and data are growing in spurts, fueling the deep application of information technology in all levels of society. It is especially important to use data mining technology to study the industry trends behind the data and to explore the information value contained in the massive data. As teaching and learning in higher education continue to advance, student academic and administrative data are growing at a rapid pace. In this paper, we make full use of student academic data and campus behavior data to analyze the data inherent patterns and correlations and use these patterns rationally to provide guidance for teaching activities and teaching management, thus further improving the quality of teaching management. The establishment of a data-mining-technology-based college repetition warning system can help student management departments to strengthen supervision, provide timely warning information for college teaching management as well as leaders and counselors' decision-making, and thus provide early help to students with repetition warnings. In this paper, we use the global search advantage of genetic algorithm to build a GABP hybrid prediction model to solve the local minimum problem of BP neural network algorithm. The data validation results show that Recall reaches 95% and F1 result is about 86%, and the accuracy of the algorithm prediction results is improved significantly. It can provide a solid data support basis for college administrators to predict retention. Finally, the problems in the application of the retention prediction model are analyzed and corresponding suggestions are given.

1. Introduction

Personalized learning refers to an educational process in which appropriate learning resources and learning methods are selected based on the learners' cognitive level, learning ability, and their own qualities, so that they can make up for the shortcomings of their existing knowledge structures and achieve the best development [1–5]. From ancient times to the present, personalized learning has been the goal of people exploring and practicing educational learning. As early as the Spring and Autumn period and the Warring States Period, Confucius, a famous educator in ancient times, proposed the idea of "teaching according to one's ability and teaching without discrimination," which was the beginning of the idea of personalized learning and is still highly respected even after more than two thousand years [6]. In 2016, the U.S. Department of Education launched the National Education Technology Initiative (Future Ready Learning: Reimagining the Role of Technology in Education), and clearly defined "personalized learning refers to teaching that is optimized for each learner's needs in terms of learning pace and teaching methods, and requires that the learning objectives, learning content and learning methods in the learning process should be different and adjustable according to the needs of learners" 11]. At the same time, governments and educational authorities at all levels also attach great importance to the development of personalized learning [7]. In traditional classroom education, students' learning activities are completely developed by teachers, and the practice of personalized learning can only rely on teachers' teaching experience due to the constraints of time and space. Therefore, it is difficult to develop a personalized learning program for each student in such an educational learning model. In recent years, as the process of education

informatization continues to deepen, more and more Internet information technologies have flooded into the educational teaching process, which has led to a significant innovation in the development of education, and the emerging online learning model has gradually developed and become an integral part of educational learning [8].

In this context, a variety of educational products and online learning systems have emerged in contemporary society. On the one hand, these learning systems break through the time and space constraints of traditional classroom teaching, provide learners with global cutting-edge courses, share learning resources, and allow students to learn what they need anytime and anywhere [9]. On the other hand, they create an open and free environment for learners to choose their own learning content and learning methods at their own pace, avoiding the "one-size-fits-all" learning style in the traditional classroom learning process. In fact, after years of development, the online learning mode has become quite large in scale [10]. According to the 44th Statistical Report on the Development of China's Internet released by China Internet Network Information Center (CNNIC), as of June 2019, the scale of online education users reached 232.2 million, up 31.22 million from the end of 2018, accounting for 27.2 percent of the overall Internet users. The scale of cell phone online education users reached 1.99 billion, up 5.3 million from the end of 2018, accounting for 23.6% of cell phone Internet users [11]. In particular, at the beginning of 2020, influenced by the pneumonia epidemic caused by the new coronavirus, the Ministry of Education advocated "stopping classes without teaching, stopping classes without learning," and various enterprises and schools launched online classroom learning, creating a high expectation and reliance on such an educational learning model in the whole society [12]. About 250 million students are using online platforms for online education on an unprecedented scale, which ensures the normal educational learning order. These phenomena show that the development of education needs to shift from offline to integration with online, from scale-scale personalization, and from knowledge-centered to learner-centered. Although the online education model brings new possibilities for truly practicing personalized learning, the highly free and open nature of its market also leads to increased mobility, increased fallibility, and increased diversity of learners' learning activities. On the one hand, these changes pose a greater challenge to the traditional research on personalized learning [13]. On the other hand, with the expansion of the types of individual learning activities, learners have new demands for the personalized services that can be provided. Fortunately, along with the application of information technology and products, a large amount of learning data has been collected concomitantly with the use of various learning systems by students [14]. For example, MOOC platforms record the activity behaviors of university students learning subject courses; online learning systems and electronic terminals leave learning records of students' knowledge point practice; in addition, the popularity of electronic marking technology has enabled the preservation of a large number of offline learning resources [15]. Therefore, how to use these learning data and mine valuable information so as to reveal the natural laws of educational learning, develop and practice personalized learning, and provide learners with customized learning contents and methods has become one of the important research problems supporting the development of education, as shown in Figure 1. At the student level, the study of educational psychology shows that the learning process is dynamic and complex. Human brain memories and knowledge forgetting constantly change the level of student knowledge. Specifically, two types of educational research will explain this phenomenon in detail and provide a basic theory of knowledge-level diagnosis. In fact, the analysis and research of educational learning data have received the attention of researchers from interdisciplinary fields such as education, psychology, statistics, and computer science and have gradually formed an independent research direction, Educational Data Mining (EDM), which involves many researches. The research has made initial progress, but it is not yet possible to use it as a research tool. Although initial progress has been made, the research and application services for data-driven personalized learning methods are still in the exploration stage [16].

Based on the above background, this thesis investigates data mining methods and application research for personalized learning to provide data-driven solutions. The research objects are learners, learning resources, and personalized learning mechanisms. Learners are broadly defined as students who are involved in learning activities (including online students and offline students whose academic activities are recorded). Learning resources are the practice questions (including daily homework questions, test questions, and general classroom exercises) that have been recorded by various products and systems and are available to students for learning. Broadly speaking, learning strategies are defined as those that can support and ensure learners to perform personalized learning recommendation strategies, focusing on online practice question recommendation algorithms, and so forth. This dissertation will be an exploratory study of the technical methods and applications to achieve personalized learning, using machine learning, data mining, and artificial intelligence technologies as technical tools, combined with knowledge from the field of pedagogy (learning theory, cognitive diagnosis theory, etc.).

2. Related Work

With the development of computer technology and artificial intelligence technology, online learning systems are becoming more and more popular. Typical online learning systems include large-scale online open courses and online grading systems. These systems can provide rich learning resources (e.g., courses, exercises, etc.) that allow students to learn independently. Although online learning systems offer great convenience to students, the high degree of freedom makes it very easy for students to lose enthusiasm for learning, leading to the phenomenon of high dropout rates [17]. For this reason, online learning systems are dedicated to the study of personalized learning services so as to

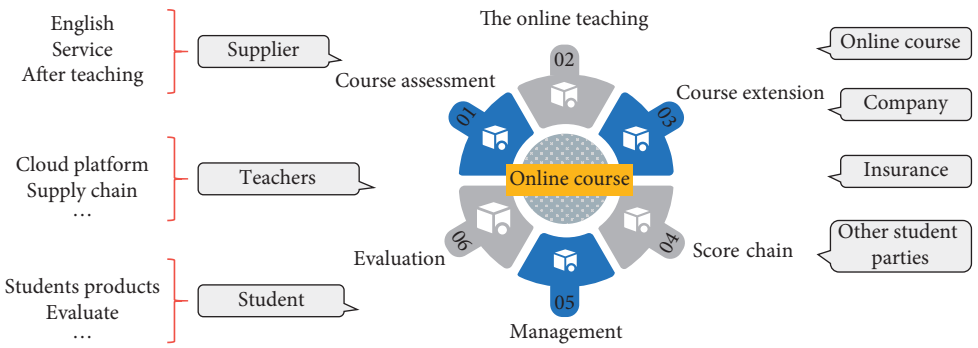

Figure 1: Research directions in online education.

improve the learning experience of students. The core task of personalized learning is knowledge-level diagnosis, that is, assessing students' mastery of different knowledge concepts, which has a positive effect on a variety of application tasks (e.g., personalized recommendations). In this paper, we use a sample task describing the knowledge-level diagnosis. The figure shows Ml and two students learning two knowledge concepts (M: "functions" and fc2: "inequalities") through the learning process. In order to solve this problem, in the field of educational psychology, relevant methods (item response principle and cognitive diagnostic model) portray students' learning status as a composite ability value 1611 or as a set of binary vectors of knowledge mastery, while, from a data mining perspective, drawing on score prediction tasks, matrix decomposition models project students in a hidden space to portray their knowledge level. Although these two types of approaches have been effective, they both ignore important learning factors that exist in the student learning process. Existing research shows that students' learning processes are usually influenced by learning factors from both the knowledge and student levels. On the one hand, in practical situations, students usually perform more consistently on practice problems containing the same knowledge concepts [18].

In March, student Ul got both force and ¯ correct, while student 2 got both wrong. As you can see, both practice problems contain the concept of "function." This phenomenon indicates that students have a better grasp of "functions" than student 7. To the best of our knowledge, although existing work has initially explored knowledge-level correlations, such as prior-order hierarchies, there is a lack of research that directly explores such topic-related properties at the knowledge level [19]. Therefore, this chapter hopes to model knowledge association properties and thus enhance the effectiveness of knowledge-level diagnostics. On the other hand, at the student level, studies in educational psychology have shown that its learning process is a dynamic and complex one. The human brain's memory and forgetting of knowledge make students' knowledge levels continuously change. Specifically, two types of pedagogical studies elaborate on this phenomenon and provide a foundational theory for knowledge-level diagnosis. First, the learning curve theory shows that students are able to acquire relevant knowledge with continuous trial and practice [20]. Second, the forgetting curve theory shows that students' memory for knowledge becomes worse over time, thus reflecting a decreasing trend in their knowledge level. If the student practiced a large number of topics between March and May, his knowledge level continued to improve. On the other hand, Student A forgot the concepts of "function" and "inequality" because he never studied them in April and May. Considering this dynamic nature, preliminary research work in both data mining and educational psychology has been conducted. The results show that adding temporal characteristics is effective. However, the existing work still suffers from the following problem: data mining algorithms, such as tensor decomposition, only add temporal information features to the hidden space, but they cannot explain the changes in student learning with explicit knowledge concepts [21].

In the study of educational psychology, the knowledge tracking model models the factors of student level learning and forgetting as additional parameters, but these approaches cannot describe the influential process of the educational theory described in the modeling process. Therefore, this chapter hopes to guide the modeling process by applying explicit educational theory to more appropriately track and explain the causes of changes in students' knowledge levels. In summary, the diagnostic task of the knowledge level learned by this white paper has the following technical challenges. In order to cope with the above problems, this chapter proposes a knowledge tracking model that incorporates learning elements called an exercise-correlated knowledge proficiency tracking (EKPT) that dynamically tracks and explains changes in students' knowledge levels from the viewpoint of probabilistic modeling. Specifically, the model incorporates a topic knowledge correlation matrix (Q matrix) that first maps the topic into the knowledge space. In the knowledge space, each topic is represented by a knowledge vector, and each dimension of the vector represents an explicit knowledge concept (e.g., "function"). Considering the correlation factor at the knowledge level, the model finds the set of neighboring topics with the same knowledge concept for each topic and then aggregates the knowledge vector information of the neighboring topics for them, so that the topics with knowledge correlation characteristics have similar distance in the space. Second, the model projects students into the

same knowledge space at each moment according to their learning performance, where each student is represented by a horizontal vector, and each dimension of the vector represents their mastery of the corresponding knowledge concept at that moment. The model then combines "learning curve theory" and "forgetting curve theory" to quantify the learning factors of students' memory and forgetting during the dynamic learning process, thus capturing the changes in their knowledge. Finally, EKPT is applied to three pedagogical tasks, namely, knowledge prediction, score prediction, and visualization of diagnostic results, to validate the effectiveness of the proposed model. This chapter conducts a large number of experiments on four real data sets, and the experimental results show that EKPT has an accurate prediction effect. It also explains the degree of influence of dynamic factors in the learning process of different students.

3. Educational Data Mining

3.1. Achievement Data Mining Algorithm. Naive Bayes Classifier is a simplified algorithm based on Bayesian theory, mainly based on Bayes' theorem and independence theorem. Compared with other mining algorithms, the advantage of Naive Bayes is that it is simple and easy to understand and master, and the classification effect is very good. It is mainly used to analyze and study the correlation between values and obtain the decision relationship between output and input values. The key of the plain Bayesian algorithm is to find the joint distribution Q (X, Y) of Y and X values, and the probability of the final class to which they belong is derived. It is known that, in the context of knowledge of probability theory, the conditional probability can be expressed by the following formula:

$$Q(Y,X) = \frac{P(X,Y)}{q(X)}. \tag{1}$$

$Q(X, Y)$ denotes that the probability of events X and Y occurring simultaneously is equal to the probability of y given the occurrence or the probability of X given the occurrence. Bayes' theorem is that after the probability of occurrence of a certain condition is known, the probability of other associated events after interchange can be obtained; for example, $P(Y|X)$ can be obtained, which has been known, so $P(Y)$ is called the prior probability and $P(Y|X)$ is called the posterior probability of Y, which can indicate that the probability of occurrence of Y can be judged after knowledge. In daily life and work, $P(X|Y)$ can be obtained directly or indirectly, but the acquisition of $P(Y|X)$ is a more difficult process, so $P(X|Y)$ can be used as an intermediate step, that is, a bridge to obtain $P(Y|X)$. Bayes' theorem can be expressed by the following equation:

$$P(Y,X) = \frac{P(X,Y)/q(X)}{k + P(Y)}. \tag{2}$$

The plain Bayesian approach ultimately gives results in the form of probabilities, so the classification of the test set samples into different types is done with reference to the magnitude of the probabilities obtained by the samples. Suppose that X is a probability space consisting of several mutually independent events, denoted by $X = \{a_1, a_2, a_3...a_n\}$, a finite set representing different classes, and the objective function $f(x)$ takes values in the set V. Then the plain Bayesian model predicts the maximum probability category of the new sample attribute values $\{a_1, a_2, a_3...a_n\}$ after the available category data, denoted as V_{MAP}, which is calculated as follows:

$$V_{\mathrm{map}} = \frac{\{a_1, a_2, a_3...a_n\}}{\max P(X,Y)}. \tag{3}$$

Based on Bayes' theorem, it can be deformed to

$$V_{\mathrm{map}} = \frac{\{a_1, a_2, a_3...a_n\}}{\max P(X,Y)} = P\left(X, Y|v_j\right)p\left(v_j\right). \tag{4}$$

According to the above formula, the frequency of occurrence of the data to be classified in the sample can be obtained. However, because the features in the probability space are independent of each other when the target value is known, estimating $P(a_1, a_2, a_3...a_n)$ is not feasible. Therefore, it can be transformed to the following:

$$V_{np} = \frac{\arg v_j\{a_1, a_2, a_3...a_n\}}{\prod P(X,Y).} \tag{5}$$

The plain Bayesian approach has good classification efficiency; it can handle multiple classification tasks, and the model works better especially when the data volume is small; however, there are also cases where the model effect is limited when the data volume is large or the data is strongly correlated, and the model is relatively simple due to the inaccurate handling of missing values, and it is commonly used for text classification, and its processing process can be shown in Figure 2.

3.2. Decision Tree Algorithm to Mine Data Correlation. Decision trees use sample attributes as nodes and sample attribute values as branches. It is a common classification method, and the main learning process is to select relatively important attributes as the middle nodes of the decision tree one by one and to branch with the feature values to build a classification tree with leaf nodes corresponding to specific categories, so that new samples can be classified into different categories according to the attribute values, as shown in Figure 3. The decision tree algorithm first constructs the tree (Tree Building) followed by optimization of the tree (Tree Pruning).

The basic student admissions data is an indication of the overall ability of the candidates before they enter the university. The records in the database of 31845 student candidates admitted in the class of 2006–2016 were obtained through the Student Affairs Office. Due to the large amount of data, only the first-year performance records of 2893 students of class 2016, about 20,000 entries, were taken as the data set in the initial model training phase. There are many information table items in the candidate records, and the data need to be sorted for better analysis and comparison. In the process of sorting, it was found that the records of

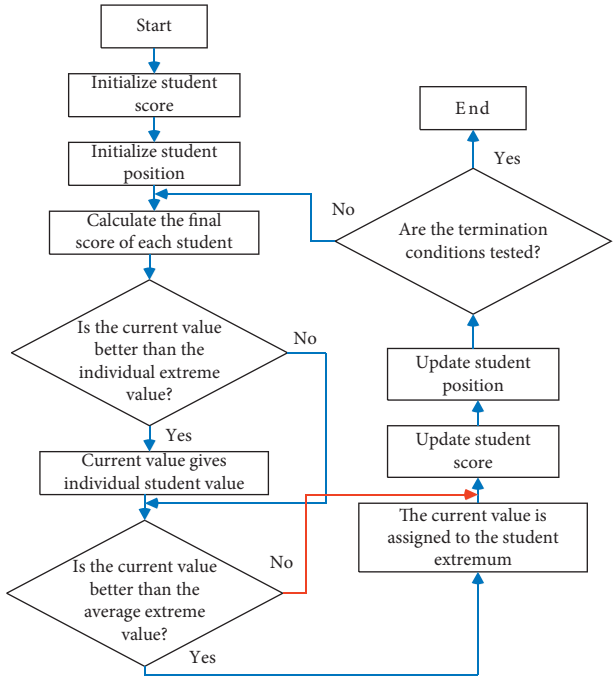

Figure 2: ELT achievement data mining algorithm.

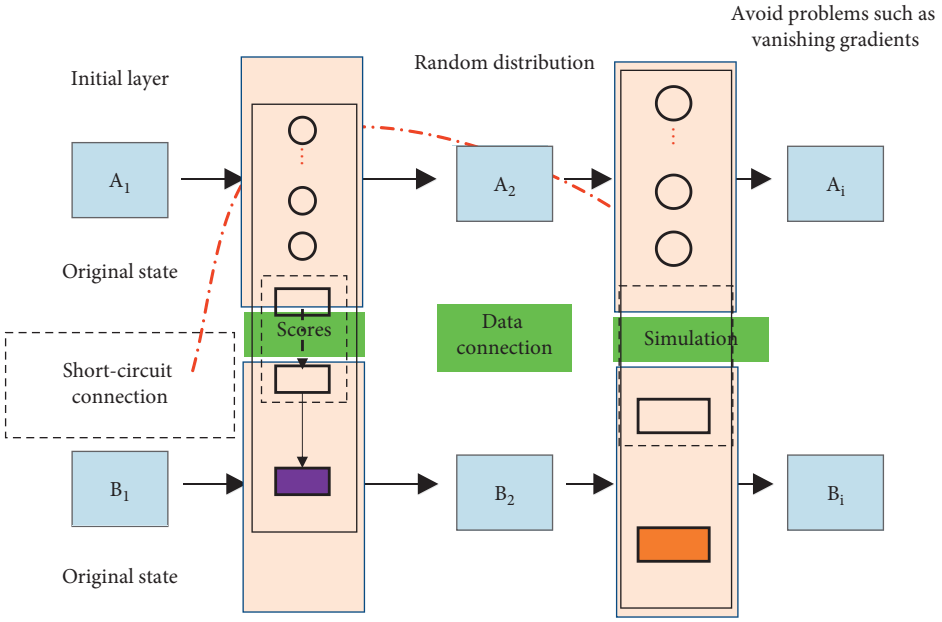

Figure 3: Decision tree representation.

candidates' personal information in the entrance examination admission system were numerous and mixed, and not all of them were recorded in the same table, which made it impossible to extract the required information at one time. Therefore, the information of candidates needs to be stitched and parsed to finally obtain the needed information. According to the data of student admission information of the current year's college entrance examination, the information including student's name, gender, ID number, candidate category, admission major, region, college entrance examination results, and political appearance was obtained as shown in Figure 4.

Student on-campus network access is all done through individual student accounts assigned by the Information Center. The gateway server records a large number of networks. The gateway server records several network logfiles. On one hand, all HTTP requests made by users during gateway access are recorded, and, on the other hand, important information such as DNS is also recorded. Students' network operation information can be accurately reflected,

including access time, student ID, target URL, target IP, student login IP, and MAC address of the login device. Students will leave records in the campus network log whenever they operate, such as surfing online, catching up on TV shows, and playing games. These records are combined by correlating student IDs with online logging behavior. After screening and universal processing operations on the logs based on previous investigation algorithms, more than 796,000 actual web access requests were obtained between September 2016 and September 2019. Web access information for current students on campus was obtained with the permission of the Information Center. However, the logs contained too much content (including text files, Java Script files) cache files used to send messages between the browser and the server, many additional records that popped up automatically when pages were clicked, various types of web page information, and so forth. The information was too much and had to be cleaned and merged.

3.3. Analysis of Neural Network Postprocessing Results. Data sets are usually divided into three subsets: training set, testing set, and verification set. Partitioning the data is the core process of machine learning experiments, ensuring that the data analysis process, model training, and data prediction can proceed more smoothly. Data is usually collected manually or semiautomatically, and each piece of input data has a corresponding output. What machine learning does is to learn the information contained in these already collected data and successfully predict the output when new input data is available. In implementing machine learning, the training set is used to train the model, giving the model inputs and corresponding outputs and letting the model learn the relationship between them. The validation set is used to estimate the training level of the model, such as the classification accuracy of the classifier, the prediction error, and so forth. The best model can be selected based on the performance of the validation set. The test set is the result of the input data on the final model, which is the output of the trained model on the simulated "new" input data. The test set can only be used to test the performance of the model, not to train it. It is certainly difficult to implement the algorithm in a way that requires the data to be averaged for sampling. In this case, random sampling becomes the practical method of operation. Even though this method can achieve relatively uniform sampling, its randomness and tractability are inevitable. Although there is a method to achieve a better operation in conducting the test followed by a variable selection method to achieve the prediction of the test, the model is not the most realistic reflection of the good or bad. A more common way to divide a data set is used in the following models: 60% for training, 20% for verification, and 20% for testing. This ratio can also be adjusted according to the size of the data set and the signal-to-noise ratio of the data.

In the data collection stage, a large amount of information is obtained about students' basic enrollment, historical grades, and other school behavior data, but not all attributes are correlated with the repetition prediction; for example, there is no correlation between students' contact phone number and students' academics, which can be determined through the calculation of correlation coefficients, and finally those attributes that are not very correlated with each other are removed through attribute statute processing in feature selection. Feature selection is the process of actively selecting a subset of features. When analyzing multiple data sources, the high dimensionality of the data combined with the large number is the biggest difficulty. This ability of machine learning algorithms to generalize can be affected by complex and lengthy features, and the efficiency of the algorithm is greatly reduced. For this reason, when using the data to study student retention prediction, the redundant and unrelated parts should be removed according to their characteristics to further improve the efficiency of the algorithm, model generalization, and interpretability. In the course of the study, the features were selected by calculating the Pearson correlation coefficients of each characteristic attribute of the sample and the repetition results. The selection of feature attributes for student historical achievement data, basic enrollment data, and web log data was all performed by using Pearson correlation also known as product difference correlation (or product moment correlation) to determine the feature set for sample optimization, which is calculated as follows:

$$P_{x,y} = \frac{N \sum_{i=1}^{n} X * Y}{\sqrt{\sum_{i=1}^{n} (X + Y)}}. \tag{6}$$

In the formula, X denotes the original feature attributes in the sample, and Y denotes the retention result.

The predictive model training was selected from the enrollment table, student achievement data, and web log data of 2893 students enrolled in the class of 2016. During the study, the deeper meaning of the items in each table was analyzed, and the field parsing and data table articulation was achieved for the three types of data according to different perspectives and methods, resulting in three types of statistical results. When further examining their characteristics, the characteristics with low correlation were removed and multiple methods were used to analyze student retention. It is easy to see that the number and rate of course repeats are directly proportional to whether students take the course seriously, where the high average score is better reflected. The number of failed courses, the number of credits failed, the failure rate, and the grade point average should be selected as important characteristics in the repetition prediction. This is because the number of failed subjects is a key indicator in the evaluation of students' repetition. The hyperparameters of the BP deep neural network are set as follows: epochs are 100; minibatch is 300; optimization function is Adam; learning rate is 0.0001; and loss is categorical cross entropy. The loss is shown in Figure 5, the horizontal coordinate is the number of training times, the vertical coordinate is the loss of the model training, the red curve is the loss of the training set, and the blue curve is the loss of the test set. The model is trained 100 times and it converges after 80 epochs.

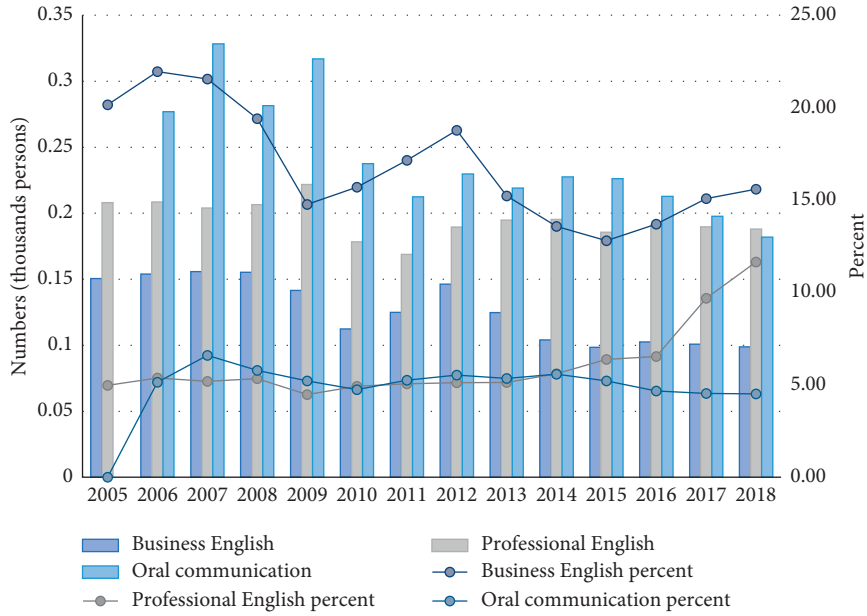

Figure 4: Student performance in English.

4. Results and Discussion

In order to be able to analyze and judge the data more clearly, the results of the above analysis are presented in the form of a bar chart as shown in Figure 6. Although the accuracy of each of these different models is well over half, it is not yet possible to determine that this is the optimal situation. The best result for the BP neural network is 83.55%, while the remaining three models are below 80%, which means that the students who need to repeat a grade are not effectively predicted in the remaining three models. The F1-measure is a comprehensive evaluation of Precision and Recall, and, for this reason, Precision and Recall are linked to each other for a comprehensive examination. Even though the above models perform differently in both Precision and Recall, the F1-measure is around 71%, and Precision and Recall only barely get some data analysis in this study, but they do not really get the best and optimal prediction, only the standard. Thus, new methods need to be tried to analyze the relationship between students' school performance and academic performance from multiple perspectives and levels.

The prediction of students' repetition was conducted by using students' historical grade data, and the results showed that simply using grade data for prediction has good results. However, there may be many factors affecting students' academic performance during their school years, and, in order to overcome the overly homogeneous type of prediction data and increase the diversity of prediction data features, further analytical exploration with students' web logs data will be attempted for better accuracy. According to the data preprocessing method previously introduced, the data related to the web logs of the class of 2016 students in the specified time period of about 22,00 GB, more than 6 billion web logs are processed and processed, and the analysis and investigation are done in the previous way with foresight; and the table of students' Internet information is obtained as shown in Figure 7.

The overall results of feature set I in predicting student repetition are poor compared to those of feature set II, which is sufficient to illustrate the two following points. First, the correlation between video time and grade repetition is low, and the results of feature screening do improve if video time is not considered. Second, it is not difficult to find that, after correlation analysis, the features related to web log data have the same low correlation with retention. This indicates that the feature data obtained by statistically processing the web behavior data do not significantly reflect the differences between repeating and nonrepeating students. Therefore, after adding the indicators of online behavior as features, the final prediction results decreased rather than increased. Adding students' online behavior data, combining the results obtained from previous predictions using only grade data, and combining features to select and apply different features for prediction, although the categories of features were increased, the effect was not improved, and the expected results of this study could not be obtained. In summary, it is necessary to consider the introduction of basic student profile data to predict student retention and observe the experimental effects. Combined with the above analysis, it was considered whether adding the basic student profile data would also result in the same situation as the Internet data. Two combinations of features were selected to train the model with the aim of observing the effect of different sets of features on the prediction effect of grade repetition.

(1) Feature set I: The statistical data on the types of websites visited, time spent on the Internet,

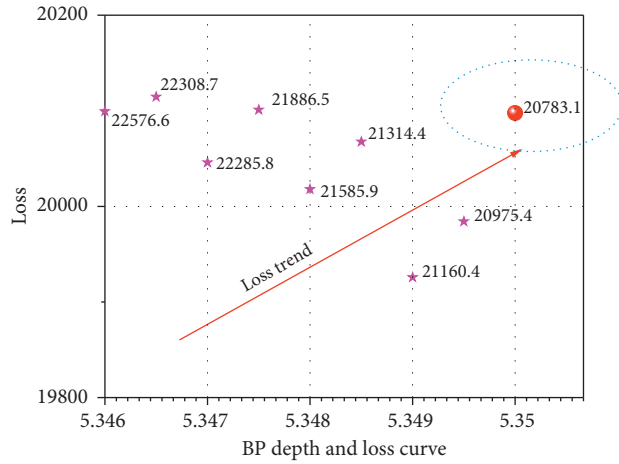

Figure 5: BP depth neural network model loss curve.

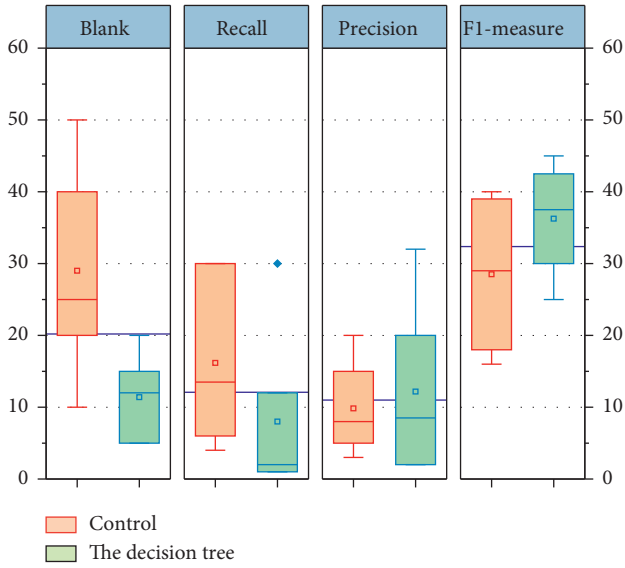

Figure 6: Comparison of the prediction effects of the four types of models.

frequency of Internet access, video time, and game time of students' Internet access data were added, and a total of 16 indicators were predicted by adding 11 pieces of feature data such as historical achievement indicators and basic student profile indicators, which were previously used for prediction.

(2) Feature set II : According to the previous analysis, the correlation between students' online behavior data and grade repetition was not high. In the previous study, it was found that the prediction results were still unsatisfactory after the inclusion of online behavior data. Therefore, only the length of time spent on the Internet was retained in the Internet behavior data, plus historical student achievement data and basic student enrollment data for prediction. The results of the logistic regression analysis of students' enrollment information and grade repetition showed that the failure rate, the score of failed subjects, and the length of time spent online were significantly and positively correlated with grade repetition, and the mean score and high school entrance exam results were significantly and negatively correlated with grade repetition as shown in Figure 8. When multiple regressions were conducted considering enrollment information, Internet access information, and historical grades, there was a highly significant correlation between failure rate and failure score, and the coefficient of failure rate was estimated to be 2.22, with a p value of $2.41E-08$. The coefficient of failure score was estimated to be 0.12, with a p value of $4.63E-14$.

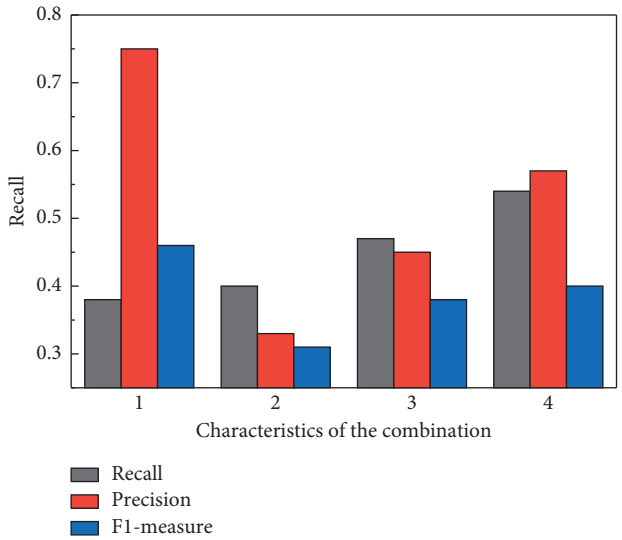

Figure 7: Student Internet access factor collection.

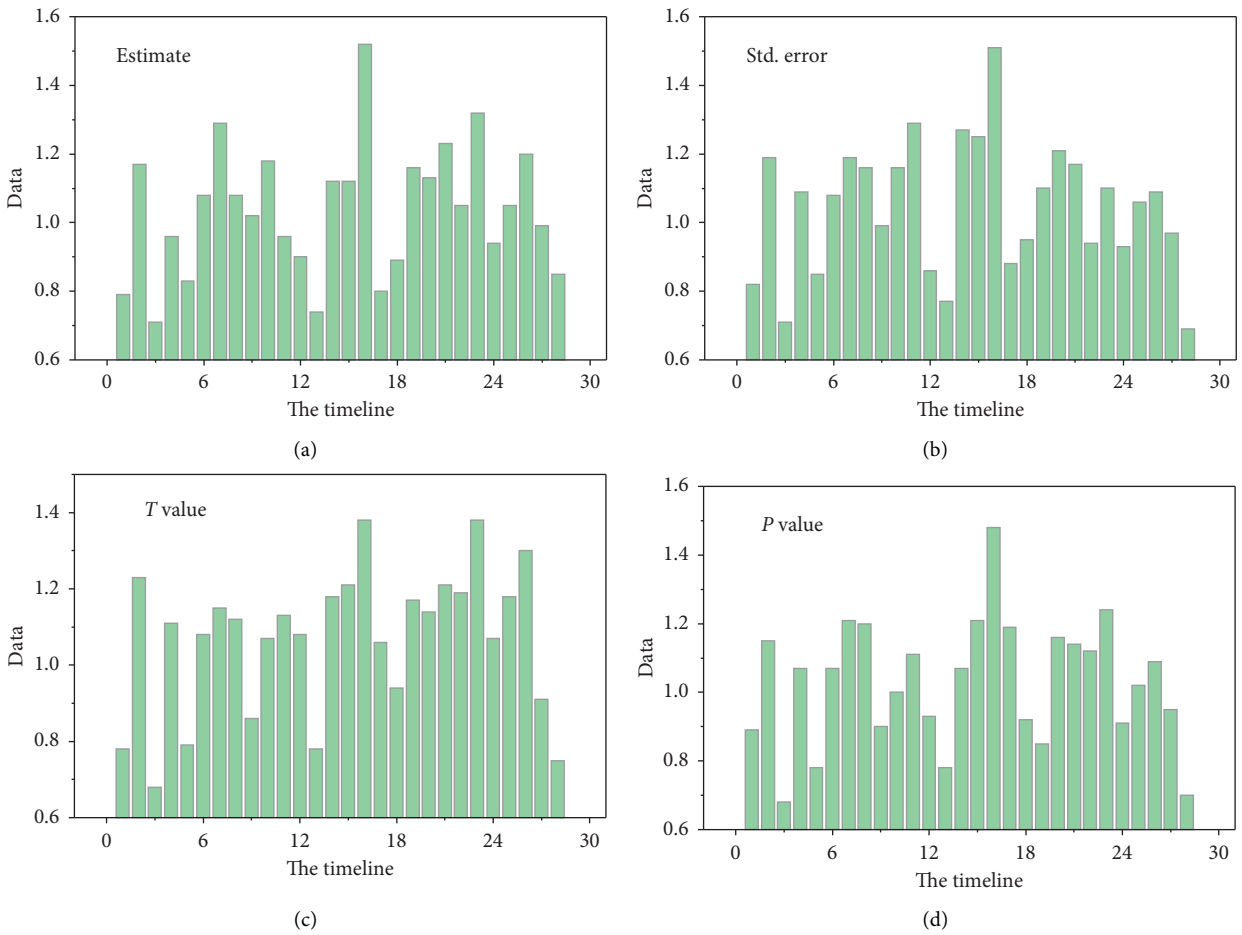

Figure 8: Results of logistic regression analysis of students' English performance.

5. Conclusion

In this study, the behavioral data of school students were collected, and four different machine learning methods were used to train models to predict whether students would be at risk of repeating a grade after their sophomore year. Through comparative analysis, we can conclude that the single data source prediction using only students' historical grades is more effective, and the F1-measure of all four machine prediction models is over 70%, although, among them, the BP neural network classifier is more effective in prediction, and Recall can reach 87.71% and Precision can also reach 62.07%. Using the basic student enrollment data and historical performance data, we can achieve the prediction of grade retention based on the combination of multiple features, and the experimental effect is significantly improved compared to the single data source. When the data source was extended to more than 30,000, the prediction effect of the BP neural network model was still the best, but the accuracy of the prediction decreased significantly. Therefore, in order to improve the accuracy of prediction, deeper mining and analysis of the BP neural network model are needed.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Z. Qian, E.-K. Lee, D. H.-Y. Lu, and S. M. Garnsey, "Native and non-native (L1-Mandarin) speakers of English differ in online use of verb-based cues about sentence structure," *Bilingualism: Language and Cognition*, vol. 22, no. 5, pp. 897–911, 2019.

[2] A. Basal, "Learning collocations: effects of online tools on teaching English adjective-noun collocations," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 342–356, 2019.

[3] L. Gao and Y. Qi, "The application of online vocabulary testing mode in college English teaching," *Canadian Social Science*, vol. 16, no. 2, pp. 7–11, 2020.

[4] Z. R. Eslami and X. Yang, "Chinese-English bilinguals' online compliment response patterns in American (Facebook) and Chinese (Renren) social networking sites," *Discourse, Context & Media*, vol. 26, pp. 13–20, 2018.

[5] S. Kong, "Practice of college English teaching reform based on online open course," *English Language Teaching*, vol. 12, no. 5, pp. 156–160, 2019.

[6] K. Stergiou, "The most famous fish: human relationships with fish as inferred from the corpus of online English books (1800–2000)," *Ethics in Science and Environmental Politics*, vol. 17, pp. 9–18, 2017.

[7] M. E. Ebsworth, T. Mcdonell, A. Defazio, and C. Cai, "Hypertext versus footnotes," *IALLT Journal of Language Learning Technologies*, vol. 47, no. 1, pp. 81–115, 2017.

[8] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774–1785, 2018.

[9] J. Chen, K. Li, Z. Tang et al., "A parallel random forest algorithm for Big data in a spark cloud computing environment," *Ieee Transactions on Parallel and Distributed Systems*, vol. 28, no. 4, pp. 919–933, 2017.

[10] X. Liang and J. Pang, "An innovative English teaching mode based on massive open online course and google collaboration platform," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 15, pp. 182–192, 2019.

[11] R. Wang, "Massive open online course platform blended English teaching method based on model-view-controller framework," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 16, pp. 188–196, 2019.

[12] J. Huang, "The dining experience of Beijing Roast Duck: a comparative study of the Chinese and English online consumer reviews," *International Journal of Hospitality Management*, vol. 66, pp. 117–129, 2017.

[13] P. Cuiqiong, "NCoV-based considerations on online- English teaching and traditional classroom-English teaching in China," *Ira International Journal of Education and Multidisciplinary Studies*, vol. 16, no. 2, pp. 96–101, 2020.

[14] J. Wu and B. Chen, "English vocabulary online teaching based on machine learning recognition and target visual detection," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 2, pp. 1745–1756, 2020.

[15] J. D. Forrester, A. K. Nassar, P. M. Maggio, and M. T. Hawn, "Precautions for operating room team members during the COVID-19 pandemic," *Journal of the American College of Surgeons*, vol. 230, no. 6, pp. 1098–1101, 2020.

[16] S. M. Lundberg, G. Erion, H. Chen et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[17] N. Arunkumar, K. Ramkumar, V. Venkatraman et al., "Classification of focal and non focal EEG using entropies," *Pattern Recognition Letters*, vol. 94, pp. 112–117, 2017.

[18] A. Riggs, "The role of stylistic features in constructing representations of Muslims and France in English online news about terrorism in France," *Perspectives*, vol. 28, no. 3, pp. 357–375, 2020.

[19] D. W. Mustikasari, "Developing massive open online course (MOOC): need analysis of teaching materials for madrasah English teachers," *Register Journal*, vol. 10, no. 2, pp. 170–184, 2017.

[20] C. Voyant, G. Notton, S. Kalogirou et al., "Machine learning methods for solar radiation forecasting: a review," *Renewable Energy*, vol. 105, pp. 569–582, 2017.

[21] M. Zhang, "Design and implementation of English online teaching system based on Big data," *Solid State Technology*, vol. 63, no. 3, pp. 2383–2392, 2020.