

Application of Machine Learning Technology in College English Teaching Evaluation

Ping Zhou

College of Foreign Studies, Hunan University of Humanities, Science and Technology,
Hunan 417000, China.

1185898551@qq.com

Abstract

Through online English learning platform for teaching and learning, data mining technology is used to analyze the students' learning data and establish relevant models to explore the correlation between English tests and various elements, which is of great value to students' learning and teachers' teaching. A portrait and analysis of undergraduates is constructed by using the data produced by the college English teaching and assessment system. Through the portrait of the users, it is possible to know the related attributes and learning habits of the students, which provides a basis for the later experimental design. The logistic regression model, enhanced decision tree model and the fusion model of the two models are used to predict the students' achievement and passing, respectively, and the factors influencing the students' passing are analyzed according to the prediction results.

Keywords

Data mining, College English, user portrait, prediction.

1. Introduction

In recent years, with the emergence of the Internet +, all walks of life have undergone earth-shaking changes, and "Internet + education" is gradually becoming a trend of development. Online education platform has attracted more and more attention, and more schools and students are willing to learn through online English learning platform. In college English teaching, teachers teach students English listening, speaking, reading and writing through online English platform, which makes it easier and more efficient for students to learn (Li, 2018).

Data mining technology has been widely used in all walks of life and achieved some results. Through online English learning platform for teaching and learning, data mining technology is used to mine and analyze students' learning data, and model training results are used to explore the correlation between English tests and various factors, which is of great value for students' learning and teachers' teaching. Therefore, it is necessary to apply data mining technology to college English teaching and learning. The research area belongs to one aspect of educational data mining. According to the different application fields of data mining, education data mining can be divided into the application of data mining in teaching and research, the application of data mining in the process of administration, and the application of data mining in the process of scientific research and management (Tsai, Ouyang & Chang, 2016).

Data mining is to mine useful and potential knowledge from a large number of data. Machine learning is one of the core research fields of artificial intelligence. It is one of the main methods to solve data mining problems and it is to train through a large number of data sets and to complete the task of data mining through a variety of related algorithms. As one of machine learning, deep learning has made great progress in artificial intelligence, human-computer

interaction and other fields, which has brought a new wave of machine learning. At present, deep learning technology has been widely applied in image recognition, speech recognition and other fields. The data of students' English exercises and examinations in college English teaching and assessment system are used to mine and analyze. Machine learning model can achieve good results, so machine learning related models are mainly used to mine and analyze.

2. Literature Review

In foreign countries where English is the second language, data mining technology is used to mine and analyze students' English learning by using the relevant data of English teaching and learning in universities. For example, in Korea, under the computer aided language mode, using the feedback data of students' online learning of English and the multiple linear model, the factors affecting students' learning attitudes are excavated and analyzed (Hartnett, 2016); in Japan, face-to-face teaching and autonomous learning are used, and according to the feedback data of students' learning of English, teacher interviews of students' survey data, final assessment data and teacher feedback data, students' enhanced autonomous learning is mined (Ask, 2017). In German, on the online English writing platform, students use data such as writing time, submission times, content changes and revision times to mine and analyze the factors affecting English writing, and put forward suggestions to improve English writing according to the influencing factors (Joorabchi, English & Mahdi, 2016). In Egypt, using the English online learning platform, the impact of various factors on learning satisfaction is mined through the questionnaire data filled by students (Mishra & Sahoo, 2016). The Nordic Digital Learning Research Project focuses on student achievement, teaching and learning processes, and knowledge sharing and communication (Liu, 2017).

Based on the above-mentioned literature, it is found that the application of data mining technology in college English teaching mainly uses the feedback data such as test results or questionnaires to mine and analyze. It also discusses how to mine and analyze the data according to the students' English training process or learning path. In accordance with learning path mining and examination results strongly-related factors, the portrait is mined and analyzed through college English teaching and assessment system of student learning data, and the relevant model is used to mine test-related factors.

3. Prediction and Analysis of Students' Passing Marks

3.1. Data Sets and Feature Processing

Students are trained using the college English teaching and assessment system to predict whether they will pass the exam. The research data comes from the teaching data generated by the system. The selected research data include undergraduates who have used the system for English training since 2011. The information involved mainly includes the following three aspects:

First, student user portrait: Student portrait data is student portrait data, mainly for students' basic information, answer behavior and other data.

Second, data on English training subjects: English training subject data is the basic information of each topic and the data when students do it. Specific data are: test paper mode, question type, question type 1, question type 2, question type duration, unit number, course number, difficulty, weight of the question, etc.

Test paper mode: there are two forms of test paper modes: exercises and examinations.

Question type: there are 18 forms of question types, such as long dialogues, short dialogues, judgments and choices.

Question type 1: there are four forms, namely listening, speaking, reading and writing.

Question type 2: it is divided into two forms: subjective questions and objective questions.

Question type duration: the length of time required to answer each question, which is calculated in seconds.

Difficulty: the difficulty coefficient of the students' selected papers.

Third, environment for students in English training: The environment data of students' English training include the time of answering questions, the equipment information used in answering questions, the way of surfing the Internet and the IP address.

The time of answering questions: According to the students' answer time, it is divided into several categories: according to the first ten days of each month, the middle ten days, and the last ten days of each month, it is divided into three categories; according to Monday to Sunday, it is divided into seven categories; according to morning, afternoon and evening, it is divided into three categories; according to the time of the day, it is divided into five periods.

The equipment information used in answering questions: Mainly for host device related information.

The way of surfing the Internet: According to the way of surfing the Internet, it can be divided into wired, WiFi and other forms.

IP address: According to the IP address, the geographical location of students in English exercises can be judged.

The collected data is saved in CSV format, where the data in the CSV file is separated by commas and the table data is stored in plain text.

The purpose of data preprocessing is to remove the features or attributes having low or even no correlation with the target variables. By preprocessing the data, the correlation between each dimension of the data set and the target variables is high when the model is put into training. In this way, the data mining work can provide cleaner and more effective data, reduce the amount of machine learning model algorithm for data processing, improve the efficiency of the algorithm for data processing, and improve the evaluation index of the final results. In this experiment, according to the type of question that appeared before, fill in the blank value manually; in the "pass or not", according to the passing score of each type of question, write a Python script to judge whether the type of question is larger than the pre-set passing score of the question, then fill in the field data; there is a blank value in the "oral self-rating number" or "mutual-rating number between students" which is filled in by means of average.

3.2. Students' Performance Passing Prediction

After data processing and feature processing, the ratio of positive and negative samples in the sample data is 9:1, that is, the ratio of passing sample data to failing sample data is 9:1, and the ratio of positive and negative samples is unbalanced. In the experiment, the method of up-sampling is used to deal with the problem of data imbalance, that is, the negative sample data is duplicated many copies to make the unbalanced data transformed into the balanced dataset. After processing, there are 14000 valid data. 11300 of them are randomly selected as training sets to establish classifier, namely learning model, and the remaining 2700 data are used as test sets to test the correctness of the model. The ratio of training set to test set in the sample is 4:1. The characteristics of the input model in this experiment include gender, the time for answering questions, scores, passing marks, test paper mode, question type, question type 1, question type 2, question type length, question difficulty, question weight, passing of CET-4 or CET-6, CET-4 or CET-6 scores, degree of interest in English, number of corrections to the answers, number of English training sessions, the average score of students' mutual evaluation, the number of students' self-rating, the final test score, the completion rate of homework, the number of watching videos, and the number of listening repetitions, a total of 22-dimensional data.

3.3. Experimental Design of Logical Regression and Enhanced Decision Tree Model Fusion Prediction Performance Passing

Fusion algorithm is the combination of multiple model algorithms through a certain way, which plays an important role in the field of recommendation and advertising. In practical application, because there are so many application scenarios of algorithm P pairs, the business data of each application scenario is also different. From the point of view of model algorithm theory, under the condition of different traffic data, different algorithms are suitable for different scenarios, and the situation that one algorithm performs better than other algorithms in all cases does not exist. Therefore, the fusion method can combine the algorithms having excellent performance in their respective scenarios to form a powerful algorithm. Linear weighting, cross-over fusion and waterfall fusion are three commonly used model fusion methods (Stupans, Mcguren & Babey, 2016).

In this experiment, linear weighted model fusion algorithm is adopted, that is, enhanced decision tree model and logistic regression model are linear weighted. The final weighted prediction results are:

$$p = lr_model * \alpha + gbdt_model * \beta, \alpha + \beta = 1$$

In the above formula, lr_model and gbdt_model refer to logical regression model prediction results and enhanced decision tree model prediction results, respectively. And when the model is fused, the two models are: pass rate of logical regression model when it iterates for 8 times and enhanced decision tree model error about 0.29. The experiments are carried out for different α values and the specific area under concentration time curve (AUC) is:

Table 1. Logical regression and enhanced decision tree fusion training result AUC value

α	0.1	0.2	0.3	0.6	0.8	0.9
AUC	0.615325	0.621744	0.632115	0.624388	0.619375	0.618257

From the AUC value, it can be found that the effect of the method of model fusion is better than that of using logistic regression model alone and enhanced decision tree model.

4. Prediction Results and Analysis

4.1. Relationship between Prediction Results and Characteristics

Logical regression model, enhanced decision tree model and two model fusion methods are used to predict the students' passing test. The correlation between the predicted results and the characteristics is as follows.

Table 2. Importance of characteristics in enhanced decision tree model

Characteristics	Degree of importance (frequency)	Characteristics	Degree of importance (frequency)
Score for each question	84	Repetition of listening	54
Student examination score	92	Average score of student assessment	51
Homework completion	81	Student self-rating score	48
English training times	79	Length of question	46
Interest in English	75	Question type	39
CET-4 and CET-6 achievement	74	Passing score	36
CET-4 and CET-6 passing result	71	Answer modification times	36
Gender	67	Time for answering questions	34
Difficulty of question	63	Test paper mode	34
Weight of question	62	Question type 1	31
Watching video frequency	56	Question type 2	29

4.2. Result Analysis

According to the sample data put into the classifier and the variables related to the prediction of students' passing test, the main factors affecting students' passing test are: the score of each question type, the score of students' examination, the completion of homework, the number of English training, the degree of interest in English, and the situation of CET-4 and CET-6. The factors are summed up and analyzed as follows:

First, the students' scores in each question type have the greatest influence on the students' passing of the exam and play a decisive role, followed by the students' final training results. A large number of samples used in the experiment are that a student participates in English training or English exams many times. When a student has no special weak items in listening, speaking, reading and writing, the students will be able to pass the exam. When the students can reach the passing level in all types of questions, the probability of a student passing the exam is also very high. In addition, students who pass the CET-4 or CET-6 and have higher scores have a higher probability of passing the exam. It can be seen from the data that whether the scores of each question type, final score, or CET-4 or CET-6 performance, to a certain extent, can reflect the students' original English foundation. Therefore, the students' original English foundation has a great impact on passing the exam.

Second, the homework completion and the number of participations in English training have a great influence on their passing the examination. In the experiment, the students whose have a high rate of completing their homework or participated in English training more frequently have a higher probability of passing the examination. The students who have completed their homework and participated in English training many times reflect that students spend a long time in English learning. It can also reflect the degree of students' hard work and students' attitude to English learning. Therefore, the high rate of student completion and the number of training times have a great impact on the passing of examinations.

Third, the degree of students' interest in English has an effect on the prediction of passing grades. The experimental prediction results show that the more students are interested in English, the more likely they are to pass the examination. In the modern "student-centered and teacher-oriented" teaching theory, it also emphasizes the importance of cultivating students' interest in learning. By cultivating students' interest in learning, we can not only stimulate students' learning motivation and arouse students' learning enthusiasm, but also promote students to actively participate in learning and cultivate students' awareness of autonomous learning. Interest in English can to some extent reflect students' emotional and psychological factors in the process of English learning. For example, in spoken English, some students' English pronunciation is influenced by geographical factors, such as local accent, and no one can understand them when they speak English. In the long run, students are no longer willing to speak English on their own initiative and even hate learning English. Therefore, the level of English learning interest can have a greater impact on English learning effect.

Fourth, from the results of the data, it can be seen that gender factor also has a great impact on English achievement. By observing and analyzing the sample data, it is found that under the same number of girls and boys, the number of girls passing the exam is more than that of boys. In the predicted results, the probability of girls passing the exam is higher than that of boys. This is positively related to female students' more positive attitude towards learning than male students. Variables "difficulty of question" and "weight of question" in the sample, the scope of change is not particularly large, the difficulty is moderate, and the weight of various types of questions fluctuates little, so the impact of the forecast results is not very great. In practice mode, watching video and listening repeatedly can improve the accuracy of the answers to the questions, which is very helpful to practice. But in the examination mode, video and listening cannot be repeated. Therefore, the influence of these two factors on the test results comes from the usual practice. For "student self-rating" and "student mutual-rating", the two variables are

assessed from the perspective of students' oral or writing. The students' mutual-rating scores have more influence on the predicted results than the self-rating scores. "Test paper mode", "Question type 1", and "Question type 2" have no effect on students' passing the exam.

5. Conclusion

With the rapid development of the Internet, "Internet + education" has become a trend. College English teaching and assessment system has made great progress in all aspects of English listening, speaking, reading and writing through online learning. After years of accumulation, the system has stored a large number of valuable teaching data, and these source data provide a basis for mining work. Through the in-depth study of data mining, machine learning and other related theory and technology, the teaching data generated in the system is comprehensively analyzed.

The teaching data produced in the teaching and evaluation system since 2011 are collected, the data are analyzed and processed, and individual portraits and group portraits are established for students. Through the user portrait, the crowd attributes of students and learning habits can be known, which provides a basis for the late experimental design to predict the passing results of students. The research environment of students' passing test situation prediction is set up, and Python language is used to write a program to extract, clean, and integrate data as well as other related work of feature engineering. Logistic regression model, enhanced decision tree model and two models after fusion model are used to achieve the prediction of students' passing test results.

References

- [1] Li, M. (2018). Investigation on application of association rule algorithm in English teaching logistics information. *Cluster Computing*, (4), 1-7.
- [2] Tsai, Y. R., Ouyang, C. S., & Chang, Y. (2016). Identifying engineering students' English sentence reading comprehension errors: applying a data mining technique. *Journal of Educational Computing Research*, 54(1).
- [3] Hartnett, J. L. (2016). Using data mining to teach applied statistics and correlation. *Teaching of Psychology*, 43(2), 142-146.
- [4] Ask, U. (2017). Business intelligence practices: adding evidence from organizations in the Nordic countries. *International Journal of Business Intelligence Research*, 4(2), 1-18.
- [5] Joorabchi, A., English, M., & Mahdi, A. E. (2016). Text mining stack overflow. *Journal of Enterprise Information Management*, 29(2), 255-275.
- [6] Mishra, B. K., & Sahoo, A. K. (2016). Evaluation of faculty performance in education system using classification technique in opinion mining based on gpu, 36(1), 1-23.
- [7] Liu, X. (2017). The application of data mining technology in the teaching evaluation in colleges and universities. *Journal of Computational and Theoretical Nanoscience*, vol. 14, issue 1, pp. 7-12, 14.
- [8] Stupans, I., MCGuren, T., & Babey, A. M. (2016). Student evaluation of teaching: a study exploring student rating instrument free-form text comments. *Innovative Higher Education*, 41(1), 33-42.