Research and Application of AdaBoost Based Prediction of Student's Academic Achievement

Ming-Wen GAO[1,a], Shuang ZHANG[1,b,*]and Qing-He HU[2,c]

[1]College of Software, Northeastern University, China

[2] College of Information Science and Engineering, Northeastern University, China

[a]137097338@qq.com, [b]zhangs@swc.neu.edu.cn, [c]huqinghe@ise.neu.edu.cn

*Corresponding author

Keywords: Teaching quality, Learning effect, Prediction of academic achievement, AdaBoost.

Abstract. Student's academic achievement is a major indicator of teacher's teaching quality and student's learning effect. But exam result is not enough to evaluate and predict student's academic achievement. The factors of prediction should be scientifically selected from the whole teaching process. Based on analyzing the shortcomings of current research, the author proposes an AdaBoost based, multiple-indicator prediction model of student's academic achievement. Experiment results show that the prediction model has good predictive performance.

Background

Education is the foundation of a country's comprehensive national strength. Teaching quality is an important factor to measure a school level, and students' academic achievement is the major factor to evaluate the teaching effect of teachers and students' learning quality. To rapidly and accurately predict teaching quality with advanced evaluation strategy is a main method to guide teachers to improve teaching method and student's learning effect, which is an important subject of educational administration departments, educational research institutes, primary schools.

Related Works

For far, there have been some researches on student's achievement prediction using computer techniques. Muhammed Salman Shamsi and Jhansi Lakshmi [1] studied the reasons for the high dropout rate of primary education in India. They used Naive Bayes, LibSVMware, J48, Random-Forest and JRip algorithms based on data mining technology to predict students' performance in grades and dropouts, compared the accuracy of various algorithms, and finally came to the conclusion that random forest Naive Bayes and JRip3 algorithm has higher prediction accuracy. But the indicators they studied did not include students' psychological factors, family factors, gender, health status and family type. Nguyen Thai-Nghe [2] applies the recommendation system to the field of educational prediction, and uses the methods of collaborative filtering and matrix decomposition to carry on the regression calculation. The experimental results show that the recommendation system also has a good performance in the field of student performance prediction. Shaobo Huang [3] developed and compared four mathematical models to predict the performance of engineering dynamics. The four models selected by the author include multiple regression model, multi-layer perceptual network model, radial basis function network model and support vector machine model. The experimental data were selected from the four pre-test courses of Engineering dynamics, i.e. statics, calculus I, calculus II and physics. A total of 2907 data points were collected from 323 students. The results of the four pre-test courses were used as input. The experimental results showed that the experimental results of the four mathematical models had little difference. The result of SVM model is the best. Guoqing Xuan [4] collects the relevant information and test results of the students in Grade Three of Junior High School, and uses cross-coverage algorithm to analyze and predict the students' scores. The author collected 408 students' scores and related information, which are divided into two groups. The first 300 samples are training samples, and the last 108 samples are test data. When the experiment adopted five attributes of activity health, aesthetic performance, learning attitude, practice innovation and seventh grade preliminary examination results, the accuracy of cross-coverage algorithm reached 75%; when the experiment adopted four attributes of activity health, aesthetic performance, learning attitude and practice innovation, the accuracy of cross-coverage algorithm reached 74.1%. Rahel Bekele and Wolfgang Menzel [5] demonstrated the application of Bayesian method in the field of education. They selected eight high school students' data sets, including students' gender, group work attitude, interest in mathematics, achievement motivation, self-confidence, personality type, English performance and mathematical performance, and used Bayesian network. The classifier predicts the performance of students, and the final experimental data show that the accuracy of the classifier is 64%.

Objective

At present, the study on the influencing factors of students' academic performance considers only a category of factors taken all other factors the same. That is, the researches did not consider the comprehensive factors of student’s academic performance. In addition, most of the current researches on student performance prediction methods are based on student’s scores, with relevant statistical data as indicators for performance prediction, or pre-curriculum performance as a predictive indicator to predict the results of the follow-up courses. It is unscientific to predict the future changes of students' academic achievement by only the scores factors. The objective of this paper is to establish a prediction model based on Adaboost, combined with multi-dimensional factors, including student’s performance, historic scores and other factors to predict student’s academic achievement.

Student’s Academic Achievement Prediction Index System

Taking full account of the requirement of quality education for students to develop in an all-round way, and combining the representative, measurable and holistic principles of index selection, this paper uses the market survey method, combines the research status of literature access, synthesizes the survey results of front-line staff and students, and summarizes the students' academic achievement forecast. The indicators of the test model mainly include four aspects: learning situation, daily performance, individual factors and teacher factors.

(1) Learning Situation

Learning situation refers to a comprehensive evaluation of students' mastery of the knowledge taught by teachers and their enthusiasm for learning. It includes students' learning attitude, average class score, passing rate, test results and pre-test results.

(2) Daily Performance

Students' daily performance includes students' performance in class and between classes, as well as their performance in life. Daily performance can be summarized as student classroom performance, student classroom test results, student disciplinary punishment records, student leave records

(3) Individual Factors

Individual factors refer to the unique indicators of each student, which are summarized in this paper as the relationship between parents, students' physical condition, students' psychological status.

(4) Teachers' Factors

Students' academic achievements will inevitably be influenced by teachers' factors. Teachers' teaching ability is the main embodiment of teachers' factors.

The overall structure of the indicators selected for students' academic influence is shown in Fig.1.

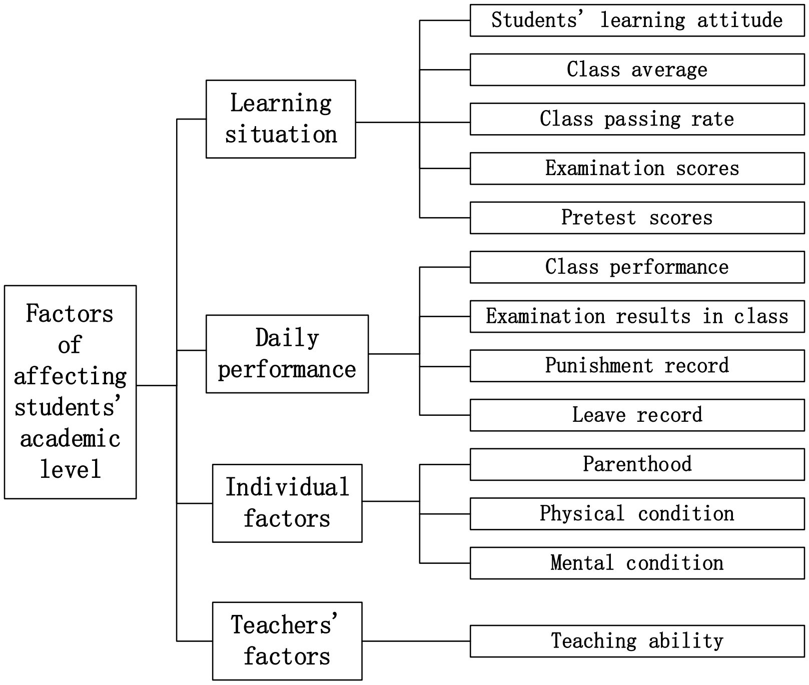

Figure 1. The overall structure chart of students' academic achievement prediction index

Solution of Academic Achievement Prediction Model Based on Adaboost.R2

Input and Output Feature Vectors Determination

The predictive model of students' academic achievement is established in this paper. The factors influencing students' academic achievement are students' examination results, learning attitude, average class score, passing rate, students' pre-test results, students' classroom performance, students' classroom test results, students' disciplinary punishment records, students' leave records, students' parents' relationship, etc. Students' physical condition, students' psychological state and teachers' teaching ability. When the prediction model is trained, the 13 indexes are used as the input eigenvectors of the AdaBoost algorithm, and when the prediction model is trained, the other 12 indexes are used as the input eigenvectors of the AdaBoost algorithm.

In this paper, the experimental data are normalized by using MATLAB data processing software, and the experimental environment is MATLAB R2016a. This paper will randomly extract 70 data from 339 student data as test data, and the remaining 269 data as training data. In the prediction process, the student's test scores will be output as the eigenvector of the algorithm. The index is a number in the [0, 1] range. The larger the number, the higher the student's score.

AdaBoost.R2 Algorithm Process

AdaBoost.R2 is a meta-algorithm, in the process of dealing with the problem, each time will consider to absorb the opinions of more than one expert, and then according to the opinions of each expert to give different weights, the final strong classifier by all the weak classifiers weighted sum. The schematic diagram of the AdaBoost.R2 algorithm is shown in Fig.2.

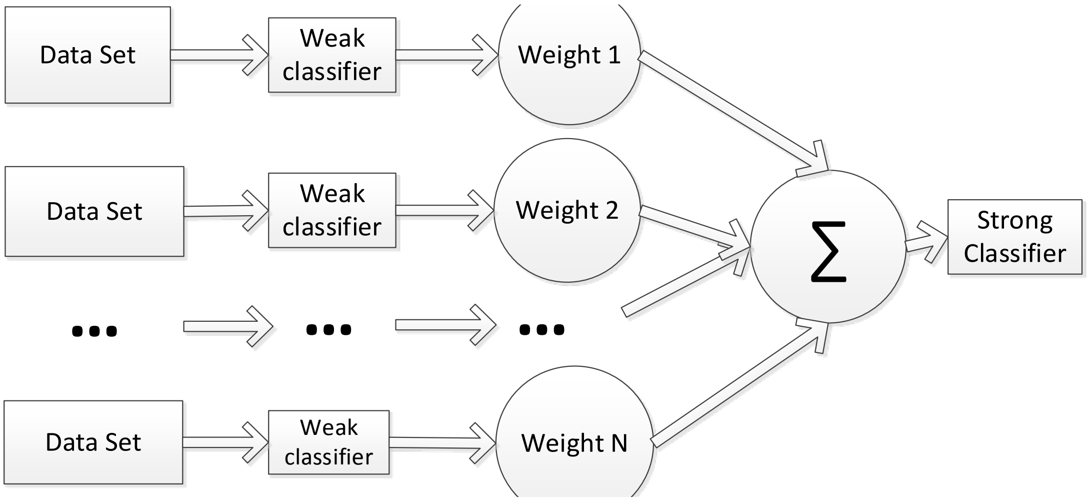

Figure 2. AdaBoost.R2 algorithm training process

Firstly, AdaBoost. R2 will construct a weak classifier. If the classifier selected in this paper is a decision tree algorithm, it is a single-level decision tree classifier. This weak classifier is equivalent to a domain expert. Although decision tree classifier is used, its prediction effect is not very good, usually only better than the result of random conjecture. The processed training data samples are input into the weak classifier, and each of the 13 training samples is given a weight. These weights form a vector. Here, the vector is defined as a vector M . When AdaBoost. R2 algorithm begins to train the model, the weight of each index is set to the same value. There are 269 training data in this paper, so the weight of each sample vector is set to $\frac{1}{269}$. When initialization, the average loss function is $\overline{L}_t = 0$, $t = 1, 2, \cdots, T$ m, T is defined as the number of iterations of user defined weak classifiers.

Firstly, a weak classifier is trained on the training data set provided, and the error rate of the weak classifier is calculated. In this paper, the error rate is defined as ε ,

$$\varepsilon\square = \frac{\text{Number of sample vectors with misclassification}}{\text{The number of all sample vectors}} \tag{1}$$

After each iteration, the weak classifier generated will have a weight value α ,

$$\alpha = \frac{1}{2}\ln(\frac{1-\varepsilon}{\varepsilon}) \tag{2}$$

Next, the loss of each sample is calculated by the loss function set by the user, and the regression model which is established by the first iteration $f_t(x) \to y$ can calculate the loss of the sample:

$$l_t(i) = \left| f_t(x) - y_i \right| \tag{3}$$

Then the weak classifier is trained on the original training data set again, but different from the first training, the weight of each sample vector in the training data set is adjusted when the weak classifier is trained for the second time, and the weight of the paired sample vectors decreases when the weak classifier is trained for the first time. Lower, the weight of the sample vectors that are mistaken will be increased. The weight vectors M of the training data set are updated, and the weights acquired after updating are used for the next iteration.

AdaBoost.R2 Algorithm Tuning

AdaBoost.R2 algorithm has 5 parameters, namelybase_estimator, n_estimators, learning_rate, loss, random_state. base_estimator is a weak classifier to enhance overall performance selection. This weak classifier must support sample weights; n_estimators is the largest number of iterations for weak classifiers. If the trainer trains well, the number of iterations for weak classifiers may be less than the set maximum number of iterations. S represents the loss function type of regression model, and the loss function of AdaBoost regression model has three kinds, including ’linear’, ’square’, ’exponentia’; random_state is used to generate a random number.

In the process of experiment, this paper combines the value of three indicators to measure the performance of the algorithm; the three indicators are mean square error, average absolute error, goodness of fit. Mean square error (MSE) is the expectation of the difference between the predicted value of the model index and the true value of the index. The smaller the MSE value, the smaller the prediction error. In this experiment, the error of 70 prediction samples is $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{69}$, and the MSE of the 70 prediction samples is:

$$MSE = \frac{1}{70}\sum_{i=1}^{70}\varepsilon_i^2 \tag{4}$$

The mean absolute error (MAE) is the expectation of the absolute value of the difference between the predicted value of the model index and the true value of the index. The smaller the MAE value, the smaller the prediction error. Similarly, the error of 70 prediction samples in this experiment is as $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{69}$, and the MAE of the 70 prediction samples is:

$$MAE = \frac{1}{70}\sum_{i=1}^{70}\left|\varepsilon_i\right| \tag{5}$$

Goodness of fit refers to the fitting degree of the regression line learnt from the prediction model to the actual observed value of the sample. The range of goodness of fit is [0, 1], the closer the value is to 1, the better the fitting degree is. On the contrary, the closer the goodness of fit is to 0, the worse the fitting degree between regression line and actual value is.

In this article, decision tree is used as weak classifier, and setting up base_estimator = Decision Tree Regressor, random_state is set to none. Combined with all the experimental results, we find that when loss='square', learning_rate=1, n_estimators=300, goodness of fit reaches 0.9015717, mean square error reached 0.00543621, the average absolute error reached 0.05356242, at this point, the performance of the algorithm is optimal.

Analysis of Experimental Results

The student's academic achievement prediction model proposed by the paper have the attributes of studying situation (including studying attitude, class's average score, class's pass rate and student's score of examination), daily performance (including student's performance in class, score of quiz in class, punishment records and leave records), individual factor (including student's closeness with his parents, physical and mental state) and teacher's teaching ability. The data comes from a grade one of a Hebei middle school. The proposed Adaboost based student's academic achievement prediction model has goodness of fit of 0.902 and mean square error of 0.074. Comparison shows that the proposed prediction model has better performance than other models.

Summary

This paper studies the current situation of elementary education and the problem of middle school students' exam results prediction, establishes a multi-index student academic achievement prediction index system, and combines the AdaBoost algorithm in the field of machine to predict students' academic achievement efficiently and accurately, which provides effective help for educators.

References

[1] Shamsi M S, Lakshmi J. Student performance prediction using classification data mining techniques [J]. 2016.

[2] Thai-Nghe N, Drumond L, Krohn-Grimberghe A, et al. Recommender system for predicting student performance[J]. Procedia Computer Science, 2010, 1(2):2811-2819.

[3] Huang, Shaobo, Fang, et al. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models [J]. Computers & Education, 2013, 61(1):133-145.

[4] Guoqing Xuan. A Research on Prediction of Academic Record Based on Neural Network and Cover Algorithm [D]. Anhui University, 2011.

[5] Rahel Bekele, Wolfgang Menzel. A Bayesian Approach to Predict Performance of a Student (BAPPS): A Case with Ethiopian Students[C]. IASTED International Conference on Artificial Intelligence and Applications, part of the 23rd Multi-Conference on Applied Informatics, Innsbruck, Austria, February 14-16, 2005.