Application of improved adaboost iteration in small sample for prediction of oral performance

Hongmei Guan
Applied Technology College of Dalian Ocean University, Dalian, China

Corresponding author and e-mail: Hongmei Guan, atlixiang@dlou.edu.cn

Abstract. Adaboost as a kind of iterative algorithm, is a kind of important integrated learning technology, can be randomly surmised with slightly higher prediction accuracy than using weak learning enhancement study for high prediction precision of apparatus, it’s is very difficult in strongly constructing learning, it provides an effective new way of thinking and new methods for the design of the learning algorithm. As a kind of algorithm framework, Boosting can be applied to almost all the popular machine learning algorithm to further improve forecasting precision of the original algorithm, and its application is very extensive and has a great influence. Based on Adaboost, 90 students’ oral English at a university as weak classifier samples, via the iterative calculation, get a high reliability of strong classifier, and it’s concluded that the prediction result with student can evaluate the prediction accuracy of the algorithm for results compared with the actual results.

1. Introduction

English as the current and the widest range of language use. English as mother tongue country is about more than 40, 20% of the world countries, and the number is on the increase. For more than 80% in the existing information on the Internet are in English with English as the carrier. As a language of communication is so important, for our country the importance of the contemporary college students, English as a language, learning to hear it is particularly important. What is the method of study is the problem needed to be solved. By Analyzing the students' oral English and listening ability, can better improve the students' ability of oral English translation.

Common grade forecast analysis methods include decision trees, clustering analysis, association rules, and correlation analysis, etc. As Wu Qiang [1] and others will combine the decision tree and LMBP neural network algorithm, using the decision tree algorithm white-box high comprehensibility, black box and the LMBP neural network classification accuracy is high, after 15 iterations, the final accuracy can reach 84.73%.And by using clustering analysis algorithm by Zhou Xia [2] and others, this paper analyzes the factors which influence on college English curriculum requirements, the results show that the "performance" and " TingKeLv correlation reached 91%," and "evaluation" and "performance" correlation reached 83%, thus the relationship of the communication between teachers and students can be given out, the accuracy of teachers’ evaluation is higher, so that they can better help college students of curriculum.

Bagging: bootstrap together, also known as the bagging method, is selected from the original data sets S time get S after a new data set of a technology. Boosting is a technology similar to the bagging. Both in boosting and bagging, the use of multiple classifier types is consistent. But in the midst of
boosting, different classifier is obtained by training the serial, every new classifiers have been trained according to the performance of the classifier for training. Boosting can obtain a new classifier by focusing on the existing classifier fault points of the data.

Due to boosting classification is based on the results of all the classifier weighted summation of the results, thus boosting and bagging is different. Classifier of bagging weights are equal, and boosting the classifier in weight is not equal, each weight represents its corresponding classifier in the iteration of the degree of success. Boosting method has multiple versions, AdaBoost is one of the most popular version.

This article adopts the method of modified AdaBoost iteration, analyze the factors affecting students' oral English, translation test and compared with the actual results. Xie Xingyu [3] and others, forecast the student achievement by using the modified TrAdaboost algorithm based on the analysis of students consumption habits and previous achievements as well as the information such as the library entrance guard, which is of great significance for university educational administration and management personnel.

2. Principle
Adaboost is an iterative algorithm, and its core idea is different for the same training set training classifier (weak classifier), then put these weak classifier together, constitute a stronger final classifier (strong classifier). Its goal is to create the final classifier of the following:  

\[ F(x) = \text{sign}\left(\sum_{m=1}^{M} \theta_m f_m(x)\right) \]  

(1)

\( \text{sign} \) function. The sign function for function symbols here. Through the judgment of real Numbers of plus or minus, if enter a positive number, the output 1; If negative, the output 1,; If 0, 0. Namely Adaboost target to judge the binary classification criterion of the \( \{-1,1\} \). Function image as shown below:

![Figure 1. Sign function image.](image-url)
The weak classifier $f(x)$. In the model $f_m(x)$ is a weak classifier. Here need to assume that Adaboost by $M$ after a weak classifier weighted summation. Each weak classifier and $f_m(x)$ by original data give a forecast results, then according to the weak classifier by weight $\theta_m$ weighted summation. From this we can see that Adaboost goal is find out the weak classifier model parameters and the corresponding weights.

2.1. Adaboost solving steps are as follows:
First, assume that n input data and the weight of each data are the same, that is:
$$\omega(x_i, y_i) = \frac{1}{n}$$  (2)

After that, we for each weak classifier $(1, \ldots, M)$ are the following:
A weak classifier training, make its minimum classification error. The weak classifier error calculation formula is as follows:
$$\varepsilon_m = \sum_{y_i \neq f_m(x)} \omega_m$$  (3)

The meaning of the formula for the model error is every fault points and the sum of the weight of the sample. And when the model for the first weak classifier, the meaning of this formula for calculating the current weak classifier by wrong sample number divided by the total number of samples, and the error of the weak classifier is obtained.

according to the error of the current weak classifier, calculation of the classifier. Computation formula is as follows:
$$\theta_m = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_m}{\varepsilon_m} \right)$$  (4)

This formula means: if the weak classifier is greater than 0.5, the accurate value of the right of the value is positive (at this point $\varepsilon_m < 0.5$). Otherwise, negative.
to update the data according to the model. Formula is as follows:
$$\omega_{n+1}(x_i, y_i) = \frac{\omega_m(x_i, y_i) \exp [-\theta_m y_i f_m(x_i)]}{Z_m}$$  (5)

$Z_m$ as the regularization coefficient, here is to ensure that the weight of all the data and to 1.
2.2. Adaboost process can also be said the diagram below:

![Flow chart of Adaboost iteration.](image)

Figure 2. Flow chart of Adaboost iteration.

3. Experimental design
Obtained 90 students by a university life clockwise for oral English four final data, and according to the course score (30%) including the usual dialogue, question and answer (35%) and performance (35%) of three parts. Student scores as shown in Table 1:

| Serial number                  | ordinary times points | Question and answer | Performance dialogue |
|--------------------------------|------------------------|---------------------|----------------------|
| (The first semester)1          | 29                     | 29                  | 25                   |
| (The first semester)2          | 20                     | 25                  | 25                   |
| (The first semester)3          | 25                     | 26                  | 29                   |
| (The first semester)4          | 26                     | 18                  | 25                   |
| (The first semester)5          | 25                     | 19                  | 25                   |
| (The first semester)6          | 22                     | 19                  | 30                   |
| ...                            | ...                    | ...                 | ...                  |
Existing data using the Adaboost principle into the python code running calculation, first of all, by using the first four oral English scores as weak the classification of the term, and after several rounds of training, to forecast the next semester and compared with the actual results. Among them after a weak classifier multiple iterations and weighted distribution after the new strong classifier of predicted scores as shown in Table 2:

| Serial number    | ordinary times points | Question and answer | Performance dialogue |
|------------------|-----------------------|---------------------|----------------------|
| (The first semester)1 | 28.6                  | 30.1                | 25.2                 |
| (The first semester)2 | 20.1                  | 24                  | 23.1                 |
| (The first semester)3 | 24.9                  | 27.8                | 32.1                 |
| (The first semester)4 | 25.1                  | 21.2                | 24.9                 |
| (The first semester)5 | 25.5                  | 20.3                | 24.3                 |
| (The first semester)6 | 23.9                  | 18.5                | 27.7                 |

... ... ... ... ... ... ...

At the end of this term oral English scores as shown in Table 3:

| Serial number    | ordinary times points | Question and answer | Performance dialogue |
|------------------|-----------------------|---------------------|----------------------|
| (The first semester)1 | 28.6                  | 30.1                | 25.2                 |
| (The first semester)2 | 20.1                  | 24                  | 23.1                 |
| (The first semester)3 | 24.9                  | 27.8                | 32.1                 |
| (The first semester)4 | 25.1                  | 21.2                | 24.9                 |
| (The first semester)5 | 25.5                  | 20.3                | 24.3                 |
| (The first semester)6 | 23.9                  | 18.5                | 27.7                 |

... ... ... ... ... ... ...

4. Results and discussion
After Adaboost training will predict the result and the actual final result comparison, here is more intuitive to show the result comparison, with the method of radar map.
At ordinary times points as shown in Figure 3:
Figure 3. Forecast and actual result contrast figure (ordinary times points).

Question and answer as shown in Figure 4:
Figure 4. Forecast and actual result contrast figure (q&a).

Dialog as shown in Figure 5:
As shown in the above, the prediction result and the actual result of coincidence degree is higher, visible Adaboost algorithm for performance forecast, the feasibility and the accuracy is higher.

Dimension 1 (ordinary times points): in the dimension, the score for performance prediction and the actual error is less than 1 samples (91.11%).

Dimension 2 (q&a): in the dimension, the score for performance prediction and the actual error is less than 1 samples (93.33%).

Dimensions 3 (performance dialogue): in the dimension, the result of forecast and actual error score less than 1 sample was 90.00%.

5. Conclusions

By Using Adaboost forecast the study result accuracy is higher, the algorithm using adaptive iteration based on earlier classifier properly with all kinds of extreme cases, and it can be seen that the above experimental results have good effect on the convergence. Through analyzing the importance of characteristic value, the algorithm can calculate weights of factors affecting the relationship between
precipitation grades. These conclusions can better help the school to understand students' weak spot, in order to "suit the remedy to the case", so as to achieve no resource waste, at the same time can be treated as a "special".

References
[1] Wu Qiang Fang Rui, Han Bin etc. Based on decision tree - the LMBP neural network analysis and prediction model of student performance study [J]. Journal of chengdu university of information engineering, 2018, 33 (3): 274-280.
[2] Zhou Xia Zhang Ying snow, lil 'g. College English curriculum requirements analysis based on clustering analysis research [J]. Journal of modern electronic technology, 2019, 42 (16): 154-158.
[3] Xie Xingyu ying-lu zhang. TrAdaboost algorithm based on improved student performance ranking prediction [J]. Computer and modernization, 2016, (2): 122-126.