Prediction of Poor Students' Classification Based on Adaboost Algorithm Integrated Learning Model

Wang Li¹ and Guie Jiao²,*

¹ Shanghai Ocean University, No. 999 Hucheng Ring Road, Pudong New District, Shanghai, China
² Shanghai University/Shanghai Jian Qiao University, No.1111 Hucheng Ring Road, Pudong New District, Shanghai, China

*Corresponding author: jiaoguie@gench.edu.cn

Abstract. Aiming at the problem of identifying the poor students in colleges, based on the improved classification imbalance, this paper proposes to establish the AdaBoost algorithm integrated learning model for the first time to solve the problem of identifying the poor students in colleges. In this model, decision tree is used as weak learner, and principal component analysis algorithm and balance cascade algorithm are used to reduce dimension and undersampling data. At the same time, the paper focuses on the comparison of the algorithms that choose the effect of sample set classification as the weight(SAMME.R) and the size of the prediction probability of sample set classification as the weight(SAMME). Experiments have proved that the AdaBoost ensemble learning algorithm model is better than the single classifier algorithm in predicting the results, which can ensure the fairness and objectivity of the recognition results, and has a strong application value for the recognition of poor students in colleges.

Keywords: Ensemble Learning, AdaBoost, Classification Prediction, Undersampling Algorithm

1 Introduction

Deprived students refer to students whose families are too poor to rely on their financial ability to complete their studies. Current college students applying for subsidies for deprived students often need the certification of the relevant departments. Students should take the initiative to apply from the school who evaluate and determine the qualification of students. However, there are always deviations in the manual identification. For example, there are dishonest students' applications, and some deprived students are reluctant to submit applications because of self-esteem. This causes many students who need subsidies to be unable to get help. Therefore, it is fair and accurate to evaluate poor students and leave the allowances to the students who really need them. It is about the fairness of the society.

As a member of Boost algorithm[1-2], there are many researches on Adaboost algorithm. Literature[3] designed a new multi-category ensemble learning algorithm Adaboost-BHC. Reference[4] proposed the
weighted voting of base classifiers that could not achieve ideal results by Adaboost algorithm, which effectively improved the recognition accuracy. Literature\cite{5} once proposed the combination of PCA algorithm and random projection to further increase the speed of the Adaboost algorithm. literature\cite{6} proposed "attention mechanism". Add different weights to the data set, so as to effectively reduce the impact of noise, these data can be used more effectively. Reference\cite{7} randomly selects feature subsets from the training set, and uses genetic algorithms to perform operations on the selected feature subsets. This part of the subsets iterates out the representativeness, and then inputs the base classifier to reduce the base classifier speed and the loss of error is minimal. But this method is only suitable for the case of small sample set. Reference\cite{8} uses a random sampling method, the integrated training is processed as an additive model of gradient descent, and the samples with low weights are discarded, thereby speeding up the running speed. Reference\cite{9} improved performance optimization on this basis. Reference\cite{10} introduced Hoeffding bounds to ensure that the distribution probability of the subset data and the total data set is similar.

For the problem of identifying and granting deprived students, using college student data, compared with other methods of this problem in this paper. The Adaboost algorithm is first applied to the problem of deprived students, choosing a decision tree with only one decision node as the weak classifier. At the same time, the Balance Cascade algorithm is introduced to sample the data to reduce the unbalanced error of the data category. Experiments have proved that the method proposed in this paper is effective, and the results are relatively ideal, which plays a positive role in identifying deprived students.

2 Experiment And Result Analysis
The experiment is based on student consumption data, and uses ensemble learning to classify the poor and non-poor students. The main flow chart of the experiment is shown in Figure 1 below.

![Fig 1 Experiment process](image)

2.1 Experimental Results And Analysis
Using the Adboosting algorithm to predict the data, the accuracy rate is shown in Figure 2 below.
Fig 2 Classification and Prediction Results

In the figure, category A is for non-poor students and category B is for poor students. The prediction accuracy rate is 92.25%, and the model prediction results are very accurate.

Figure 3 Function decision value

Figure 3 is the size of the function decision value of category A and category B. The decision value can be simply interpreted as a confidence level. The larger the absolute value of the decision value, the greater the confidence that the model predicts for the category, and the more likely it is to be determined as the category.

Fig 4 Comparison of SAMME and SAMME.R
Figure 4 is a comparison of two different weighting algorithms. It can be clearly seen that in this experiment, the prediction error rate of the SAMME algorithm is greater than the SAMME.R algorithm. After the first 40 weak classifier iterations, the SAMME.R algorithm has reduced the error rate to about 25%. Its iteration speed is far superior to the SAMME algorithm.

![Figure 4](image)

**Fig 5 Error rate of each sub-model**

Figure 5 is the error rate of each weak classifier of the Adaboost algorithm in this experiment. Blue is the weak classifier using SAMME algorithm, red is a weak classifier using the SAMME.R algorithm. A total of 600 weak classifiers. The actual error rate of the SAMME algorithm in each sub-model is also higher than the actual error rate of the sub-model of the SAMME.R algorithm.

![Figure 5](image)

**Fig 6 Each sub-model weight of SAMME algorithm**

Figure 6 is about the weak classifier weights of two different algorithms. The SAMME algorithm focuses on the classification effect of the previous sample set, and its sub-model weights vary from 0.1 to 0.4. The SAMME.R algorithm uses the classification prediction probability of the previous sample set as the weight, and the sub-model weight values are all 1.

![Figure 6](image)
Fig 7 Logistic regression

Figure 7 shows the classification of logistic regression experiments, where the blue dots are the predicted non-poor students. It can be seen that the classification situation is very bad, and it is seriously affected by the outliers, which cannot be used as a reference in the identification of poor students.

![Logistic regression](image)

Fig 8 Decision tree

Figure 8 is about the classification of decision trees. Dark areas are poor areas. Because of the classification imbalance treatment, the accuracy is slightly higher than that of logistic regression. However, the result is still very unsatisfactory, and is still seriously affected by outliers.

2.2 Experimental Comparison And Model Analysis

2.2.1 Algorithm Evaluation
This article uses Precision, Recall, and F1-score to comprehensively evaluate the pros and cons of the following three algorithm models. The comparison is shown in Table 1.

|                | Logistic regression | Decision tree | Adaboost     |
|----------------|---------------------|---------------|--------------|
| Precision      | 44.36%              | 67.94%        | 92.25%       |
| Recall         | 78.33%              | 75.43%        | 92.15%       |
| F1-score       | 56.64%              | 71.48%        | 92.20%       |

2.2.2 Model comparison analysis
The accuracy rate in Table 1 represents that the predicted result is the proportion of positive examples in the positive sample. The higher the value of this ratio, the higher the accuracy of the model. The recall rate is the proportion of the sample that represents the true positive example and the prediction result is the positive example. The higher the value, the stronger the ability to distinguish the positive sample. The F1 value is a value based on the accuracy rate and the recall rate. This value indicates that the accuracy rate and recall rate are equally important. The higher the value, the more stable the model.

From the accuracy analysis, we know that the accuracy of model 1 and model 2 are not good. Model 3 uses the Adaboost algorithm to integrate the basic learner, and combines the learner with feature extraction, selects the best features, and classifies the targets. Its prediction performance is more ideal, and it can identify poor students who really need help. From the analysis of the recall rate,
we can see that the recall rate of the three models exceeds 75%, which indicates that the majority of the student subsidies that have been issued by the poor students in the model. The recall rate of Model 3 reached 92.15%, which is the best among the three models. In contrast to the stability of the model, this paper uses F1 values, that is, the accuracy rate and recall rate are regarded as equally important data. The F1 value of Model 3 is much higher than that of Model 1 and Model 2, and it can be seen that Model 3 is the best compared to Model 1 and Model 2.

2.3 Analysis And Summary
(1) This experiment applies the ensemble learning Adaboost algorithm to the classification of poor students for the first time. The basic weak classifier of the algorithm uses a decision tree of single node. Set of 600 weak classifiers combined as the final strong classifier. Compared with the previous method for predicting poor students, the accuracy of the prediction in this experiment is higher. Compared with the previous method for predicting poor students, the accuracy of the prediction in this experiment is higher. It provides a more effective and credible reference for the identification of poor students.
(2) The experiment focuses on comparing two different weight acquisition algorithms of SAMME.R algorithm and SAMME algorithm. The predicted error rate of each algorithm, the actual error rate of each sub-model and the weight of the sub-model are compared respectively. In this experiment, the SAMME.R algorithm performed better.
(3) This experiment improves the classification imbalance caused by the various methods proposed in the previous literature. Compared to a single decision tree algorithm or basic logistic regression, the abnormal data was removed. The accuracy is effectively improved, and the classification effect is more excellent.

3 Conclusion
This paper proposes a prediction model for poor students based on ensemble learning, which combines the work of Adaboost algorithm and identification of deprived students. Considering some correlations between data dimensions, PCA is used to reduce the dimensions of the original data to 6 dimensions. Then use the Balance Cascade algorithm, which is one of the undersampling algorithms based on the Adaboost algorithm, to deal with the data imbalance problem. Finally, the Adaboost algorithm using the decision tree as the basic weak classifier. Compared with a single algorithm classifier, a better classification effect is achieved. Which has helped the identification of poor students.

The next research direction can be placed on the comparison of different weak classifiers. The base classifier of AdaBoost algorithm can choose other algorithms as weak classifiers. For example, support vector machine (SVM) can be selected as a weak learning machine to classify and predict the poor students' data, and the final set becomes the final strong classifier. In addition, the combination of features among data-related dimensions, the selection and construction of the best feature value, the prediction of the target can also be the next research direction.

References
[1] Freund Y, Schapire R E A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1999, 55: 119-139.
[2] Ohnbar E, Trivedi M M. To boost or not to boost? On the limits of boosted tree for object detection[C]. International Conference on Pattern Recognition Cancun, Mexico, 2016: 3350-3 355
[3] Goo Jun, Ghosh J. Multi-Class Boosting with Class Hierarchies[C]/Proc of the 8th International Workshop on Multiple Classifier Systems, 2009: 32-41.
[4] Nock R, Nielsen F A Real Generalization of Discrete Adaboost[J]. Artificial Intelligence, 2007, 171(1): 25-41.
[5] Paul B, Athithan G, Murty M N, et al. Speeding up AdaBoost classifier with random
projection[C].International Conference on Advances in Pattern Recognition.

[6] Tolba M, Moustafa M. G. Adaboost: Accelerating Adaboost feature selection with genetic algorithms[C]. International Conference on Evolutionary Computation Theory and Applications, Porto, Portugal, 2016: 156-163

[7] Lin Y, Shen S, Liu Z, et al. Neural Relation extraction with selective attention over instances[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, (1): 2124-2133.

[8] Friedman J H. Greedy function approximation: A Gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1891-1892.

[9] Friedman J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis, 2002, 38(4): 367-378.

[10] Pedro M Domingos, Geoff Hulten. Mining high-speed data streams[C]. International Conference on Knowledge Discovery and Data Mining, Boston, Massachusetts, USA, 2000: 71-80.