Students’ Classification Model Based on Stacking Algorithm

Fei Pan, Yingchun Yuan and Yufei Song
College of Information and Technology, Hebei Agricultural University, Baoding, Hebei, 071000, China
289776537@qq.com

Abstract. A model for Students’ Classification which aims to help college teachers teach in accordance with students’ aptitude is proposed based on the daily practice data of a course. Based on the Stacking method, this classifier model ensembles eXtreme Gradient Boosting algorithm, Random Forest algorithm and the Logistic Regression algorithm. The experimental results show that this model can divide effectively students into different classes and the accuracy can be significantly improved by comparing with other classifier models. And it can serve as reference and guidance for college students’ classifier on other courses.

1. Introduction
With the continuous development of educational informatization, the teaching management data of colleges and universities has increased rapidly. How to obtain valuable knowledge and information from them with scientific methods to provide better services for college teachers? This issue has promoted the development of data mining research in the field of education. Generally speaking, teachers are more eager to analyze the data of students' daily learning to grasp the students' course learning effects timely and accurately. On this basis, teachers can also carry out different teaching for students with different levels of study to help them progress, such as providing extended ability training for students with excellent daily performances, giving warnings for students with poor curriculum, and using extracurricular tutoring. This kind of data mining can not only help teachers improve the teaching effect of the course, but also achieve the goal of precise education under the support of educational big data. For this purpose, this paper combines the results of entrance examination scores, daily training scores, mock exam scores and final test scores of the 2018 undergraduate "Computer Foundation" of Hebei Agricultural University, and builds a Stacking model for students’ classification based on XGBoost algorithm, RF algorithm and LR algorithm. According to the degree of mastery of the course, the students are divided into three categories: poor, good and excellent. The model is designed to help college teachers find out the poor students and take effective aids for them. It can also help outstanding students to expand their training, and finally realize the teaching of students in accordance with their aptitude and precise education.

2. Related Research
The application of data mining in the field of teaching has attracted extensive attention in recent years. Since 2005, a number of international conferences on data mining and artificial intelligence have added seminars on “educational data mining”. Based on rough set theory, P. G. JansiRani [3] extracted the dominant attributes of students' learning and found the rules to improve students' performance. C. Romero and S. Ventura[4] summarized and analyzed the development of educational data mining during the decade from 1995 to 2005, and summed up four directions of data mining application in the field of...
teaching: Statistics and visualization, classification, clustering and outlier detection; Wu Xiping [5] used the improved AprioriTid algorithm to analyze the relationship between students' daily scores, college entrance examination scores and teacher information, and found the main factors that affect the students’ college entrance examination scores, thus promoting the improvement of teaching effect and teaching efficiency, however, in the course of his research, he did not consider the completion of student work, daily practice scores, etc. Yang Guojing [6] used the association rule algorithm to analyze the daily course data of students, and used the relevant rules between courses to guide students, at the same time, he used decision tree algorithm to analyze the relationship between the interest of students, students’ learning time and teaching methods to improve the teaching effect, but the sample for data mining is not diverse and small, the results are not completely consistent with the actual data. In summary, education data mining mainly uses a single algorithm such as Bayesian Network, Association Rules, clustering algorithm and logistic regression to model the relationship between the course and students' performance. However, it caused a large difference between the prediction and the actual result because of the few samples.

In college teaching, we usually classified students according to the test scores, thus transforming the students’ grade into a two-category or multi-category problem. In the existing research results, compared with a single classifier such as naive Bayes and decision tree, integrating different classifiers may make the final model more accurate and stable. We call it Ensemble algorithm. Ensemble algorithm is mainly divided into Bagging and Boosting. The commonly used Bagging method is Random Forest algorithm (RF), and the Boosting method has XGBoost algorithm. Among them, the Random Forest algorithm uses the CART decision tree as a weak learner, randomly selects some features and samples from the data set, then constructs multiple decision tree models in parallel. By voting method, the category with the most votes will be taken as the final output. However, when the data noise is relatively large, the algorithm will cause over-fitting result, which makes the model inaccurate. The XGBoost algorithm builds an optimal model by minimizing the loss function, and combines multiple CART trees iteratively to improve model performance. The Stacking algorithm refers to an integrated learning method that combines multiple classification models by a meta-classifier, which stacks the primary learner and the secondary learner. The primary learner is trained on the complete training data set, and then the secondary learner re-trains the data on its output feature set. The Stacking algorithm can improve the accuracy of the algorithm and can effectively avoid the risk of over-fitting, so it has received extensive attention in recent research.

3. Description of Student Classification Model Based on Stacking Algorithm

By using the data of the daily practice and the examination which are about “Computer Foundation”, this paper proposes a student classification model based on Stacking algorithm. According to the difference in the mastery of the course, students are divided into three categories: poor, good and excellent. Based on the Stacking idea, the above three methods are integrated to improve the accuracy of the model, it can avoid the over-fitting result and enhance the generalization ability of the model.

The process is as follows:

1. Input training set D, test set T;

2. Algorithms A and B in base learner (XGBoost algorithm and RF algorithm) are subjected by 5-fold cross-validation based on D, then randomly using the 4 fold data for 5-times trainings, and using the remaining 1 fold for 5-times predictions, the results are retained in $D_A$ and $D_B$;

   At the same time, we predicted 5 times with 4 fold data in the test set T, then added the 5 predicted to find the mean value, the result is retained in $T_A$ and $T_B$;

3. Splicing $D_A$ and $D_B$ together as secondary training set, the number of rows is same as D, and the number of columns is 6; $T_A$ and $T_B$ are merged into a new matrix as secondary test set, the number of matrix row is same as the test set T, and the number of columns is also six columns.

4. The model in the secondary learner is re-trained using the secondary training set and the secondary test set, and the predicted results of the A and B models are used as features (6 features), we give weights to the features to make the final result most accurate.
The algorithm framework is shown in Figure 1.

![Figure 1. The framework of Stacking algorithm.](image)

In Figure 1, the XGBoost model and the RF model in the base learner are trained on the complete training set, and the 5-fold cross-validation of the two models is performed to reduce the risk of overfitting and improve the accuracy, then we combined the output results of each fold to obtain the probability value of the prediction result of the base model. The secondary learner LR algorithm trains the prediction probability mosaic results of the two base layer models as features (2*3=6 features), and gives the weights to the two algorithms in the base learner to make the final prediction more accurate. The following is the model used in the Stacking algorithm.

### 3.1 XGBoost Algorithm

The XGBoost algorithm combines multiple classification regression trees (CART trees) to make the combined model generalization ability stronger. The XGBoost model can be expressed as formula (1):

\[
\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)
\]  

(1)

The K is the number of decision trees, \( f_k \) is the kth decision tree, and \( \hat{y}_i \) represents the final prediction of the sample \( x_i \).

The XGBoost algorithm is an additive model, rather than simply repeating the combination of several CART trees. It uses the error generated by the last prediction (model composed of t-1 trees) as a reference to establish the next tree (t tree) and reduce its loss function. The prediction result obtained by adding the t-th tree to the model can be expressed as:  

\[
\hat{y}_i^t = \sum_{k=1}^{K} f_k(x_i) - \hat{y}_i^{t-1} + f_t(x_i).
\]  

How to find the newly added model \( f_t \) to minimize the objective function becomes the target of the XGBoost algorithm. Its calculation formula is as shown in (2) (3):

\[
\text{Obj}^t = \sum_{i=1}^{m} l(y_i, \hat{y}_i^t) + f_t(x_i)) + \Omega(f_t) + \text{constant}
\]  

(2)

\[
\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2
\]  

(3)
Among them, $l(.)$ is the loss function, that is the error between the predicted value and the true value in training, $\Omega(\cdot)$ is the regular term penalty function of model complexity, $\gamma$ is the number of leaves, $\lambda$ is a fixed coefficient, $\omega_j^2$ is the L2 modulus square of $\omega$.

3.2 RF Algorithm
The random forest algorithm evolved on the basis of Bagging's thought. It uses CART decision tree as the base learner and improves the decision tree model. The CART decision tree selects an optimal feature among all $n$ features on the node to divide the left and right subtrees of the decision tree, but the random forest algorithm randomly selects some samples and features (less than $n$) on the nodes, and then selects an optimal feature among these partial sample features to divide the left and right subtrees and enhance the generalization ability of the RF model.

Assume that the model input is sample set $D = \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)\}$, Weak classifier iteration $T$ times, The final output is a strong classifier $f(x)$.

For $t=1, 2, \ldots, T$: First, the $t$-th random sampling is performed on the training set, and the samples are sampled $m$ times to obtain a sample set $D_t$ containing $m$ samples. Then train the $t$-th decision tree model $G_t$ with $D_t$. When training the decision tree model, a part of the sample features are randomly selected from the nodes, and then an optimal feature is selected to divide the left and right subtrees. The category in which the $T$ weak learners voted the most was used as the final prediction result.

3.3 LR Algorithm
In the Stacking algorithm, the secondary learner generally uses a higher stability and simpler algorithm to reduce the risk of model overfitting. Therefore, this paper uses LR algorithm as a secondary learner to build the overall model.

The LR algorithm adds a Sigmoid function to the linear regression algorithm to convert the original numerical result into a probability between 0 and 1. If the probability value is greater than or equal to 0.5, it is divided into category 1; less than 0.5 Classified as category 0.

Among them, the Sigmoid function is expressed as:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{4}$$

Among them, $\theta^T x = \sum_{i=1}^{\eta} \theta_i x_i$, $\theta_i$ is the factor that affects the target value. In the classification task, the probabilities belonging to category 1 and category 0 can be expressed as:

$$P(y = 1|x; \theta) = h_\theta (x) \tag{5}$$

$$P(y = 0|x; \theta) = 1 - h_\theta (x) \tag{6}$$

This makes it possible to intuitively get the classification results we want. The LR algorithm is fast and easy to understand, and has high stability, so it is used as a secondary learner of the Stacking algorithm.

4. Students’ Classification
We divided the research of student classification based on Stacking algorithm into four stages: data collection; data preprocessing; model design; model evaluation. The model framework is shown in Figure 2.
4.1 Data Collection
The experimental data in this paper are from the entrance examination scores, daily training scores, comprehensive exercises, mock exam scores, examination scores, sex, etc. of the 2018 undergraduate "Computer Foundation" course of Hebei Agricultural University. Among them, daily training results include Excel scores (including Excel format and fill, Excel fill, Excel processing, Excel data processing, Excel calculation, Excel intermediate, Excel advanced, etc.), Word scores (including subsections, emails, etc.), Windows scores (including beginner, intermediate, advanced, etc.), Internet scores (IP settings) and single choice scores (algorithm). A total of 7662 data were collected from the entire data set, including 19 variable values.

4.2 Data Preprocessing
(1) Data consolidation. Use the Tableau Prep tool to read each Excel transcript separately, combine the data according to the uniqueness of the student number field, and write it into a new CSV file.
(2) Handling missing values. The general rule is to filter out data with more missing values (more than 40% of the total). That is, for a student, delete the characters if they are missing more than 40% of the scores at the same time, and the remaining missing values are filled with the average value.
(3) Data normalization. The total scores of the exercises in this experiment are different, which has a great influence on the experimental results. Therefore, the practice subjects are normalized to the hundred percentage point system.
(4) Data reduction. Daily practice data is too detailed, so we reduce the dimensions by merging data. For example, Word scores include two sections of sub-sections and mail exercises, it can be averaged as Word practice score. Other training scores, such as Excel, contain more parts, it can be weighted average as the final Excel score. The final 5 practice scores are obtained. A total of nine features are retained after data reduction with other data.
(5) Feature ranking. In order to improve the efficiency of the model and obtain the importance of the factors affecting the students' classification, this experiment performs the feature sorting work. Use the XGBoost algorithm to measure the importance of variables.

The characteristic importance of the XGBoost model is shown in Table 1.

| Number | Features          | Importance |
|--------|-------------------|------------|
| 1      | Entrance exam     | 1325       |
| 2      | Mock exam         | 1194       |
| 3      | Comprehensive practice | 1152   |
| 4      | Word score        | 1111       |
| 5      | Excel score       | 948        |
| 6      | Windows score     | 506        |
| 7      | Internet score    | 382        |
| 8      | Algorithm score   | 313        |
| 9      | Sex               | 130        |

It can be seen from the table that the entrance examination, the mock exam and the comprehensive exercise are several important factors influencing the final student classification results, indicating that the student's entrance basis may have a greater impact on the classification, and the daily training of the students determines different classification results. In the training of single subjects, Word and Excel account for a large proportion, indicating that these two subjects are more important in the basic computer courses. Sex is the least important, indicating that men and women have roughly the same situation in computer science and have a lower impact on student classification.

After the data preprocessing process, 5551 effective experimental data were obtained. According to the final test scores, the student classification is determined, and 0-70 is divided into poor students, and the code is set to 0; 70-90 is divided into medium students, the code is set to 1; 90 points or more is divided into outstanding students, the code is set to 2. From the processed valid data, 70% of the randomly extracted data is used as the training set, and the remaining 30% is used as the test set; The students' bottoming, practice, simulation and other grades are taken as sample x, and the classification result based on the final test score is used as the label value y to establish a classification model to detect the classification effect.

4.3 Model Design
The specific steps of the model design are as follows:

(1) The XGBoost algorithm cuts the optimal feature of the tree model from the root node according to the information gain until the maximum tree depth we set is reached (If the tree is too deep, it is easy to have a partial sample overfitting), generate a CART tree; Then build multiple CART trees in the direction of the gradient of the loss function gradient, and establish the base layer XGBoost model of the Stacking algorithm.

(2) The RF algorithm uses the CART tree as a model in Bagging to generate m training sets. For each training set, we construct a CART tree, randomly extract a part of features from all features, and find the optimal solution among the extracted features, then split. This establishes a base layer RF model based on Stacking algorithm.

(3) The LR algorithm is used as a secondary learner to further train the secondary data sets obtained by the base layer model to improve the accuracy of the algorithm.
In the training process, the accuracy of the algorithm can be significantly improved by adjusting different parameters. After 5-fold cross-validation, the final parameters of the XGBoost algorithm are determined as: Learning rate is 0.3, the maximum depth of the tree is 9, the subsample and the colsample_bytree are both 0.8, best when the number of iterations is 23, the regularization weight is 5; The final parameters of the RF Algorithm are determined as: the maximum depth of the tree is 6, the number of trees is 100. The final parameter of the Stacking algorithm is determined as: use_probas is True, average_probas is False.

4.4 Classification Model Performance Evaluation

This paper uses the "accuracy" indicator to evaluate the student classification model. Accuracy is the ratio of the number of samples with the correct classification of the model to the total number of samples. In general, the higher the accuracy, the better the classification of the model.

In order to verify the superiority of the algorithm, RF algorithm, XGBoost algorithm, LR model, CART tree and KNN model are selected for comparison. The experimental results are shown in Table 2.

Table 2. Comparison of classification accuracy with 5-fold cross-validation.

| Algorithm | 1     | 2     | 3     | 4     | 5     | Average   |
|-----------|-------|-------|-------|-------|-------|-----------|
| Stacking  | 75.06%| 73.62%| 75.55%| 75.55%| 74.23%| 74.80%    |
| XGBoost   | 74.81%| 74.00%| 74.90%| 75.29%| 73.58%| 74.52%    |
| RF        | 71.47%| 70.40%| 71.43%| 69.76%| 70.88%| 70.79%    |
| LR        | 69.15%| 68.51%| 68.73%| 69.50%| 68.52%| 68.89%    |
| CART      | 68.38%| 67.10%| 69.24%| 69.24%| 70.32%| 68.85%    |
| KNN       | 66.20%| 65.68%| 65.38%| 66.80%| 69.55%| 66.72%    |

Table 2 shows that the classification accuracy of the Stacking algorithm is significantly better than the XGBoost, RF, LR, CART, and KNN algorithms. It is seen from the average that the classification accuracy of the XGBoost algorithm reaches 74.52%, the RF algorithm reaches 70.79%, and the LR algorithm reaches 68.89%. After the integration of XGBoost, RF, and LR algorithms by the Stacking model, the prediction accuracy is improved by about 0.28% on the basis of XGBoost, by about 4.01% on the basis of RF, and by about 5.91% over LR. In summary, the student classification model based on Stacking algorithm has higher accuracy.

5. Conclusion

Based on the daily learning data of the 2018 undergraduate students, this paper builds the Stacking integrated student classification model which is combined with Bagging and Boosting ideas. Then we compare it with integrated models such as XGBoost, RF algorithm and other single classification models like LR, CART and KNN algorithm, the result shows that the student classification model based on Stacking algorithm has higher accuracy. It is suitable for processing data in this article.

In the future research, we will continue to learn the parameter optimization method of Stacking algorithm integration, and try to integrate other models to further improve the accuracy and calculation speed of the classification model and improve the classification model.

Acknowledgments

This project is supported by the ‘Science and Technology Research Project of College and University in Hebei Province’ (No. ZD2015054); Research on Cloud Resource and Task Scheduling for Big Data Processing System (No. 2018KFKT003); The Graduate Innovation Funding Project of Hebei Province (No. CXZZBS2017059).
References

[1] Yang G J, Xu X, Zhao F Q. (2019) Predicting User Ratings with XGBoost Algorithm. Data Analysis and Knowledge Discovery, 01: 118-126.

[2] Cao W, Li C, He T T, et al. (2018) Predicting Credit Risks of P2P Loans in China Based on Ensemble Learning Methods. Data Analysis and Knowledge Discovery, 10: 65-76.

[3] P. G. Jansiri Rani, R. Bhaskaran. (2010) Extraction of Dominant Attributes and Guidance Rules for Scholastic Achievement Using Rough Set Theory in Data Minging. International Journal of Computer Science Issues, 7(3): 28.

[4] Romero C, Ventura S. (2007) Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1): 135-146.

[5] Wu X P. (2010) The Analysis of College Student Achievement Based on Association Rules Minging Technology. Southwest JiaoTong University.

[6] Yang G J. (2015) Data Mining Based Research on Analysis to the Data of the College Teaching. Hebei Normal University.

[7] Cheng X F, Zhang X Y, Wang N. (2018) Identifying Trending Topics in Q&A Community with CART Decision Tree. Data Analysis and Knowledge Discovery, 2(12): 52-59.

[8] Luan H B. (2017) Research and Application of Data Mining in College English Teaching and Evaluation. Beijing University of Posts and Telecommunications.

[9] Breiman L I, Friedman J H, Olshen R A, et al. (1984) Classification and Regression Trees (CART). Encyclopedia of Ecology, 40(3): 582-588.

[10] Li Y J, Guo H X, Li Y N, et al. (2016) A Boosting Based Ensemble Learning Algorithm in Imbalanced Data Classification. Systems Engineering-Theory & Practice, 36(1): 189-199.

[11] Zhang T. (2018) Research on Correlation Analysis of University Students’ Score Based on Data Minging. Beijing University of Posts and Telecommunications.

[12] L. Guo, J. H. Lin, Q. Guo, et al. (2016) Identify multiple influential spreaders in term of the distance-based coloring. Physics Letters A, 380: 837-842.

[13] P N Tan, M Steinbach, V Kumar. (2006) Introduction to Data Mining. Addison Wesley/ Pearson, 1: 320-332.

[14] Zhang X. (2015) Study of an improved Apriori algorithm for data mining of association rules. International conference on Applied Science and Engineering Innovation.

[15] Sen B, Ucar E, Delen D. (2012) Predicting and analyzing secondary education placement-test scores: a data mining approach. Expert Systems with Applications, 39(10): 9468-9476.

[16] Zhang X D. (2017) Application Research of Educational Data Mining Technology in University Computer Foundation. Inner Mongolia Normal University.

[17] Xia Y, Liu C, Da B, et al. (2018) A Novel Heterogeneous Ensemble Credit Scoring Model Based on Stacking Approach. Expert Systems with Applications, 93: 182-199.

[18] Zhang Y, Cao J. (2016) Decision Tree Algorithms for Big Data Analysis. Computer Science, 43(S1): 374-379, 383.

[19] Li F. (2017) Study on the Application of Data Mining in Academic Achievements Analysis. LanZhou University.

[20] Qiao F, Ge Y H, Kong We. (2017) MapReduce based distributed improved random forest model for graduates career classification. Systems Engineering–Theory & Practice, 37(5): 1383-1392.