Prediction Model of Graduate Enrollment Rate Based on Improved Random Forest

Yumin Zhen, Bincheng Xu, Hui Zhong

School of Software Engineering, Beijing Jiaotong University
{18126362, 18126341, 18126363}@bjtu.edu.cn

Abstract. With the improvement of people's knowledge level, the number of graduate students is also increasing. In this paper, the research direction of proceed from actual demand of the students to apply for to attend graduate school, through the data mining technology, the students of various grades of information as a feature vector, university graduate student's acceptance rate prediction, enables the student to about their chances of ever attaining a university admits there is a fair understanding, so as to provide decision support for their entrance to choose. In recent years, in order to better predict student performance, many researchers have proposed many prediction models, such as bayesian classification model, random forest, neural network and so on. We will also use the random forest algorithm based on the prediction research, and make appropriate improvements. In terms of how to improve the accuracy of model prediction, we will conduct research and optimization and improvement from the following two aspects:(1) the sample distribution of the data set is not balanced and the number is small. How to improve it?(2) how to improve the classification accuracy of random forest based on the complexity of data samples?By comparing the traditional random forest with the optimized random forest, it is proved that the optimized random forest can improve the prediction accuracy of graduate admission rate.

Keywords: Prediction of admission rate; Random Forest; Data Set Optimization

1 Introduction

Up to now, there are many academic research topics based on the prediction of student performance, such as improving the teaching quality through the prediction of student performance [1], improving the student retention rate and so on[2,4,20,21]. However, as far as we know, there is little research on the prediction of overseas graduate enrollment rate. In addition, with the expansion of people's horizons and the increasing number of people studying abroad, it is of great significance for applicants to predict the probability of being admitted to universities through the application materials submitted for overseas study.

There have been many models for predicting classification regression and so on, and random forest is one of them. In this study, we found that the decision tree with poor classification performance would be generated when the decision forest was used, which had a bad impact on the final voting result and the prediction performance of the model. Therefore, we will use an improved random forest method. The decision trees with good classification performance in the random forest model are selected for similarity calculation, and a new random forest model is formed according to the decision trees with different similarity. Before generating a new random forest model, we compared the accuracy of the random forest model composed of decision trees at different depths, and took the corresponding depth with the highest accuracy as the depth parameter of our improved random forest algorithm. In addition, SMOTE technology was adopted to analyze a small number of class samples and add new samples to the data set manually according to the small number of class samples, so as to finally realize the improvement of random forest optimization and improve the prediction ability of postgraduate admission.

The paper is organized as follows. In the next section, we will review the relevant work. Section 3 introduces our data set. We report our findings in sections 4-7. Finally, we discuss the conclusions and future work.

2 Related Works

With the boom of machine learning and data mining technology, prediction model algorithm has been widely used in various fields, especially the integrated learning technology of Stochastic Forest algorithm, which can predict regression problems and classify problems. For example, disease forecast [3], stock forecast [5], traffic forecast [6] and so on. At the same time, according to our research and development, compared with the basic prediction model, random forests tend to have higher performance and more accurate classification accuracy in prediction problems, and can process high-dimensional data [1,7,8].

There are many excellent algorithms in ensemble learning, and we compared random forest with other algorithms simply through research. In terms of the selection of optimal variables, Bagging selects the optimal partition value by traversal of all prediction variables. The random forest selects the best segmentation variable by reducing the correlation between trees through randomness [7]. In contrast, the classification intensity of random forest is higher than bagging, which can reduce variance more effectively. On the other hand, AdaBoost generally has a great advantage in classification accuracy, and can perfectly fit the training data [9], but the training time of the model is long, and it is easy to lead to the phenomenon of "overfitting". Although XGBoost is an upgraded GBDT, it also supports linear classifiers, which can generate candidate segmentation points more efficiently [4]. But the XGBoost algorithm is computationally more complex in the validation process. Through the comparison of the above algorithms, our group decided to use the random forest as the main model to predict the graduate enrollment rate.

Random forest has great advantages over other algorithms, but there are still many areas for improvement. In terms of algorithm optimization, b. Ravi Kiran [10] improved the generalization error of decision tree by combining OOB unsampled samples, but this idea may be incompatible with the smooth weighted average decision function provided by random forest. In terms of model optimization, Yiyi Liu [12] importance variable weighted random forest is put forward, on the basis of information feature extraction using weighted sampling strategy, can increase under the condition of weak signal and big noise prediction accuracy, but at the same time there are continuous variables and classification or classification variables on the number of levels at the same time, the variable importance rating estimation inaccuracy problems.

Although there are many improvements based on random forest [11,13~16], few people have paid attention to the research on the problem of unbalanced samples using random forest algorithm. The imbalance of data classification is one of the common problems in data mining. Therefore, it is necessary to effectively improve the random forest prediction rate for unbalanced samples.

3 Dataset

For this study, we collected a set of data from Kaggle, which collected 500 pieces of data on UCLA graduate student enrollment. The data set was provided by Mohan S Acharya[22] to help students place their personal data on university shortlists, and the predicted output gives them a fair idea of their access to a particular university. A description of all the characteristics and their values used in this study is shown in table 1.

Table 1． The characteristics and values of the data.

Characteristics	Description	Value
GRE Score	Results of postgraduate entrance examination	0~340
TOEFL Score	TOEFL score	0~120
University Rating	The level of a student's undergraduate school	0~5
SOP	student's score of the statement of purpose	0~5
LOR	the letter of recommendation strength	0~5
CGPA	student's score of the undergraduate GPA	0~10
Research	Whether the student has research experience	0 or 1
Chance of Admit	the value of Chance of Admit	0~1

Among these 8 characteristic attributes, we take the Chance of Admit attribute as predict labels. Meanwhile, in order to improve the accuracy of prediction, we classify the value of Chance of Admit into different levels(Table 2).

Table 2． The level of the Chance of Admit(value).

Chance of Admit(value)	Level
0.9＜value≤1	1
0.8＜value≤0.9	2
0.7＜value≤0.8	3
0.6＜value≤0.7	4
0.5＜value≤0.6	5
value≤0.5	6

4 Preliminary Studies

We did a simple visual analysis of the features in the data. When conducting data statistics on the characteristics of GRE Scores(Fig.1.), we found that most students' GRE Scores were concentrated in the Scores from 310 to 325, with relatively few students having high Scores and low Scores. When conducting data statistics on the characteristics of University Rating(Fig.2.), we find that students' undergraduate schools mainly belong to schools with grade 3 or above. When conducting data statistics on the characteristics of CGPA(Fig.3.), we find that only a few students have a CGPA below 8.0, and there are more students with high grades. Based on the statistics of these three characteristic attributes, we made a comparative analysis of GRE Scores and CGPA(Fig.4.). According to the figure, we can find that people with higher CGPA usually have higher GRE Scores maybe because they have strong learning ability. We also compared the characteristics of University Rating, CGPA and Research(Fig.5.). We found that people with higher University Rating usually have higher CGPA. Meanwhile, with the same University Rating, students with research experience tend to have higher CGPA than those without research experience.

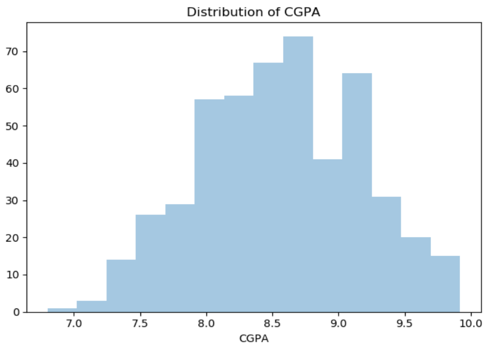


Fig.1. Distribution of GRE

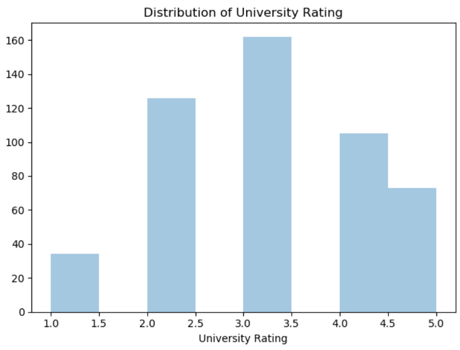


Fig.2. Distribution of University

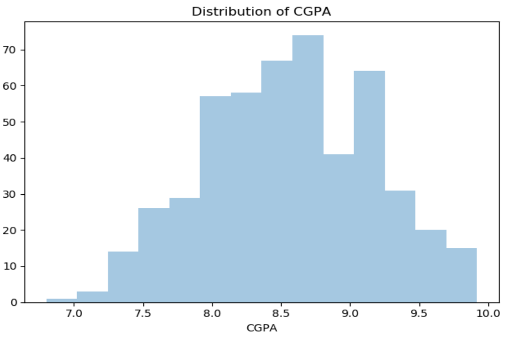


Fig.3. Distribution of CGPA

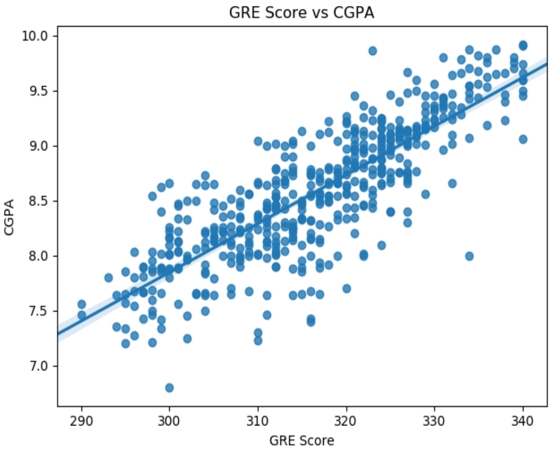


Fig.4. GRE Scores VS CGPA

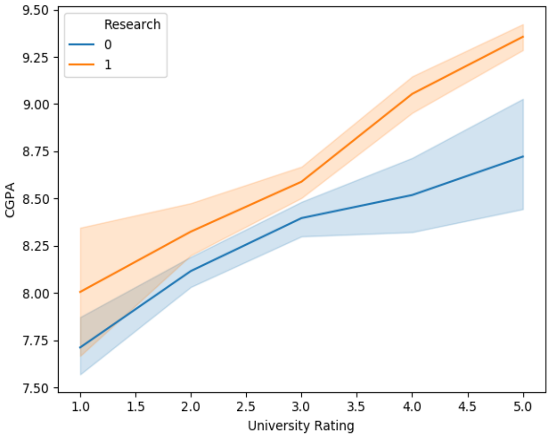


Fig.5. The relationship between University Rating, CGPA and Research

In addition， we divided the Chance of Admit tag of native data, marking the feature below 0.6 as 0 and the feature above 0.6 as 1. However, we found that the classified data presented unbalanced phenomenon. In the 500 data sets, the number of positive samples was 403, while the number of negative samples was only 97. The difference in the number of positive and negative samples was too large (Fig.6.). When model training is carried out for data with unbalanced sample size, the prediction of a small number of samples will be biased, resulting in a decline in the accuracy of the model. Therefore, we will through SMOTE technique to deal with a few sample class imbalance problem.

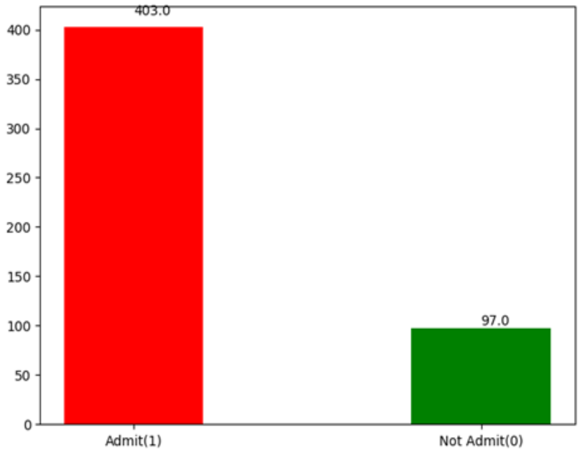


Fig.6. The original data.

5 Method

5.1 SMOTE

SMOTE(Synthetic Minority over-sampling TEchnique) is an algorithm to synthesize new sample data and "oversample" a small number of classes by using existing samples and their nearest neighbors. The idea of Smote algorithm is as follows: set a sampling ratio according to the unbalanced proportion of samples to determine the sampling ratio N. For each small number of samples a, select a number of samples randomly from its k-nearest neighbor, and assume that the selected nearest neighbor is b. For each randomly selected neighbour b, respectively, and the original sample according to the following formula to build a new sample: c=a+rand(0,1)∗|a−b|.

According to the data after SMOTE processing(Fig.7.), the number of negative samples has been greatly improved, and the total number of data sets has increased from 500 to 800.

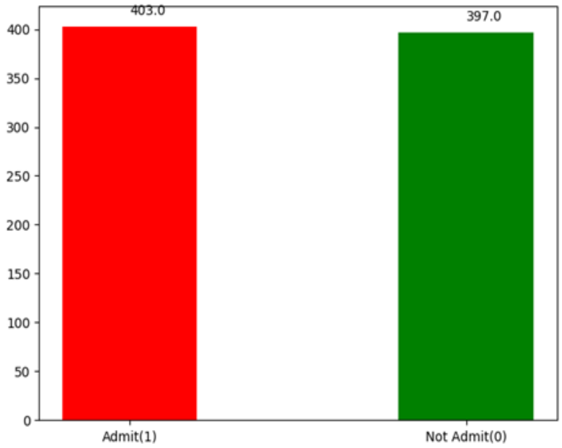


Fig.7. The optimized data.

5.2 Random Forest Algorithm

Random forest is an algorithm that integrates multiple trees through the idea of ensemble learning. Its basic unit is decision tree. Random forest is embodied in two aspects: random selection of data and random selection of features to be selected.

Bagging algorithm is used to conduct K times of put back random sampling of the original training set, so as to obtain K training subsets, each training subset corresponding to a tree.In the process of generating the decision tree, for each node, M features are selected from the feature set as the feature subset each time. When splitting features, the optimal features are selected from the feature subset as the node.All the decision trees generated are combined to form a random forest.

Intuitively, every decision tree is a classifier. Test each decision tree with test set data, then N trees will have N classification results for an input sample. The random forest integrates all the classified voting results and designates the category with the most votes as the final output.

5.3 Classification Performance And Correlation

For dichotomies, we usually calculate ROC curve and AUC according to the confusion matrix as the evaluation index of the model(Table.3).

Table 3. The confusion matrix.

Pred_lable/True_lable	Positive	Negative
Positive	TP	FP
Negative	FN	TN

The horizontal axis of ROC curve is FPR and the vertical axis is TPR. True Postive Rate (TPR) represents the proportion of actual positive instances in positive classes predicted by the classifier to all positive instances. False Postive Rate represents the proportion of actual negative instances to all negative instances of positive classes predicted by the classifier. The calculation formula is as follows:

$$FPR= FP/(FP+TN)$$

$$TPR=TP/(TP+FN)$$

AUC is often used to evaluate binary classification problems. It is defined as the area under the ROC curve enclosed by the coordinate axis, and the value is usually between 0.5 and 1. The reason why AUC value is used as the evaluation standard is that in many cases, the ROC curve cannot clearly indicate which classifier has better classification effect, while as a value, the higher AUC value indicates that the classifier has better classification effect.

In the process of implementing the random forest model, we need to calculate the AUC value of each decision tree, sort the AUC in descending order, select some decision trees with high AUC, and form a new random forest. Since the training samples generated by each decision tree are random and the selection of node features is also random, there is a certain correlation between decision trees. The greater the correlation between any two trees in the forest, the greater the error rate[11].

The correlation is obtained through similarity. The method to calculate the similarity in this experiment is as follows:Each tree is stored as a dictionary structure, and each node has corresponding index and value values to represent features and partition values. The vector inner product between the two nodes of the parent node and the child node is calculated and stored in the list. By comparing the same number in the two lists, the similarity of the two trees can be obtained. The parent-child inner product computation is as follows:

We assume that $the\ parent\ node\!: paren\vec{t} = (index_p, value_p)^T$， $the\ c\hbar ild\ node\!: chil\vec{d} = (index_c, value_c)^T$ then the inner product calculation formula is as follows:

$$Inner\ product = parent^T \bullet child$$

Therefore, by setting a threshold value, if they are within a certain degree of relevance, they are considered to be similar. Delete the tree with low AUC from the two similar decision trees and keep the tree with high AUC. This reduces the correlation between the trees. Finally, the new random forest is composed of the remaining trees.

5.4 Parameter tuning

The depth of the decision tree sometimes affects the model of the random forest. If it is too large, it is easy to overfit, and if it is too small, some hidden information of features will be ignored. At the same time, the size of different size sample data sets, feature subsets and the number of decision trees will all affect the relationship between trees in the forest and affect the classification effect. In this paper, the sample size of the data set is small and the features are few, resulting in greater uncertainty. Therefore, we plan to increase the accuracy of prediction by adjusting depth parameters in real time.

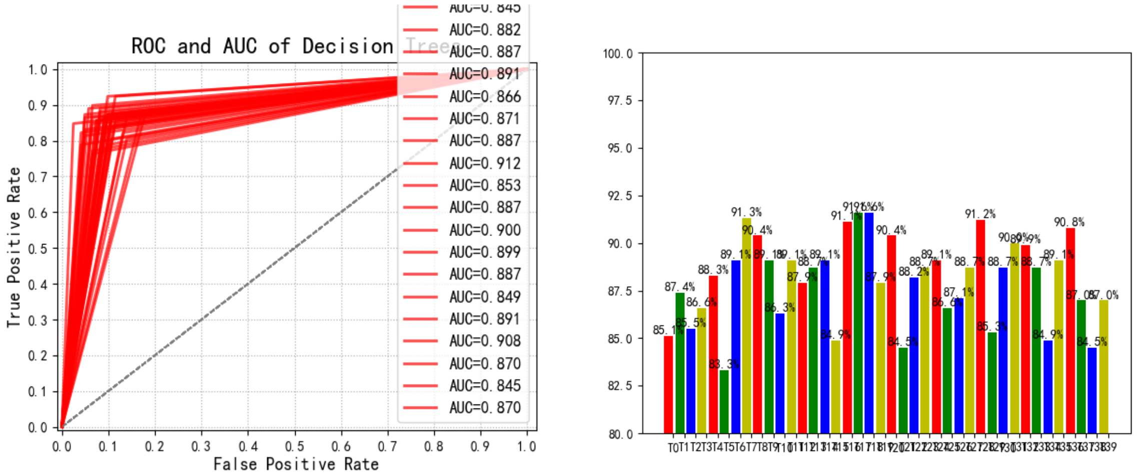
The method we adopted is to select the optimal depth value of the tree as the final depth value before formally generating the random forest model. In other words, the traditional random forest algorithm is used to generate different random forest models for different depth parameters, and the depth values with higher model accuracy and smaller depth are taken as the parameters of the prediction algorithm in this paper. This ensures that the parameters used each time the model is generated are optimal.

6 Experiments

6.1 Experimental Settings

We implemented it in python. The experimental steps are as follows:

Firstly, import the data set and use SMOTE to add a few class samples for it to make the sample balanced. By analyzing and comparing different decision tree depths, the optimal depth value is selected to complete parameter tuning. Then the AUC of each tree is calculated and sorted in descending order, and the optimal 2/3 trees are selected to form the new random forest. After that, the threshold value is set according to the descending order of AUC values. Starting from the high AUC decision tree, the value of similarity is carried out with the later decision tree successively, and the two trees are considered to be similar. The decision tree with low AUC value is deleted, and the decision tree with high AUC is retained to form a new random forest(Fig.8.).



(a). ROC of each decision tree.

(b). AUC distribution for each decision tree.

Fig.8.Description of ROC and AUC

6.2 Experimental Results And Analysis

Accuracy is our most common evaluation index, generally speaking, the higher the accuracy, the better the classifier. The classification index we selected is the accuracy rate.

The calculation formula is as follows:

Accuracy = （TP+TN）/(P+N)

In the process of the experiment, we set the number of different decision trees as 20, 30, 40, 50, 60, 70 and 80 respectively. We first tested the improved random forest, and then tested the improved random forest. By comparing the correct rate, we found that the improved random forest had the highest correct rate of 3.96% improvement. Overall, the improved random forest classification performance is better(Table 4.).

The experimental data are represented by graph, which can show the comparison results more intuitively. For the prediction problem of graduate enrollment, the improved random forest has an improved accuracy rate compared with the previous random forest(Fig.9.).

Table 4. The comparison of classification accuracy.

Number of decision trees	Random forests of the original one（%）	The improved random forests（%）
20	90.60	94.56
30	91.95	93.31
40	92.62	94.56
50	93.29	94.98
60	92.62	94.14
70	92.62	94.14
80	93.29	95.82

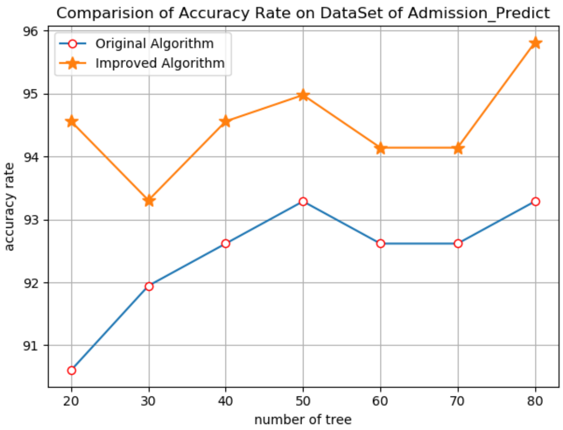


Fig.9.The comparision of Accuracy Rate on DataSet of Admission_Predict

7 Conclusion

In this paper, the data set of UCLA graduate student enrollment was analyzed and predicted. On the basis of improving the random forest, SMOTE processing was conducted on the data and the depth parameters of the random forest were adjusted, so as to obtain the random forest model with higher prediction accuracy. Therefore, we can draw a conclusion that in the process of model training, the accuracy of the model can be improved more effectively through data preprocessing and model parameter optimization. Of course, there are still some improvements to be made in this paper, such as small data set, long model training time and algorithm optimization in future work.

References

[1] SONG Yuan1, ZHU Li-qin2. Research on student performance evaluation based on random forest. Journal of Qiqihar University: Natural Science Edition, 2017, 33(6):1-5.

[2] Awaji, Mansour. Evaluation of Machine Learning Techniques for Early Identification of At-Risk Students [Ph.D. Thesis]. ProQuest Dissertations and Theses, Nova Southeastern University, 2018.

[3] Ming Chen, Xudong Zhao. Fatty Liver Disease Prediction Based on Multi-Layer Random Forest Model. In: CSAI '18, Proceedings. Shenzhen, China, pp.364-368, December 2018.

[4] Sims, Michael S. Predicting Four-Year Graduation: A Sequential Modeling Approach [M.S. Thesis]. ProQuest Dissertations and Theses, California State University, Long Beach, 2018.

[5] Sharma, Nonita, Juneja, Akanksha. Combining of random forest estimates using LSboost for stock market index prediction. In: I2CT 2017, pp.1199-1202, December 2017.

[6] Adetiloye, Taiwo, Awasthi,Anjali. Predicting Short-Term Congested Traffic Flow on Urban Motorway Networks. Canada: Elsevier Inc, 2017.

[7] Xu Jie. Three Essays on Improving Ensemble Models [Ph.D. Thesis]. ProQuest Dissertations and Theses, The University of Alabama, 2015.

[8] CUI Ren-jie. Data Mining Application on Prediction of Students' Major Performance. Computer Engineering & Software, 2016(1):24-27.

[9] Olson, Matthew. Essays on Random Forest Ensembles [Ph.D. Thesis]. ProQuest Dissertations and Theses, The University of Pennsylvania,2018.

[10] Predić, Bratislav, Dimić, Gabrijela. Improving final grade prediction accuracy in blended

learning environment using voting ensembles. Serbia: John Wiley and Sons Inc, 2018.

[11] WANG Ri-sheng, XIE Hong-wei, AN Jian-cheng. Improvement of Random Forests Algorithm Based on Classification Accuracy and Correlation. Science Technology and Engineering, 2017,17(20):67-72.

[12] Yiyi Liu,Hongyu Zhao. Variable importance-weighted Random Forests. Quant. Biol., 2017, 5(4): 338-351.

[13] Angshuman Paul, Dipti Prasad Mukherjee. Reinforced random forest. In: ICVGIP '16,Proceedings.Guwahati,Assam,India, pp.1-8,18-22, December 2016.

[14] Jingjing Tian, Ping'An Li. An Intrusion Detection Algorithm of Dynamic Recursive Deep Belief Networks. In: ICIT 2017, Proceedings.Singapore, Singapore, pp.180-183,27-29, December 2017.

[15] Qingping Huang, Yujiao Li, Song Liu. Short term load forecasting based on wavelet decomposition and random forest.In:SmartIoT '17, Proceedings.San Jose, California, pp.1-6,14-14,October 2017.

[16] Ameni Bouaziz, Célia da Costa Pereira. Interactive generic learning method (IGLM): a new approach to interactive short text classification. In: SAC '16, Proceedings.Pisa, Italy, pp.847-852,04-08, April 2016.

[17] Ma Xiaojuan. Research on the Classification of High Dimensional Imbalanced Data based on the Optimization of Random Forest Algorithm. In: BDET 2018, Proceedings. Chengdu, China, pp.60-67, 25 - 27, August 2018.

[18] Xia, J., Zhang, S., Cai, G. Adjusted weight voting algorithm for random forests in handling missing values(2017). Pattern Recognition, 69, pp. 52-60.

[19] B. Ravi Kiran,Jean Serra. Cost-Complexity Pruning of Random Forests. In: ISMM 2017: Mathematical Morphology and Its Applications to Signal and Image Processing, pp.222-232. 12 April 2017.

[20] ZHANG Qizeng, DAI Hanbo. Student performance prediction model based on data preprocessing technology. Journal of Hubei University:Natural Science Edition, 2019,41(1):101-108.

[21] CAO Xinyu, CAO Weiquan, LI Zheng. Prediction method of college students′ scores toward uncertain missing data. Modern Electronic Technique, 2018,41(6):145-149.

[22] Mohan S Acharya, Asfia Armaan, Aneeta S Antony. A Comparison of Regression Models for Prediction of Graduate Admissions. In: IEEE International Conference on Computational Intelligence in Data Science 2019.