**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.** By answering these questions, we not only unveil useful patterns and strategies for literature recommendation, but also identify some challenging problems for future development.

随着大数据技术的不断发展成熟，覆盖领域越来越广泛，教育行业对数据挖掘与分析的需求也在不断扩大。本文从学生申请赴美读研的实际需求出发，通过数据挖掘技术，将学生的各类成绩资料作为特征向量，进行大学研究生的录取率预测，使得学生能够对自身获得某所大学承认的机会有一个公平的认识，从而能为他们的升学选择提供决策支持。近些年来，为了能够更好地预测学生成绩，研究者们提出了不少预测模型，如贝叶斯分类模型、k-Means聚类、随机森林、神经网络等。在本文中我们也将通过SMOTE数据处理技术，以及在优化后的随机森林算法进行预测研究。通过传统随机森林和优化后的随机森林进行对比，证明优化后的随机森林可以提高研究生录取率的预测精度。

随着生活水平的不断进步，人们的知识水平也有所提高，研究生的数量也在不断增长。本文的研究方向是从学生申请赴美读研的实际需求出发，通过数据挖掘技术，将学生的各类成绩资料作为特征向量，进行大学研究生的录取率预测，使得学生能够对自身获得某所大学承认的机会有一个公平的认识，从而能为他们的升学选择提供决策支持。近些年来，为了能够更好地预测学生成绩，已有不少研究者们提出了不少预测模型，如贝叶斯分类模型、k-Means聚类、随机森林、神经网络等。而我们也将采用随机森林算法的基础上进行预测研究，并做出适当改进。针对于如何提高模型预测的精确度，我们将从以下两个方面进行研究优化和提升：

（1）数据集的样本分布不均衡，数量较少，如何改进？

（2）基于数据样本的复杂性，又该如何提高随机森林的分类精确度？

关于数据集，我们选取了一组来自于Kaggle的数据集中针对印度研究生入学的数据。这个数据集给出了学生的成绩、大学等级、推荐信重要度等信息，我们将最终的入学机会从高到低将学生群体分成等级A、等级B、等级C、等级D、等级F等不同等级类别，这也是一种获得入学机会的表示方法。但由于数据集存在样本分布不均衡的情况，所以我们将对数据进行简单清洗的基础上，采用少数类样本合成(ＳＭＯＴＥ)过采样化技术平衡数据集，即对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中，以提升预测能力。

关于随机森林的优化，虽然许多学者对随机森林算法进行了广泛的研究，并且取得了许多显著的研究成果，现有的对随机森林算法的改进主要集中在决策树的算法、投票方法、数据集预处理、特征选取等的优化或改进。但是在使用决策森林时，会生成分类性能较差的决策树，从而影响分类的精确度。因此，我们将随机森林生成的每一棵决策树，通过AUC值比较，选取AUC值最高的决策树进行相似度比较，把最终性能最高，相似度最低的决策树作为新的随机森林预测模型。

通过传统随机森林和优化后的随机森林进行对比，我们期望经过优化后的随机森林可以提高研究生录取率的预测精度。

摘要：

随着人们的知识水平的提高，研究生的数量也在不断增长。本文的研究方向是从学生申请赴美读研的实际需求出发，通过数据挖掘技术，将学生的各类成绩资料作为特征向量，进行大学研究生的录取率预测，使得学生能够对自身获得某所大学承认的机会有一个公平的认识，从而能为他们的升学选择提供决策支持。近些年来，为了能够更好地预测学生成绩，已有不少研究者们提出了不少预测模型，如贝叶斯分类模型、随机森林、神经网络等。而我们也将采用随机森林算法的基础上进行预测研究，并做出适当改进。针对于如何提高模型预测的精确度，我们将从以下两个方面进行研究优化和提升：

（1）数据集的样本分布不均衡，数量较少，如何改进？

（2）基于数据样本的复杂性，又该如何提高随机森林的分类精确度？

通过传统随机森林和优化后的随机森林进行对比，证明优化后的随机森林可以提高研究生录取率的预测精度。

翻译：

With the continuous improvement of living standards, people's knowledge level has also improved, and the number of graduate students is also growing. The research direction of this paper is to forecast the admission rate of graduate students based on the actual needs of students applying for postgraduate study in the United States. Through data mining technology, students'performance data are taken as eigenvectors, so that students can have a fair understanding of their own opportunities to be recognized by a university, thus providing decision support for their choice of admission. 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. On the basis of Stochastic Forest algorithm, we will make prediction research and make appropriate improvements. In order to improve the accuracy of model prediction, we will study and optimize from the following two aspects: (1) the sample distribution of data sets is not balanced, the number is small, how to improve? (2) Based on the complexity of data samples, how to improve the classification accuracy of random forests? 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:** Academic access data; Literature recommendation; User Study

**1 Introduction**

迄今为止，学术上基于学生成绩预测的研究课题有很多，如通过预测的学生成绩来提高教学质量[1], 提高学生保留率[2]，提升学生毕业率[3]等等，但据我们调查所知，目前对于预测海外研究生入学率方面的研究却很少。而且随着人们视野的不断扩大，出国留学的人数不断增多，通过递交的留学申请材料，预测被大学录取的几率，这对申请人而言有着非常重要的意义。

一直以来，关于预测分类回归等问题的算法模型有很多，随机森林就是其中之一。在本研究中，我们发现在使用决策森林时，会生成分类性能较差的决策树，对最后的投票结果以及模型的预测性能都带来不好的影响。因此，我们将使用一种改进的随机森林方法。选取出随机森林模型中分类性能好的决策树进行相似度计算，根据相似度不同的决策树组成新的随机森林模型。在生成新的随机森林模型之前，我们通过比较不同深度的决策树所组成的随机森林模型的准确率，把准确率最高的对应深度作为我们改进随机森林算法的深度参数。除此之外，我们还采用SMOTE技术，对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中，最终实现随机森林优化改进，并且提升研究生录取的预测能力。

  论文组织如下。在下一节中，我们将回顾相关的工作。第3节介绍了我们的数据集。我们在第4-7节报告了研究的内容和结果。最后，我们讨论了结论和未来的工作。

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 [2], and improving the student graduation rate [3], etc. 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**

随着机器学习和数据挖掘技术的热潮兴起，预测模型算法被广泛应用到各个领域，特别是随机森林算法这类集成学习技术，即可以做回归问题预测，也能实现分类问题的预测。如疾病预测【5】、股票预测【8】、车流量预测【9】，也可以预测学生成绩【10】等等。同时，根据我们的调研发现，与基本的预测模型相比，随机森林在预测类问题中往往有更高性能和更准确的分类精确度，并且可以处理高维度的数据[1,4,5]。

集成学习中有许多优秀的算法，我们通过调研将随机森林与其他的算法进行简单的对比。在最优变量选取上，Bagging通过遍历所有的预测变量，选取最佳的拆分值。而随机森林通过随机性降低树与树之间的相关性来选取最佳分割变量[4]。相比之下，随机森林的分类强度要高于bagging，能更有效地降低方差。另一方面，在分类精确度上，AdaBoost一般拥有较大的优势，能完美地拟合训练数据[6]，但模型的训练时间长，并且容易导致“过度拟合”现象。虽然XGBoost作为升级版的GBDT，还支持线性分类器，能够更高效地生成候选的分割点 [3]。但是XGBoost算法在验证过程在计算上较为复杂。通过以上几种算法的比较，我们小组决定把随机森林作为预测研究生入学率的主要模型。

随机森林相对其他的算法，有着很大的优势，但仍然存在许多需要改进的地方。一方面，在算法优化上，B. Ravi Kiran【7】通过结合OOB未采样样本改善决策树的泛化误差，但这个想法可能与随机森林提供的平滑加标平均决策函数不相容。另一方面，在模型优化上，Yiyi Liu[8]提出变量重要性加权随机森林，在信息量特征提取的基础上使用加权抽样策略，能提高在弱信号和大噪声情况下的预测精度，但同时存在连续变量和分类变量或者分类变量的水平数不同时，变量重要性评分估计存在不准确问题。

虽然有很多基于随机森林的改进【1，2，3，5，6】[7]，但是很少有人注意到使用随机森林算法对于不平衡样本的问题研究。数据分类不平衡是数据挖掘研究中经常遇到的问题之一。因此，针对不平衡样本，并且真正有效的提升随机森林预测率是很有必要的。

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 **[5],** stock forecast **[8]**, traffic forecast **[9]**, students' scores can also be predicted **[10]**, 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, 4, 5].

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 [4]. 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 [6], 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 [3]. 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 **[7]** 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 [8] 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 **[1,2,5,3,6]** [7], 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**

在本文研究中，我们收集了一组来自于Kaggle的数据集中加州大学洛杉矶分校研究生招生的500条数据。该数据集是由Mohan S Acharya提供，建立目的是帮助学生将他们的个人资料列入大学的候选名单，预测的输出使他们对他们获得某所大学的机会有了一个公平的认识。本文研究中使用的所有特征及其值的描述如表1所示。

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.

|  |  |  |
| --- | --- | --- |
| 特征 | 描述 | 值 |
| GRE Score | 研究生入学考试成绩 | 0~340 |
| TOEFL Score | TOEFL考试成绩 | 0~120 |
| University Rating | 学生本科学校的等级 | 0~5 |
| SOP | 学生目的陈述 | 0~5 |
| LOR | 学生的Letter of Recommendation Strength | 0~5 |
| CGPA | 本科GPA成绩 | 0~10 |
| Research | 是否拥有研究经历 | 0/1 |
| Chance of Admit | the value of Chance of Admit | 0~1 |

在这8个特征属性中，我们把Chance of Admit属性作为预测标签。同时，为了能提高预测的准确性，我们把Chance of Admit的值做了二分类处理，如表2所示。

Among these 8 characteristic attributes, we take the Chance of Admit attribute as the prediction tag. Meanwhile, in order to improve the accuracy of the prediction, we did a binary processing of the value of Chance of Admit as shown in table 2.

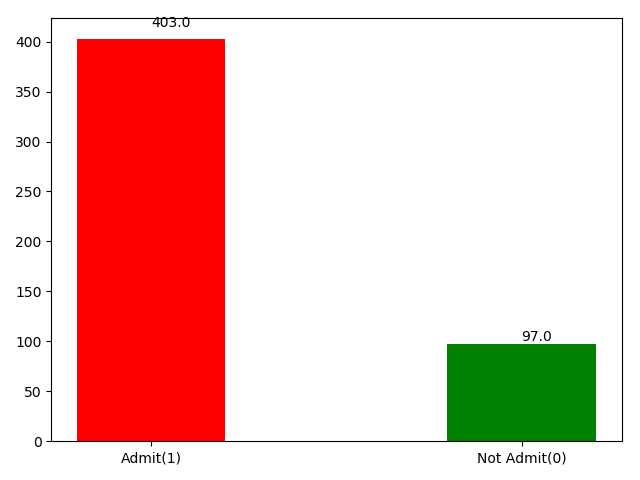
Table 2．The level of the Chance of Admit(value).

|  |  |
| --- | --- |
| 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（初步研究）**

我们对原生数据的Chance of Admit标签进行了划分，把低于0.6的特征标记为0，把高于0.6的特征标记为1。但我们发现分类后的数据呈现不平衡的现象，500条数据集中，正样本的数量有403条，而负样本的数量只有97条，正负样本的数量差距过大(如图1所示)。对于样本数量不平衡的数据进行模型训练时，会对少数类样本的预测会存在偏差，从而导致模型的准确度下降。因此，接下来我们将通过SMOTE技术来处理少数类样本不平衡问题。

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.1.). 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.1. The original data.

我们对数据中的特征做了简单的可视化分析。对“GRE Scores’特征进行数据统计时（Fig.1.），我们发现多数学生的GRE Scores主要集中在310到325的分数段，高分段和低分段的学生相对较少。对“University Rating”特征进行数据统计时（Fig.2.），我们发现学生本科学校的所属等级主要是3等以上的学校。对“CGPA”特征进行数据统计时（Fig.3.），我们发现只有少数学生的CGPA是在8.0以下，高分段的学生较多。基于这三个特征属性的统计，我们把“GRE Scores’与“CGPA”进行了比较分析（Fig.4.），通过图表我们可以发现，people with higher CGPA usually have higher GRE scores maybe because they have strong learning ability.同时，我们也把“University Rating”、“CGPA”和"Research"特征进行了对比（Fig.5.），we find that people with higher University Rating usually have higher CGPA。University Rating相同的情况下，有研究经历的学生的CGPA往往比没有研究经历的学生的CGPA高。

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.

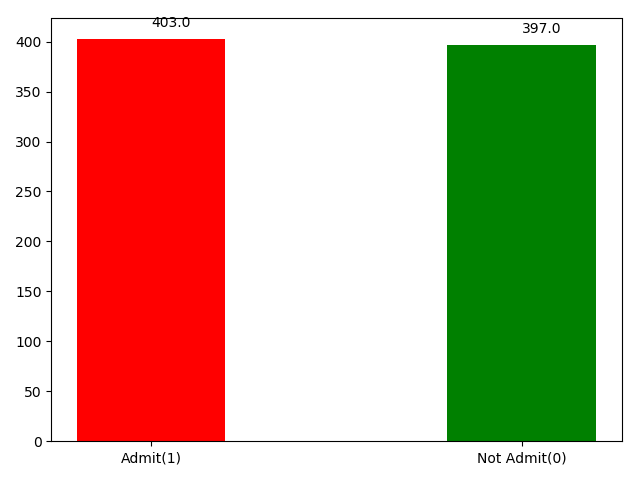
**5 Method**

SMOTE

SMOTE(Synthetic Minority Over-sampling TEchnique)是一种利用已有样本以及其近邻，合成新样本数据对少数类进行“过采样”的算法。Smote算法的思想如下：根据样本不平衡比例设置一个采样比例以确定采样倍率N，对于每一个少数类样本a，从其k近邻中随机选择若干个样本，假设选择的近邻为b。对于每一个随机选出的近邻b，分别与原样本a按照如下的公式构建新的样本:c=a+rand(0,1)∗|a−b|。

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|。

根据SMOTE处理后的数据显示，负样本的数量得到很大的提高，数据集总数量由原来的500条增加到800条。

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

1. 随机森林算法

随机森林就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树。随机森林的随机体现在两个方面：数据的随机选取以及待选特征的随机选取。

使用Bagging算法对原始训练集进行K次有放回的随机抽样，由此得到K个训练子集，每个训练子集对应一棵树。

在生成决策树的过程中，对于每个节点，每次都从特征集合中选取M个作为特征子集，分裂特征时都从特征子集中选取最优特征作为该节点。

将生成的所有决策树组合在一起，就形成了随机森林。

从直观角度来解释，每棵决策树都是一个分类器。采用测试集数据对每棵决策树进行测试，那么对于一个输入样本，N棵树会有N个分类结果。而随机森林集成了所有的分类投票结果，将投票次数最多的类别指定为最终的输出。

（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.

（3）分类性能和相关性

对于二分类的问题，我们通常根据混淆矩阵，计算ROC曲线和AUC来作为模型的评价指标。

AUC常被用作评价二类分类问题。它被定义为ROC曲线下与坐标轴围成的面积，值常常介于0.5到1之间。使用AUC值作为评价标准是因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，当AUC值越大时，说明该分类器分类效果更好。

样本为正负两类，则有以下术语：

True positives(TP): 被正确地划分为正例的个数，即实际为正例且被分类器划分为正例的实例数；

False positives(FP): 被错误地划分为正例的个数，即实际为负例但被分类器划分为正例的实例数；

False negatives(FN):被错误地划分为负例的个数，即实际为正例但被分类器划分为负例的实例数；

True negatives(TN): 被正确地划分为负例的个数，即实际为负例且被分类器划分为负例的实例数。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实  际  类  别 | 预测类别 | | | |
|  | 1 | 0 | 总计 |
| 1 | TP | FN | TP+FN |
| 0 | FP | TN | FP+TN |
| 总计 | TP+FP | FN+TN | TP+FN+FP+TN |

ROC曲线的横轴为FPR,纵轴为TPR。真正类率(True Postive Rate)TPR代表分类器预测的正类中实际正实例占所有正实例的比例。负正类率(False Postive Rate)代表分类器预测的正类中实际负实例占所有负实例的比例。计算公式如下：

FPR= FP/(FP+TN)

TPR=TP/(TP+FN)

在实现随机森林模型的过程中，我们需要计算出每棵决策树的AUC值后，对AUC进行降序排序，选择AUC高的一些决策树，组成新的随机森林。由于每次生成决策树的训练样本随机，节点特征选择也随机，因此决策树之间具有一定的相关性。森林中任意两棵树的相关性越大，错误率越大。

而相关性是通过相似度得到的，在这次实验中计算相似度的方法是：

每棵树存储为字典结构，每个节点有对应的index和value值表示特征和划分值，计算父节点与子节点两个节点间的向量内积，存储在list中，比较两个list里相同的数量。就能得到两棵树的相似度。

其中父子节点内积计算如下：

假设，则内积计算公式为：

因此，通过设定阈值，若在一定相关度内，则认为他们相似，删去相似的两棵决策树中AUC低的树，保留高AUC的树。这样就减少了树之间的相关性。最后，由剩下的树组成新的随机森林。

1. Classification Performance And Correlation

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.

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 then the inner product calculation formula is as follows:

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.

（4）参数调优

决策树的深度有时会影响随机森林的模型，过大容易过拟合，过小会忽略一些特征隐含信息。同时对于不同大小的样本数据集、特征子集的大小、决策树的个数都会影响森林中树与树之间的关系，影响分类效果。在本文中，数据集样本数不多，特征少，所面临的不确定性会较大。因此，我们拟通过实时的调优深度参数来增加预测的正确率。

我们采用的方法是：在正式的生成随机森林模型之前，选择树的深度最优的值作为最终的深度值。即针对不同的深度参数，使用传统随机森林算法来生成不同的随机森林模型，取模型正确率较高而深度较小的深度值作为本文预测算法的参数。这样可以保证，每次生成模型时使用的参数是最优的

(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（实验设置）**

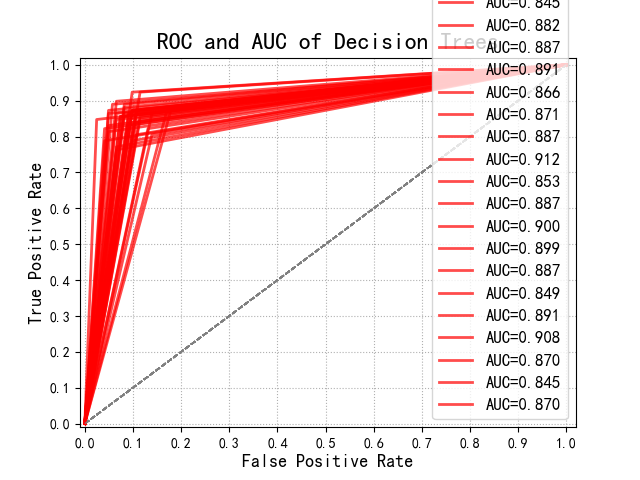
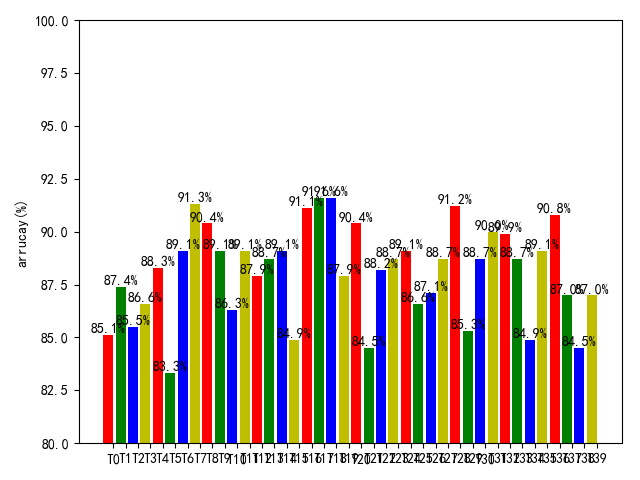
我们采用python语言实现。实验步骤如下：

首先导入数据集，对其采用SMOTE增加少数类样本，使样本均衡。通过分析比较不同的决策树深度，选择最优的深度值，完成参数调优。然后计算每棵树的AUC，进行降序排列，选取最优的2/3棵树组成新的随机森林。之后根据AUC值降序排列，设定阈值，从高的AUC决策树开始，依次与后面的决策树进行相似度的值，则认为两棵树相似，删去AUC值低的决策树，保留AUC高的决策树，形成新的随机森林。

(a). ROC of each decision tree.

(b). AUC distribution for each decision tree.

Fig.8.Description of ROC and AUC



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

6.2 实验结果及分析

正确率是我们最常见的评价指标，通常来说，正确率越高，分类器越好。我们选用的分类指标为正确率。

计算公式如下：

accuracy = （TP+TN）/(P+N)

在实验过程中，我们设置了不同的决策树棵数，分别为20，30，40，50，60，70，80，首先对改进前的随机森林进行测试，然后对改进后的随机森林进行测试，通过对正确率的比较，发现改进后的随机森林正确率最高提升了3.96%。综合来看，改进的随机森林分类性能更好。

将实验数据用图表示出来，能更加直观的展示对比结果。针对研究生入学预测问题，改进的随机森林相较之前的随机森林在正确率上有所提升。

表：模型正确率比较（Comparison of Classification Accuracy）

|  |  |  |
| --- | --- | --- |
| 决策树数目 | 改进前的随机森林（%） | 改进后的随机森林（%） |
| 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 |

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

**7 Conclusion**

本文对加州大学洛杉矶分校研究生录取的数据集进行了分析和预测，我们在改进随机森林的基础上，对数据进行了SMOTE处理以及对随机森林的深度参数进行调优，从而得到预测准确率更高的随机森林模型。因此我们可以得出结论，在进行模型训练时，通过数据预处理以及模型参数的优化能更加有效的提升模型的准确率。当然，本文还存在一些需要改进的地方，例如数据集较小，模型训练时间长，未来的工作还需要实现算法的优化。

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] Sayah J Y, Kime C R. Test scheduling in high performance VLSI system implementations. *IEEE Trans. Computers*, 1992, 41(1): 52-67. [journal paper]

[2] Geddes K O, Czapor S R, Labahn G. Algorithms for Computer Algebra. Boston: Kluwer, 1992. [book]

[4] Kwan A W, Bic L. Distributed memory computers. In *Proc. the 6th Int. Symp. Parallel Processing*, March 1992, pp.10-17. [conference paper].

[5] Harris M J. Real-time cloud simulation and rendering [Ph.D. Thesis]. Department of Computer Science, The University of North Carolina at Chapel Hill, 2003. [thesis]

[7] Gipp B, Meuschke N, Gernandt A. Decentralized trusted timestamping using the crypto currency Bitcoin. arXiv:1502.04015, 2015. https://arxiv.org/abs/1502.04015, May. 2018. [ar-Xiv document]

随机森林的应用：对生物研究中的各种预测问题具有强大的功能[变量重要]

随机森林的改进：

1性能改进：当特性的数量增加时，树与树之间的相关性增大，导致了模型性能的下降。于是作者提出了一种软“特征选择”策略——变量重要性加权随机森林，通过根据特征的信息量抽样特征，加权抽样策略，使得模型更加依赖于信息特征，从而提高了在弱信号和大噪声情况下的预测精度，边缘检验的思想 ，虽然变量重要性给出了变量的重要性排序，但仅通过变量的影响度进行采样，仍可能漏掉一些具有分类能力的变量问题，特别是对于变量数目巨大但具有分类能力的变量比例很低的数据的变量筛选情况。而且但作者没有考虑到当同时存在连续变量和分类变量或者分类变量的水平数不同时，变量重要性评分估计存在不准确问题。[变量重要]

2分类预测精度：基于二次规划的方法；添加adabs类，利用加权重采样提高分类器强度[THREE ESSAYS ON] 从分类性能和相关性的角度对随机森林算法进行优化，虽然提升了随机森林算法的分类精确度，但在树的生成过程中速度要比传统的随机森林慢。[基于分类精度]

数据类优化：

Ｎｉｔｅｓｈ Ｖ. Ｃｈａｗｌａ [３] 等人提出了一种合成少数类样本过采样技术(ＳＭＯＴＥ)ꎬ他

们利用少数类样本合成新的样本并加入数据集中，能提高分类器性能。但smote

随机森林与其他算法相比的优越性：

1Bagging和Random Forest使用相同的权值来组合完全生长的树，但Bagging通过遍历所有的预测变量，选取最佳的拆分值，而随机森林通过随机性降低树与树之间的相关性来选取最佳分割变量，相比之下，随机森林的分类强度要高于bagging。[THREE ESSAYS ON]。

崔仁桀利 用 C 4.5算法构造学生专业成绩预测模型，但当变量数目增加时，C4.5算法的性能会大大降低[2]  崔仁桀 . 数据挖掘在学生专业成绩预测上的应用 J]. 软件， 2016(1):24-2。

陈勇将遗传神经网络应用于大学生成绩分析 [3] ，实现了精确的分值预测，并引入遗传算法来解决 BP神经网络收敛速度慢、训练时间长的问题。然而该方法的实验并不完备，仅在 16条成绩数据上开展神经网络训练与预测，实验结果不具统计可信性，也没有足够丰富的实证分析来佐证其模型方法的推广能力。CHEN Yong. Research and implementation of result prediction based on genetic neural network [J]. Modern electronics technique，2016，39（5）：96⁃100.

确定了预测响应的重要特征。

•抗过度拟合

•允许我们按顺序逐项建模。。使用XGBoost作为我们的模型的一个潜在缺点是，验证过程在计算上相当昂贵。随着决策树和CART的出现，出现了两个主要问题。

首先，当自变量数量较大时，需要非常复杂的树来解释数据中所有的变异性，这使得树结构中节点的解释变得复杂，而这些树具有很高的偏差和很低的方差。XGBoost与RF非常相似，因为它使用一组相对较浅的树(复杂度较低)来优化偏差和方差之间的内在平衡，最终更好地理解数据的底层分布。[PREDICTING FOUR].所以我们小组选取随机森林作为此次研究的主要模型。2Breiman (2001b)认为CART和类似的技术在解释方面表现良好，但在预测方面表现不佳。他给CART的可解释性打了a +分，但就其对未来数据的预测性能而言，只打了B分。[THREE ESSAYS ON]

分类预测模型方法：

1利用集成技术是提高早期预测模型性能的有效途径。集成分类器结合多种分类算法的输出对新实例进行分类。来自不同学科的研究人员在广泛的领域探索和应用集成方法。研究发现，与基本分类器相比，集成技术能够不断地获得更好的预测性能

课题方面：查阅了许多论文，发现在学术研究上主要是关于为了提高学生保留率进行成绩预测预警的研究方向[Evaluation of Machine Learning Techniques for Early Identification of At-Risk Students]，研究重点领域是对学生毕业率的深入分析 [PREDICTING FOUR-YEAR GRADUATION:] 释诸如学生保留率和毕业率。随着机器学习的开始，我们看到了奥尔克等人(2016)的研究，预测学生辍学是主要的兴趣[PREDICTING FOUR].。 而对于为了提高大学的研究生入学率、学生申请研究生就读的录取率方面的研究很少。

随机森林算法是机器学习领域中的一种集成学习方法。将其与多个决策树的分类结果相结合，形成一个全局分类器。随机森林算法与其它分类算法相比有很多优势,分类效果优势反映在分类精度和泛化误差小,处理高维数据的能力,训练过程的优势快速和简单的并行学习算法。由于数据分类的不平衡特性，它在现实世界中广泛存在[Research on the]

即通过随机采样[7]来增加正类的数量。该方法的缺点是容易造成分类算法过拟合[Research on the]

迄今为止，学术上基于学生成绩预测的研究课题有很多，如通过预测的学生成绩来提高教学质量[1], 提高学生保留率[2]，提升学生毕业率[3]等等，但据我们调查所知，目前对于预测海外研究生入学率方面的研究却很少。而且随着人们视野的不断扩大，出国留学的人数不断增多，通过递交的留学申请材料，预测被大学录取的几率，这对申请人而言有着非常重要的意义。