Research and Analysis on the Prediction of College Enrollment based on Random Forest

Lei Yang [a,*], Liwei Tian [b], Liang Yu [c] and Yungui Chen [d]

Guangdong University of Science and Technology, Dongguan 523083, China

[a,*]gdstyl@qq.com, [b]656453927@qq.com, [c]840493518@qq.com, [d]370064596@qq.com

Abstract. The registration rate of new students has always been a concern of all colleges, and it is a difficult problem to accurately predict the number of new students before they are registered. At present, there are no researchers using machine learning method to predict the registration of new students, because the intuitive feeling is that whether students register or not, which is a very subjective thing, affected by many subjective factors. At present, the traditional methods are used to predict the number of new students in Colleges, that is, telephone inquiry and tuition payment status inquiry. According to the historical enrollment data of a university, this research uses the method of random forest to study it. The results show that whether the freshmen register or not can be predicted, and the enrollment data of universities over the years is valuable.

Keywords: Prediction; enrollment; random forest.

1. Introduction

For universities, judging whether a student is registered or not usually involves two traditional methods. The first is to call the inquiry, and the second is to judge according to the student's payment. Unfortunately, many students do not receive calls or are unwilling, to tell the truth (possibly enrolled in multiple schools), and many students do not pay tuition in advance. Therefore, the usual practice of universities is to make a rough estimate based on the registration rate of previous years and estimate a total number of people, and this vague estimation is a very big risk to the decision-making of universities.

So far, no researchers have used machine learning to study this data. We collected data sets for freshmen admission and registration from a university in Guangzhou. This data set is preprocessed to make it recognizable by the machine. In this data set, the data for the first 3 years is used for the training set and the data for the 4th year as the test set. The random forest algorithm is used to machine learning this data set. The results show that freshman registration prediction is feasible.

The random forest is a classifier that uses multiple decision trees to train and predict samples[1]. This algorithm is closely related to the size of the forest (ie the number of decision trees) and has the characteristics of random data[2]. Therefore, we optimize and improve the random forest algorithm from two aspects. On the one hand, more trees are used to build a larger forest[3]. On the other hand, the number of calculations is increased and use the method of calculating the mean multiple times as the final learning result[4]. Specifically, 10,100,1000,10000 trees are used to construct a random forest, and each number of forests calculated 10 times and finally took the mean as its final result. The results are exciting and the predicted accuracy can reach more than 60%.

In order to evaluate this work more accurately, the evaluation criteria of F-Measure is introduced[5]. The results show that the F1 values are close to 0.66, which proves that this work is effective.

2. The Dataset

2.1 The Data Source &Size

The data used in this study came from a university in Guangzhou, with a total of 4 years of admission data and registration data from 2009 to 2012. These data have been authorized by universities to be used only for research purposes.

The entire dataset contains 4 years of data. The total amount of data is 10,382, of which 5,989 are registered and 4,393 are not registered. The data from the first 3 years is used as the training set and the data from the 4th year as the test set. The total number of samples in the training set is 6,599, and the total number of samples in the test set is 3,783.

2.2 Data Preprocessing

Before machine learning data, the data must be preprocessed. This is because most of the raw data collected is missing, noisy, repetitive, ambiguous or incomplete[6]. These data cannot be directly used by the program, so the raw data needs to be preprocessed. It becomes data that can be used by the program.

At the same time, many data types are not directly involved in the operation, such as name, email address, secondary name, and other text-type data, so many data types need to be converted so that the program can perform operations[7].

A typical conversion is to modify the data of the fixed telephone to 0 and 1, that is, 0 means that there is no fixed telephone in the student's home, and 1 means there is[8].

The same data is processed for the data in the mobile phone column. In addition, we have processed the text type numbers and changed them to numeric numbers.

After the above processing, the dataset is reduced from the original 38 to 17 columns, the null value is padded to 0, and the processed dataset is shown in Table 1.

Table 1. Processed dataset

	y	a_1	a_2	a_3	...	a_{14}	a_{15}	a_{16}
	Reg	Gender	Score	Course	...	Senior	Nation	Poli
x_1	1	1	465	0	...	1	1	3
x_2	1	1	463	0	...	1	1	13
x_3	1	2	462	0	...	1	1	13
...	...	...	...	...	...	...	...	...

The symbol y means whether registered or not, 1 for registration, 0 for no registration, and the other symbols are defined as follows:

Attribute set A = $\{a_1, a_2, a_3, \ldots a_{15}, a_{16}\}$

Sample $x_i = \{a_1^i, a_2^i, a_3^i, \ldots a_{15}^i, a_{16}^i\}$

Samples set $X = \{x_1, x_2, x_3, \ldots x_{10381}, x_{10382}\}$

Training set $D = \{(x_1,y_1),(x_2,y_2),\ldots,(x_{6599},y_{6599})\}$

e.g. (x1,y1) = (1 465 0 0 0 7 1421 1 1402011 1 1 514021 4 1 1 3, 1)

3. Proposed Method

A random forest is a classifier that uses multiple decision trees to train and predict samples[9]. In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, i.e. have a low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance[10]. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

Because the data used by the random forest algorithm is not a complete training set, it has randomness, so the result of machine learning also has randomness[11]. In order to stabilize the random as much as possible, we calculated it 10 times and then took the mean value as its final result.

In order to further improve the accuracy of our prediction, we continue to increase the number of trees, using 10, 100, 1000 and 10000 trees to construct random forest, the results show that more trees, more better[12]. The above procedure is detailed in Algorithm 1.

Algorithm 1 Random forest Generate

INPUT: Training set *D*;Test set *T*;Number of trees *N* ;

OUTPUT: Mean value of prediction results

Procedure: function RF(*D*, *N*)

```
for K=10 do
    for i=N do
        Randomly generate this training set D';
        Generating a fully growing decision tree from training set D';
        Predict test set T;
    end for
    Vote on N predictions, with the minority subordinate to the majority;
    Store the prediction results;
end for
```

4. Performance Metric

For the binary classification problem, the sample can be divided into four cases: true positive, false positive, true negative, and false negative according to the combination of its real category and the classifier prediction category. The definitions TP, FP, TN, and FN respectively represent the corresponding samples[13]. The "confusion matrix" of the classification result is shown in Table 2.

Table 2. Confusion matrix of the classification

Actual	PredictedPositive	PredictedNegative
Positive	TP(True Positive)	FN(False Negative)
Negative	FP(False Positive)	TN(True Negative)

According to the above confusion matrix, accuracy, precision and recall can be defined[14]. Accuracy is the correct proportion of all predictions and is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} . \tag{1}$$

Precision is correctly predicted as the proportion of Positive which is all Positive and is defined as

$$Precision = \frac{TP}{TP + FP} . \tag{2}$$

Recall is correctly predicted as the proportion of Positive, which is all practically positive, is defined as

$$Recall = \frac{TP}{TP + FN} . \tag{3}$$

Another more common metric is F-Measure, which is the Precision and Recall weighted averaging, is defined as

$$\frac{1}{F_\beta} = \frac{1}{1+\beta^2} \cdot \left(\frac{1}{P} + \frac{\beta^2}{R} \right). \tag{4}$$

The simplified formula is

$$F_\beta = \frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R}. \tag{5}$$

In the formula, $\beta > 0$, It measures the relative importance of recall to precision. When $\beta = 1$, it is the standard F1 score. It is considered that recall and precision are equally important. When $\beta > 1$ means more emphasis on recall, and $\beta < 1$ is more emphasis on precision. In our work, we have a value of 1 for β. Substituting equations (2) and (3) into equation (5).

$$F_1 = \frac{2P*R}{P+R} = \frac{2TP}{2TP+FP+FN}. \tag{6}$$

The F1 score combines the results of Precision and Recall[15]. When the F1 value is high, the classification model is ideal. Obviously, the F1 score ranges from 0 to 1.

5. Performance Evaluation

5.1 Dataset Partition

In this section, MATLAB tools are used to learn about the data. The data of the first 3 years is regarded as the training set, and the data of the 4th year is regarded as the test set. First, the algorithm based on random forest based algorithm 1 is used, then the performance indicators of the previous chapter are used to compare and analyze the learning results. After partition, the data set is shown in Table 3.

Table 3. The dataset

Dataset partition	Total sample size	Not registered	Registered
Complete dataset	10382	4393	5989
Training set (first 3 years)	6599	2889	3710
Test set (4th year)	3783	1504	2279

5.2 Random Forest

In this paper, the research platform is 2018a MATLAB, using the RF_MexStandalone-v0.02 package for classification study[16].

Based on the random forest algorithm which is shown in Algorithm 1, 10, 100, 1000 and 10000 trees are used to construct different random forests. We calculated 10 times for each number of forests and finally took the average value as its final result. The experimental results are shown in Table 4.

Table 4. Performance metric of random forest with different number of trees

Number of trees	Recall	Precision	Accuracy	F1
10	64.537%	65.955%	58.565%	0.65234
100	64.520%	67.549%	59.952%	0.65999
1000	64.866%	67.915%	60.373%	0.66355
10000	65.213%	68.047%	60.596%	0.66600

Compared with a single decision tree, the learning effect of the random forest is much better, and each metric is greatly improved[17,18]. At the same time, it can be seen that the more trees, the better learning effect of random forest, as shown in Fig. 1. Because of the random characteristics, the accuracy of the random forest is about 60%, and the F1 score is about 0.66.

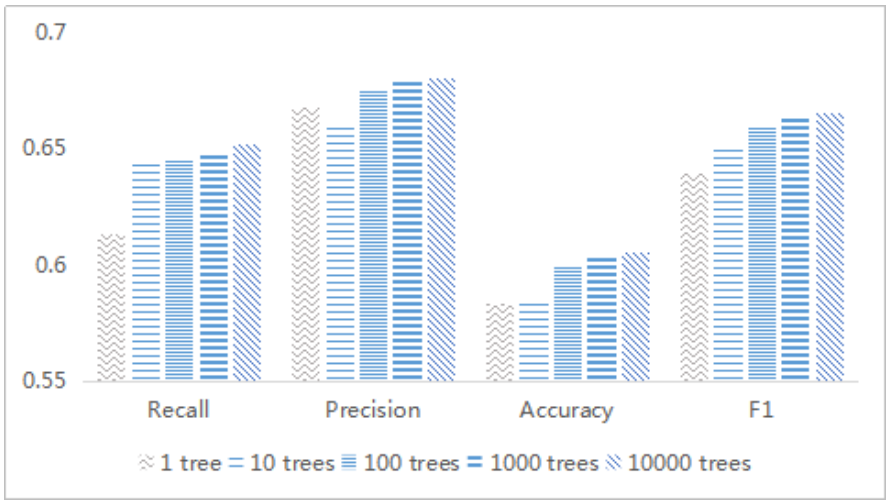

Fig. 1 More trees more better

6. Discussion

Single classifier classification easily leads to fitting problems, the combination of multiple classifiers for collective decision-making can have a better resist noise capability and more accurate than a single classifier. In addition, according to our experimental experience, the performance of random forest algorithm is limited by the performance of PC. The more training data, the more trees, the more memory we need. With 8G memory, the data in the training project can only be about 18000 trees at most. In this experiment, core i5-6400 CPU and 8G memory are used to complete 10000 trees at one time, which takes about 270 seconds. With the increasing number of trees, the improvement of performance metric is becoming more and more limited, and even it may decline. The reason may come from "random", which can be avoided by multiple calculations adopted in this paper.

7. Conclusion

In this paper, the random forest is used to predict freshman registration. The results show that this is feasible, the accuracy rate is reached 60%, and the F1 score is close to 0.7. For this data set, there is still further improvement in the performance of this classification, and further research would be done in future work.

Acknowledgements

This paper was financially supported by Natural Science General Projects from Guangdong University of Science and Technology NO. GKY-2019KYYB-31, NO. GKY-2019KYYB-36 and NO. GKY-2019KYYB-42, Innovation and Improve School Project from Guangdong university of Science and Technology NO. GKY-2015CQPT-2.

References

[1]. Yu B, Wang H Z, Shan W X, et al. Prediction of Bus Travel Time Using Random Forests Based on Near Neighbors, Computer-Aided Civil and Infrastructure Engineering, Vol. 33 (2018) No. 4, p. 333-350.

[2]. Qin J, Luo Y, Xiang X, et al. Coverless Image Steganography: A Survey, IEEE Access, Vol. (2019) .

[3]. Lin W W, Wu Z M, Lin L X, et al. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis, Ieee Access, Vol. 5 (2017) No. p. 16568-16575.

[4]. Wang X, Zhou C, Xu X Application of C4.5 decision tree for scholarship evaluations, Procedia Computer Science, Vol. 151 (2019) .

[5]. Breiman L Random forests, Machine Learning, Vol. 45 (2001) No. 1, p. 5-32.

[6]. Ye J, Hou L-D Improvement and Application of Decision Tree C4.5 Algorithm, DEStech Transactions on Computer Science and Engineering, Vol. (2018) .

[7]. Mu X Y, Remiszewski S, Kon M, et al. Optimizing decision tree structures for spectral histopathology (SHP), Analyst, Vol. 143 (2018) No. 24, p. 5935-5939.

[8]. Mu Y, Liu X, Wang L, et al. A parallel tree node splitting criterion for fuzzy decision trees, Concurrency and Computation: Practice and Experience, Vol. 31 (2019) No. 17.

[9]. Oshiro T M, Perez P S, Baranauskas J A. How many trees in a random forest?[C]. International workshop on machine learning and data mining in pattern recognition, 2012: 154-168.

[10]. Resende P a A, Drummond A C A Survey of Random Forest Based Methods for Intrusion Detection Systems, Acm Computing Surveys, Vol. 51 (2018) No. 3, p. 36.

[11]. Robinson R L M, Palczewska A, Palczewski J, et al. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets, Journal of Chemical Information and Modeling, Vol. 57 (2017) No. 8, p. 1773-1792.

[12]. Sarica A, Cerasa A, Quattrone A Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review, Frontiers in Aging Neuroscience, Vol. 9 (2017) No. p. 12.

[13]. Mi C R, Huettmann F, Guo Y M, et al. Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence, Peerj, Vol. 5 (2017) No. p. 22.

[14]. Zhou Z, Qin J, Xiang X, et al. News Text Topic Clustering Optimized Method Based on TF-IDF Algorithm on Spark, Computers, Materials & Continua, Vol. 61 (2019) No. 3, p. 217-231.

[15]. Li H, Qin J, Xiang X, et al. An efficient image matching algorithm based on adaptive threshold and RANSAC, IEEE Access, Vol. 6 (2018) No. p. 66963-66971.

[16]. Han T, Jiang D X, Zhao Q, et al. Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery, Transactions of the Institute of Measurement and Control, Vol. 40 (2018) No. 8, p. 2681-2693.

[17]. Biau G, Scornet E A random forest guided tour, Test, Vol. 25 (2016) No. 2, p. 197-227.

[18]. Cootes T F, Ionita M C, Lindner C, et al. Robust and accurate shape model fitting using random forest regression voting[C]. European Conference on Computer Vision, 2012: 278-291.