Research on machine learning strategy based on voting model

Shipei Du[a*], Wenhui Ding[a], Dongjie Yang[a], Lei Yang[b]
[a]School of Finance and Economics, Guangdong University of Science and Technology, Dongguan, China; [b]School of Computer Science, Guangdong University of Science and Technology, Dongguan, China

ABSTRACT

In this research, machine learning algorithms such as decision tree, random forest, and BP neural network are used to predict a certain dataset, and then a voting prediction model is built based on the above three machine learning algorithms. To verify the performance of this voting model, we introduced confusion matrix and F1 score to evaluate the effectiveness of machine learning. The experimental results show that the performance of the machine learning strategy based on the voting model outperforms that of a single machine learning algorithm and that adjusting the voting weights of a single algorithm can also affect the performance of the whole model. This result is well worth further study.

Keywords: Machine learning, voting model, decision tree, neural network

1. INTRODUCTION

In universities, predicting the number of new students is a very difficult but essential work. This paper attempts to predict whether freshmen will register through machine learning algorithm. It is hoped that this study can play a guiding role in the enrollment management of college freshmen.

We collected a new student registration data set, which is the enrollment data of a Chinese University in recent years. These raw data contain a lot of useless noise information. We have done a lot of work on this data set, including data cleaning, data conversion and so on. Finally, the data set can be used for machine learning.

We built a huge decision tree to predict the data set. In addition, we also used two other machine learning algorithms: random forest and BP neural network to machine learn the data set, and achieved good results.

In order to improve the performance of machine learning, we also designed a voting model and used it to predict. The voting model is composed of three machine learning algorithms: random forest, decision tree and BP neural network. We used hard voting to predict. Through our experiments, the effectiveness of the model is verified, and the performance has been significantly improved.

Colleges are progressively inquisitive about recognizing the variables that maximize their enrollment. These components permit enrollment administration directors to recognize the candidates who have a higher propensity to select at their teach and appropriately to way better designate their rewards. Ahmad Slim et al. used the methods of logistic regression (LR), support vector machine (SVM), and semi-supervised probability to verify the real data of applicants from the University of New Mexico[1]. The comes about appears that a little set of components related to freshman and college characteristics are profoundly connected to the applicant's choice of enrollment.

Wanjau et al. proposed a common system for mining freshmen information enlisted in Science, Innovation, Building, and Science (STEM) utilizing performance-weighted outfit classifiers[2]. The result appears that utilizing outfit models not as it gave way better prescient exactnesses on understudy enrollment in STEM but moreover gives way better rules for understanding the components that impact understudy enrollment in STEM disciplines.

2. MACHINE LEARNING STRATEGY BASED ON VOTING MODEL

We use decision tree, random forest and BP neural network algorithm to machine learn the dataset respectively, and then on this basis, we build a voting model, and use this voting model to machine learn the dataset again. Our experiments

* dushipei@gdust.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125062X
 · 0277-786X · doi: 10.1117/12.2662004

show that the performance of the voting model is better than that of a single machine learning algorithm.

2.1 Decision tree

A decision tree could be a prescient examination show of a tree structure that reflects the mapping between objects and their trait values[3]. It comprises a root node, department node, and leaf node. The latter is the beginning point of the complete decision tree and is found at the best. The department node may be a modern quality shaped by isolating an upper node, speaking to an information subset of information. The leaf node speaks to the classification result[4]. The decision tree judges from the root node and chooses the node concurring to the property esteem of the upper node in a top-down way until the leaf node forms a modern lesson[5]. Each way of the decision tree from the root node to the leaf node could be a prescient way that outwardly speaks to the relationship between properties and comes about.

The original decision tree is used to predict the test set, and the confusion matrix of the predicted classification results is shown in Table 1.

Table 1. Confusion matrix of original decision tree.

Actual	Predicted Positive	Predicted Negative
Positive	1551(TP)	950(FN)
Negative	871(FP)	1041(TN)

The following results can be calculated, as shown in Table 2.

Table 2. Performance metrics of original decision tree.

Recall	Precision	Accuracy	F1
62.02%	64.04%	58.74%	0.63

2.2 Random forest

Random forests are a classifier that employments different decision trees to prepare and foresee tests. In specific, trees that are developed exceptionally profound tend to memorize profoundly unpredictable designs: they overfit their preparing sets, i.e. have low bias, but exceptionally tall change. Random forests are a way of averaging different profound decision trees, prepared on diverse parts of the same preparing set, with the objective of decreasing the variance[6]. This comes at the cost of a little increment within the predisposition and a few misfortune of interpretability, but for the most part, enormously boosts the execution within the final model.

Each tree in the random forest algorithm grows to the greatest extent, and there is no pruning process. The training set of each tree is randomly selected[7]. If there is no random sampling, the training set of each tree is the same, then the classification result of the finally trained tree is exactly the same. In addition, the feature extraction of the sample is also randomly selected. Assuming that the feature dimension of each sample is N, we specify a constant n <N, randomly select n feature subsets from N features and select the best from these n features each time the tree is split. The "randomness" in the random forest refers to this two randomness. The introduction of these two randomness is very important to the classification performance of random forests[8]. Because of their introduction, random forest is not easy to fall into overfitting and has good anti-noise ability. Therefore, the results of machine learning are also random. In arrange to stabilize the haphazardness as much as conceivable, we calculated 10 times and at last, took the cruel as its last forecast result.

We use 10, 50 and 100 trees to create the random forest respectively, and use these models to predict the test set. In order to avoid the influence of random factors, we calculate each model 10 times and finally take the average value as the final result. The experimental data are displayed in Tables 3 and 4.

Compared with a single decision tree model, the effect of machine learning of the random forest model is much better, and all measurement indexes are greatly improved. From the learning results, the more trees, the better the learning effect of random forest, but it does not grow all the time as displayed in Figure 1.

Table 3. Performance metric of different random forest.

Number of trees	Recall	Precision	Accuracy	F1	Cost time (s)
10	71.11%	64.59%	62.82%	0.6769	0.6412
	70.04%	65.66%	63.12%	0.6778	0.5959
	71.35%	65.02%	62.60%	0.6804	0.5975
	70.16%	65.10%	62.86%	0.6753	0.6006
	71.10%	66.23%	63.54%	0.6858	0.6006
	69.74%	65.16%	62.51%	0.6737	0.6053
	71.13%	65.23%	62.31%	0.6805	0.5881
	70.48%	66.18%	63.44%	0.6825	0.6037
	69.16%	64.91%	62.11%	0.6697	0.5975
	69.44%	65.40%	62.59%	0.6736	0.5928
50	70.60%	66.95%	63.86%	0.6873	2.9890
	71.11%	67.21%	64.36%	0.6910	2.9578
	71.10%	67.05%	64.39%	0.6901	2.8938
	70.68%	66.11%	63.45%	0.6831	2.8954
	71.22%	65.55%	63.71%	0.6826	2.9157
	72.63%	65.33%	64.15%	0.6878	2.9125
	71.10%	65.61%	63.84%	0.6825	2.9016
	72.49%	66.98%	64.24%	0.6963	2.8969
	71.61%	67.64%	64.91%	0.6957	2.8985
	73.88%	66.30%	64.77%	0.6988	2.8876
100	73.60%	66.36%	64.83%	0.6979	5.8017
	71.90%	65.89%	63.86%	0.6877	5.8220
	68.36%	68.31%	64.05%	0.6833	5.7798
	70.46%	67.52%	64.84%	0.6896	5.7970
	70.14%	67.65%	64.05%	0.6887	5.8313
	73.05%	67.26%	65.09%	0.7004	5.8344
	69.69%	66.45%	63.59%	0.6803	5.7908
	71.61%	66.86%	64.39%	0.6915	5.8142
	70.24%	67.09%	64.23%	0.6863	5.8922
	71.61%	66.62%	63.61%	0.6903	5.7954

Table 4. Average value of different random forest.

Number of trees	Recall	Precision	Accuracy	F1	Cost time (s)
10	70.37%	65.35%	62.79%	0.6776	0.6023
50	71.64%	66.47%	64.17%	0.6895	2.9149
100	71.06%	67.00%	64.25%	0.6896	5.8159

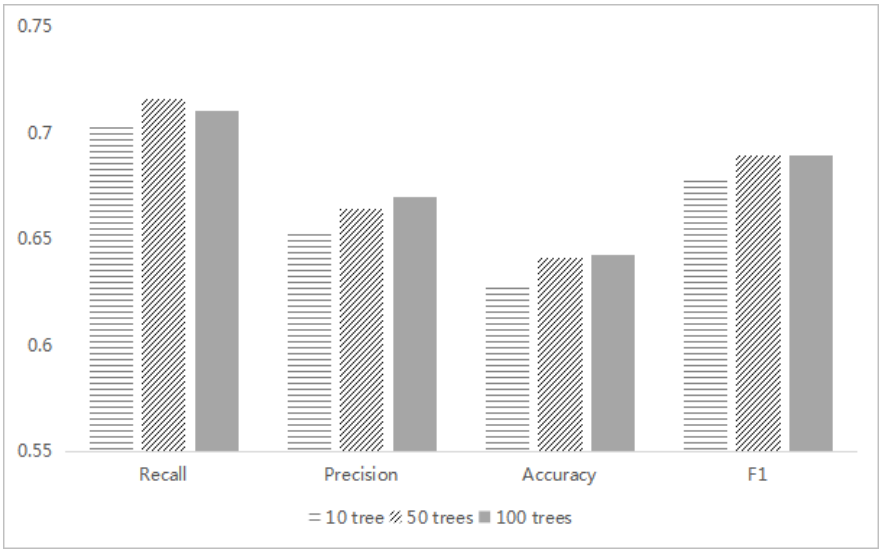

Figure 1. Effects of the number of trees in random forests.

2.3 BP neural network

BP (back propagation) neural network is a multi-layer feedforward neural network trained according to the error back propagation algorithm[9]. It adds several layers (one or more layers) of neurons between the input layer and the output layer. These neurons are called to hide cells, they are not specifically related to the exterior world, but their state changes can influence the relationship between input and yield. Each layer can have several nodes[10].

The BP neural network model is used to predict the data set. The final predicted performance indicators are displayed in Table 5.

Table 5. Performance metric of BP neural network.

Recall	Precision	Accuracy	F1
68.57%	64.68%	61.07%	0.6656

Finally, we compare the performance indexes of the above three algorithms, as displayed in Figure 2. It can be found that the random forest algorithm performs best on this dataset, and each performance index exceeds the other two algorithms.

3. THE VOTING MODEL

For the same problem, different machine learning algorithms may give different prediction results. In this case, which algorithm is selected as the final result? At this time, we can concentrate a variety of algorithms to make different algorithms predict the same problem, and finally adopt the principle that the minority obeys the majority to select the final prediction result, we construct a voting model to predict the enrollment of college freshmen. Hard voting is to choose the result with the largest number of votes as the final prediction result. Finally, a label is returned[11].

Compared with the single machine learning model, the performance of the voting model is significantly improved, as displayed in Figure 3.

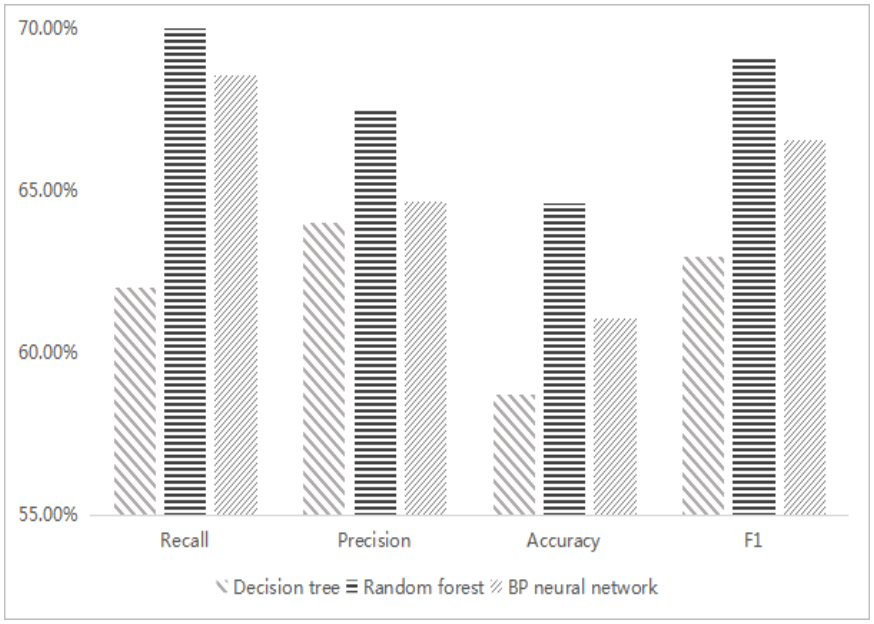

Figure 2. Comparison of the three algorithms.

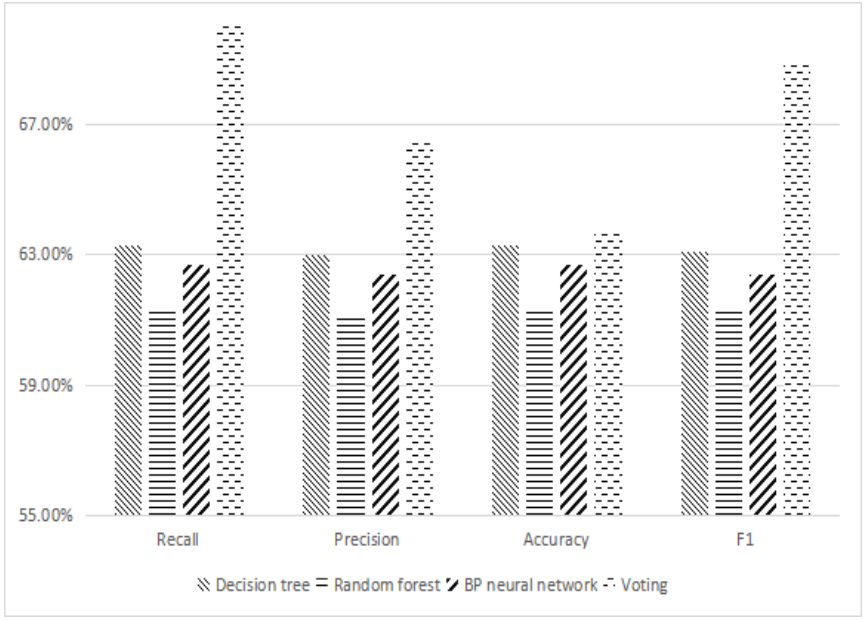

Figure 3. Compared with the single machine learning model.

4. CONCLUSION

In this paper, we want to predict the registration intention of college freshmen through machine learning. Our work shows that this work is predictable. We use three machine learning algorithms: random forests, decision tree, and BP neural network to learn the dataset. At the same time, we also propose a voting model to improve the performance of machine learning. In the future, we plan to work on the following topics. The number of basic classifiers of the voting

model needs to be improved. More machine learning algorithms will be used to vote to see whether the performance of machine learning can be improved.

ACKNOWLEDGMENTS

This research was funded by Provincial Key Platforms and Major Research Projects of Guangdong Universities, Young Innovative Talents Project (Humanities and Social Sciences), grant number 2018WQNCX206, and Social Sciences Project of Guangdong University of Science and Technology NO. GKY-2020KYYBW-70.

REFERENCES

[1] Naghibi, S. A., Ahmadi, K. and Daneshi, A. "Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping," Water Resources Management, 31(9), 2761-2775 (2017).
[2] Wanjau, S. K. and Muketha, G. M., "Improving student enrollment prediction using ensemble classifiers," International Journal of Computer Applications Technology and Research, 7(3), 122-128(2018).
[3] Tran-Nguyen, M. T., Bui, L. D. and Do, T. N., "Decision trees using local support vector regression models for large datasets," Journal of Information and Telecommunication, 4(1), 17-35(2020).
[4] Trabelsi, A., Elouedi, Z. and Lefevre, E., "Decision tree classifiers for evidential attribute values and class labels," Fuzzy Sets Syst., 366(1), 46-62 (2019).
[5] Jia, Z. C., Han, Q. Y., Li, Y. Y., Yang, Y. L. and Xing, X., "Prediction of web services reliability based on decision tree classification method", CMC-Computers Materials & Continua, 63(3), 1221-1235(2020).
[6] Su, Y., Weng, K. L., Lin, C. and Zheng, Z. M., "An improved random forest model for the prediction of dam displacement", IEEE Access, 9(1), 9142-9153(2021).
[7] Panagiotakis, C., Papadakis, H. and Fragopoulou, P., "A dual hybrid recommender system based on SCoR and the random forest," Computer Science and Information Systems, 18(1), 115-128(2021).
[8] Chen, X. H., Yu, S. Y., Zhang, Y. F., Chu, F. F. and Sun, B., "Machine learning method for continuous noninvasive blood pressure detection based on random forest," IEEE Access, 9(1), 34112-34118(2021).
[9] Tian, S. K., Dai, N., Li, L. L., Li, W. W., Sun, Y. C. and Cheng, X. S., "Three-dimensional mandibular motion trajectory-tracking system based on BP neural network," Math. Biosci. Eng., 17(5), 5709-5726(2020).
[10] Lyu, J. C. and Zhang, J., "BP neural network prediction model for suicide attempt among Chinese rural residents," Journal of Affective Disorders, 246(1), 465-473(2019).
[11] Pande, M. and Mulay, P., "Bibliometric survey of quantum machine learning", Science & Technology Libraries, 39(4), 369-382(2020).