Journal of Computing and Electronic Information Management

ISSN: 2413-1660

Application of Decision Tree in Prediction of College Freshmen Registration

Lei Yang[1, a, *], Yuewei Xia[2, b], Longqing Zhang[1, c]

[1]Guangdong University of Science and Technology, Dongguan, China

[2] Luohe Vocational Technology College, Luohe, China

[a]gdstyl@qq.com, [b]719122100@qq.com, [c]bjzlq@qq.com

Abstract: We propose a prediction method based on the historical enrollment data of a university. Decision tree algorithm is used to predict the registration of freshmen. The results show that the registration of freshmen is predictable.

Keywords: Decision Tree, Prediction, Freshmen, Registration.

1. Introduction

Machine learning, as a discipline derived from artificial intelligence and statistics, is one of the key research directions in the field of data analysis. In the past three decades, the ability of human beings to collect, store, transmit and process data has been rapidly improved. In all walks of life, a large amount of data has been accumulated in every corner of society. Many data have been favored by researchers, while some data are stored in a corner unknown.

In the real world, many colleges and universities often encounter the situation that the number of freshmen can not be accurately predicted, but can only be estimated according to the previous year's registration rate[1]. This may not be a big gap in the total number, but it involves the number of specific boys, girls, the number of a certain profession, this is a case[2]. The unpredictable problem, and once the school opened a new major, new departments, the total number of forecasts will be very different.

The usual practice of colleges and universities is to make a rough estimate based on the previous year's reporting rate, which is very vague. According to the sleeping data of a university, this paper presents a new data set, that is, admission and reporting to the data set. Our goal is to learn by machine learning, hoping to learn something[3]. First of all, we have done a lot of processing on this data set, including data cleaning, specification, transformation, so that it can be recognized by the machine, and the

data of the first three years as a training set, and the data of the fourth year as a test set.

Secondly, we use decision tree algorithm to learn it. The results are encouraging, with accuracy approaching 60%. In order to evaluate our work more accurately, we introduce F-Measure evaluation criteria[4]. The results show that the value of F-Measure is close to 0.7, which proves that our work is effective.

Fig. 1 Use Gymnasium to Provide Rest Space for Parents of Freshmen

2. The dataset

Regarding the admission data, this part of the data comes from the Candidate Information Base of Guangdong Enrollment Office. The basic information includes the student's examinee number, name, gender, identity card number, professional code, professional name, birth date, political outlook, telephone and so on. There are 38 columns of data[5].

Among them, gender names include two values for men and women, plus 20 and 15 values for sub-items. Political features include three values for Party members, League members and the masses. Applicants for examinations include three values for liberal arts, science and 3+ certificates. Graduation categories include three values for general high schools, vocational high schools and other secondary technical schools, and so on.

Regarding the registration data, this part of the data comes from the school educational administration system. The basic information includes students' names, gender, ID number, birth date, major and so on. When a student's information exists in the educational system, we think that the student is registered. Because there is not too much student information stored in the school's educational administration system, we use the student's ID number to match uniquely. We compare the admission data with the data in the educational administration system using ID number.

3. Proposed method

3.1 Decision tree

Decision tree is a prediction and analysis model of tree structure, which reflects the mapping relationship between the object and its attribute values. It consists of root node, branch node and leaf node. Root node is the starting point of the whole decision tree and is located at the top. Branch nodes are new attributes formed by segmentation of upper nodes, representing data subset and leaf nodes representing classification results.

The decision tree starts from the root node and chooses the node according to the attribute value of the upper node in a top-down manner until the leaf node to form a new classification. Each path from the root node to a leaf node in the decision tree is a predictive path, which intuitively expresses the relationship between attributes and results.

3.2 Model evaluation

To evaluate a classification model, the accuracy is usually used to evaluate it. The higher the accuracy is, the better the model is. Correctness is indeed a very good and intuitive evaluation index, but sometimes high accuracy does not represent a model.

For the two-classification problem, the samples can be divided into four cases: true positive, false positive, true negative and false negative according to the combination of their real classes and classifier prediction classes. If TP, FP, TN and FN are used to represent their corresponding sample numbers respectively, then there is TP + FP + TN + FN = total number of samples. The confusion matrix of classification results is shown in Table 1.

Table 1. The Confusion Matrix

The truth	The prediction	
	positive	negative
positive	TP(true positive)	FN(false negative)
negative	FP(false positive)	TN(true negative)

The accuracy, precision and recall are defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Precision and recall are contradictory indicators. Generally speaking, when the precision is high, the recall is often low, while when the recall is high, the precision is often low. For a specific classifier, we can not improve all the indicators at the same

time. Of course, if a classifier can correctly classify all the instances, then all the indicators have reached the optimum, but such classifiers often do not exist. Usually, it is hoped that the precision of the classifier should be as high as possible on the premise of a certain recall.

Usually, the performance of the model is evaluated by the F1 score. When the F1 score is higher, the performance of the classification model is better. The range of the F1 score is 0 to 1, and the F1 is defined as:

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

4. Experiment

4.1 Experiment Settings

Before training, in order to save computing resources, we put the data of 4 years into a data table. The first 6599 rows are the data of the first three years, and the last 3783 rows are the data of the fourth year. As shown in Table 2.

Table 2. The Dataset

Dataset	Total	Unregistered	Registered
Full dataset	10382	4393	5989
Training set (first 3 years)	6599	2889	3710
Test set(the 4th year)	3783	1504	2279

4.2 Decision tree

The training results are shown in Figure 2. The decision tree has 60 layers and the root node is x13. The corresponding characteristics are current or past students.

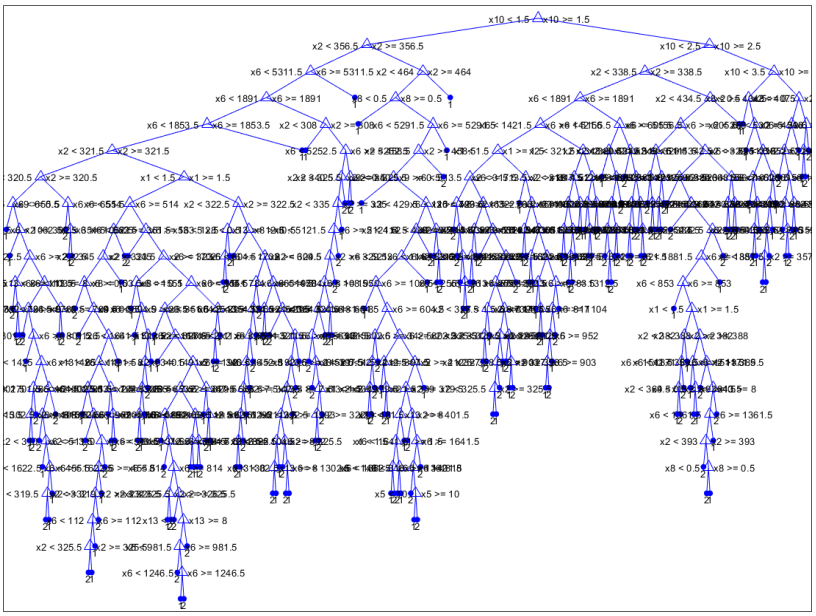

Fig. 2 The Decision Tree

Prediction of data for the fourth year, and the confusion matrix of classification results is shown in Table 3.

Table 3. Confusion Matrix of Decision Tree Classification Results

The truth	The prediction	
	positive	negative
positive	TP (1399)	FN (880)
negative	FP (696)	TN (808)

According to formula (1) (2) (3) (4), the evaluation can be calculated: Accuracy= 58.34%, Precision=66.78%, Recall=61.39%, F1=0.64 .

5. Conclusion

We use the decision tree algorithm to predict the registration of freshmen. The research proves that this is feasible. In addition, if this prediction result is used, the report data of the fourth year is equivalent to polluted after the intervention of the students. Whether we can continue to use the data of the fourth year to forecast the fifth year needs to be further explored in the future.

Acknowledgements

This paper was financially supported by Guangdong Provincial Innovation and Entrepreneur-ship Training Program Project NO.201713719017, College Students Innovation Training Program held by Guangdong university of Science and Technology NO.1711034, 1711080,and NO.1711088.

References

[1] Lihong Y, "Application of Decision Tree in Analysis and Forecast of Higher Vocational Students' Reporting Rate", Journal of Henan Business College, 2007, Vol. 01 (01), p113-116

[2] Guangbiao D, "Application of Predictive Big Data Analysis in College Enrollment", Microcomputer Applications, 2017, Vol. 33 (11), p20-23

[3] Sihong L, "Application and Research of Decision Tree Algorithms in College Enrollment Decision System", Journal of Xi'an Academy of Arts and Sciences (Natural Science Edition), 2015, Vol. 18 (03), p67-70

[4] Yang J, "Research and Application of Decision Tree Algorithms", Computer Technology and Development, 2010, Vol. 20 (02), p114-116+120

[5] Aihui H, "Improvement and Application of Decision Tree C4.5 Algorithms", Science and Technology and Engineering, 2009, Vol. 9 (01), p34-36+42.