2016 4th International Conference on Enterprise Systems

PREDICTION OF THE SCHOLARSHIP USING COMPREHENSIVE DEVELOPMENT

Wenjuan Wei
School of Computer Science
Xi'an ShiYou University
Shaanxi, China 710065
Email:wenjuan1510@gmail.com

Jiaxin Han
School of Computer Science
Xi'an ShiYou University
Shaanxi, China 710065
Email:jxhan@xsyu.edu.cn

Jie Kong
School of Computer Science
Xi'an ShiYou University
Shaanxi, China 710065
Email:jiekong@xsyu.edu.cn

Haiyang Xia
Shool of Computer Science
Xi'an ShiYou University
Shaanxi, China 710065
Email:haiyangxia15@gmail.com

***Abstract*—In major colleges and universities, in order to mobilize students enthusiasm for studying and participating in extracurricular activities, all colleges make an evaluation on students comprehensive quality and set different rewards regulations for the various level. The main way is to provide financial incentives, they distribute scholarship for students of meeting requirements. The Decision Tree algorithm is frequently used all the time, however, because of the tree node selected is based on attribute's mutual information, which will lead to some crucial attribute lost their decisive role. Therefore, in this paper, we focused on predicting whether students obtain scholarship on the comprehensive quality of students with Naive Bayes algorithm.**

Keywords—scholarship prediction; classification;Naïve Bayes; ROC curve

I. INTRODUCTION

Scholarships were set up by schools reward students that have excellent achievements. Its a set of incentives and benefits, the original intention is encourage students and improve school's learning culture. Currently, every university has their own methods to realize scholarship assessment. Nevertheless, for some students it is unfair to apply the exiting scholarship system. From that point of view, we consider an optimal way to assess and distribute scholarship are needed. And after determining the evaluation method of scholarship, we hope that, given the set of the impact factors may predict whether the student will get scholarships.

Therefore, in this paper, we make a comprehensive evaluation of grades, morality and the enthusiasm of participate in activities. This paper identifier the different levels in the four types of scholarship result, the one is those students who have failed to get scholarship we marked as class N, surplus is those students who have get first, second or third level scholarship we marked as F, S and T in turn.

Classification is a technique where an algorithm learns a classification model on a labeled data-set and which to predict the labels of new [9], it is a supervised learning, the main target is mining the relation between data characteristics and labels. In this article, we proposed a classification model to achieve prediction of the scholarship distribution. We collected and analyzed the factors associate with the scholarships, evaluate the prediction model and observe if the prediction model can be applied to a future scholarship forecast.

This paper is organized as follows: In section 2, we discuss the development of the data mining (DM) and educational data mining (EDM), then introduces the application of the EDM in scholarships prediction. Section 3 includes the main algorithms in this work. Finally, Section 4 describes the experimental setup, and analyzes the implementing results. Section 5 concludes the study.

II. BACKGROUND

In this section, we will discuss the development of DM, after that, we will introduce the application of EDM in scholarship classification.

A. DM and EDM

With the rapid development of internet technology, the exchange of the information between people is no longer limited by time and space. With the maturity of database technology and the popularity of data application, the amount of the data that accumulated by people are growing rapidly at an exponential rate. How to deal with the mass of data is become troublesome issues, so data mining technology is born at the right moment [9]. DM, often called knowledge discovery in database (KDD), is known for its powerful role in uncovering hidden information from large volumes of data [10][13]. Its advantages have landed its application in numerous fields, within the educational research which commonly known as EDM [14].

EDM is an emerging discipline, it refers to apply data mining methods and algorithms to exploring significative data from education settings, then as the most innovations

978-1-5090-3790-2/16 $31.00 © 2016 IEEE
DOI 10.1109/ES.2016.30

in educational systems[11][15], EDM is a set of powerful methods for finding rules, analyzing patterns[12] and making predictions from the students' behaviors and achievements[8]. It depends on continuous improvement to adapt the rapid development of current education all the time.

B. *EDM for scholarship prediction*

EDM can be used not only to learn the model for the learning [17] or student modeling [16], but also to evaluate and to improve e-learning systems [18] by discovering useful learning information from learning portfolios. With the numbers of major colleges and universities student rapid rise, scholarship has always been a major university council hot spot, and each student pays more and more attention to scholarships awarded objects. Though, EDM has become a hot topic, the little research were applied the forecast for scholarship. For existing system and research, the main evaluation standard is grades, so there are more and more disputes on it. Therefore, to solve this problem this title proposed, we need to use EDM to find a reasonable way to achieve a scholarship strategy on a maximum degree, and make the all-round development (or defined as comprehensive development) of morality, intelligence physique and aesthetic students on every students, so that scholarship can play its own meaning, ensure each student would get his deserved reward,.

III. METHODOLOGY

A. *Data collection and preprocessing*

We use a real-word data sets conduct our experiments. The data sets we used are provided by a university of china. We need to make a preprocessing of the original data. Firstly, we removed missing data and reduced the dimension of the data. Secondly, deleted unnecessary information. Then we transformed the data-set into we need. Finally, made a statistic on the data. The processed data is shown in Table1. In finished data-set, GPA is the ratio of the total credits grade points and total grades, named grade point average, the value of retake is "Y" means the student is have one or more times failure in exam, conversely, the "N" is represent success in an exam on each class. Times are meant to the times of the student taking part in activities. The value of Demerits is "Y" means the student has break the rules of school , on the contrary, "N" is not.

TABLE I. PART OF THE TRAINING SAMPLE DATA

GPA	Retake	Times	Demerits	class
1.24	Y	8	N	N
3.26	N	8	N	F
2.84	N	3	N	N
2.95	N	8	N	S
3.15	N	8	Y	N
…	…	…	…	…

As can be seen from table 1, if the students' GPA, retake information, times of participating activities, demerits information are shown to us, we could know who will be given a scholarship.

Generally, the binary method is used in nature most [7], nevertheless, the scholarship prediction in this article belongs to multi-class predict. Multi-class classification is one of the major challenges in real word application [7]. Therefore, during the prediction in this title, we should convert the multi-class predicts to binary predict.

B. *Classification Algorithm*

In this section we will give the details of the Naive Bayes algorithm.

3.2.1 Naive Bayes algorithm. The classification principle of Bayesian classifier is to use a prior probability of an object and calculate the probability of posterior, then choose the maximum posterior probability as the object class. Given a class variable y and a dependent feature vector x1 through xn, Bayes' theorem states the following relationship:

$$P(\mathrm{y} \mid x_1,\ldots,x_n) = \frac{P(y)P(x_1,\ldots,x_n \mid y)}{P(x_1,\ldots,x_n)} \qquad (1)$$

Where P(y) is prior probability, $P(x_1,\ldots,x_n|y)$ is class-condition probability, $P(x_1,\ldots,x_n)$ is evidence factor for normalization.

However, class condition probability $P(x_1,\ldots,x_n|y)$ is a joint probability of all attribute, it is difficult to get from the limited training set immediately. So in order to avoid the issue, Naive Bayes classifier algorithm utilizes "attribute condition independence assumption". For the categories we all known, it is assumed that all properties are independent [6]. In other words, assume each attribute have an influence on classification result independently. Based on attribute condition independence assumption, the formula (1) is instead by

$$P(y \mid x_1,\ldots,x_n) = \frac{P(y)P(x_1,\ldots,x_n \mid y)}{P(\mathrm{x}_1,\ldots,\mathrm{x}_n)} = \frac{P(y)}{P(x)}\prod_{i=1}^{n} P(x_i \mid y) \quad (2)$$

In (2), n is numbers of attributes, x_i is values that x in the i-th attribute. Because of the P(x) is same for all class, therefore, formula (3.2) is transformed to Naive Bayes classifier formula(3.3).

$$P(y \mid x) = \arg\max_{y \in Y} P(y)\prod_{i=1}^{d} P(x_i \mid y) \qquad (3)$$

Conspicuous, Naive Bayes algorithm is estimates class-prior probability P(y) based on training set D, then calculate the condition probability $P(x_i|y)$ for each attribute. For items which are required to classify, to get the probability of each category appears under every class, the largest probability is signified the item is defined as the class, which is the ideological foundation of Naive Bayes algorithm. Implementation steps as follows:

(1)Determining characteristic property, obtain training sample.

(2)Calculating the probability of each class $P(y_i)$.

(3)Calculating the condition probability of each characteristic properties $P(y_1|x),P(y_2|x),...,P\{y_n|x\}$.

(4)If $P(x|y_k)=\max\{P(x|y_1),P(x|y_2),...,P(x|y_n)\}$,则 $x \varepsilon y_k$.

In summary, calculate every condition probability is the main operation. We should give a sorted sample set, statistics prior probability and the conditional probability of each feature attributes.

C. *Model Assessment*

Cross-validation is a measurement of assessing the performance of a predictive model, and statistical analysis will generalize to an independent data-set[4]. When handling the small data sets, the data can be divided into K disjoint data subsets and the every data subset is the same large. Each time K-1 data are set as the training sample, the surplus one data set as a validation set, it is K-precisions, and calculate the average K-test indicators. Because of the k-fold's importance in the method, so the cross-validation usually called k-fold cross-validation.

D. *Performance Measurement*

3.4.1 Precision and Recall. We always use the classification model to evaluate the process of classification, we will employ TP(true positive), FP(false positive), TN(true negative) and FN(false negative) to represent the relationships between real class and prediction class. And TP+FP+TN+FN is equal to the sample's total number.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (5)$$

It is appears in the confusion matrix. However, precision and recall ratio is a couple of contradictory measures, in general, high precision with low recall rate and low precision are meant recall rate is high. As a consequence, we'd better have an integrated consideration of them, the most common method is shown in equation 3.6.

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (6)$$

As shown in (6), F1 is precision and recall weighted harmonic mean, when F1 is higher, the test method is more effective.

3.4.2 ROC curve. ROC curve is a method that we evaluate the general classification model, the full name is Receiver Operating Characteristic. It has been widely used in the machine learning community to measure the performance of classifiers[2]. According to the result of the classification to sort sample, in the sequence, then see the every sample as a positive example to predict, and calculate two values each time, respectively, they plotted as abscissa and ordinate and obtained ROC curve. ROC curve ordinate is "True Positive Rate" and the abscissa is "False Positives Rate."

$$TPR = \frac{TP}{TP + FN} \qquad (7)$$

$$FPR = \frac{FP}{TN + FP} \qquad (8)$$

In reality, the ROC curve is a step function. For a good classification model, the ROC curve should be close to the top left corner of the unit square. And it means the curve is above the diagonal which the positive and negative points connection.

It is no doubt that ROC curve capability is intuitive clearly demonstrated, but it may not accurate enough, as a result, we often want to the model be evaluated by a digital standard in practical applications. Therefore, we always use the values of AUC (the Area Under ROC), that is the below area of ROC curve. It has been an important performance measure in many learning task[3]. Usually AUC values ranged between 0.5 and 1.0, and larger AUC values means to the true positives greatly exceed false positives[1], it represent better performance.

IV. Experiment and Analysis

We design the experiment to test the effective of our scholarship prediction method. We will utilize decision tree and Naive Bayes algorithm to test the classification model in the experiment. The questions we are trying to answer are: (1) What is the key factor. (2) Can we predict whether the student obtains a scholarship from one of the factors.

A. *Experiment setup*

The purpose of the study is to accomplish the scholarships prediction through the student's comprehensive development. The GPA, whether or not rebuilt, whether or not demerit and times of participating activities all plays an important role in the classification model, all of the characteristic properties has their own standards. Finally, we applied 10-fold cross-validation, confusion matrix and AUC to evaluate the model.

B. *The result of experiment*

A perfect classification model is: if a student has obtained the scholarship level is 'N', 'F' or 'T', and the results of the model prediction is coincident with the fact. However, for our classification model, it always has many mistakes, so we ought to know the numbers of correct and incorrect prediction. Confusion matrix put all of the messages into a table, it is a standard form that is use to denote the evaluation of accuracy. The matrix main diagonal elements are correctly classified samples, other data are incorrectly classifying.

We use the 10-folds cross-validation under Decision Tree and Naive Bayes to test the effectiveness of our prediction method. Using the test we can obtain the confusion matrix shown in Table 2 and Table 3.

TABLE II. CONFUSION MATRIX WITH DECISION TREE

a	b	c	d	Classified as
131	3	5	2	a=N
0	13	0	0	b=F
1	0	8	0	c=S
2	0	0	1	d=T

TABLE III. CONFUSION MATRIX WITH NAIVE BAYES

a	b	c	d	Classified as
137	1	3	0	a=N
0	13	0	0	b=F
1	0	8	0	c=S
2	0	0	1	d=T

TABLE II and TABLE III give many messages of the classification to us. For instance, we can calculate precision, recall and F1 about all of the classes.

According to the result of calculating, histograms of the result are shown Fig. 1 to Fig. 4.

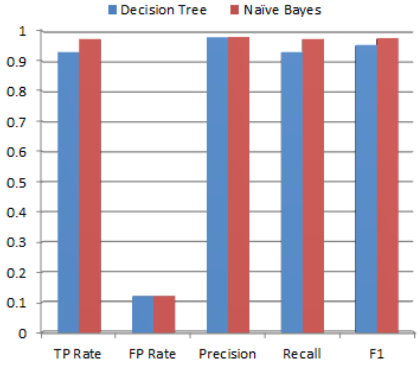

Fig. 1 . Detailed Accuracy of the Class N

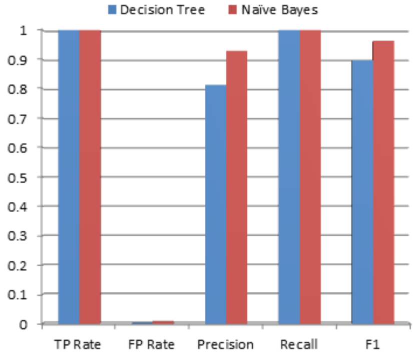

Fig. 2 . Detailed Accuracy of the Class F

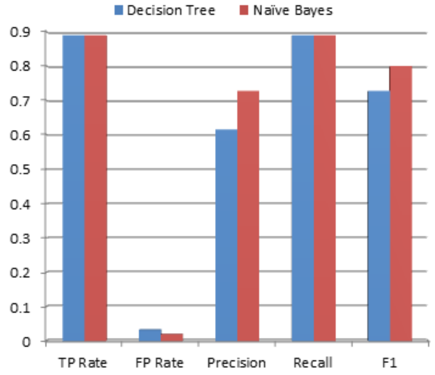

Fig. 3 . Detailed Accuracy of the Class S

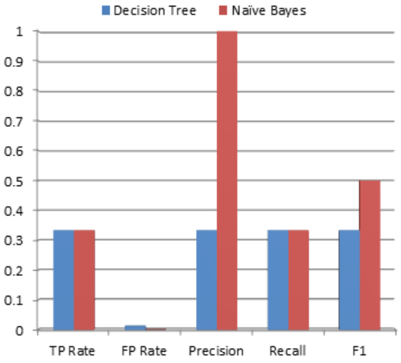

Fig. 4 . Detailed Accuracy of the Class T

Fig. 1 to Fig. 4 show the TPR, FPR, Precision, Recall and F1 of the four class compared decision tree and Naïve Bayes the better result could be observed, and we also could find the Naïve Bayes algorithm is better than Decision Tree algorithm.

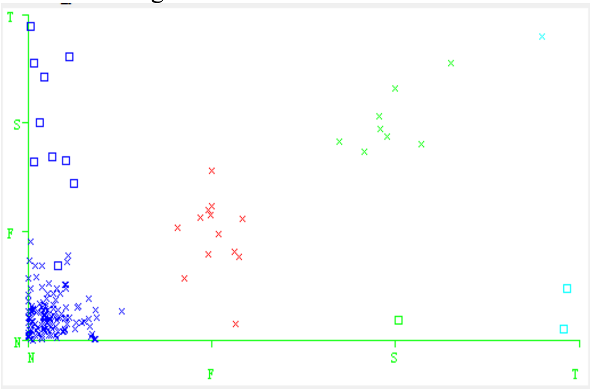

Fig. 5 . Classifier Errors Visualize with Decision Tree

07

ibnu triyuono

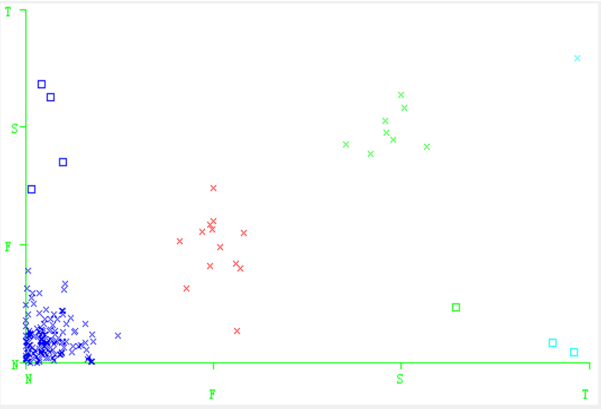

Fig. 6 . Classifier Errors Visualize with Naive Bayes

The above figures show us the result of we use classify algorithm to evaluate the prediction model. In figures, the abscissa is actual results obtained scholarships, the ordinate is a result of the prediction model, ’×’ represents correct classification, ’□’ represents incorrect classification. From figures, for class of N, F, and S class, we can find that the numbers of ‘□’ is far less than ’×’, though, in T class, ‘□’ is more than ’×’, However, as a whole, the quantity of correct predictions is far more than error predictions.

In addition to, from the experiment, we observed that all of ROC curves are closer to the upper left corner.

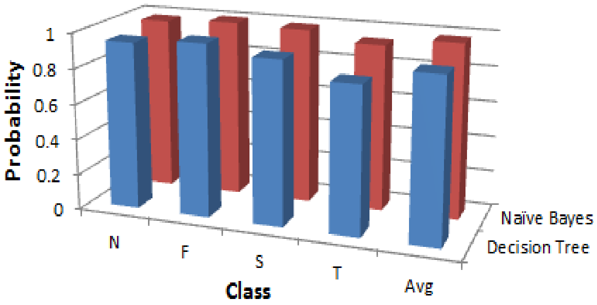

Fig. 7 . the ROC Area of all Class

From the Fig.7, we can see all of the AUC are between 0.5 and 1, and the classification model has a better classification effect.

Therefore, it is obvious that utilize 10 cross-validations based on Naive Bayes algorithm have obtained a better result than Decision Trees’.

V. Conclusion

This paper focus on applications of data mining and machine learning to achieve the goals of scholarships prediction, we use the method of supervised learning based on student comprehensive development level of moral, intellectual, physical, aesthetics and education to achieve scholarships forecast. According to some input attributes, we also can find that whether the student can get scholarships and what the scholarship level is. Experiments are based on real data sets show the effectiveness of our prediction model. By using classification methods, we have given an overall assessment models.

Through compare with other method, we can consider that the classification model is provided a fair way to scholarships assessment. And every student could accord the current situation of themselves, to pay their attention to their weak link. In addition, we also can get some information from the result when the student don’t get a scholarship, we can find out the reason why he is not, first, if he is failure in the exam or the GPA is not up to standard, the teacher should take care of his study, if he is recorded a demerit, the teacher should pay more attention to his ideological education, and if the student has less times of participating activities, the teacher should arouse the enthusiasm of the student. On the contrary, we think he is a qualified college student.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments. This research is supported by the International Cooperation and Exchanges in Science and Technology Plan Project in Shannxi under the Grant No. 2016kw-047.

Reference

[1] James Michael Lampinen. ROC analyses in eyewitness identification research. Journal of Applied Research in Memory and Cognition, 2016:21-33.

[2] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (3) (2005) 299–310.

[3] C. Cortes, M. Mohri, AUC optimization vs. error rate minimization, in: S. Thrun, L. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems, vol. 16, MIT Press, Cambridge, MA, 2004, pp. 313–320.

[4] Ping Jiang , Jiejie Chen. Displacement prediction of landslide based on generalized regression neural networks with K-fold cross-validation. Neurocomputing, 2016.

[5] Zhi-Hua Zhou. Machine Learning[M].Nanjing University. Beijing: Tsinghua University Press, 2016.

[6] Arpit Bhardwaj, Aruna Tiwari, Harshit Bhardwaj and Aditi Bhardwaj. A Genetically Optimized Neural Model for Multi-class Classification. Expect System With Application, 2016.

[7] Alejandro Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. Expert System with Applications, 2016.

[8] Sašo Karakatic, Vili Podgorelec.(2016). Improved classification with allocation method and multiple classifiers. Information Fusion.

[9] Man Wai Lee, Sherry Y. Chen, Kyriacos Chrysostomou, Xiaohui Liu. Mining students behavior in web -based learning programs. Expert Systems with Applications, 2009.

[10] JEONG, BISWAS. Mining Student Behavior Models in Learning-by-Teaching Environments. Proceedings of the 1st International Conference on Educational Data Mining, Québec, Canada, 2008.

[11] Janice D. Goberta, Yoon Jeon Kima, Michael A. Sao Pedroa, Michael Kennedyb & Cameron G. Bettsc. Using educational data mining to

assess students' skills at designing and conducting experiments within a complex systems microworld. Thinking Skills and Creativity, 2015.

[12] Witten, I.H. and Frank, E. 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kauffman, San Francisco, CA

[13] Baker, R.S.J.d.: Data Mining for Education. In: McGaw, B., Peterson, P., Baker, E. (eds.) To appear in International Encyclopedia of Education, 3rd edn. Elsevier, Oxford (2010)

[14] Ha, S., Bae, S., & Park, S. (2000). Web mining for distance education. In IEEE international conference on management of innovation and technology (pp. 715–719).

[15] Tang, T., & McCalla, G. (2002). Student modeling for a web-based learning environment: A data mining approach. In Eighteenth national conference on artificial intelligence, Menlo Park, CA, USA (pp. 967–968).

[16] Hamalainen, W., Suhonen, J., Sutinen, E., & Toivonen, H. (2004). Data mining in personalizing distance education courses. In World conference on open learning and distance education, Hong Kong. Hammouda,

[17] Zaiane, O., & Luo, J. (2001). Web usage mining for a better web-based learning environment. In Proceedings of conference on advanced technology for education, Banff, Alberta (pp. 60–64).

[18] Hwang, W., Chang, C., & Chen, G. (2004). The relationship of learning traits, motivation and performance-learning response dynamics. Computers & Education Journal, 42(3), 267–287.

Prediction of the Scholarship Using Comprehensive Development

Wei, Wenjuan; Han, Jiaxin; Kong, Jie; Xia, Haiyang

01	ibnu triyuono	20/10/2019 8:58	Page 1
02	ibnu triyuono	20/10/2019 8:59	Page 1
03	ibnu triyuono	20/10/2019 9:00	Page 2
04	ibnu triyuono	20/10/2019 9:01	Page 2
05	ibnu triyuono	20/10/2019 9:01	Page 3
06	ibnu triyuono	20/10/2019 9:03	Page 3
07	ibnu triyuono	20/10/2019 9:05	Page 5