Study on College Student Credit Evaluation and Prediction Based on RF Algorithm

Jiong Mu, Lijia Xu, Haibo Pu
College of Information Engineering and Technology
Sichuan Agricultural University
China
lijiaxu13@163.com

Journal of Digital
Information Management

ABSTRACT: In the increasingly serious environment of employment of college students, the college student credit evaluation model taking into account of the "*basic personal situation*", "*on-campus situation*" and "*economic situation*" is employed in this paper for the purpose of improving the quality of college student credit education more efficiently, and in addition, a college student credit prediction mechanism based on the improved RF algorithm is put forward, seen from the test result, the accuracy of college student credit prediction of the algorithm is relatively high, and capable to make student credit education more targeted.

Categories and Subject Descriptors:
I.1.2 [Algorithms]: K.3 [Computers and Education]

General Terms: Random Forest Evaluation, E-Evaluation

Keywords: Credit Evaluation, Random Forest Algorithm, Prediction Mode

Received: 17 August 2013, Revised 9 October 2013, Accepted 18 October 2013

1. Introduction

Integrity and faithfulness is a fine tradition of Chinese civilization and the basic contents of the civic virtues. As an ethical requirement, faithfulness is the foundation of all morals and the precondition of harmonious campus building. However, the fates of college students are more uncertain as they are currently encountered with the severe challenges at the social transition period, particularly with the constant growing pressure of employment. During the period of drastic social changes, college students' judgment may deviate from faithfulness. Taking into account their own benefits, some college students may give up their ethics.

At present, the credit crisis among college students is seeing a rising trend , then, how to efficiently strengthen the credit education of college students and how to build up an effective college student credit evaluation system are urgent issues to settle. However, many researches only stay at the suggestions on credit evaluation system building, few can put forward operational plans or methods. Through the analysis on the factors impacting student credit, a complete prediction system for credit evaluation is put forward in this paper, trying to carry out advance screening of students that lack of credit so as to prevent the loss of credit to the largest extent.

2. Construction of College Student Credit Evaluation System

The learning situation and behaviors of college students and the way they get along with people are closely related to their consciousness of credit. When constructing the college student credit evaluation system, three procedures should be taken into account: firstly, the basic personal information, mainly including each student's name, age, education background, place of enrollment, economic conditions of his/her parents, etc; secondly, the on-campus situation, mainly including learning situation, behaviors and habits of faith-keeping and the situation that he/she gets on with teachers and other students, these reflect the student's knowledge and awareness of credit on some aspects; finally, the main source of income and consuming capability of each student at school, mainly including the financial supports from his/her families, his/her income from on-campus work, the amount of scholarship and stipend from the government, and the consumption situation of the student at school, etc . This procedure reflects the main economic conditions of the student, and it's also one of the direct factors to be considered in the credit evaluation system. Such student information was collected from each administration department, including such personal information as ideology, morality and behaviors, living and learning situation, and the information of organization and disciplinary and as well as credit consumption records, and track the credit situation of loan repayment. Such information and behavior records can be obtained from the student union, class teachers, student affairs office, dean's office and other administration departments, and screened strictly to ensure the objectivity and authenticity, and meanwhile, new information be added or modifications made in time to ensure the information in the credit evaluation system is true and integral so as to achieve fair, impartial and effective application.

3. Brief Introduction to RF

As for college student credit system evaluation and prediction, many researchers established prediction models by classification algorithm of machine learning. There are many classification algorithms of machine learning, and support vector machine, decision tree, neural network and among others are widely used . Random forest (RF) algorithm is a combined classifier algorithm put forward by Breiman in 2001. It adopts classification and regression tree (CART) as an element classifier and produces different discrepant training sample sets by bagging method, and randomly selects characteristics for attribute split of internal node when constructing a single tree. The combination of Bagging method and CART algorithm plus the attribute splitting make RF more noise tolerant and has higher classification performance. Breiman proved that there is an upper error limit in RF algorithm; therefore, RF algorithm is employed in this paper to predict the college student credit evaluation model composed of high-dimensional data.

The basic unit of RF algorithm is a decision tree, and the structure of the decision tree is decided by a random vector. RF is a classifier composed of multiple decision trees $\{h(x, \theta_k)\}$, therein, $\{\theta_k\}$ are mutually independent, and the vectors are in identical distribution. The final tag of the final input vector X depends on the comprehensive decision of all decision trees. When constructing k trees, k random vectors need to be generated, and these random vectors $\theta_1, \theta_2,...\theta_k$ are independent and in identical distribution. K classifiers $(h_1(X), h_2(X),..., h_k(X))$ and random vectors X and Y are given to define the edge function:

$$mg(X, Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \qquad (1)$$

Therein, $I(\bullet)$ is an indicator function. The edge function presents the degree that the vote of the correct classification Y of vector X exceeds the average vote of any other class. It can be seen that the greater the edge the higher the degree of confidence of classification.

Generalization error of classifier:

$$PE^* = P_{X,Y}(mg(X, Y) < 0) \qquad (2)$$

With the increase of quantity of trees, for any random vectors θ_k, PE^* shows the following trend:

$$P_{x,y}(P_\theta(h(x, \theta) = Y) - \max_{j \neq y} p_\theta(h(x, \theta) = j < 0) \qquad (3)$$

It indicates RF will not over-fit. It is a key feature of this algorithm, and the generalization error PE^* approaches to an upper limit with the increase of trees, therefore, RF has very good expansibility in case of unknown prediction.

The growth process of decision tree is shown below:

Steps of RF algorithm: (1) a training sample sub-set is built up by bootstrap method, and a binary recursive survival tree is established for each sample set by top-down recursion. (RF decision tree constructed by this method is a binary tree in a single structure). The split test on each decision node in the decision tree is generated in one random set, and the split is decided by the quantitative standard splitting threshold Δ. (3) Have the survival tree grow as much as possible, and each decision tree needn't pruning. Until the sample quantity of each end point isn't smaller than d_0 (> 0), the corresponding conclusion is achieved at each leaf node (tag). (4) Multiple paths from the root node to each leaf node of the decision tree compose multiple rules of classification.

4. Establishment of Prediction Model

4.1 Data Attributes

Each entry of the data set includes three procedures, 20 attributes in total, and the details are shown in Table 1.

The information of each college student provided in Table 1 is classified into three key categories: basic personal information which mainly reflects the objective social relations of the college student; "*On-campus situation*" mainly reflects the behaviors of the college student; "*Source of income*" reflects the financial income and expenditures of college student. The three aspects constitute a model for college student credit evaluation. In this model, some non-numeric attributes are numeralized after graded evaluation, and each entry is represented by a specific numerical value . The ages of college students are similar, 18-23 for undergraduates, 24-26 for postgraduates, and 27 and above for doctorial

Class	Test variables	Remark
Basic personal information	Age	1, 2 and 3 indicate undergraduate, postgraduate and doctor respectively
	Sex	0 and 1 indicate male and female respectively
	Education background	1: undergraduate; 2: postgraduate; 3: doctor; 4: post doctor
	Place of recruitment	0 and 1 indicate urban and rural respectively
	Working status of parents	2: Either parent works; 1: One parent works; 0: Neither parent works
	Economic condition of family	Classified by income: poor, medium, good, excellent
On-campus situation	Marital status of parents	0 indicates divorced, and 1 indicates not divorced
	Learning situation	According to result ranking: poor, medium, good and excellent
	Subjects retaken	Number of subjects retaken: 1, 2, 3, 4 (and above)
	Cheating in exam	0 indicates "*Yes*", and 1 indicates "*No*"
	Payment of tuition fees	Not paid off, paid off (but with delay), paid off on time
	Fulfillment of obligations and commitments	Not fulfilled, partially fulfilled, fulfilled passively, fulfilled positively
	Borrow and return Awards	No return, partial return, return upon urging, return on time Adding 3, 2 and 1 point(s) for national, provincial and school level awards respectively
	Punishment	Number of punishments recorded
Source of income	Supply from family (monthly)	RMB 300 (and below); 300-600; 600-900; above 900
	Income from on-campus work	1 and 0 indicate with income and without income respectively
	Stipend from the government (monthly)	Below RMB 100; 100-300; 300-500; above 500
	Scholarship	Below RMB 100; 100-300; 300-500; above 500
	Monthly expense	RMB 300 (and below); 300-600; 600-900; above 900

students. Therefore, the ages are classified into three ranges, 1, 2 and 3 represent undergraduates, postgraduate and doctorial students. In addition, subjective factors exist in some items, taking "*fulfillment of obligations and commitments*" for example, each student is evaluated by the comments of teachers and the class committee, and the average score of evaluation is taken as the evaluation result.

4.2 Normalization

After the nomalization of each attribute shown above, the difference between the values of attributes is relatively remarkable due to the difference of value ranges. To avoid deviation of analysis result caused by big difference in numeric values, each attribute value is normalized in advance. The maximum value in this item is divided by the actual score of each item provided in the table, thereby, the value of each attribute drops within [0,1].

5. Simulation Experiment

5.1 Data Source

According to the data required in the above established prediction model, some newly graduated students (graduated in 2007, 2008 and 2009, including undergraduates, graduates and doctorial students) from Sichuan Agricultural University and their current working units are chosen for questionnaire, 4,367 questionnaire forms are distributed and 3,672 of them were collected, and 2,116 with more complete information were selected, and subsequently, the credit situations of the students at school and recently graduated students are analyzed based on the feedback from the students' affairs office of the university, the working units of the graduated students and the lending banks, and integral information for 1,816 students is finally achieved. The learning information of the 1,816 students is obtained from the dean's office of the university, and the authenticity of their families, source of income, etc are checked. We take the complete information of the 1,816 entries as sample data of the experiment, and each data entry is taken as a complete record which contains 20 attributes and one tag, and the tag takes negative or positive value. In this model, positive sample indicates "*bad credit*" and negative sample indicates "*good credit*". Subsequently, the 1,816 samples are classified into two parts, one part is taken as the training set of classifier, the training set totally includes 902 sample entries, including 108 positive samples and

794 negative samples; the remaining samples are used as checking set which contains 914 samples in total, therein, 102 are positive and 812 are negative, and are used to validate the classification efficiency of the algorithm.

5.2 Selection of Standard Splitting Threshold and Parameter

In the RF algorithm, the splitting threshold Δ of leaf node has influence on both the classification effect and complexity of the algorithm, but it's also directly related to mixing ratio α. In this algorithm, after comprehensive consideration of the characteristics of samples and the accuracy of algorithm classification and through multiple experiment comparisons, it's found: when $\Delta = 0.31$ (corresponding to $\alpha = 0.2$), the accuracy of the algorithm is high at this time, and the cost is low compared with the established forest. Therefore, during the algorithm splitting process, the judgment condition selected for leaf node splitting is: $\Delta = 0.31$ is taken as the basis for leaf node splitting.

5.3 Measures for Sample Balance Treatment

The ratio of positive and negative samples for this survey is about 1: 8, and it's unbalanced. As each tree in the RF depends on the random vectors of independent samples with the same distribution, the final classification effect of RF algorithm may be affected if the ratio of positive and negative samples in the sample set is too big. Therefore, weighting method is adopted for pre-processing of the imbalance of samples to achieve balanced samples , and the positive samples and negative samples are multiplied by different weighting values respectively. If the ratio of positive and negative samples is $m : n$, then $\zeta = m / (m + n)$, and the weight of positive samples is selected as: $(1 - \zeta)$; and that of negative samples is as ζ.

5.4 Experiment Result

Comparative method is adopted in this experiment, the above data are used to compare with the experiment results by different algorithms, and typical classification algorithms are adopted in this paper: support vector machine (SVM) and K nearest neighbor (KNN) are compared with the improved RF algorithm.

Evaluation index adopts ROC curve which is a comprehensive index reflecting the continuous variables of sensitivity (SE) and specificity (SP). On the ROC curve, the true positive rate (SE) is the ordinate and the false positive rate (1 – SP) is abscissa. Theoretically speaking, the curve is a diagonal (opportunity line) drawn from the origin to the top right corner; ROC curve is usually located above the opportunity line, the further from the opportunity line the higher the accuracy of prediction; the area under ROC curve (Area Under Curve, AUC) can reflect the accurate degree of the diagnostic experiment, and the value of this index drops within 0.5 –1, and the value 0.5 – 0.7 indicates lower accuracy of prediction; 0.7 – 0.9 indicates intermediate accuracy; and value above 0.9 indicates higher accuracy.

Firstly, training and test are conducted with the above data by RF algorithm, and it's found: the prediction accuracy is relatively high if 62% positive samples are taken when 20% of the total number of students is achieved, and 84% positive samples are taken when reading 40% users. And the improved RF algorithm is superior to the original algorithm in respect of time complexity and spatial complexity.

The comparison results achieved by three algorithms are provided in Table 2, the accuracy of positive samples capturing is 89.67% and the area under ROC curve (Area Under Curve, AUC) is 0.88. The improved RF algorithm shows the best performance in this test, and the accuracy of positive sample capturing is as high as 95.38% and AUC (Area Under Curve) reached 0.95, this shows that the improved RF algorithm greatly improved the classification of positive samples, and indicates classifier constructed thereby has higher performance.

Classification algorithm	Accuracy (%)	AUC
KNN	91.25	0.9
SVM	89.67	0.88
IRF	95.38	0.95

Table 2. Testing result of each algorithm

6. Conclusion

The study result indicates the credit evaluation model is capable to provide a full image of the credit condition of students. The improved RF algorithm is capable to more effectively predict the credit of college students (among 20 participating factors), this indicates the evaluation model and algorithm are effective.

Seen from a deeper perspective, establishment of a standard and scientific college student credit system with strong operability is a key factor for improving the quality-oriented education of colleges and universities. Student credit can be predicted by "*personal information*", "*on-campus situation*" and "*economic situation*", for the students with "*bad credit*", an early-warning mechanism should be established to enhance the cultivation of students' consciousness of "*credit* ", the students should be managed, educated and guided properly, only by this way can the college students become faithful elites, and can they boost the establishment and development of the credit system in the entire society.

7. Acknowledgements

This paper was funded by provincial science and technology department. Project name: On the Elements of Successful Agricultural Enterprise E-commerce Implementation in Earthquake-stricken Area of Sichuan Provinceÿ2010ZR0134.

References

[1] Wang, Yanhong. (2012). The Study of College Students Credit Management Strategy in China. *Journal of Population and Economy*, (1) 170-171.

[2] Ma, Xin ., Wang, Xue., Yang, Yang (2012). Prediction of Degradation for Undergraduate Using Random Forest. *Journal of Jiangsu University of Science and Technology*, 26 (1) 86-90.

[3] Yuan, BenXin (2012). Construction and Counter measures of College Students Credit File - College Students Credit File System in Guangdong Province as an Example. *Journal of Ideology and Education Research*, (3) 14-16.

[4] Wu, Ran., Zhang, Yajing. (2012). Construction of College Students' Personal Credit System. *Journal of Everyone*, (9) 316-317.

[5] Lessmann, S., V O B, S. (2009). A Reference Model for Customer-Centric Data Mining with Support Vector Machines. *European Journal of Opertional Research*, 199 (12) 520-530.

[6] Breiman, L. (2001). Random Forests. *Journal of Machine Learning*, 45 (1) 5-32.

[7] Figini, S., Fantazzini, D. (2009). Radiom Survival Forests Models for SME Credit Risk Measurement. *Methodology and Computing in Applied Probability*, 11 (1) 29-45.

[8] Coussement, K. Poel D. V. D. (2009). Improving Customer Attrition prediction by Integrating Emotions form Client/Company Interaction Emails and Evaluation Multiple Classifiers. *Expert Systems with Applications*, 36 (3) 4626-4636.

[9] Wang, Aiping., Wan, Guowei., Chen, ZhiQuan., Li , Sikun (2011) . Incremental Learning Extremely Random Forest Classifier for Online Learning. *Journal of Software*, 22 (9) 2059-2074.

[10] Kubat, M., Matwin, S. (1997). Addressing the Curse of Imbakanced Training Sets:one-sided Selection. Place: San Francisco: Morgan Kaufmann Publishers.

[11] Burez, J., Poel, D. V. D. (2009). Handing Class Imbalance in Customer Churn Prediction. *Expert Systems with Applications*, 36 (3) 4626-4636.

[12] Bradley, A., P. (1997). The Use of The Area Under The ROC Cunre In The Evaluation of Machine Learning Algorithms. *Journal of Pattern Recogn*, (30) 1145-1159.

Author Biographies

Jiong Mu, Born in April 1971, Sichuan province, China. She graduated from Sichuan Normal University in 1993, and obtained a master's degree in computer applications from Sichuan University in 2007. Now She is an Associate Professor in Sichuan Agricultural University. Her main research is about computer used in e-commerce, intelligent algorithm etc.

Lijia Xu. Born in December of 1973, Sichuan province, China. She graduated from Sichuan Engineering Institute in 1996, and then obtained a master degree from Beijing University of Technology in 2004. She got a doctor degree in automation engineering in University of Electronic Science and Technology of China in 2009. Now she is a Professor in Sichuan Agriculture University. Her main research interests include intelligent signal processing, intelligent algorithm and fault diagnosis etc.

Haibo Pu, Born in June of 1973, Sichuan province, China. He Graduated from Chengdu University of Technology in 1996. Now he is a Lecturer in Sichuan Agriculture University. His main research interests include wireless sensor network, embedded system and intelligent algorithm etc.