Student Retirement Analysis Using Decision Tree Techniques

Pattariya Supaudon[1], Nipaporn Chutiman[2], Bungon Kumphon[3*]

Received: 18 February 2013 Accepted: 15 June 2013

Abstract

This study presents the work of data mining in predicting the retirement feature of students by applying decision tree technique to choose the best model for prediction. Three widely used measures the quality of tree are recall rate, precision and F-measure. The results show 94.96 % of correction to predict the student's retirement.

Keywords: reitrement, data mining, decision tree, rule

Introduction

Data mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data[1]. Decision tree analysis is a popular data mining technique that can be used in many areas of education. It offers comprehensive characteristics analysis of students and contains rules to predict the target variables[2]. One critical question in any educational institution is the following What are the risk factors or variables that are important for predicting the results (pass/fail) of students?

Although many risk factors that affect results are obvious, subtle and non-intuitive relationships can exist among variable that are difficult, or impossible to identify without applying more sophisticated analysis.

Modern data mining models such as decision trees can more accurately predict risk than current models, educational institutions can predict the results more accurately, which in turn can result in quality education. An indicator of potential weaknesses in the higher education system may be a large number of dropouts in the first years of studies. The strategic goal of educational institute should therefore be planning, management and control of education processes with the purpose of improving the efficiency of studying. The retirement trends have to be recognized and the causes (course, previous knowledge, assessment) isolated. Also, the typical student profile is to be determined in order to plan the number of potential students in lifelong learning programs or those that need additional motivation. It is possible to follow the retirement trend throughout several years in order to check the effectiveness of corrective activities.

Graduation, especially timely graduation is an increasingly important policy issue[3]. College graduates earn twice as much as high school graduates and six times as much as college dropouts[4] (Murphy and Welch, 1993)[4]. In addition to the financial rewards, the spouses of college graduates are more educated and their children do better in schools and colleges. Graduation rates are considered as one of the institutional effectiveness[4]. Student's retirement due to different reasons; academic trouble, academic preferences, their financial position, parental income, parent occupation, grade at the first year[5,7]. (Pattarapong (2010), Aorathai (2007), Kao and Thomson (2003)). The remainder of this paper is organized as follows. Section 2 discusses data and methodology of decision tree. Section 3 presents the data analysis, and some conclusions are stated in the last Section.

[1,2,3] *Mahasarakham University Mathematics Department Science Faculty Maha Sarakham, Thailand*

E-mail: pattariya_planoi@hotmail.com

** Coressponding author: E-mail: kbungon@gmail.com*

Data and Methodology

The secondary data, with ten variables as in Table 1, were employed from the registrar section, Chalerm Phrakiat Sakhon Nakhon Provience Campus, Kasetsart University. The sample of 7,333 students during 2004 – 2011 academic years was split into two groups as training data set and testing data set. The four ratios between training data set (5,833 students) and testing data set (914 students),X, were 50:50 (X_1), 60:40 (X_2), 80:20 (X_3) and 90:10 (X_4).

Table 1 The influence variables

Variable	*Type*	*Description*
Faculty (Fc)	nominal	A = Natural resource and Agro-industry B = Science and Engineering C = Liberal Arts and Management Sciences
Type of entrance (Te)	nominal	1 = University self-admit 2 = Entrance
Gender (Ge)	nominal	Male, Female
Father occupation (Fo)	nominal	0 = not indicate 1 = Government service 2 = State enterprise 3 = Employee 4 = Business 5 = Agriculture
Mother occupation (Mo)	nominal	
GPAX from high school (Gh)	ordinal	1 = less or equal 2.00 2 = 2.01 – 2.50 3 = 2.51 – 3.00 4 = 3.01 – 3.50 5 = greater or equal 3.51
GPAX at the first semester (Gf)	ordinal	
Parent relationship (Pr)	nominal	1 = stay together 2 = separate, divorce 3 = father or mother deceased 4 = father and mother deceased 5= not indicate
Scholarship (Sc)	nominal	0 = no 1 = yes
Student status (Ss)	nominal	0 = retire 1 = not retire (NR)

Decision trees are a highly flexible modeling technique. For instance, to build regression models and neural networks models, the missing values have to be inserted into training data while decision trees can be built even with missing values. Decision trees are intended for the classification of attributes regarding the given target variable. Decision trees are attractive because they offer, in comparison to neural networks, data models in readable, comprehensible form – in fact, in the form of rules. They are used not only for classification but also for prediction. The tree techniques provide insights into the decision making process as shown in Fig. 1. This model, make use of the software Weka the J4.8 algorithm (J4.8 implements a later and slightly improved version called C4.5) for predictive data mining. The condition to choose the attribute or variable in the tree, for the first node, is $max\ \{information\ gain\}_i$ where i is the number of attributes. Then the second value would be the second node, respectively. Sometimes, the over fitting can occur because of the complicated nodes and branches in the tree or the small size of training data set or creating decision rules that work accurately on the training set based on insufficient quantity of samples. Pruning tree (reduced error pruning: REP) or clustering are used to solve that problem.

Data Analysis

The parameters in this study are binary splits, number of folds (N) and the number of leaf (M) as M=2, 4, 6; N = 3, 4, 6. Defined T_iX_j; i and j=1, 2, 3, 4 as

T_1 is model building with REP, binary, M=2, N=3,

T_2 is model building with REP, not binary, M=4, N=6,

T_3 is model building with REP, not binary, M=2, N=3,

T_4 is model building with REP, not binary, M=6, N=4,

X_j's are the four ratio set of data as mentioned in Section 2. So, sixteen models were studied to predict the drop out of the student.

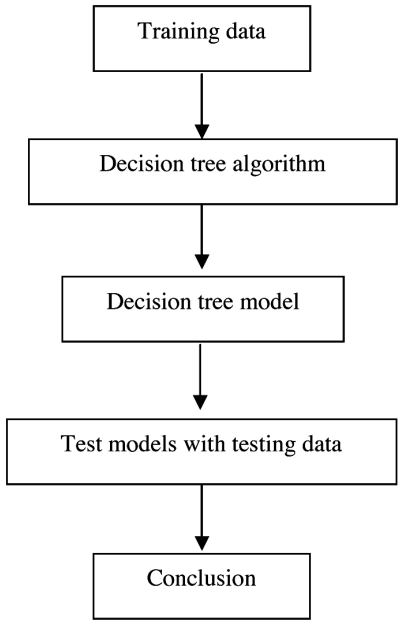

Figure 1 Decision tree diagram

The recall rate (R), precision (P) and F-measure are three widely used measures for finding the quality of tree which can define as

$$R = \frac{C' \cap C}{C}, \quad (1)$$

$$P = \frac{C' \cap C}{C'}, \quad (2)$$

$$F = \frac{2PR}{P+R} \quad (3)$$

where C is the set of samples in the class and C' is the set of samples which the decision tree puts into the class. Table 2 shows the results of the quality for decision tree. Two candidates model --- viz. T_2X_1 and T_4X_1 with the best results of R, P and F are considered. T_4X_1 is selected as a predictive model because of the smaller size of tree compared to another. The best tree is Show in Figure 2 and Table 3 is a testing result compared to the true data (testing data set). The benefit is the rule to predict the student's retirement with 94.96% of correction. For example, if Gf > 2.01 imply that "not retire". If Gf ≤ 2, not get scholarship, Fc = A, Mo>4, Gh > 3 and Te ≤ 1 mean "retire".

Table 2 Model and the measures of quality for decision tree for training data set

Model	Precision		Recall		F-measure	
	Retire	NR	Retire	NR	Retire	NR
T_1X_1	64.1	84.4	57.3	87.2	51.7	90.0
T_1X_2	65.0	84.2	52.5	90.0	58.1	87.0
T_1X_3	63.5	85.2	55.2	89.1	59.0	87.1
T_1X_4	57.3	84.7	52.1	87.2	54.5	86.0
T_2X_1	66.7	84.8	52.7	91.0	58.9	87.8
T_2X_2	61.5	87.6	65.9	85.4	63.6	86.5
T_2X_3	64.8	85.4	55.5	89.6	59.8	87.5
T_2X_4	56.3	84.8	52.8	86.6	54.5	85.7
T_3X_1	64.8	84.4	51.5	80.4	57.4	87.3
T_3X_2	64.8	84.2	52.5	89.9	58.0	87.0
T_3X_3	62.3	83.4	48.2	90.0	54.3	86.6
T_3X_4	56.8	83.4	46.5	88.4	51.1	85.8
T_4X_1	67.8	84.2	49.9	91.8	57.5	87.9
T_4X_2	64.5	85.7	58.0	88.7	61.1	87.1
T_4X_3	63.4	85.7	56.9	88.7	60.0	87.2
T_4X_4	60.2	83.8	47.2	89.7	52.9	86.7

Table 3 The measures of quality for decision tree for testing data set

correc-tion	Recall		Precision		F-measure	
	Retire	NR	Retire	NR	Retire	NR
94.96	25.00	99.20	65.00	95.60	36.10	97.40

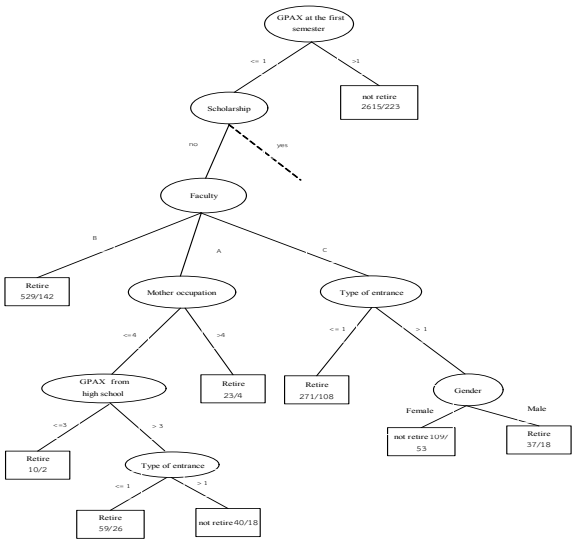

Figure 2 show the tree diagram for analysis.

Conclusion

This study introduced the data mining approach to modeling retirement feature and some implementation of this approach. The key to gaining a competitive advantage in the educational industry is found in recognizing that student databases, if properly managed, analyzed and exploited, are unique, valuable assets. The obtained data should, in the earliest stage, be used to raise awareness on the possibilities and need to use the data mining models and methods at the institution in which this research has been carried out. Data mining uses predictive modeling, database segmentation, market basket analysis and combinations to more quickly answer questions with greater accuracy.

The future research will be directed towards the design of an applicative solution to allow observation and classification of each student at the university into a particular retirement and dropout category depending on his/her characteristics. The fuzzy module based on certain attributes of students' previous knowledge is the interesting idea.

Acknowledgment

Financial support was provided by the graduate studies, Mahasarakham University. The authors also thank anonymous reviewers for their constructive comments and suggestions.

References

1. Quadril, M. N. and Kalyankar, N. V. (2010) Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques. Global Journal of Computer Science and Technology, Vol. 10 Issue 2 (Ver 1.0), April 2010 p. 2-5.
2. Shyamala, K. and Rajagopalan, S.P. (2007) Mining Student Data to Charecterisze Drop out Feature using Clustering and Decision Tree Technique. International Journal of Soft Computing 2(1), p. 150-156.
3. DesJardins, S.L., D.A. Ahlburg and B.P. McCall, 2002. A temporal investigation of factors related to timely degree completion. J. Higher Education, 73:555-581.
4. Murphy, K and F. Welch, 1993. Inequality and relative wages. Ameri. Economic review, 83: 104-109.
5. P.Pongpatarakant. Factors Anaslysis of Undergraduate Student's Retirement Using Filtering by Committee Machine. Computer Science, Loei Rajabhat University. Prooeeding of NCCIT2010, 2010.
6. L. Aorthai and et al,2007 "A Study of effecting on Dismissal of The Ubon Rajathanee University, *(references)*
7. Kao Grace and Jiennifer S Thomson (2003).Racial and stratification in educational achievement and attainment. Annaual Review of Sociology,29:417-442.