Analysis of Graduation Thesis Information Based on Decision Tree

Mengyi Li[1, a], Jiuru Dai[2,b] and Zhigang Zhang[1,c]

[1]Department of Mathematics and Physics, University of Science and Technology Beijing

30 Xueyuan Road, Beijing, 100083, China

[2]School of Mathematical Sciences, Peking University, 5 Yiheyuan Road, Beijing, 100871, China

[a]lmylw1120@163.com, [b]daijiuru@sina.com, [c]zzgcyf@263.net

Keywords: graduation thesis achievements, multivariate statistical analysis, decision tree.

Abstract. In order to improve the quality of graduation thesis, many colleges and universities carried out reforms on the thesis work according to their own actual situation. Through data mining method, this paper explores the main factors affecting the result of graduation thesis, and then classifies and predicts the achievements of graduation thesis, so as to give reasonable proposal and more scientific reform.

Introduction

The achievements of graduation thesis is a visual representation of quality, so the study of factors affecting thesis achievements has a vital role to improve the quality of thesis. This paper will use data mining methods to explore the main factors affecting graduation thesis achievements, further classify the achievements and predict the performance of new students graduation. Then extra attention and guidance should be given to the students with low predicted achievements.

Research Thought and Process

This paper takes the data of students who are in Department of Mathematics and Physics of Beijing University of Science and Technology (USTB) as an example to research (the grade includes 08, 09 and 10). The first step is data preprocessing. Through data collection, induction, cleaning and transforming, a data table which can be used in data mining can be gained. Making the basic statistical analysis for data, the main factors that affect graduation thesis achievements will be gained preliminarily. Then use multivariate statistical analysis method to do further research.

Next, classify and predict graduation thesis achievements by decision tree in data mining. This paper uses MATLAB and SPSS to accomplish six kinds of decision tree algorithm, including ID3, C4.5, CART, CHAID, Exhaustive CHAID and QUEST. Finally, compare the accuracy and efficiency of different algorithms.

In summary, according to the main factors that affect graduation thesis achievements and the classification rules from data mining, colleges can strengthen the management of controllable factors and get the basis of teaching reform, so as to improve the quality of graduation thesis.

Data Preprocessing and Multivariate Statistical Analysis

Data Preprocessing. Data preprocessing includes data collection, induction, cleaning and transforming.

Data collection: mainly includes the data of the students, teachers and graduation thesis. Specific factors are: students' grade, major, gender, college entrance examination scores, weighted average scores, teacher title, teacher education, teacher age, student employment, innovation credits and graduation thesis achievements.

Data induction: taking the student id as the retrieval attribute, the collected data can be arranged to be a complete data table.

Data cleaning: filling the vacancy of the data, cleaning up the noise of data source to make the data keep consistent.

Data transforming: transforming the text information into digital information, the continuous data into discrete data, in order to fit for data mining.

Multivariate Statistical Analysis. Based on the preliminary data table, using the multivariate statistical analysis method to study, including hypothesis testing, variance analysis, correlation analysis, regression analysis and factor analysis. This paper only takes hypothesis testing and variance analysis as examples.

Hypothesis testing: firstly, the null hypothesis and alternative hypothesis are put forward, and test statistics can be constructed to obtain the rejection region and acceptance region, and then judge the correctness of the null hypothesis. With the help of SPSS software, it can be verified in turn if different factors have significant impact on graduation thesis achievements. The result shows that the main factors are major, gender, teacher title and weighted average score.

Variance analysis: for the significance test of two and more than two samples' mean. Use the ratio of square sum within groups and between groups divided by the respective degree of freedom to construct F distribution. Compare the F value and the critical value to draw a conclusion. With the help of SPSS software, the result shows that the main factors are major, gender, teacher title and weighted average score, which is same as the result of hypothesis testing.

Decision Tree

Decision tree is a tree structure, and its generation algorithm is divided into two steps: the first step is the tree generation, which involves repeated splitting the training set. The second step is the tree pruning, that is removing some data which may be noise or unusual.

ID3 algorithm chooses the attribute which has maximum information gain value in current sample set as the test attribute. The information gain value is higher, the uncertainty is less. The average depth of decision tree to generate is minimum [1]. The calculation method of information gain value [2] is as follows:

Supposing *S* is the set of *s* samples, the class attribute has *m* different values. Define *m* different classes $C_i(i=1,\cdots,m)$, and suppose the sample number of class C_i is s_i. For a given sample set, its total information entropy is:

$$I(s_1,s_2,\cdots,s_m)=-\sum_{i=1}^{m}P_i\log_2(P_i) \tag{1}$$

In which, P_i is the probability that arbitrary sample belongs to C_i, $P_i=\frac{s_i}{s}$ [3].

Suppose attribute *A* has *v* different values. *S* can be divided into *v* subsets $\{S_1,S_2,\cdots,S_v\}$ by attribute *A*, in which, S_j includes the samples in *S* that have value a_j on *A*. If choosing *A* as testing attribute, these subsets are corresponding to new leaf nodes growing from the nodes of representative sample set *S*. Supposing s_{ij} is the sample numbers whose class is C_i in subset S_j, then the information entropy of samples divided by *A* is:

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\cdots+s_{mj}}{s}I(s_{1j},\cdots,s_{mj}) \tag{2}$$

In which, $I(s_{1j},\cdots,s_{mj})=-\sum_{i=1}^{m}P_{ij}\log_2(P_{ij})$, $P_{ij}=\frac{s_{ij}}{|S_j|}$ is the probability that the class of samples is C_i on S_j. Finally, using attribute *A* to divide sample set *S*, the information gain value is $Gain(A)=I(s_1,s_2,\cdots,s_m)-E(A)$ [4].

C4.5 algorithm is different from ID3, which chooses testing attribute based on information gain ratio. The information gain ratio is equal to the ratio of information gain value and split information. C4.5 algorithm improves the disposal of continuous attributes and the vacancy of attribute value, also has carried on the processing to tree pruning [5]. The calculation method of information gain ratio is as follows:

For sample set *T*, supposing *A* is a discrete attribute with *s* different values, the algorithm that uses *A* to divide sample set to obtain information gain value is same as ID3. The split information is:

$$Split(T) = -\sum_{i=1}^{s} \frac{|T_i|}{|T|} \times \log_2(\frac{|T_i|}{|T|}) \tag{3}$$

According to the method above, solve the information gain ratio of all attributes in the current candidate set of attributes, and compare them to find the attribute with maximum information gain ratio.

Beside ID3 algorithm and C4.5 algorithm, the decision tree algorithms used in this paper include CART, CHAID, Exhaustive CHAID and QUEST. They are accomplished by MATLAB and SPSS software.

Generating Decision Tree by MATLAB. First of all, use MATLAB to accomplish ID3 algorithm. The decision attributes and classification attribute are shown in Table 1.

Table 1 The Decision Attribute and Classification Attribute of ID3 Algorithm

Decision Attributes	Major (M)	Gender (G)	Teacher Title (TT)	Weighted Average Score (WAS)
Classification Attribute	Graduation Thesis Achievements (GTA)			

The data table to generate decision tree has 455 records, supposing C_1 is corresponding to [0,80) of graduation thesis achievements, which has 96 students, C_2 is corresponding to [80,90) of achievements, which has 288 students, C_3 is corresponding to [90,100] of achievements, which has 101 students. The information entropy of classification attribute is: $I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} P_i \log_2(P_i) =$

$-\frac{90}{455}\log_2(\frac{90}{455}) - \frac{269}{455}\log_2(\frac{269}{455}) - \frac{96}{455}\log_2(\frac{96}{455}) = 1.3843$. And then by the formula (2) to calculate entropy of each attribute: when $A = M$, $v = 3$, $S_1 = 151, S_2 = 150, S_3 = 154$, so: $H(M) = \frac{151}{455}(-\frac{18}{151}$ $\log_2\frac{18}{151} - \frac{97}{151}\log_2\frac{97}{151} - \frac{36}{151}\log_2\frac{36}{151}) + \frac{150}{455}(-\frac{29}{150}\log_2\frac{29}{150} - \frac{85}{150}\log_2\frac{85}{150} - \frac{36}{150}\log_2\frac{36}{150}) + \frac{154}{455}(-\frac{43}{154}\log_2\frac{43}{154}$ $-\frac{87}{154}\log_2\frac{87}{154} - \frac{24}{154}\log_2\frac{24}{154}) = 1.3612$.

In the same way, $H(G) = 1.3562$, $H(TT) = 1.3648$, $H(WAS) = 1.1499$. Therefore, the information gain value of each attribute can be calculated: $\text{Gain}(M) = 0.0231$, $\text{Gain}(G) = 0.0281$, $\text{Gain}(TT) = 0.0195$, $\text{Gain}(WAS) = 0.2344$, so the root node is weighted average score.

According to the algorithm above, calculate the node below root node in turn. The final decision tree is shown in Fig.1.

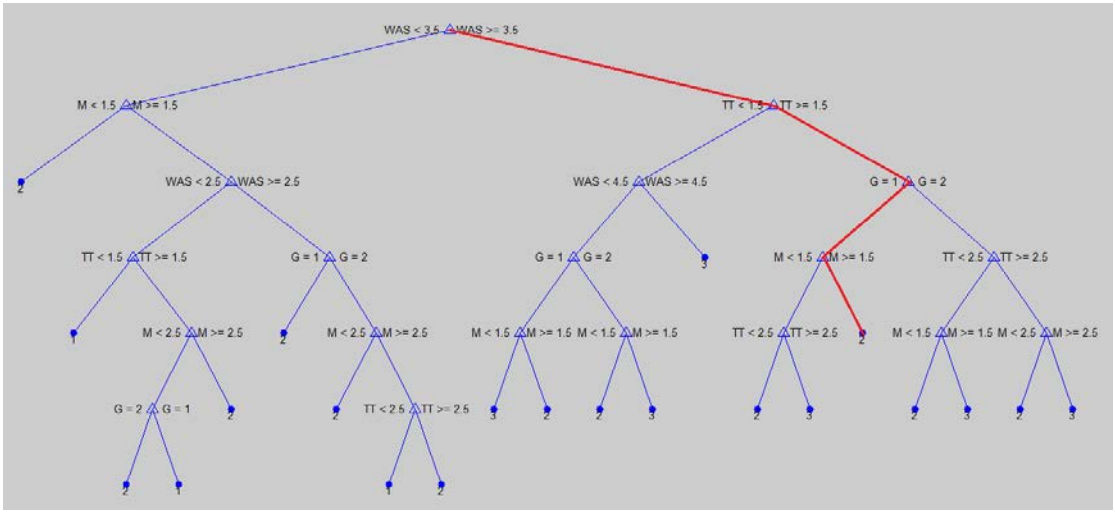

Fig.1 The Decision Tree Generated by ID3 Algorithm

Prediction: supposing a student majors in information and computing science, whose gender is male, teacher title is associate professor, weighted average score is in [80,90), that is, M=3, G=1, TT=2, WAS=4. Then according to the red line, the predicted graduation thesis achievements is 2, which is in [80,90).

By the decision attribute and classification attribute in Table 1, using C4.5 algorithm to calculate information gain ratio of each attribute, and taking the attribute with maximum information gain ratio as the branch node to generate decision tree.

When $T = M$, $s = 3$, $|T_1| = 151$, $|T_2| = 150$, $|T_3| = 154$, then: $Split(M) = -\frac{151}{455}\log_2\frac{151}{455} - \frac{150}{455}\log_2\frac{150}{455} - \frac{154}{455}\log_2\frac{154}{455} = 1.5849$.

In the same way, $Split(G) = 0.8631$, $Split(TT) = 1.5729$, $Split(WAS) = 1.4408$. Therefore, the information gain ratio of each decision attribute can be calculated: $Gain.ratio(M) = \frac{0.0231}{1.5849} = 0.0146$, $Gain.ratio(G) = \frac{0.0281}{0.8631} = 0.0326$, $Gain.ratio(TT) = \frac{0.0195}{1.5729} = 0.0124$, $Gain.ratio(WAS) = \frac{0.2344}{1.4408} = 0.1627$, so the root node is weighted average score.

According to the algorithm above, calculate the node below root node in turn to generate decision tree.

Finally, CART algorithm chooses attribute of each node by calculating GINI coefficient of each decision attribute. Using MATLAB can generate decision tree and predict.

Generating Decision Tree by SPSS. CHAID, Exhaustive CHAID, CART and QUEST algorithms of decision tree can be accomplished by SPSS.

CHAID: chi-square automatic interaction test. In each step, it selects independent variables (predictor variables) that have the strongest interactions with the dependent variable. If the category of each predictor variable is not significantly different from the dependent variable, it will merge these variables.

Exhaustive CHAID: improved CHAID algorithm. It tests all possible separations of each predictor variable.

CART: a complete binary tree algorithm. It splits the data into several homogeneous segments with the dependent variable as much as possible. In all cases, the terminal nodes in which the dependent variable values are same are homogeneously “pure” nodes.

QUEST: a kind of fast, unbiased and effective decision tree. It can avoid the bias that other methods may produce in face of predictor variables with lots of categories. Only when the dependent variable is a nominal variable the QUEST can be appointed.

Taking the CHAID algorithm as an example, according to the principle of chi-square automatic interaction test, firstly, the weighted average score is selected as the root node because it has maximum information gain value in four predictor variables. Then divide it into five child nodes which respectively denotes the score is in [0,70), [70,75), [75,80), [80,85) or [85,100]. Next, according to each child node, select one attribute with maximum information gain value in major, gender and teacher title as the name of the child node, and then divide it in terms of the attribute value until the leaf node.

Because the figure of final decision tree generated by SPSS is too big, only the decision tree below the first child node is shown in Fig.2.

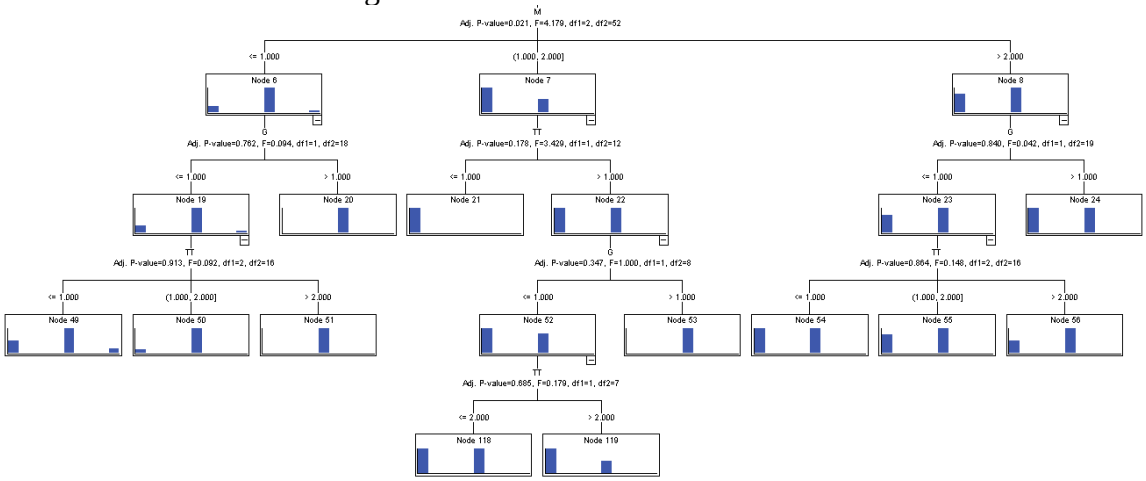

Fig.2 The Decision Tree below the First Child Node Generated by CHAID Algorithm

Comparison of Different Algorithms. Table 2 shows the prediction accuracy, classification accuracy and efficiency of each algorithm.

Table 2 The Comparison of Different Algorithms

Software	MATLAB			SPSS			
Algorithm	ID3	C4.5	CART	CHAID	Exhaustive CHAID	CART	QUEST
Prediction Accuracy	70.0%	73.3%	70.0%	70.0%	73.3%	63.3%	60.0%
Classification Accuracy				67.7%	67.3%	65.3%	64.8%
Efficiency	1.056s	0.070s	0.028s				

On the prediction accuracy, C4.5 algorithm and Exhaustive CHAID algorithm have the highest level, 73.3%, and the QUEST algorithm is the lowest, which is 60%.

On the efficiency, the time of ID3 algorithm, C4.5 algorithm and CART algorithm to generate decision tree are respectively 1.056s, 0.070s and 0.028s, so CART has the highest efficiency and ID3 has the lowest efficiency.

On the simplicity of description, the trees generated by ID3, C4.5, CHAID and Exhaustive CHAID are easier to describe. The IF-THEN rule is formed on every path from the root node to leaf node, and the decision attribute appears only once.

On the robustness of the model, the robustness is a complement of model prediction accuracy, which is the ability of classifying data in the presence of noise and data vacancy. Because the ID3 algorithm can't handle the null value and continuous data, its robustness is the worst.

On the processing scale, the CART algorithm is more suitable for large-scale data due to its high efficiency and fast computing speed.

Summary

This article mainly uses the decision tree method of data mining to explore the main factors that affect graduation thesis achievements, as well as to classify and predict graduation thesis achievements, and obtains good results.

There are two reasons why the prediction accuracy is in 60%-80%. On the one hand, some data information of decision attributes is not obtained, such as the time students spend on the graduation thesis. On the other hand, the data mining method is only used one kind. Therefore, the prediction accuracy will be improved if more information of decision attributes can be considered and various data mining methods are used.

Main factors that affect graduation thesis achievements can be found through data mining, so as to strengthen education and supervision on these factors. Also, the decision tree can classify and predict graduation thesis achievements, and teachers should pay more attention and guidance to the students whose predicted achievements are not good.

Acknowledgements

This article is funded by the Beijing University of Science and Technology Education Research Project (JG2015M60).

References

[1] Ming Yang and Zaihong Zhang: Microcomputer Development Vol. 5 (2002), p. 6-9.

[2] Liming Wang: Research on Decision Tree Induction and Pruning Algorithm (Wuhan University of Technology, 2007).

[3] Aihui Huang and Xiangtao Chen: Computer Engineering and Science Vol. 31, No. 6 (2009), p. 109-111.

[4] Jin Chen, Delin Luo and Fenxiang Mu: Proceedings of 2009 4th International Conference on Computer Science and Education, p.127-130.

[5] Qiuhua Xie: Journal of Sanming University Vol. 29, No. 4 (2012), p. 34-39,100.