Application of C4.5 algorithm in graduate enrollment

Huang Juan

Shanghai University of Engineering Science, Shanghai 201620, China.

jnhj3032@163.com

Keywords: C4.5 algorithm, Decision Tree, Enrollment, The rate of first wish

Abstract. In order to improve the rate of first wish, we propose that the C4.5 decision tree classification algorithm is applied to a college graduate enrollment. The candidate information is processed, decision attributes are selected and decision tree is constructed. From extracting rules, it can acquire the relationship between candidates first wish, hometown area, candidates sources, the first test scores and graduate institution. The result shows that the algorithm can classify graduate institution well. It can support admissions officers to more effectively develop brochures and target for admission propaganda, thereby raising the rate of first wish.

Introduction

In graduate enrollment process, due to limitations school level, academic settings, whether the doctoral and other factors, some colleges and universities may appear the low phenomenon of the rate of first wish. Therefore, on the basis of ensuring the quality of enrollment, how to improve the rate of first wish has become some colleges and universities concerned content.

In this paper, combining with the data mining technology, select C4.5 decision tree algorithm to analyze M University swap candidates [1] . The results of the analysis can assist M University clear propaganda focus, effectively launch advocacy work, and ultimately achieve the target of improving the school's the rate of first wish.

C4.5 algorithm Introduction

C4.5 algorithm is based on the well-known classification algorithm ID3 algorithm improved a more complete decision tree classification algorithm, is a classic decision tree algorithm[2-4].

C4.5 algorithm

The main processing steps of C4.5 algorithm are as follows:

(1) The information entropy of categories

Let S be a set of training sample, which contains s training samples, and there are a total of m sample class Ci , (i=1,2,…,m), let si be the number of samples that Ci class is in the S set. For a given sample classification desired information as required is shown in equation (1) [5,6]:

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

which pi is the probability of the sample belonging to Ci, usable si / s estimation.

(2) The conditional entropy of categories

Let attribute A has v different values {a1, a2 , ⋯, av}, the set S can be divided into v subsets by attribute A {S1, S2,⋯, Sv}, which the set Sj is a subset of the set S, the subset Sj samples have the same value aj on the attribute A （j=1,2,⋯,v）. A classification based on the attribute A required for the conditional entropy is shown in equation (2) [5,6]:

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

which sij is the number of samples belonging to class Ci in a subset Sj,

© 2015. The authors - Published by Atlantis Press

$I(s_{1j},\ldots,s_{mj}) = -\sum_{i=1}^{m} P_{ij}\log_2(P_{ij})$.

$P_{ij} = \frac{s_{ij}}{|s_j|}$ is the probability of the sample belongs to class C_i, which the sample is the element of subset S_j.

(3) Information Gain

The information gain of attribute A is shown in equation (3) [5,6]:

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A) \quad (3)$$

(4) Information Gain Ratio

The information gain ratio is shown in equation (4) [5,6]:

$$Gainratio(A) = \frac{Gain(A)}{I(s_1, s_2, ..., s_v)} \quad (4)$$

which v is the number of branch for the node, si is the number of records for the i-th branch.

Tree pruning

In the process of creating the decision tree, due to data noise and other reasons, may generate abnormal branching. Therefore, tree pruning is proposed. After tree pruning, the decision tree can improve the correct classification ability.

There are two common tree pruning method: One is first pruning (prepruning); another is after pruning (postpruning) [7-9].

Establish decision tree using C4.5 algorithm

Data preprocessing

The data used in this paper is the enrollment and admissions data of M university for one year. Based on the previous research objectives, the swap candidates of N college are used as training data set to extract rules. After initial processing, candidates graduated college, the first wish of candidates, the code of Candidates Birthplace province, Candidates source, the first test score are applied to classify the candidates. And the above five attributes are numbered as a1, a2, a3, a4, a5. The discretization results of specific attributes as follows:

The discretization of candidates graduated college (No.a1) and the first wish of candidates (No.a2) are shown as follows: Ministry of Education, 985, 211, Ordinary colleges;

The discretization of the code of Candidates Birthplace province (No.a3) is shown as follows: North, Northeast, East, Central, South, West, special region;

The discretization of Candidates source (No.a4) is shown as follows: fresh graduate, within 3 years after graduation, within 5 years after graduation, more than 5 years;

According to Equation (5), the first test score (No.a5) is converted into a standard score. Then the first test score is divided into A, B, C, D and E five grade.

$$Z\_grade = \frac{\text{first test score-the average of first test score}}{\text{standard deviation of first test score}} \quad (5)$$

After the discretization, the duplicate objects are deleted, then the data preprocessing has been finished.

Combined the research topic of this article, attribute Candidates graduated college (No. a1) is the classification attribute. And others attributes that may have a potential impact on classification attribute as the decision attribute [10].

The establishment model of decision tree

The decision tree model of Candidates graduated college (No.a1) is generated by the C4.5 algorithm while data preprocessing is completed. First, the required information entropy for a given sample classification is calculated according to equation (1):

$$I(3,2,1,5,46)$$
$$= -\frac{3}{57}\log_2\frac{3}{57} - \frac{2}{57}\log_2\frac{2}{57} - \frac{1}{57}\log_2\frac{1}{57} - \frac{5}{57}\log_2\frac{5}{57} - \frac{46}{57}\log_2\frac{46}{57}$$
$$= 1.0530939$$

Then calculate the information gain ratio of each decision attribute. Take the first wish of candidates (No.a2) for example. Expects entropy can be calculate separately when it is 211, 985, Ordinary colleges, Ministry of Education, and draw its information gain ratio finally.

1） When a2 is 211, by the equation (3):

$$I(0,0,0,0,1) = -\frac{1}{1}\log_2\frac{1}{1} = 0$$

2） When a2 is 985, by the equation (3):

$$I(0,1,0,1,10) = -\frac{1}{12}\log_2\frac{1}{12} - \frac{1}{12}\log_2\frac{1}{12} - \frac{10}{12}\log_2\frac{10}{12} = 0.8166891$$

3） When a2 is Ordinary colleges, by the equation (3): $I(0,0,0,0,0) = 0$

4） When a2 is Ministry of Education, by the equation (3):

$$I(3,1,1,4,35)$$
$$= -\frac{3}{44}\log_2\frac{3}{44} - \frac{1}{44}\log_2\frac{1}{44} - \frac{1}{44}\log_2\frac{1}{44} - \frac{4}{44}\log_2\frac{4}{44} - \frac{35}{44}\log_2\frac{35}{44}$$
$$= 1.0894363$$

Thus, desired information entropy of decision attribute a2 can be calculated by the equation (2):

$$E(\text{a2}) = \frac{1}{57}I(0,0,0,0,1) + \frac{12}{57}I(0,1,0,1,10) + \frac{0}{57}I(0,0,0,0,0) + \frac{44}{57}I(3,1,1,4,35)$$
$$= 1.0129029$$

Information gain of decision attribute a2 can be calculated by equation (3):

$$Gain(\text{a2}) = I(3,2,1,5,46) - E(\text{a2}) = 0.040191$$

$$Spliti(\text{a2}) = -\frac{1}{57}\times\log 2\frac{1}{57} - \frac{12}{57}\times\log 2\frac{12}{57} - \frac{44}{57}\times\log 2\frac{44}{57} = 0.86386$$

Information gain rate of decision attribute a2 can be calculated by equation (4):

$$Gainratio(\text{a2}) = \frac{Gain(\text{a2})}{Spliti(\text{a2})} = 0.0465$$

Information gain rate of the code of Candidates Birthplace province (No. A3), candidates source (No. A4) and the first test score (No. A5) is 0.1477, 0.0999 and 0.1249 according to the equation (1) to (5).

Therefore, the information gain rate of decision attribute A3 is maximum, so the attribute as the root node of the decision tree, and each attribute value leads to a branch. Each branch nodes are further divided according to the equation (1) ~ (5) respectively until the decision tree construction completed.

When the decision tree construction completed, this paper adopts the post pruning method in order to prevent over training. The pruned decision tree is shown in fig 1.

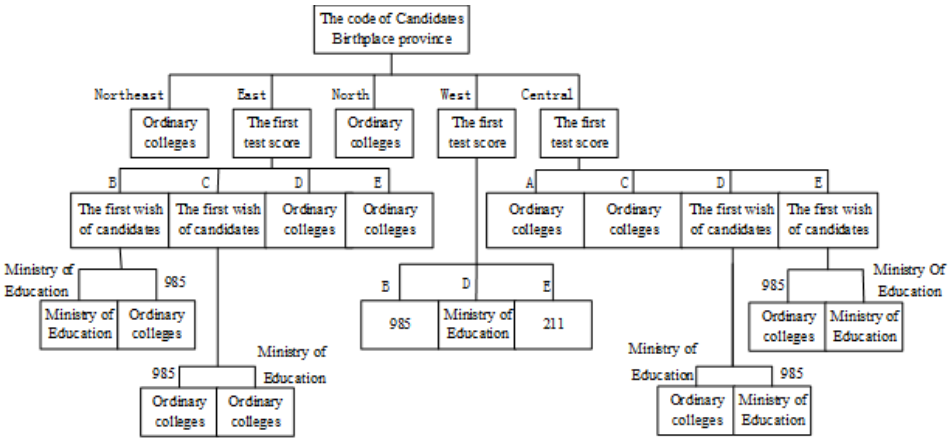

Fig.1: Class decision tree of undergraduate university

Generate classification rules

Due to space limitations, only a part of the classification rules are listed here:

IF the code of Candidates Birthplace province = 'East' AND First test score ='B' AND the first wish of candidates ='985' THEN Candidates graduated college =' Ordinary colleges' 50%

IF the code of Candidates Birthplace province = 'East' AND First test score ='D' THEN Candidates graduated college =' Ordinary colleges' 90%

IF the code of Candidates Birthplace province = 'East' AND First test score ='C' AND the first wish of candidates = 'Ministry of Education' THEN Candidates graduated college =' Ordinary colleges' 50%

IF the code of Candidates Birthplace province = 'West' AND First test score ='B' THEN Candidates graduated college ='985'33.3%

IF the code of Candidates Birthplace province = 'West' AND First test score ='D' THEN Candidates graduated college = 'Ministry of Education' 33.3%

IF the code of Candidates Birthplace province = 'Center' AND First test score ='D' AND the first wish of candidates = 'Ministry of Education' THEN Candidates graduated college =' Ordinary colleges' 60%

Result analysis

Firstly, the universities of these swap candidates are the vast majority of the ministry of education, 985 universities. However, the swap candidates of the university are graduated from mostly common colleges and Universities. Secondly, there are strong dependence between the attribute of candidates birthplace and the attribute of graduate colleges. The admission of candidates are mainly concentrated in North and East China, central region in the selection of adjustment; Thirdly, from the decision tree can also be found the candidates in central and western is relatively good. Although their geographical location are not as good as the bustling Eastern, their undergraduate college level is relatively high.

Conclusions

Through the above analysis, we can draw the following reference policy recommendations:

Firstly, intensify propaganda to provide preferential policies for the outstanding students. The result shows the excellent students in common colleges and universities are the publicity object of M university. Secondly, pay attention and encourage the western region of the candidates to candidate M university. Thirdly, pay attention to consider the geographical factors in the recruitment process.

In this paper, some interesting results are obtained. This helps to improve the school registration rate in ensuring a high quality of students. However, due to data limitations in time and space, there are limitations on making the guidance on decisions propaganda. Therefore, the quality of data should be improved to ensure the accuracy of the classification rules and provide better reference for graduate admissions M University.

Acknowledgements

Fund Project: Training foundation program for young teachers of university in shanghai (ZZGJD12012).

References

[1] Gao Yang, Liao Jiaping, Wu Wei. ID3 Algorithm and C4.5 Algorithm Based on Decision Tree [J]. Journal of Hubei University of Technology, 2011, 26(2):54-56.
[2] Huang Aihui. C4.5 Algorithm of Decision Tree Improvement and Application [J]. Science Technology and Engineering, 2009, 9(1):34-36, 42.
[3] Qu Shouning, Lu Jian. Application of C4.5 Classification Algorithm in the Survey and Evaluation of Postgraduate Intellectual Education [J]. Journal of University of JiNan, 2009, 23(3):253-256.
[4] Song Hui, Zhang Liangjun. Application of C4.5 Algorithm in Intelligent Evaluation for Air Quality [J]. Science Technology and Engineering, 2011, 11(20): 4848-4850.
[5] Wang ning. Research and Design of High School Test and Evaluation System Based on Data Mining [D].Changchun: Jilin University, 2009.
[6] Li Xiaowei, Chen Fucai, Li Shaomei. Improved C4.5 decision tree algorithm based on classification rules [J]. Computer Engineering and Design, 2013, 34(12): 4321-4325, 4330.
[7] Lu Zhengsong. Research of graduate fellowships evaluation by decision tree classified data mining [J].Computer Engineering and Applications.2012, 48(26):139-143.
[8] Chang Xu，Li Yijie，Liu Wanjun. Decision Tree Pruning Optimization Algorithm of CDC and REP Combination [J]. Computer Engineering. 2012, 38(14):32-34.
[9] Li Daoguo, Miao Duoqian, Yu Bin. Research and Improvement of Decision Tree’s Prune Algorithm [J].Computer Engineering.2005, 31(8):19-21.
[10]Huang juan. Applicationand Research of Data Mining in the Graduate Enrollment Quality [D].Shanghai: Donghua University.2010.