WEP Warwick Evans Publishing

International Journal of Global Economics and Management

ISSN: 3005-9690 (Print), ISSN: 3005-8090 (Online) | Volume 2, Number 3, Year 2024
DOI: https://doi.org/10.62051/IJGEM.v2n3.33
Journal homepage: https://wepub.org/index.php/IJGEM/index

Analysis of the Impact of Graduate Scholarships on the Employment System Using the C5.0 Algorithm

Yu Song

Management School of Xi'an Engineering University, Xi'an 710600, China

ABSTRACT

In order to explore the relevant mechanisms of graduate students obtaining scholarships and employment, this study starts from the perspective of data science and utilizes the C5.0 algorithm to collect 9 basic attributes of students, thereby determining the input variables of the decision tree. Employment is taken as the output variable. The employment situation of the graduates in 2021 is used as the training set, and the situation of the graduates in 2022 is used as the test set. This opens the black box of the mechanism of scholarship acquisition and student employment. The study indicates that the publication of papers, whether a student cadre, and academic conference papers are the three most critical indicators for employment

KEYWORDS

Scholarship policy; C5.0 Decision Tree; Information entropy

1. INTRODUCTION

The 20th report emphasizes, "We must adhere to the priority development of education, accelerate the construction of an education powerhouse, insist on nurturing talents for the Party and the nation, comprehensively enhance the autonomy of talents, and, with the further deepening of graduate education, focus on improving the quality of education, and strive to cultivate top-notch innovative talents." At the same time, it stresses the implementation of the employment priority strategy, strengthening employment priority policies, improving the employment promotion mechanism, and promoting high-quality and full employment. Against this backdrop, the proportion of graduate employment in the total employment continues to increase, leading to a continuous improvement in the quality of employment talents in our country. Meanwhile, the scholarship system for graduate students has gradually improved its processes, deepened its quality, and expanded its coverage. However, the current research has not clarified the relationship between the design of graduate scholarships and student employment. The commonly used methods in current research include questionnaire surveys, theoretical exploration [1], factor analysis to determine the main factors affecting employment [2], T-tests [3], and Logit models [4]. Such methods not only have a lagging effect but also limit the application scope of the model, as the questionnaires often focus on individual schools or a few schools in a region. The reason for this is that the factors influencing graduate students' receipt of scholarships and the mechanisms of employment are not yet clear, and the internal mechanisms need to be explored. Therefore, given the continuous accumulation of relevant student scholarship and employment data today, there still exists a situation of "rich data, scarce knowledge." This paper aims to depart from the perspective of data mining. By employing the C5.0 algorithm, it intends to unveil the black box of the scholarship system design and the relationship between student employment, starting from the data, and deciphering the transmission mechanism between student scholarship acquisition and employment. This will form a new graduate scholarship system with

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

academic achievement as its core and employment orientation. This study extends the theoretical boundaries of the scholarship system and student employment and provides a new methodological perspective for the design of student scholarship systems.

2. THE PRINCIPLE OF THE C5.0 DECISION TREE ALGORITHM

The C5.0 algorithm is built upon the foundation of the classic decision tree algorithm ID3, and later extended by Ross Quinlan (1987) into the C4.5 algorithm. It dynamically defines a discrete attribute based on numeric variables, thus categorizing continuous attribute values into discrete intervals. From this, two new attributes are proposed: separation information and information gain ratio. This allows us to strongly oppose the restriction that original features must be categorical, as we subsequently use information gain ratio as the basis for our branching. The C5.0 algorithm, also proposed by Quinlan [5], achieves smaller rule sets than its predecessors while consuming less memory, and simultaneously attains superior accuracy. This algorithm has been applied in various scenarios such as agricultural data analysis [6], telecom retail industry [7], and the Chinese stock market [8], garnering strong support and widely regarded as a stable and interpretable algorithm.

2.1. Classification of Data

In the defined training set "S," a training tuple represents an individual data tuple within it, belonging to the set of categories $\{c_1, c_2, \dots, c_k\}$. The quantity of tuples in the training set S belonging to category C_i is denoted by $freq(C_i, S)$. Now, assuming the training tuples are vectors with n metric attributes, represented by an n-dimensional vector $A = (A_1, A_2, \cdots, A_n)$, the training set "S" can be partitioned by one of the metric attributes in A into n different subsets $T_1, T_2, \cdots, T_n$.

2.2. information gain ratio

The core of the C5.0 algorithm lies in using the rate of decrease in information entropy as the primary criterion for selecting splitting attributes. Based on the information gain ratio of each metric attribute within the samples, the attribute with the highest value is chosen as the splitting attribute [9]. Thus, the problem is transformed into how to determine the information gain ratio.

(1) For a random variable X, its information entropy is defined as:

$$Entropy(S) = -\sum p(x_i)\log_2\big(p(x_i)\big),\ (i = 1,2,\dots,k)$$

$freq(C_i, S)/|S| = x_i,\ p(x_i)$ represents the probability of event x_i occurring, and $\sum p(x_i) = 1$.

(2) The conditional entropy of attribute A, denoted as H(A|S), is defined as the class entropy resulting from partitioning the training set S according to attribute A.

$$H(C \mid A) = -\sum_{i=1}^{n} \big((|A_i| \mid |A|) \times Info(A_i)\big)$$

(3) The information gain of attribute A is:

$$Gain(A) = H(C) - H(C|A)$$

(4) The information gain ratio of attribute A is:

$$GainRatio(S, A) = \frac{Gain(A)}{SplitInformation(S, A)}$$

$(SplitInformation)$

$$SplitInformation(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$xsx_i x_i' = \frac{x_i - x}{s} x_i' = \frac{x_i}{\sum_{j=1}^{n} x_j'}$$

2.3. Pruning of Decision Trees

The Post-Pruning method prunes decision trees from leaf nodes upwards, aiming to prevent overfitting of the training sample's feature descriptions by the decision tree. If the weighted error of a leaf node to be pruned exceeds the error of its parent node, then pruning is permissible; otherwise, pruning is not carried out [10]. The error e_i is estimated as:

$$e_i = f_i + Z_{\frac{\alpha}{2}} \sqrt{\frac{f_i(1-f_i)}{N_i}}$$

In the equation, f_i represents the error rate, calculated as the number of misclassified instances in node i divided by the total number of instances in that node; N_i is the total number of instances in node i; $Z_{\frac{\alpha}{2}}$ is the critical value, with $Z_{0.4}$ providing the best classification performance for the model.

2.4. Adjusting Sample Weights

The Boosting algorithm is utilized to train the scholarship-employment dataset multiple times while adjusting the weights of the samples, assigning greater weight to misclassified samples to increase the likelihood of correct classification during iterations. The implementation process of the Boosting algorithm is outlined as follows:

Step 1: Weight Initialization

At $M = 1$, the weights are initialized with a uniform distribution for the accident samples. The weight distribution is given by $w_b^1 = 1/n$, where n represents the number of samples in the training set, M denotes the iteration count, and $b = 1, 2, \cdots, n, M = 1, 2, \cdots, 10$.

Step 2: Calculate the classification error rate of the λ-th model on the scholarship-employment samples.

$$\theta_b^\lambda = \begin{cases} 1, \text{Sample } b \text{ is misclassified by the } \lambda\text{-th mode} \\ 0, \text{Sample } b \text{ is correctly classified by the } \lambda\text{-th model} \end{cases}$$

$$p_b^\lambda = \frac{w_b^\lambda}{\sum_{b=1}^{n} w_b^\lambda}$$

$$e^\lambda = \sum_{b=1}^{n} p_b^\lambda \theta_b^\lambda$$

In the equation: θ_b^λ is a function ranging from $0 \sim 1$. Specifically, when $\theta_b^\lambda = 1$, it indicates that sample "b" is misclassified by the λ-th model, whereas $\theta_b^\lambda = 0$ indicates that sample "b" is correctly classified by the λ-th model. w_b^λ represents the weight of the scholarship-employment sample "b" in

the M-th iteration, while p_b^λ signifies the normalized weight value of sample "b" in the M-th iteration. e^λ denotes the error of the λ-th model.

Step 3: Adjust the weight of scholarship-employment data samples, $w_b^{\lambda+1}$, based on e^λ, and normalize it.

$$w_{\mathrm{b}}^{\lambda+1} = \begin{cases} w_b^\lambda \dfrac{e^\lambda}{1-e^\lambda}, \text{Sample classified correctly} \\ w_b^\lambda, \text{Sample classified incorrectly} \end{cases}$$

Normalization of sample weights.

$$w_b^{\lambda+1} = \frac{w_{\mathrm{b}}^{\lambda+1}}{\sum_{b=1}^{n} w_{\mathrm{b}}^{\lambda+1}}$$

Step 4: The iteration of the scholarship-employment training samples is completed, resulting in the C5.0 decision tree model.

2.5. Evaluation of decision tree effectiveness

The classification accuracy rate P_A is the proportion of samples that are correctly classified out of the total samples.

$$P_{\mathrm{A}} = n_{\mathrm{C}}/n_{\mathrm{T}}$$

In the formula: n_{C} represents the number of samples classified correctly, and n_{T} represents the total number of samples.

3. EXPERIMENTAL RESEARCH

3.1. Experimental Foundation and Attribute Selection

In order to explore the factors affecting the acquisition of graduate scholarships and employment, specific experiments were conducted using MATLAB and SPSS Modeler software as the simulation experiment platform. Firstly, graduate basic data was extracted from the graduate management information database, and various student theses, awards, etc., were manually compiled from each college to obtain the predetermined data set. The attribute list of this data is shown in Table 1.

Table 1. Students attribute

Attribute Categories	Attribute Names
Basic Attributes	Student ID
	Gender
	Student Cadre
	Student Union Member
Research Attributes	Publication of Papers
	Publication at Academic Conferences
	English Proficiency (CET-4/6)
Awards Attributes	Academic Scholarships
	National Scholarships
Evaluation Criteria	Employment

3.2. Data Acquisition

(1) In the basic attributes section, student ID and gender are obtained from the graduate basic information database, while whether a student is a cadre or a member of the student union is manually obtained from ledgers of each college.

(2) In the research attributes section, data regarding paper publications, academic conference presentations, and English proficiency (CET-4/6) are obtained from the library's achievement certification database.

(3) Awards attributes are manually obtained from ledgers of each college.

Based on the acquisition of the aforementioned data, the relevant attributes are merged using the graduate student ID as the key to form a dataset suitable for analysis.

3.3. Training Set and Test Set

After organizing the data related to awards and employment for the graduating class of 2021, the classification rules for obtaining this dataset were examined. Data related to awards and employment for the graduating class of 2022 were then used to validate the accuracy and confidence of these rules. After data preprocessing, a total of 621 sets of data for the graduating class of 2022 were obtained. Among them, gender, student cadre, student union membership, English proficiency (CET-4/6), receipt of academic scholarships, and receipt of national scholarships are categorical variables. There were 34 recipients of first-class scholarships, 51 recipients of second-class scholarships, 84 recipients of third-class scholarships, and 16 recipients of national scholarships. Papers and academic conference papers are numerical data. For papers, scoring principles are defined as follows: CSCD journals receive 40 points, Chinese core journals and scientific core journals receive 20 points, and general journals receive 5 points. The scoring principle for academic conference papers is: poster presentations receive 5 points, while other presentations receive 3 points. The school's training program stipulates that graduate students in the second year must intern with their supervisors, providing favorable experimental conditions for considering only the factors affecting the acquisition of scholarships and their correlation with employment. Relevant parts of the dataset are shown in Table 2.

Table 2. Test dataset (partial)

Student ID	Gender	Student Cadre	Student Union	Paper	Academic Conference Paper	CET-4/6	Academic Scholarship	National Scholarship	Employment
210521103	Male	Learning Committee Member	No	80	49	CET-6	First-Class Academic Scholarship	Yes	Advanced Studies
210521130	Female	No	No	100	35	CET-6	First-Class Academic Scholarship	No	Advanced Studies
210521112	Female	No	No	80	50	CET-6	First-Class Academic Scholarship	Yes	Advanced Studies
210521126	Female	No	No	100	35	CET-6	First-Class Academic Scholarship	No	Employed
210521189	Female	No	No	80	35	CET-6	First-Class Academic Scholarship	No	Employed
210521120	Female	No	No	60	49	CET-4	First-Class Academic Scholarship	No	Employed
210521132	Male	No	No	60	35	CET-4	Second-Class Academic Scholarship	No	Advanced Studies
210521154	Female	No	No	60	35	CET-4	First-Class Academic Scholarship	Yes	Unemployed
210521168	Male	Class Monitor	No	60	35	CET-4	First-Class Academic Scholarship	No	Employed
210521156	Female	No	No	60	49	CET-4	First-Class Academic Scholarship	No	Employed
210521130	Female	No	No	100	35	CET-6	First-Class Academic Scholarship	No	Advanced Studies
210521112	Female	No	No	80	50	CET-6	First-Class Academic Scholarship	Yes	Advanced Studies
210521126	Female	No	No	100	35	CET-6	First-Class Academic Scholarship	No	Employed
210521189	Female	No	No	80	35	CET-6	First-Class Academic Scholarship	No	Employed

Experimental Results Utilizing the C5.0 decision tree algorithm, with branching based on selecting attributes with the maximum information gain as leaf nodes, the decision tree execution steps described above are iterated. Eventually, a classification decision tree and its rules for the graduating class of 2021 are generated. These are then applied to the employment test samples of the graduating class of 2022 to achieve validation. The eight attributes, including basic attributes, research attributes, and awards attributes, are used as input fields, while the employment status of students corresponds to the output variable (target variable). The training results obtained are as follows:

From the importance graph of the variables we can see that the thesis is the most important, with a share of 46 per cent. The next is whether or not it is a student leader, with a share of 31%. The last one is academic conference papers, with a share of 23%.

Further analysis, in the decision tree generated by the C5.0 algorithm, we can intuitively observe that: the first level of branching occurs in the input variable of the thesis, out of the 532 students with a thesis of less than 45 points, of which a high number of 443 are employed, this is because there are rather more students with a score of less than 45 points, so we have to pay attention at the same time to the fact that there are also 15.35%of the students who chose to stay in the workforce. The students who scored more than 45 points on the paper, on the other hand, provided a richer value of information, 20 students were those who chose to go on to higher education, and out of these 20 students, 16 did not have any class position, while 4 were serving as members of the study committee. There were also 13 students who were inactive. It is interesting to note that these 13 non-engaged students were among those who did not hold any class positions, which is counter-intuitive. It can also be seen that for those who scored more than 45 points on the essay, the class position attribute shows that students who were class secretary and class president chose to be employed, while those who were members of the study committee chose to go on to higher education. And in the third level of the decision tree, we can observe that these 13 non-engaged students are all present in the population with academic conference paper less than 49 points, while all the students with academic conference paper score greater than 49 chose to pursue higher education, provided the rest of the input variables are the same. Based on the decision numbers, we conclude that: 1) the paper is the most important variable, and students with a paper score of more than 45 have more choices and are less likely to be unemployed; 2) graduate students who want to get a job should actively participate in classroom management; 3) students who want to go on to higher education should firstly, make sure that they publish their papers, and secondly, pay attention to how much energy is taken up by classroom management; 4) students who want to go on to higher education must be aware of the importance of the paper's publication and the importance of the paper's publication. Finally, the academic conference paper has a "threshold effect", we are not only concerned about the academic paper less than 49 points, 12 students got the opportunity to go to college, and at the same time concerned about when the academic conference paper score is more than 49 points, under the same conditions, there is a 100% chance of going to college.

3.4. Evaluation of the model

From the model evaluation results, we can observe that the correctness rate reaches 81.61%, which indicates that the model has a strong generalisation ability and a good ability to interpret the data.

4. SUMMARY

This paper is based on the current lack of clarity in the relationship between scholarships awarded to postgraduate students and employment; do students who are awarded scholarships in schools get better employment opportunities after graduation? This is a realistic question. From the perspective of data mining, we focus on the current situation of "abundant data and lack of knowledge", which is neglected in the current research. A C5.0 decision tree algorithm is introduced to obtain the information gain ratio through the concept of information entropy and generate a complete set of decision tree models, which has a good generalisation ability and can well explain the problem we are studying. Of course, due to the collection and management of the current data warehouse, higher dimensional data, such as evaluation data, can not be well adopted, and due to the specificity of the data, there is still room for further research on the correlation between internships and employment. More dimensional attributes will be considered in the future so that the scholarship-employment correlation can be further deciphered.

REFERENCES

[1] Mi Yanyan. Analysis on the Effect of Enterprise Special Scholarships on the Employment of Pharmacy Students: A Case Study of Undergraduates in the Pharmacy Major of Xuzhou Medical University [J]. Journal of Wuhan Vocational and Technical College. 2014, 13(06):113-116.

[2] Kong Xiaoli, Wu Tonghui. Research on the Incentive Effect of University Scholarships on High-Quality Employment of College Students: A Case Study of Xiyasi International College, Zhengzhou University [J]. Henan Science and Technology, 2012, (03): 32-33..

[3] Cao Tongyan, Cui Yu, Yang Lu. Study on the Correlation between Scholarship Incentive Effect and Employment Salary [J]. Contemporary Education Theory and Practice, 2014, 6(10): 97-98. DOI: 10.13582/j.cnki.1674-5884.2014.10.107

[4] Qu Yinjiao, Yue Changjun, Ji Xiaohui. Research on the Influence of Economic Assistance for College Students on Employment Quality [J]. Tsinghua University Education Research, 2018, 39(01): 84-90.

[5] Quinlan L R, Rivest R L. Inferring decision trees using the minimum description length principle [J]. 1987.

[6] Rajeswari S, Suthendran K. C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud [J]. Computers and Electronics in Agriculture, 2019, 156: 530-539.

[7] Wang S T. Integrating KPSO and C5. 0 to analyze the omnichannel solutions for optimizing telecommunication retail [J]. Decision Support Systems, 2018, 109: 39-49.

[8] Zhou L, Si Y W, Fujita H. Predicting the listing statuses of Chinese-listed companies using decision trees combined with an improved filter feature selection method [J]. Knowledge-Based Systems, 2017, 128: 93-101.

[9] Bu Xiaoyang, Cai Yan, Wang Zongwei, et al. Data Mining of Power Marketing Based on C5.0 Decision Tree Algorithm [J]. Microcomputer Applications, 2022, 38(01): 23-2..

[10] Huang Changhai, Shen Jia, Zhu Ranchao, et al. Analysis Model and Application of Causes of Maritime Traffic Accidents Based on C5.0 Decision Tree [J]. China Safety Science Journal, 2022, 32(10): 90-99.