Eng Int Syst (2022) 6: 417–422

Engineering
Intelligent Systems

A study of the Influence of Various Characteristic Factors on the Employment Choice of College Graduates Using Data Mining

Chuanchuan Tang*

Chongqing Finance and Economics College, Chongqing, Sichuan 401320, China

The employment choice of graduates is influenced by many factors. This paper mined the data of 2015–2019 graduates using the C4.5 algorithm, eliminated the factors with less relevance to employment choice using correlation analysis, and extracted the classification rules from the remaining factors. The results showed that graduates who chose to be employed in state-owned enterprises were mostly Communist Party members who had a grade point average in the top 25%, those who chose to be employed in private enterprises had better English and computer skills, and those who chose to pursue higher education had higher English proficiency and a higher grade point average. English and computer proficiency had an influence on whether graduates chose to freelance, and English proficiency, computer skills, a scholarship and a high grade point average influenced whether graduates succeeded in their employment. The results verify the effectiveness of data mining, which is beneficial in uncovering useful data and enhancing decision making regarding the provision of relevant educational services.

Keywords: data mining, college graduates, employment choice, characteristic factors, decision tree

1. INTRODUCTION

With the development of network technology, a huge volume of education-related data is being generated (Zheng and Zhou, 2021). To obtain useful information from this massive amount of data, educational data mining (EDM) has emerged (Hegazi and Abugroon, 2016), which is to analyze and mine information collected in the education system, such as a student's personal information, grades, etc. (Río and Insuasti, 2016) and is used to solve educational problems (Neto, 2016), such as improving teaching models, making curriculum recommendations, analysing student behavior (Yang et al., 2021), predicting student grades (Kumar et al., 2017), etc. Data mining methods are being increasingly used in education (Agaoglu, 2016). Kaur et al. (2015) used real-world datasets from high schools and mined the datasets using the WEKA open-element tool to compare five classifiers: multilayer perceptual, naive Bayesian, SMO, J48, and REPTree, to find classification methods that are able to accurately identify students who struggle academically due to a low learning speed. Black et al. (2021) studied 177 graduates who graduated in 2017-2019 using a random forest tree approach to identify at-risk learners and found that the method provided a positive predictive value of 63.3% to identify learners who may encounter academic challenges and provide assistance to them. Alawi et al. (2017) used a k-means clustering method on 42,484 students in the Sultanate of Oman to better understand student behavior and academic performance. Thangakumar et al. (2020) investigated the feature selection process in EDM using

*Corresponding address: No. 906, Shangwen Avenue, Longzhouwan, Banan District, Chongqing 401320, China. Email: ccy8ap@yeah.net.

a method that combined ant colony optimization (ACO) and logistic regression (LR) and found that the method provided a recall rate of 96.07% and an accuracy rate of 94.91%. With the expansion of enrollments in universities, graduates' employment choices are receiving increasing attention. It is important to understand the influence of different characteristic factors on graduates' employment choices to adjust college and university training methods and to provide employment guidance (Wang and Chung, 2021). Therefore, this study uses the decision tree algorithm in data mining to collect relevant data on graduates and the different characteristic factors that influenced their employment choice to improve the level of employment guidance provided in schools and thus improve the employment rate of university graduates.

2. DATA MINING AND DECISION TREE

Data mining refers to the mining of potential and useful information from a large amount of noisy data (Sharma and Goyal, 2015), covering disciplines such as statistics and computer science and including methods of machine learning, expert systems, etc. (Helma et al., 2018). The purpose of data mining is to discover the patterns and relationships between data and to predict future trends (Krishnaiah et al., 2015). It has extensive applications in the fields of fraud detection, customer relationship management, intrusion detection, bioinformatics, etc. (Xu et al., 2015). Common data mining methods include regression analysis, association rules (Peng et al., 2020), neural networks (Zhao, 2021), decision trees, etc. This paper focuses on the decision tree algorithm.

A decision tree is a tree structure (Decaestecker et al., 2015) which starts from the root node and ends at the leaf node to classify different categories, and a decision result can be obtained from the classification rule in the tree. ID3 is a well-known algorithm which is simple and clear in theory and has strong practicality. The disadvantage of the ID3 algorithm is that when making attribute selection, it will be biased to select the attribute with many values, but in practice, the optimal attributes are not necessarily the ones that take many values; therefore, an improvement to the ID3 algorithm, the C4.5 algorithm (Kustiyahningsih et al., 2021), has emerged.

The two most important concepts in decision trees are entropy and information gain. Entropy is a measure of uncertainty; the greater the range of values of a variable, the greater the amount of information. Information gain refers to the amount of change in entropy; the smaller the entropy, the greater the information gain, and the more stable the system. Therefore, when the ID3 algorithm divides the nodes, the node with the greatest information gain is selected, and the improvement of the C4.5 algorithm reflects that it uses the information gain rate to determine the branching. The calculation formulas are as follows:

$$\text{GainRation}\,(S,\,A) = \frac{\text{Gain}\,(S,\,A)}{\text{SplitInfo}\,(S,\,A)},$$

$$\text{SplitInfo}\,(S,\,A) = -\sum_{i=1}^{n} \frac{s_i}{s} log_2 \frac{s_i}{s},$$

where Gain(S,A) refers to information gain, which is calculated in the same way as the ID3 algorithm, and SplitInfo(S,A) refers to splitting information, which indicates the breadth and uniformity of splitting sample S with attribute A. Compared with the ID3 algorithm, the C4.5 algorithm has advantages in terms of operation speed and accuracy.

3. ANALYSIS OF THE CHARACTERISTIC FACTORS AFFECTING EMPLOYMENT CHOICE

3.1 Data Processing

This study took the data of 2015–2019 graduates of Chongqing Finance and Economics College as an example and collected students' basic information, grade point average, scholarship status, etc., from the university's academic affairs system, student management system, career guidance center, etc. Firstly, duplicate data were eliminated. For missing values, similar samples were supplemented and were eliminated when supplementation cannot be realized. After sorting, a total of 3360 samples were obtained. To facilitate the subsequent data mining, the data needed transformation.

(1) Employment options: state-owned enterprises = 1, private enterprises = 2, further education = 3, freelance = 4, not employed = 0

(2) Gender: male = 1, female = 0

(3) Political affiliation: member of the Communist Party = 1, Communist Youth League member = 2, mass = 0

(4) English proficiency: CET4 = 1, CET6 = 2, failed = 0

(5) Computer skills: Level 2 = 1, Level 3 = 2, Level 4 = 3, failed = 0

(6) Mandarin level: Level 1 = 1, Level 2 = 2, failed = 0

(7) Scholarship: college level = 1, university level = 2, national level = 3, none = 0

(8) Grade point average: top 75% = 1, top 50% = 2, top 25% = 3, bottom 75% = 0

The collated samples are shown in Table 1.

3.2 Correlation Analysis Method

Using correlation analysis (Strecker et al., 2015), the correlation coefficients of the data were counted. The characteristic factors with large correlation factors were retained and those with small correlation factors were excluded using SPSS software. The output was expressed by r. The judgment criteria are shown in Table 2.

Table 1 Sample data after processing.

Sample number	1	2	3	……	3264
Employment options	1	0	3	……	4
Gender	0	1	1	……	0
Political affiliation	1	2	1	……	2
English proficiency	2	1	2	……	1
Computer skills	3	1	2	……	1
Mandarin level	2	1	1	……	1
Scholarship	3	0	2	……	2
Grade point average	1	0	3	……	2

Table 2 Judgement criteria for degree of correlation.

Value of $\|r\|$	Degree of correlation
> 0.8	Highly correlated
0.5–0.8	Moderately correlated
0.3–0.5	Lowly correlated
< 0.3	Uncorrelated

Table 3 Results of correlation analysis.

	Employment options
Employment options	1.00
Gender	0.032
Political affiliation	0.567
English proficiency	0.872
Computer skills	0.864
Putonghua level	0.212
Scholarship	0.678
Grade point average	0.742

Table 4 Number of graduates choosing different employments.

Employment options	Training set	Test set	Total
State-owned enterprise	575	246	821
Private enterprise	552	236	788
Further education	537	230	767
Freelance	538	231	769
Unemployed	151	65	215
Total	2352	1008	3360

The initial samples in Table 1 were input into the SPSS software and the results obtained are shown in Table 3.

The different characteristic factors were analyzed as follows:

(1) gender and employment choice: the correlation coefficient was 0.032, indicating uncorrelation, so it was excluded;

(2) political affiliation and employment choice: the correlation coefficient was 0.567, indicating moderate correlation, so it was retained;

(3) English proficiency and employment choice: the correlation coefficient was 0.872, indicating high correlation, so it was retained;

(4) computer skills and employment choice: the correlation coefficient was 0.864, indicating high correlation, so it was retained;

(5) Mandarin level and employment choice: the correlation coefficient was 0.212, indicating uncorrelation, so it was excluded;

(6) scholarship and employment choice: the correlation coefficient was 0.678, indicating moderate correlation, so it was retained;

(7) grade point average and employment choice: the correlation coefficient was 0.742, indicating moderate correlation, so it was retained.

4. THE APPLICATION OF THE C4.5 ALGORITHM IN EMPLOYMENT CHOICE ANALYSIS

Seventy percent of the samples in the dataset were used as the training set and 30% as the test set, and the number of graduates choosing different employment options is shown in Table 4.

Table 5 Information gain rates of different characteristic factors.

	Information gain rate
Political affiliation	0.065
English proficiency	0.359
Computer skills	0.328
Scholarship	0.315
Grade point average	0.216

The expected information amount of the sample set for classification is calculated as follows:

$$I(575, 552, 537, 538, 151) = -\frac{575}{2352}\log_2\left(\frac{575}{2352}\right) - \frac{552}{2352} \times \log_2\left(\frac{552}{2352}\right) - \frac{537}{2352}\log_2\left(\frac{537}{2352}\right) - \frac{538}{2352} \times \log_2\left(\frac{538}{2352}\right) - \frac{151}{2352}\log_2\left(\frac{151}{2352}\right) = 2.21459096.$$

Then, following the same method, the information gain rates of different characteristic factors were calculated, as shown in Table 5.

Table 5 shows that of the five characteristic factors, the one with the largest information gain rate was English proficiency, so it was taken as the root node. Then, other root nodes and leaf nodes were generated according to the size of the information gain rate to get the decision tree and rule set. To ensure the accuracy and applicability of the results, only the rules with more than 50 samples were organized and analyzed, as follows.

(1) Graduates whose employment choice was state-owned enterprises

- Political affiliation = member of the Communist Party, English proficiency = CET6, computer skills = Level 2, Scholarship = national level, grade point average = top 25%
- Political affiliation = member of the Communist Party, English proficiency = CET6, computer skills = Level 2, scholarship= school level, grade point average = top 25%
- Political affiliation = member of the Communist Party, English proficiency = CET4, computer skills = Level 2, scholarship= college level, grade point average = top 25%
- Political affiliation = member of the Communist Party, English proficiency = CET4, computer skills = level 2, scholarship = college level, grade point average = top 25%

According to these four rules, graduates who chose to be employed in state-owned enterprises were all members of the Communist Party, had a grade point average in the top 25%, CET4 and Level 2 computer skills, indicating that graduates who were Communist Party members and had excellent grades were more inclined to choose employment in state-owned enterprises.

Based on the above rules, we found that political affiliation and grades at school (grade point average) influenced whether graduates chose to be employed in state-owned enterprises. Current students who want to be employed in state-owned enterprises need to apply for membership in the Communist Party and improve their academic performance.

(2) Graduates whose employment choice is private enterprises

- Political affiliation = member of the Communist Party, English proficiency = TEM6, computer skills = Level 3, scholarship = school level, grade point average = top 25%
- Political affiliation = communist youth league member, English proficiency = TEM6, computer skills = Level 3, scholarship = college level, grade point average = top 50%
- Political affiliation = communist youth league member, English proficiency = TEM6, computer skills = Level 4, scholarship = college level, grade point average = top 50%
- Political affiliation = communist youth league member, English proficiency = TEM6, computer skills = Level 4, scholarship = college level, grade point average = top 50%

These four rules show that graduates who chose to be employed in private enterprises all had a political profile of at least a Communist Youth League member, an English proficiency of TEM6, computer skills of Level 3 or higher, a scholarship at the college level or higher, and a grade point average in the top 50%.

The above four rules reveal that English proficiency and computer skills had a great influence on whether graduates choose to work in private enterprises, indicating that private companies had some requirements for graduates' English and computer skills. Current students who want to be employed in private enterprises need to work hard to improve their English and computer skills.

(3) Graduates whose employment choice is to go on to further education

- Political affiliation = communist youth league member, English proficiency = TEM6, computer skills = Level 2, scholarship = school level, grade point average = top 25%
- Political affiliation = communist youth league member, English proficiency = TEM6, computer skills = Level 2, scholarship = college level, grade point average = top 25%

- Political affiliation = communist youth league member, English proficiency = TEM6, computer skills = Level 2, scholarship = college level, grade point average = top 25%
- Political affiliation = communist youth league member, English proficiency = TEM6, computer skills = Level 2, scholarship = college level, grade point average = top 25%

The above four rules show that graduates who chose to go on to further education were all Communist Youth League members, had an English proficiency of TEM6, computer skills of Level 2, a scholarship at the college level or above, and had a grade point average in the top 25%.

The above four rules reveal that English proficiency and grade point average influenced whether graduates chose to go on to further education. Graduates with good grades tend to choose to go on to further education; therefore, graduates who aim to go on to further education should strive to improve their English proficiency while maintaining a high grade point average.

(4) Graduates whose employment choice is to freelance

- Political affiliation = communist youth league member, English proficiency = TEM4, computer skills = Level 2, scholarship = none, grade point average= top 50%
- Political affiliation = communist youth league member, English proficiency = TEM4, computer skills = Level 2, scholarship = none, grade point average = top 75%
- Political affiliation = communist youth league member, English proficiency = TEM4, computer skills = Level 2, scholarship = none, grade point average = top 75%

The above four rules show that most of the students who chose to freelance were Communist Youth League members, had an English proficiency of TEM4, computer skills of Level 2, had not received a scholarship, and had a grade point average in the top 75%.

The above three rules reveal that English proficiency and computer skills influenced the graduates' decision to freelance, a certain level of English proficiency and computer skills was required when graduates chose to freelance, and political affiliation, scholarship and grade point average had no significant impact.

(5) Graduates who are unable to secure employment

- Political affiliation = Communist Youth League member, English proficiency = failed, computer proficiency = failed, scholarship = none, grade point average = top 75%.
- Political appearance = mass, English proficiency = failed, computer proficiency = failed, scholarship = none, grade point average = top 75%.

The two above rules reveal that graduates who did not pass English or the computer exams, did not receive a scholarship during their school years, and had a low grade point average were more likely to be unemployed. Overall, English proficiency, computer skills, scholarships, and grade point average influenced whether graduates were able to find employment.

5. CONCLUSION

In this paper, the characteristic factors affecting graduates' employment choices were studied using the C4.5 algorithm in data mining. Through the extraction and analysis of classification rules, it was found that different factors had different effects on employment choices. For example, graduates who were members of the Communist Party and had a grade point average in the top 25% preferred to choose employment in state-owned enterprises, while students with higher English proficiency and high-level computer skills preferred to choose employment in private enterprises. The study results verified the reliability of the data mining method in the study of students' employment choices. The data mining method can be further applied in practice.

REFERENCES

1. Agaoglu, M. (2016). Predicting Instructor Performance Using Data Mining Techniques in Higher Education. *IEEE Access*, *4*, 2379–2387.
2. Alawi, S.J.S., Shaharanee, I.N.M. & Jamil, J.M. (2017). Profiling Oman Education Data using Data Mining Approach. *AIP Conference Proceedings*, *1891*(1), 1–6.
3. Black, E.W., Buchs, S.R. & Garbas, B. (2021). Using Data Mining for the Early Identification of Struggling Learners in Physician Assistant Education. *Journal of Physician Assistant Education*, *32*(1), 38–42.
4. Decaestecker, C., van Velthoven, R.F.P., Peteinf, M., Janssen, T., Salmon, I., Pasteels, J., Ham, P.V., Schulman, C.C. & Kiss, R. (2015). The use of the decision tree technique and image cytometry to characterize aggressiveness in World Health Organization (WHO) grade II superficial transitional cell carcinomas of the bladder. *Journal of Pathology*, *178*(3):274–283.
5. Hegazi, M.O. & Abugroon, M.A. (2016). The State of the Art on Educational Data Mining in Higher Education. *International Journal of Emerging Trends & Technology in Computer Science*, *31*(1), 46–56.
6. Helma, C., Cramer, T., Kramer, S. & De Raedt, L. (2018). Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *Journal of Chemical Information and Computer Sciences*, *35*(4), 1402–1411.
7. Kaur, P., Singh, M. & Josan, G.S. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science*, *57*, 500–508.
8. Krishnaiah, V., Narsimha, G. & Chandra, N.S. (2015). Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach. *Advances in Intelligent Systems and Computing*, *337*, 371–384.

9. Kumar, M., Singh, A.J. & Handa, D. (2017). Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques. *International Journal of Education and Management Engineering*, *6*(6), 40–49.
10. Kustiyahningsih, Y., Khotimah, B.K., Anamisa, D.R., Yusuf, M., Rahayu, T. & Purnama, J. (2021). Decision Tree C 4.5 Algorithm for Classification of Poor Family Scholarship Recipients. *IOP Conference Series: Materials Science and Engineering*, *1125*(1), 012048 (7pp).
11. Neto, V. (2016). Apprentices Identifying Groups with Difficulties in Programming Education Using Data Mining. *The International Journal of E-Learning and Educational Technologies in the Digital Media*, *2*(2), 59–72.
12. Peng, H., Yang, S., Liu, Q., Peng, Q., & Li, Q. (2020). Intelligent Indexing Algorithm for the Association Rules of a Multi-Layer Distributed Database. *Engineering Intelligent Systems*, *28*(4), 229–240.
13. Río, C.A.D. & Insuasti, J.A.P. (2016). Predicting academic performance in traditional environments at higher-education institutions using data mining: A review. *Ecos de la Academia*, *4*(Diciembre), 2016.
14. Sharma, M. & Goyal, A. (2015). An application of data mining to improve personnel performance evaluation in higher education sector in India. *Computer Engineering & Applications*, 559–564.
15. Strecker, R., Scheffler, K., Klisch, J., Lehnhardt, S., Winterer, J., Laubenberger, J., Fischer, H. & Hennig, J. (2015). Fast functional MRA using time-resolved projection MR angiography with correlation analysis. *Magnetic Resonance in Medicine Official Journal of the Society of Magnetic Resonance in Medicine*, *43*(2), 303–309.
16. Thangakumar, J. & Kommina, S.B. (2020). Ant Colony Optimization Based Feature Subset Selection with Logistic Regression Classification Model for Education Data Mining. *International Journal of Advanced Science and Technology*, *29*(3), 5821–5834.
17. Wang, L. & Chung, S.J. (2021). Sustainable Development of College and University Education by use of Data Mining Methods. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(5), 102.
18. Xu, Q., He, D., Zhang, N., Kang, C., Xia, Q., Bai, J. & Huang, J. (2015). A Short-Term Wind Power Forecasting Approach With Adjustment of Numerical Weather Prediction Input by Data Mining. *IEEE Transactions on Sustainable Energy*, *6*(4), 1283–1291.
19. Yang, C.Y., Chen, I. & Ogata, H. (2021). Toward Precision Education: Educational Data Mining and Learning Analytics for Identifying Students' Learning Patterns with Ebook Systems. *Educational Technology & Society*, *24*(1), 1176–3647.
20. Zhao, Y. (2021). Interactive Genetic Algorithm Based on the BP Neural Network Proxy Model. *Engineering Intelligent Systems*, *29*(1), 45–53.
21. Zheng, C. & Zhou, W. (2021). Research on Information Construction and Management of Education Management Based on Data Mining. *Journal of Physics: Conference Series*, *1881*(4), 042073 (6pp).