Data Analysis of Undergraduate Employment Based on Decision Classification Algorithm

Yu Guan and Jing Luo(✉)

Fuyang Normal University, Fuyang 236041, China
200809009@fynu.edu.cn

Abstract. When the school uses the employment situation analysis system to analyze the employment situation of undergraduate students every year, because the classification decision rules are complex, it is difficult for the system to understand the complex rules when analyzing the employment situation of graduates, which leads to the slow operation of the system. Therefore, a brand-new employment situation analysis system is designed based on the decision classification algorithm. On the system hardware, the communication circuit of CAN bus is redesigned. On the software side, we use the decision classification algorithm to set up a most concise system classification decision rule, so as to automatically generate the analysis results of the employment situation of undergraduate graduates. The proposed system is applied to the analysis of the employment situation of undergraduate graduates in a university. The results show that the time taken by the system to analyze the employment situation is 97 356 ms and 70 372 ms shorter than the previous system. It can be seen that the rules set this time are easier to be understood by the system, and the results of employment situation analysis can be quickly obtained.

Keywords: decision classification algorithm · Undergraduate employment · data analysis

1 Introduction

According to the results of the 6th population census in 2010, the population with a bachelor's degree in 31 provinces in China totaled about 115 million, and 8,765 out of every 100,000 people were undergraduates. In the 5th population census in 2000, there were 3,579 graduates out of every 100,000 people, which was nearly 2.5 times higher than that in 2000. Students' learning ability has improved, and parents pay more attention to the cultivation of their children's learning of a professional knowledge [1]. Therefore, it is expected that the proportion of undergraduate graduates will increase in 2020. According to these traditional analysis systems, several colleges and universities analyze the employment situation of students and try to formulate teaching methods suitable for the current situation of social development. After practical application, it is found that although these systems have the ability of analysis or decision-making, in the process of practical application, the processing speed of employment information of undergraduate graduates is extremely slow and always in the loading stage [2, 3].

C. F. Peng et al. (Eds.): EIMT 2023, AHSSEH 8, pp. 854–860, 2023.
https://doi.org/10.2991/978-94-6463-192-0_111

Therefore, based on the traditional employment situation analysis system, this paper designs a brand-new employment situation analysis system for undergraduate graduates. The decision classification algorithm selected this time has the characteristics of decision subdivision and clear data classification. By perfecting the original analysis system, the analysis ability of the traditional analysis system on employment situation is further improved.

2 Decision Tree Classification

The rapid development of the information industry has not only expanded the scale of databases in all walks of life, but also made the amount of data accumulated by people grow rapidly. However, relatively, a large number of data also brings a series of problems, such as data is not easy to search, information is difficult to distinguish true from false, and organizational form is inconsistent. And today's database management system is not perfect, and people's use of data is still in the simple management and processing stage. More valuable information is quietly buried without powerful access tools. This situation is summarized as "data explosion and information scarcity". In order to solve these problems, a data analysis technology for deeper analysis and processing of complex data came into being, which we call data mining technology. So in order to make the research go smoothly, this paper chooses to use the existing more accurate student characteristic information to simulate the graduate information database. The pruning of the decision tree is the process of checking, correcting and modifying the decision tree generated in the previous stage (Fig. 1). After the decision tree is generated, many branches may reflect the abnormal phenomena such as noise in the sample training set, resulting in incorrect prediction in new decision problems, or reducing the classification accuracy, or increasing the complexity of the decision model. Therefore, in order to increase the reliability of the extracted rules and the classification accuracy of the decision tree, the generated decision tree is further processed [7].

Generally, the pruning of decision trees can be divided into two methods: first pruning and then pruning. First pruning, that is, to prune the tree, the establishment of the tree is terminated in the process of tree building. Post pruning, that is, after the decision tree is generated, the leaf nodes or branches that are not generally representative are cut

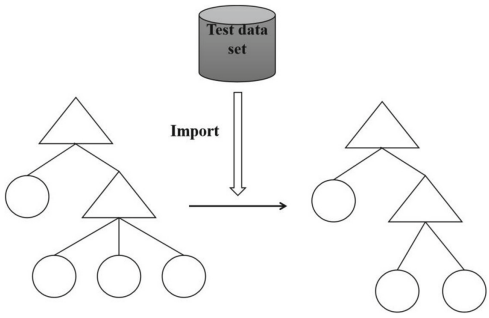

Fig. 1. Pruning process of decision tree

according to certain rules. The generated rules are usually used to test the prediction accuracy according to each tuple in the test data set. Finally, the branches that affect the accuracy of prediction will be cut.

2.1 Decision Classification Algorithm Setting System Analysis Rules

It is known that the database table automatically generated by the system is the object processed by the decision classification algorithm, so suppose the analysis system is:

$$W = (A, B, K, F, f) \tag{1}$$

where: A = a1, a2,…, an, which means that there is a non-empty finite set of n data, using the decision classification algorithm to define the employment information of undergraduate graduates and deal with the employment situation analysis problem. The following equations are the boundary equations of the data set calculated under the application of the decision classification algorithm:

$$\bar{S}(A) = \bigcup_{a \in A} \{[a]_s \cap A \neq \emptyset\} \tag{2}$$

where: S- (A) represents the upper approximation of the nonempty finite set A. Under the application of the decision classification algorithm, the system sets the approximate region for the graduate data set. At this time, the simplest decision-making rules for analyzing the employment situation of undergraduate graduates are:

$$a \in S^{-}(A), P(\mathrm{Y}|[a]) \geq \lambda \tag{3}$$

The decision classification algorithm sets the system analysis rules, and generates an automatically pop-up employment information table for undergraduate graduates according to the above rules.

3 The Application of Decision Tree Method in College Students' Employment

3.1 Data Mining Classification Implementation Process

The flow chart of classification implementation is shown in Fig. 2, and the details are as follows:

The first step is to determine the object and purpose of data mining. It is necessary to clearly define the problem and recognize the purpose of mining.

The second step is data collection. Because the implementation of data mining needs to analyze clear data information, data should be collected and stored in the database. Some data can be obtained directly, while others need to be calculated.

The third step is data preprocessing. In order to improve the quality and efficiency of data mining, it is necessary to clean up the collected data, fill in the missing values, smooth the noise data, correct the inconsistent data and convert all the data into a unified data format. The data information in multiple data sources is integrated, and according

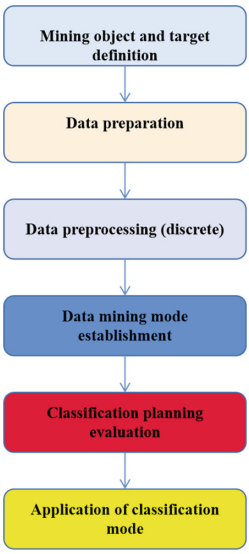

Fig. 2. Flow chart of classification implementation

to the research purpose, the data is transformed into an analytical data model by means of smooth aggregation, data generalization and standardization [8].

The fourth step is data classification mining. Select the appropriate algorithm and programming software, classify and mine the converted data, construct the classification model and find out the classification rules.

The fifth step is the analysis of classification rules. Explain and evaluate the classification results.

The sixth step is knowledge application. Using the mined classification rules, the data to be classified are tested.

3.2 Data Acquisition

Data mining needs a clear data analysis object, so it is necessary to collect graduates' employment information data first, which is a huge task and very time-consuming. The basic information of graduates collected in this paper includes the basic information of students in the "student status management system" of the academic affairs department, the information of students' grades in the "student performance management system" and the information of students' employment in the student affairs office. A total of 600 records were selected, of which 400 were training data sets and 200 were test data sets [9].

The basic information of students mainly includes the following attributes: student number, major, class, name, gender, politics, appearance, award-winning situation and practical ability, as well as some attributes such as nationality, school year, native place, ID number, home address, etc. Considering that it has no influence on employment, the basic information of students listed in this paper only selects the main attributes.

3.3 Data Preprocessing

Because the initial data may be missing or different from the actual data, the results of data mining will be unsatisfactory. This is enough to reflect the importance of data quality. Therefore, it is very important to use data preprocessing technology before data mining, which can not only improve the quality of data, but also reduce the time required for data mining [10].

4 Experimental Research

According to the analysis system of the employment situation of undergraduate graduates in this study, a comparative test experiment is put forward. The analysis system designed this time is used as the experimental group, and the two traditional analysis systems proposed before are used as the control group A and the control group B respectively. Compare the basic differences of the data obtained in the three test groups, and draw the experimental test conclusion according to the actual data. A university is selected as the experimental test data source, and three test groups are used to analyze the employment situation of undergraduate graduates last year. Table 1 shows the time taken by the three systems to get the analysis results [11].

As can be seen from the curve trend in Table 1, when the data volume of the analysis system designed this time exceeds 24 000 MB, the time it takes for the system to get the analysis results is within 80 000 ms; However, it took more than 100 000 ms for the two control groups to get the employment situation of graduates. Table 2 shows the statistical results of the average time used by the system [12].

According to the statistical results, the average time of the experimental group was 97 356 ms and 70 372 ms lower than that of the two control groups. It can be seen that the employment situation analysis system designed this time has better performance and solved the problem of slow operation efficiency of traditional analysis system [13–15].

Table 1. System processing time test results

Basic data volume of undergraduate graduates/MB	Time spent in the experimental group	Time spent in the control group, group A	Time used in the control group B
4	1.46	0.44	1.46
8	1.46	3.65	0.58
12	1.46	2.63	0.58
16	1.46	4.67	6.72
20	1.46	3.65	5.69
24	7.74	12.85	18.98
28	62.92	146.72	123.21

Table 2. Statistics of Time Average Used

Statistical items	Experimental group	Control group a	Control group b
The first group time	62087	155428	130625
Time spent in the second group	57944	159316	130150
Average time spent	60016	157372	130388

5 Conclusion

The employment situation analysis system designed this time gives full play to the calculation ability of decision-making classification algorithm, enhances data clarity, and brings great work efficiency to system analysis through simple analysis rules. However, when setting the analysis rules, the system designed this time has a huge amount of calculation, and there may be some calculation errors. Therefore, when setting the analysis rules, it should be highly concentrated to prevent distraction and lead to data errors.

Acknowledgment. This study was supported by the Key Research Projects of Humanities and Social Sciences in Universities of Anhui Province, A study on the relationship between ESOP and pay performance sensitivity (2022AH052811).

Quality Engineering Project of Department of Education, Anhui Province: Research on the Teaching Model of Bisection Class in the Intelligent Classroom Environment take financial management section for example (2021jyxm1120);

Quality Engineering Project of Department of Education, Anhui Province: Research and Practice of Financial Management in the context of new liberal arts (2021sx118).

Key Research Project of Talent Fund of College of Information Engineering, Fuyang Normal University: Research on the Intermediary Effect between Cash Dividend and Financial Performance (2022xgrcxm03).

References

1. Si, Y. . (2022). Construction and application of enterprise internal audit data analysis model based on decision tree algorithm. Discrete Dynamics in Nature and Society, 32(6), 341-346.
2. Charbuty, B. , & Abdulazeez, A. . (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends(01), 45(11), 2402–2412.
3. Dong, X. . (2021). Prediction of college employment rate based on big data analysis. Mathematical Problems in Engineering: Theory, Methods and Applications(Pt.51), 26(3), 183–192.
4. Hall, M. , Higson, H. , & Bullivant, N. S. . (2022). The role of the undergraduate work placement in developing employment competences: results from a 5 year study of employers, 9(3–4), 177–184.
5. Wang, Y. . (2021). Research on E-commerce Big Data Classification and Mining Algorithm Based on BP Neural Network Technology, 11(3), 295-299.

6. Deng, X. , Tang, G. , & Wang, Q. . (2022). A novel fast classification filtering algorithm for lidar point clouds based on small grid density clustering. Geodesy and Geodynamics, 13(1), 38-49.
7. Su, R. . (2021). Analysis of language features of english corpus based on java web. Microprocessors and Microsystems, 80(4), 103611.
8. Samsudin, E. . (2021). Modeling student's academic performance during covid-19 based on classification in support vector machine. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(5), 1798-1804.
9. Xie, W. , She, Y. , & Guo, Q. . (2021). Research on multiple classification based on improved svm algorithm for balanced binary decision tree. Scientific Programming, 13(4), 2106-2116.
10. He, H. , Sun, M. , Li, X. , & Mensah, I. A. . (2022). A novel crude oil price trend prediction method: machine learning classification algorithm based on multi-modal data features. Energy, 42(6), 1101-1111.
11. Abdulkareem, N. M. , & Abdulazeez, A. M. . (2021). Machine learning classification based on radom forest algorithm: a review. International Journal of Science and Business, 549(3), 34-52.
12. Chen, H. , Cai, M. , Huang, K. , & Jin, S. . (2021). Classification and evolution analysis of key transportation technologies based on bibliometrics. Hindawi Limited, 13(6), 2036-2052.
13. Costa, S. A. , Nogueira, V. , Bussanelli, D. G. , Restrepo, M. , Escobar, A. , & Cordeiro, R. . (2021). Impact of the undergraduate clinical teaching-learning process on caries detection and treatment decision-making. Brazilian Journal of Oral Sciences, 50(2), 385-414.
14. Gu, Z. , & He, C. . (2021). Application of fuzzy decision tree algorithm based on mobile computing in sports fitness member management. Wireless Communications and Mobile Computing, 2021(6), 1-10.
15. Yu, S. , Li, X. , Wang, H. , Zhang, X. , & Chen, S. . (2021). C_cart: an instance confidence-based decision tree algorithm for classification. Intelligent Data Analysis, 25(4), 929-948.