Research on college graduates employment prediction model based on C4.5 algorithm

Fujie Zhou¹, Lingxiao Xue¹, Zhiqiang Yan² and Yongxian Wen¹*

¹College of Computer and Information Science, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002, China
²School of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian 350108, China
*Corresponding author’s e-mail: Wenyx9681@fafu.edu.com

Abstract. Introducing C4.5 algorithm into graduates’ employment field, organizing and analyzing the collected tremendous amount of historical data, building classify Decision-Tree, generating rule set, constructing graduates’ employment prediction model. Using Cross Validation algorithm to verify the accuracy of the model, and gradually accomplish a high-accuracy and high-practically prediction model, applying this prediction model to the employment related decision making and college major structure optimizing.

1. Introduction
Workers with Highly Education were needed by the society in order to promote the all-round development of society. In recent years, with the growing group of College graduates, the number of Chinese graduates has exceeded 7 million in 2014. Higher education has changed from "elite" to "popular", and the employment situation of college graduates is becoming more and more serious. What qualities college graduates should have which could adapt to the changing employment situation and smoothly embark on the job is an important issue of concern at this stage.

In the past, because of the limited means of collecting information, not only the basic information of graduates cannot be fully known, but also the information of graduates' employment is insufficient. However, with the use of information technology to the teaching and employment management of Students in Colleges and universities, a large number of information data related to students have been collected and a huge database has been established, including the student basic information database, graduates employment information database and graduates employment unit quality tracking survey database. These huge data provide abundant research materials for the employment problem of College graduates, but because the data volume is too large and complex, a complex method is not suitable to deal with such huge data, if just only a simple query, it is difficult to get effective employment information, and these data are classification data basically, it is difficult to use common mathematical models to do a statistical analysis.

C4.5 algorithm (Quinlan, 1986) is a decision tree classification algorithm in machine learning algorithm. C4.5 uses the method based on information gain ratio to select test attributes to build decision tree. Each node of decision tree could be just one attribute of student information, which is a suitable statistical model for classification data. A complete decision model can be obtained by traversing the generated decision tree which is simple and understandable. Moreover, C4.5 algorithm adds the processing of continuous attributes and attribute value vacancies on the basis of ID3, so it is a...
more mature method for tree pruning. However, although the algorithm has been widely used in industry, commerce, medicine and other fields (Wiharto et al., 2016; Ngoc et al., 2017; Li et al., 2019), it is rarely used in the field of education.

Therefore, this study uses C4.5 algorithm to classify and analyze the data of graduate employment. The main purpose is to explore the feasibility of this method in predicting the employment of College graduates. It is hoped that the employment prediction model of college graduates can be established so as to carry out the employment work pertinently and provide decision support for the managers of colleges and universities. It can provide reference for optimizing the professional structure, alleviating the structural contradiction of employment and deepening the reform of education system.

2. The Establishment of model

2.1. Sample space
This study collected nearly 10,000 graduates' employment data from a comprehensive university in the past three years as sample space. The data are huge and informative. It mainly includes several types of databases: students' basic information database, including student number, name, gender, student origin, political outlook, academic background, major and foreign language level, computer level, academic papers, social practice experience and so on; graduates employment information database, including attributes such as school number, employment category, employment unit, competent department of the unit, unit category and subordination of the unit; graduates employment unit quality tracking survey database, including the name of the employment unit, students' evaluation, the answers to the questionnaire on graduates' employment intention and career choice concept, etc.

2.2. Data preprocessing
This research has a large amount of data, so the data preprocessing is particularly important. Data preprocessing is to generate target data by cleaning up and generalizing massive data for the use of core mining algorithms, namely knowledge base. The main task of this stage is to reduce the dimension of original data, so as to improve the quality of case data and speed up data mining. In this paper, data is preprocessed by following four steps (Quinlan, 1993):

Firstly, data integration which combines heterogeneous data in multi-file or multi-database running environment to solve the model of semantics. It mainly involves the selection of data, the conflict of data and the processing of inconsistent data;

Secondly, data cleaning which involves removing noise data and irrelevant data from source data sets, processing missing data and cleaning dirty data, removing white noise from blank data domain and knowledge background, and considering time sequence and data changes. It mainly includes repetitive data processing and missing data processing, and completes some data type conversion;

Thirdly, data transformation which is to find the characteristic of data, reduce the number of variables to find the invariants of data by dimension transformation or transformation: standardization, induction, switching, rotation and projection. Data table was transformed to be more suitable for data mining.

Fourthly, data simplification which is to find useful features of the expressed data base on the discovery target and the content of the data itself, so as to minimize the amount and size of data while maintaining the original appearance of the data as far as possible. There are two main ways to simplify data: attribute selection and data sampling for records in the database.

3. Decision tree

3.1. Decision tree generation
C4.5 algorithm was proposed by Quinlan in 1992. It is a complete decision tree generation system based on ID3 algorithm (Witten and Frank, 2005). It establishes decision tree through two steps: tree generation stage and tree pruning stage. Compared with ID3, C4.5 algorithm increases the processing
of continuous attributes and attribute value vacancies. What’s more, C4.5 algorithm is more mature on
tree pruning. Unlike ID3, C4.5 uses an information gain ratio-based approach to select test attributes.
The information gain ratio is equal to the ratio of information gain to information split.

Assuming that the sample set \( S \) is divided into \( n \) subsets according to \( n \) different values of the
 discrete attribute \( A \), the information gain rate of \( S \) divided by \( A \) is as follows:

\[
\text{GainRatio}(S,A) = \frac{\text{Gain}(S,A)}{\text{SplitInformation}(S,A)}
\]

(1)

Where

\[
\text{SplitInformation}(S,A) = -\sum_{i} \frac{|S_i|}{|S|} \log \left( \frac{|S_i|}{|S|} \right)
\]

(2)

The JAVA language was used to design and implement the main process of C4.5 algorithm in this
study. The class diagram of program design is shown in Figure 1.

In order to make the employment prediction model of college graduates verifiable and predictable,
this study used K-Fold Cross Validation method (Kantardzic, 2003) and error-based pruning (EBP).
It is an improvement of pessimistic error pruning (PEP) algorithm, but it’s still a pessimistic algorithm.
EBP algorithm is divided into three steps:

Firstly, calculate the upper confidence limit \( U_{CF} \) of the misclassified sample rate of leaf nodes. In
C4.5 algorithm, the probability of default misclassified samples obeys binomial distribution, the
confidence is 25% by default, and the probability distribution is

\[
P\left( \frac{e(t)}{n(t)} \leq U_{CF} \right) = CF
\]

(3)

The upper confidence limit \( U_{CF} \) of misclassified sample rate can be obtained by looking up
binomial distribution table. However, in practice, because of the large number of samples, it is not
suitable for the programming of pruning algorithm to prune decision tree by looking up table.
Therefore, the upper confidence limit \( U_{CF} \) of misclassified sample rate of decision tree can be
obtained by using the following function:

\[
U_{CF} = \frac{f + \frac{z^2}{2N} + \frac{f^2}{N} + \frac{z^2}{4N^2}}{1 + \frac{z^2}{N}}
\]

(4)

Where \( f = \frac{E}{N} \) E is the number of misclassified samples in the actual experiment process, \( N \) is the
total number of samples; \( Z \) is 0.69, which is confidence difference calculated based on 25% default
confidence of C4.5 algorithms; By using this formula, the upper confidence limit \( U_{CF} \) of the
misclassified sample rate of nodes can be obtained.
Figure 1. Algorithm JAVA class diagram.

The second step is to calculate the expected number of misclassified samples of leaf nodes. PE is the number of misclassified samples of leaf nodes. There are formulas:

\[ PE = N \times U_{CF} \]  \hspace{1cm} (5)

The third step is to determine whether to prune. By comparing the sum of the number of misclassified samples of leaf nodes before pruning and the size of the number of misclassified samples of upper nodes, we can determine whether to prune or not.

3.2. Rule set establishment

This decision tree is pruned by C4.5 pruning algorithm. It looks more compact, but it is still very complex. Such a large decision tree is difficult to understand because each node has a specific environment based on the test results of the prior node. In order to make the decision tree model more readable, the path to each leaf node is transformed into IF-THEN rule. The IF part includes all tests of a path, and the THEN part is the final classification. This form of rule is called decision rule. The decision rule set of all leaf nodes can classify samples as tree nodes. As a result of the starting point of the tree, the IF part of the rule is mutually exclusive and complete, so the order of elements in the IF part of the rule can be disrupted without affecting the final result.

For example:

1. IF “Graduation Choice” = employment and “Education” = undergraduate and “Computer Level” = secondary and “Award-winning Situation” = international national award THEN “Employment Category” = category A.

2. IF “Graduation Choice” = self-employment and “Working Experience” = have THEN “Employment Category” = category C.

A total of 38 rules have been found in the decision tree generated in this paper. These rules are the employment prediction models for college graduates determined by the data mining objectives. If we want to make the results simpler, we can generalize the rules. Since the preceding of a single rule can include irrelevant conditions, deleting irrelevant conditions of those rules does not affect the classification results of the rules. For example, if the rule \( R \) is:

\[ \text{If} \ A \ \text{Then} \ C \]

More general rule \( R’ \) is:

\[ \text{If} \ A’ \ \text{Then} \ C \]
Where $A'$ is the condition set which deletes conditions $X$ from $A(A = A' \cup X)$. Conditions $X$ are conditions that can be proved to be unimportant in training samples. Each sample satisfying condition $A'$ in the database can satisfy or not satisfy the extended condition $A$. All samples belong to or do not belong to category C. The results can be organized into a $2 \times 2$ contingency table, as shown in Table 1.

| Category C       | Others       |
|------------------|--------------|
| Satisfying $X$   | $Y_1$        | $E_1$        |
| Not satisfying $X$ | $Y_2$ | $E_2$ |

Rule $R$ contains $Y_1 + Y_2 + E_1 + E_2$ samples, where $R$ has $E_1 + E_2$ misclassified samples because they belong to other classes rather than C. Similarly, $Y_2 + E_2$ is the total number of samples contained in Rule $R'$, and $E_2$ are the misclassified samples. The criterion for eliminating condition $X$ from rule is based on pessimistic estimates of correctness of rule $R$ and $R'$. The pessimistic error rate of rule $R$ is estimated to be $U_{CR}(Y_1 + Y_2 + E_1 + E_2, E_1 + E_2)$; The pessimistic error rate of rule $R'$ is estimated to be $U_{CR}(Y_2 + E_2, E_2)$. If the pessimistic error rate of rule $R'$ is not greater than that of original rule $R$, then deleting condition $X$ will not cause too much influence. Of course, when generalizing rules, more than one condition must be deleted. Instead of observing all possible subsets of deleted conditions, the C4.5 algorithm performs greedy deletion: one condition with the least pessimistic error is deleted at each step. For all greedy search methods, it is not guaranteed that the minimum of each step will minimize the overall situation.

4. Model application

4.1. Verification of Model Accuracy

In order to ensure the accuracy of the model, it is necessary to evaluate the performance of the model. This study validates the accuracy of the employment prediction model with the employment data of the 2015 year graduates. The reliability parameters of each rule are obtained as shown in Table 2. we can screen out high quality and high reliability rule sets year by year.

| Rule ID | Total number of samples | number of misclassified samples | accuracy      |
|---------|------------------------|--------------------------------|---------------|
| 31      | 72                     | 11                             | 84.72%        |
| 32      | 107                    | 9                              | 91.59%        |
| 33      | 345                    | 33                             | 90.43%        |

4.2. Model Application and Prediction

The most important role of establishing the employment prediction model for college graduates is the application of the model and the prediction of the employment situation of graduates. Through the prediction, the expected employment categories of each student are determined, and targeted employment guidance is given to the students of the disadvantaged groups in employment. At the same time, decision-making support is provided for the decision-makers of colleges and universities.

In this study 38 rules were established, it include the expected employment categories under all attribute values of the samples. Each new sample can find its own expected employment categories by
these rules. The accuracy of each rule can be gradually improved through the model accuracy verification, so that we can set up new samples as shown in Table 3.

In the current severe employment environment, it is of great practical reference value for optimizing the professional structure and employment guidance through such application and prediction.

5. Conclusion

By studying the employment prediction model of college graduates, we can understand the main factors that affect the success of college graduates' employment and the quality of employment. We can forecast the employment situation of college graduates by classification, so as to provide help for the employment guidance of college graduates, and provide decision support for specialty system setting and teaching reform in colleges and universities.

(1) Visualization of the generated decision tree objects. Observe the structure of the tree with pages that generate HTML for multi-level navigation.

(2) This model is applied to the new graduates of a university in 2015. Through forecasting, the expected employment categories of each student are determined, and targeted employment guidance is given to the students of the disadvantaged groups.

(3) The rule set of the employment prediction model for college graduates can be further improved and extended to the employment analysis system of colleges and universities throughout the country.

The employment of college graduates has become an important issue related to the stable development of the country. How to establish an efficient, reliable and perfect employment guidance system for college graduates is an important issue we should consider. Using the subjective knowledge in the field of College graduates' employment and the objective historical data left over from many years to establish the model, I believe that it will certainly play a certain role and significance in the development of our college graduates' employment work.

ACKNOWLEDGEMENT

This work was supported by the key educational reform projects of Fujian Agriculture and Forestry University (project no.: 111417025)

Reference

[1] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
[2] Wiharto, W., Kusnanto, H., & Herianto, H. (2016). Interpretation of clinical data based on c4.5 algorithm for the diagnosis of coronary heart disease. Healthcare Informatics Research, 22(3), 186-195.
[3] Ngoc, P. V. , Ngoc, C. V. T. , Ngoc, T. V. T. , & Duy, D. N. . (2017). A c4.5 algorithm for english emotional classification. Evolving Systems, s12530-017-9180-1.
[4] Li, J., He, J., Liu, Z., Zhang, H., & Zhang, C. (2019). Traffic accident analysis based on C4. 5 algorithm in WEKA. In MATEC Web of Conferences (Vol. 272, p. 01035). EDP Sciences.
[5] Kantardzic, M. (2003). Data mining: concepts, models, methods, and algorithms, 2nd edition. Iie Transactions, 36(5), 495-496.