Knowledge Discovery of the Orientation of University Graduates in Beijing Area Based on Decision Tree

Si-meng LIU*

School of Humanities and Economic Management, China University of Geosciences, Beijing, China

*Corresponding author

Keywords: Data mining, Graduate, Decision tree, C5.0 algorithm.

Abstract. With the expansion of undergraduate enrolment in universities in Beijing, the pressure on the graduates is increasing. It is particularly important to study the orientation of graduates. In this paper, through questionnaire method to collect information about undergraduates in Beijing, according to the characteristics of undergraduates from the data collected, we built a decision tree model with C5.0 algorithm and pruned. The rules of classification are used to display the mining results intuitively. Then, we analyzed and verified the model to provide theoretical support for the undergraduates' choice of direction.

Introduction

With the expansion of enrollment of undergraduate students in universities in Beijing, the number of graduates is constantly increasing, and the labor market is gradually transformed into a pattern of oversupply. This has resulted in a more severe employment situation for graduates this year. Besides employment, more and more graduates are sending their future to apply for graduate students, study abroad, and self-employment. The choice of college graduates has become the focus of social concern.

In the new century, with the continuous development of information industry and network interconnection and the surge of data, the field of data mining has been continuously integrating new knowledge and technical methods and constantly diversified from multiple angles. As a kind of prediction model in data mining technology, decision tree is widely welcomed because of its simple requirement for data preparation and its easy comprehension and realization. Therefore, we can process the data of university graduates through effective data mining technology, extract valuable information to analyze the key factors that affect the employment of graduates, and accurately forecast the graduates' flow. This enables the employment guidance center of colleges and universities to carry out work in a planned and targeted manner, provide reference for the employment and career of graduates, and better achieve the goal of college training talents and serving the society.

Principles and Methods of Decision Tree

Principles. Decision tree is a tool in data mining, which is a tree structure similar to flowchart. It provides a rule-like approach to what value under what conditions is suitable for judging situations where the environmental conditions are less and the logical combination is relatively simple. The decision tree can express the logic requirement of data mining more intuitively, and it can be easily converted into SQL statement, which can be classified as similar or better.

Method of C5.0 Algorithm. C5.0 is one of the classic decision tree model algorithms that splits samples based on the fields that bring the maximum information gain. The first subset of samples is then split again, usually based on another field, and the process repeats until the sample subset cannot be split. Finally, focus on the lowest-level splits, and those subsets of samples that have no significant contribution to the model value are found and pruned.
The C5.0 algorithm uses an information gain extension called the information gain ratio as a property selection metric. It normalizes the information gain with the "split information" value. Split information is similar to Info (D), defined as follows:

\[
\text{SplitInformation}_A(D) = -\sum_v \left( \frac{|D_j|}{|D|} \log \left( \frac{|D_j|}{|D|} \right) \right)
\]  

This value represents the information resulting from the v partitions divided from the training data set D corresponding to the v outputs of the attribute test A. For each output, it considers the number of tuples with that output instead of the total number of tuples in D. It differs from the information gain, which is based on the information obtained by the same partitioning. The information gain ratio is defined as follows:

\[
\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInformation}_A(D)}
\]

Gain (D, A) = Info(D) - Info\(_A\) (D)  

Gain (D, A) tells us how much we got through the division on A. It reduces the expectation of information needs as a result of knowing A's value.

The C5.0 model is robust in the face of data omissions and problems with many input fields. It is easier to understand than some other types of models and the rules presented by the model have very intuitive explanations. Usually do not need a lot of training to estimate. C5.0 also provides the powerful technology to improve the accuracy of classification.

Investigation on the Factors Influencing the Orientation of Undergraduates in Beijing

**Focus Groups.** Focus Groups Also called in-depth group interviews. Initiated by Paul Lazarsfeld and Robert Merton, refer to the discussion of a particular topic under the guidance of the host. The moderator aims to promote interaction between group members and ensure that the discussion focuses on the research topic.

A total of 20 graduates were selected as interviewees for this interview. According to whether the GPA is the top 30% of the professions, the interviewees were divided into excellent group and general group, with 10 in each group for focus group interview. The main content of the interview there are two: First, "What is the direction choice after you graduate?", And second, "What are the factors that affect your choice of direction?". The interview results are summarized as follows.

First, there is no significant difference in the length of interviews between the excellent group and the general group. However, there are differences in the categories and frequency of the selection direction and influencing factors.

Second, both the results of the excellent group and the general group have appeared apply for graduate students, study abroad and employment. No matter which group, choose to apply for graduate students are higher than the other options.

Thirdly, excellent group and general group agree that the type of school and the expected employment unit are the important influencing factors of graduates' choice of direction. With the improvement of the standards of employing units, the social demand for graduate students is also growing, so that undergraduate employment paths are relatively narrow. In addition, the relevant policies introduced by the government also have some restrictions for undergraduates, which make undergraduates more likely to see the prospect of graduate students.

**Questionnaire.** According to the results of our focus group interviews and the results of our literature research, we have compiled the questionnaires on the direction choice and influencing factors for undergraduate graduates in Beijing.

After the small-scale test, the questionnaire was put online for data collection. After a week, the questionnaires were collected and the results were counted. In this survey, a total of 600 questionnaires were collected, of which 589 were valid. Among these 589 useful samples, 500 were selected as training samples and 89 were used as test samples. Each sample is described by 21 attributes. Since not all attributes are useful, it is necessary to exclude some attributes. The following table shows the useful attributes after screening.
Among these attributes, gender, nation, family economic status, university level, specialty, whether as a student cadre, internship experience, type of employment, professional counterpart, target employment area, target employment city type, orientation after undergraduate degree, select these 12 attributes as classification attributes, then digitize these before building decision tree. However, average household income, general point average and expected salary are interval attributes and do not need to be processed.

After preprocessing, the "sample set" is shown in the following table.

### Table 2. Sample set.

<table>
<thead>
<tr>
<th>No.</th>
<th>Gender</th>
<th>Nation</th>
<th>Average household income</th>
<th>Family economic status</th>
<th>University level</th>
<th>Specialty</th>
<th>Student cadre</th>
<th>GPA ranking</th>
<th>Internship experience</th>
<th>Type of employment</th>
<th>Professional counterparts</th>
<th>Targeted employment area</th>
<th>Targeted employment city type</th>
<th>Expected salary</th>
<th>Orientation after undergraduate degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1</td>
<td>B0</td>
<td>4</td>
<td>C0</td>
<td>D1</td>
<td>E10</td>
<td>F0</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A1</td>
<td>B0</td>
<td>1</td>
<td>C2</td>
<td>D1</td>
<td>E9</td>
<td>F1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A1</td>
<td>B0</td>
<td>1</td>
<td>C1</td>
<td>D2</td>
<td>E10</td>
<td>F0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>A1</td>
<td>B0</td>
<td>2</td>
<td>C0</td>
<td>D3</td>
<td>E4</td>
<td>F1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Application of Decision Tree Model in Undergraduate Orientation Selection Forecasting

In this dissertation, the SPSS Modeler software is used to model the data collected from the questionnaire and the C5.0 algorithm is used to construct the decision tree. In the sample of training samples, orientation after undergraduate degree as a category identification attribute, others as decision attributes.

The initial decision tree model established by C5.0 algorithm is too large and complex, less practical, and prone to over-fitting problem. Therefore, the decision tree needs to be trimmed.

According to the result of the decision tree generated after pruning, we select part of rules which support degree is above 65%, and list the classification rules as follows.
Table 3. Classification rules.

<table>
<thead>
<tr>
<th>No</th>
<th>IF</th>
<th>THEN</th>
<th>SUPPORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Type of employment= Chinese-foreign joint venture + Average household income in [&quot;10000-50000 yuan&quot;]</td>
<td>postgraduate</td>
<td>71.66%</td>
</tr>
<tr>
<td>2</td>
<td>Type of employment = Chinese-foreign joint venture + Average household income in [&quot;150000-400000 yuan&quot; &quot;400,000 yuan or more&quot;]</td>
<td>further study abroad</td>
<td>71.66%</td>
</tr>
<tr>
<td>3</td>
<td>Type of employment = public institutions + expected salary in [&quot;12000 yuan or more&quot; &quot;5000-8000 yuan&quot; &quot;8000-12000 yuan&quot;]</td>
<td>postgraduate</td>
<td>66.20%</td>
</tr>
<tr>
<td>4</td>
<td>Type of employment = public institutions + expected salary in [&quot;3000-5000 yuan&quot; &quot;under 3000 yuan&quot;]</td>
<td>work</td>
<td>66.20%</td>
</tr>
<tr>
<td>5</td>
<td>Type of employment = public institutions + gender = male</td>
<td>work</td>
<td>65.52%</td>
</tr>
<tr>
<td>6</td>
<td>Type of employment = public institutions + gender = female</td>
<td>postgraduate</td>
<td>65.52%</td>
</tr>
<tr>
<td>7</td>
<td>Type of employment = others+ targeted employment area in [&quot;others&quot; &quot;stay in the city where the university is located&quot;]</td>
<td>postgraduate</td>
<td>71.66%</td>
</tr>
<tr>
<td>8</td>
<td>Type of employment = others+ targeted employment area in [&quot;back to hometown and nearby area&quot;]</td>
<td>work</td>
<td>71.66%</td>
</tr>
<tr>
<td>9</td>
<td>Type of employment = state-owned enterprises + expected salary in [&quot;12000 yuan or more&quot; &quot;3000-5000 yuan&quot; &quot;under 3000 yuan&quot;]</td>
<td>work</td>
<td>71.66%</td>
</tr>
<tr>
<td>10</td>
<td>Type of employment = state-owned enterprises + expected salary in [&quot;8000-12000 yuan&quot;]</td>
<td>postgraduate</td>
<td>71.66%</td>
</tr>
<tr>
<td>11</td>
<td>Type of employment = university</td>
<td>postgraduate</td>
<td>82.47%</td>
</tr>
<tr>
<td>12</td>
<td>Type of employment = state organs +GPA ranking in [&quot;60-100%&quot;]</td>
<td>test a civil servant</td>
<td>65.20%</td>
</tr>
</tbody>
</table>

With 89 test data into the classification rules for testing, compared with the actual employment situation of the graduates, there are 52 identical and 37 different ones, and the correct rate is 58.43%. Therefore, the classification rules generated are relative accurate.

Conclusions

The analysis results of the decision tree show that the most important factor in the decision tree is the "type of employment", and secondly, the factors that influence the employment of graduates include "targeted employment area", "expected salary", "university level" "specialty", "GPA ranking". These factors are all important aspects affecting the direction of graduates. However, the influence of "nation" and "whether as a student cadre" on the choice of undergraduates' direction is not significant.

Compared with the previous study, this study draws a similar conclusion that the impact of individual subjective factors on the choice of students' direction is the most significant. The expectations of students for future occupations and salaries affect their options after graduation. At present, the most stable jobs are the state and government organs, public institutions and some large enterprises. However, the employment opportunities provided by these units are very limited, resulting in more and more undergraduates choosing to apply for graduate study or study abroad to enhance their own competitiveness.

Contrary to previous findings, the impact of family factors on graduates is relatively weak. Due to the fact that the state places a great deal of importance on the educational problems of students who have family financial difficulties, various related supportive policies have been promulgated, including the national student loans and all kinds of scholarships for schools. Institutionally, it is guaranteed that no student will be allowed to drop out of school for the family's financial difficulties. Graduate students who choose to apply for postgraduate studies believe that a high degree can bring high returns. When financial input such as tuition is no longer a problem for students, whether family conditions are superior or not, parents want their children to continue their postgraduate study and have a better development in the future.
References


