Position Prediction Method Based on Improved Decision Tree

Wenzhe Wang¹, Xi Yue¹ and Mengxuan Tang²

¹School of Software Engineering, Chengdu University of Information Technology, No.24 Block 1, Xuefu Road, Chengdu, China.
²The Software Engineering Technology Research Support Center of Informatization Application of Sichuan.
Email: 1585103087@qq.com

Abstract. C4.5 decision tree algorithm is one of the commonly used classification prediction algorithms. It is a tree structure. Its advantage is that the process of obtaining results is easy to understand and the calculation is small, etc. Its disadvantage is that it is easy to cause over-fit, and will be very complex when there are too many categories. In the process of analyzing skill proficiency and positions, different job positions are available for different levels of skill proficiency. When there are more job categories, the prediction effect of decision tree classification is poor and still needs to be improved. In this paper, an improved decision Tree algorithm WF_D-tree is proposed, which adds skill proficiency weight to each data in the data table. Through the calculation method of skill proficiency weight, the relative redundant data of table data are removed. Through the longitudinal calculation method of skill proficiency weight, dimension reduction of data. Experimental results show that the improved decision tree greatly reduces the running time of the decision tree algorithm in the job prediction, and also improves the accuracy of the prediction results. In terms of time, the WF_D-tree is also significantly higher than the decision tree before the improvement. In terms of accuracy, the decision tree was improved by about 11 percentage points.

1. Introduction

The rise of computer network has driven the development of all walks of life, and the development of all walks of life in turn has promoted the rapid progress of computer network. From the end of the 20th century to the beginning of the 21st century, under the influence of China’s traditional marketing ideas, some recruitment sites began to appear on the Internet in China. During the emergency period, the online talent market exploded. After the temporary suspension of the job fair, the online talent market suddenly caught fire and human resource information websites mushroomed[1].

At present, IT technologies such as the Internet, the Internet of Things, big data, artificial intelligence (AI), edge computing and high performance computing are increasingly penetrating into the work field [2], and more and more jobs are created along with them. As the number of jobs increases, the demand for skills will become more diverse and the learning of people's skills will become more variable. In view of this, educational data mining also plays a vital role. First of all, through educational data mining, people can predict and analyze jobs according to their own skills as well as their skill proficiency. Secondly, through educational data mining, people can also know what kind of job they want, what necessary skills they need to learn, as well as how much they have mastered the skills. Tubagus Muhammad Akhriza et al.[3] used data mining technology based on the theory of evolution to make the curriculum adapt to the huge and rapid changes in the industry as much as possible. Börner Katy et al.[4] showed in their research how the curriculum is coordinated between research and work. It is also found how the demand for skills from industry in academic,
educational, and industrial systems may drive skill concerns in research and vice versa. Mohamed Ezz et al. [5] proposed an adaptive recommendation system to predict the appropriate educational path for college students. To recommend students to appropriate engineering departments with high accuracy. Atanas Ivanov et al. [6] used decision tree algorithms to build regression models to predict students’ performance and assess their mathematical ability. The above researches are all the fields of big data in education, but they do not well solve the needs of adapting themselves to their job positions through their own skill proficiency.

Combined with educational data mining, this study makes an in-depth exploration of how people can predict and analyze jobs based on their own skills and skill proficiency through data mining.

2. Decision Tree C4.5 Algorithm

As a classification algorithm, the goal of decision tree algorithm is to distribute N samples with P-dimensional characteristics into C categories. C4.5 algorithm is a kind of decision tree algorithm, which adds information gain rate on the basis of ID3 algorithm. C4.5 algorithm is also a machine learning method based on information theory. Its core idea is to build a decision tree recursively on the whole data set by analyzing the training data set [7]. The detailed process for establishing the decision tree is as follows.

**Step1.** First of all, the data set is preprocessed. If the data set is continuous, the data set should be discretized. After that, the information gain rate of all attributes is calculated, and then the attribute with the maximum information gain rate is selected as the attribute of the root node of the decision tree.

(1) Assuming that the training set is D, there are k categories, which are respectively \{C_1, C_2, \ldots, C_k\}, each row of data corresponds to a data category C_i. Set |D| as the sample number of training set, and |C_i| as the sample number of category C_i in training set D. Then the probability of each category in the total category is P_i = |C_i|/|D|. Information entropy of category is represented by Info(D), as shown in Equation (1) below:

$$Info(D) = - \sum_{i=1}^{k} P_i \log_2 (P_i)$$

(2) Suppose the attribute A_j (j=1,2, \ldots, m) there are m different values \{a_{1j}, a_{2j}, \ldots, a_{mj}\}, the training set D can be divided into V different classes according to attribute A. After partitioning, the information entropy of the attribute can be obtained and expressed as Info_A(D). As shown in Equation (2) below:

$$Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} \times Info(D_j)$$

(3) By subtracting equation (1) and equation (2), the information Gain of training set D can be divided according to the attribute A, which is expressed by Gain(A). As shown in Equation (3) below:

$$Gain(D) = Info(D) - Info_A(D)$$

(4) The information gain rate USES the split information value to normalize the information gain. The classification information is similar to Info(D) and is represented by SplitInfo(D). As shown in Equation (4) below:

$$SplitInfo_A(D) = - \sum_{j=1}^{V} \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right)$$

This value represents the information generated by dividing the training data set D into V partitions corresponding to the V outputs of the attribute A test. From Equations (3) and (4), the information gain rate is expressed by GainRatio(D). As shown in Equation (5) below:
Select the attribute with the maximum gain rate as the split attribute.

**Step 2.** Repeat Step 1 in the child nodes to recursively build branches of the tree. Each time a child node is selected, the attribute with the maximum information gain is selected as the branch node. The value of all the sample attributes of the leaf node stops at the same time.

**Step 3.** During the construction of the decision tree, many branches reflect the anomalies in the training data due to the noise and outliers in the data. Pruning method is used to deal with the problem of over-fitting data. All pruning methods use statistical measures, subtracting the least reliable branches.

**Step 4.** According to the established decision tree, the new data are classified and predicted.

### 3. Improvement of C4.5 Algorithm

ID3 algorithm uses information gain rate as the selection standard of attribute splitting. C4.5 algorithm is an improvement on ID3 algorithm, using information gain rate as the selection standard of attribute splitting. WF_D-tree is an improvement on the C4.5 algorithm. Here are the steps to improve:

1. Firstly, the weight value of this position is calculated horizontally and a threshold row_value is set. The threshold is set by calculating the average weight and the value needs to be less than the average weight value, and the value needs to be obtained near the average weight value. Less than this threshold, this data will be regarded as relatively redundant data, and this data will be discarded.

2. Secondly, the obtained data set is reduced vertically again. The weight value of the column attribute is calculated, and a threshold cow_value is set. Through the judgment of the threshold value, the attributes of the data set are reduced. The setting range of this threshold is from greater than 0 to the average weight value, which is established when the test is most accurate.

### 4. Analysis of Experimental Results

#### 4.1. Preparation of Experimental Data

**4.1.1. Data set Introduction**

The experimental data set was obtained from 51job.com, and the obtained data set is shown in Figure 1 below:

![Figure 1. Data acquisition](image)

After data preprocessing, such as removing pause words, missing values, abnormal data, irrelevant data, duplicate data, etc., the data in Figure 1 is obtained. The first column represents the ID, the second column represents the job, and the third column represents the skill requirements and proficiency for the job.
4.1.2. Data Preprocessing

Word segmentation was performed on the third column content column through natural language processing to obtain the skills-> proficiency. The word segmentation tool USES Jieba word segmentation. The word segmentation tool is not ideal, so it is necessary to remove no words, ambiguous words, adverbs, special symbols and so on according to the actual situation. After screening, we have to match skills to proficiency. Therefore, the skills in the job position need to be matched with the skill proficiency. Obtained, skill -> proficiency -> job position relationship table. As shown in Figure 2:

![Table 1: Skill-Proficiency-Job Position Relationship](image)

**Figure 2.** The data processing

For the job title, we need to merge the jobs. The same post can be expressed in different ways. In order to reduce the difficulty of subsequent work, so job positions consolidation is needed. As shown in Figure 3:

```python
for job_title in data:
    if job_title in positions_to_merge:
        print(f'position: {job_title}
Software development engineer
The front-end development engineer
Operation and Maintenance engineer
Image algorithm engineer
Demand engineer
Technical architect
The product manager
ERP engineer
Java Development Engineer
.NET development engineer
Software engineer
Software engineer
The front-end development engineer
Software engineer
Back-end development engineer
Machine vision engineer
```

**Figure 3.** Post the name

4.1.3. Data Conversion

The data is converted into a format as shown in Figure 4 below:

![Table 2: Data Conversion](image)

**Figure 4.** Data conversion

0 means: no; 1 represents: proficient; 2 represents: proficient; 3 represents: familiar with; 4 for: Understand
4.2. The Process of Training a Model

4.2.1. The Environment
This experiment is completed under the operating system of Windows10. The computer has 12 gigabytes of memory and a 64-bit operating system. Python version 3.5.

4.2.2. The Training Process
C4.5 algorithm is first used to analyze the experimental data. Get the time of the algorithm, as well as the prediction accuracy of the results.

The improved algorithm is used again to analyze the experimental data. Get the time of the algorithm, as well as the prediction accuracy of the results.

4.2.3. Analysis of Experimental Results
The following is the comparison and display of test data results before and after improvement, as shown in Table 1:

|                  | C4.5  | WF_D-Tree | Improvement |
|------------------|-------|-----------|-------------|
| Number of data   | 350   | 350       |             |
| Time consuming   | 12.173| 8.594     | Reduce 3.579|
| Accuracy         | 0.693 | 0.801     | Improve 0.108|

The experimental results are as follows: in this experiment, when the experimental data is 350, the effect difference is the most obvious. The results before and after the improvement are better than those before. The improved algorithm is about a third faster in time than the unimproved one. Accuracy was also improved by nearly 11 percentage points. Effectively improve the operation efficiency, as well as the prediction accuracy of the results. It also proves the feasibility of the scheme.

5. Conclusion
The WF_D-tree algorithm proposed in this paper is based on the C4.5 algorithm to add weight value to data elements, and to carry out the processing of relatively redundant data and the reduction of attributes in the form of table calculation method. The problem of accurate classification of C4.5 algorithm in the case of high-dimensional and multi-classification data is improved. Compared with the improvement of the original C4.5 algorithm, the experimental results of this method also have some obvious improvement.

In addition to The C4.5 classification algorithm, other classification prediction algorithms are also blooming. As a scholar, he should work hard, learn more and make more progress to integrate other algorithms and data structures.

6. Reference
[1] Liu Jiayin. Current Situation and Development of Human Resource Information Websites in China -- Based on investigation and Analysis of Recruitment Websites [J]. Intelligence Science,2004(06):728-733.
[2] HongXueHai Cai Di. ”Internet +” oriented OT merge with the IT development and research [J/OL]. China engineering science: 1-6 [2020-08-01]. http://kns.cnki.net/kcms/detail/11.4421.G3.20200728.1130.002.html.
[3] Tubagus Mohammad Akhriza,Yinghua Ma,Jianhua Li. Revealing the Gap Between Skills of Students and the Evolving Skills Required by the Industry of Information and Communication Technology[J]. International Journal of Software Engineering and Knowledge Engineering,2017,27(5).
[4] Börner Katy, Scrivner Olga, Gallant Mike, Ma Shutian, Liu Xiaozhong, Chewning Keith, Wu Lingfei, Evans James A. Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. [J]. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(50).

[5] Mohamed Ezz, Ayman Elshenawy. Adaptive recommendation system using machine learning algorithms for predicting student’s best academic program [J]. Education and Information Technologies: The Official Journal of the IFIP Technical Committee on Education, 2020, 25(4).

[6] Atanas Ivanov. Decision Trees for Evaluation of Mathematical Competencies in the Higher Education: A Case Study [J]. Mathematics, 2020, 8(5).

[7] Yao yafu, xing liutao. Improvement and application of decision tree C4.5 continuous attribute segmentation threshold algorithm [J]. Journal of central south university (natural science edition), 2011, 42(12): 3772-3776.