Improvement and Application of Decision Tree C4.5 Algorithm

Jie YE¹ and Li-duo HOU²

¹School of Computer and Engineering, Guizhou University of Commerce, Guiyang 550014, Guizhou, China
²School of Computer Science and Technology, Guizhou University, Guiyang 550025, Guizhou, China

*Corresponding author

Keywords: Decision tree, C4.5 algorithm, Information entropy, Pruning, Engineering quality supervision.

Abstract. The C4.5 algorithm as the most popular decision tree algorithm, there are still some deficiencies. C4.5 algorithm uses post-pruning to solve the over-fitting problem, but increase the modeling overhead, in response to this problem the idea of combining over-fitting branches ahead of time in the process of creating a decision tree is proposed and improved on the original C4.5 algorithm, which solves the problem of long decision-making time and large models of the decision tree. Experimental verification shows that the improved algorithm has a significant improvement in model simplification and model accuracy. The improved algorithm is applied to the "Engineering Quality Decision Support System" to provide supervision and decision-making support for the supervisory department.

Introduction

In recent years, the rapid development of data mining technology, decision tree algorithm as a branch of data mining classification algorithm is favored, because of its excellent data analysis efficiency, intuitive and easy to understand, high precision characteristics. In the current decision tree algorithm, the most widely used is the C4.5 algorithm.

However, the C4.5 algorithm uses post-pruning to solve overfitting. Therefore, a complete decision tree needs to be constructed, which results in the problem that the complexity of the decision tree model is too large. When the amount of data is large, the full growth decision book will consume a lot of time, but the messy branches in the decision tree that are prone to overfitting will be pruned after being grown, thus increasing the overhead of the algorithm. Based on the discovery of these cluttered branches during the growth of the decision tree, and stopping its growth in advance, but also need to avoid the problem of "visual limitations" of the pre-pruning, an improved C4.5 algorithm with variable parameters was proposed.

C4.5 Decision Tree Algorithm

C4.5 Decision Tree Algorithm Principle

The core idea of the C4.5 algorithm is to use information gain as an attribute selection criterion at all levels of the decision tree. By selecting the largest attribute of information gain as the test attribute of the sample partition, the purpose of classifying the test data set with the minimum amount of information is achieved.

Suppose there are n samples in the set S, it is assumed that the classification attribute C has m different values, namely Ci (i=1, 2,..., M). Let ni be the number of samples in Ci. For class attribute C, the information entropy of the sample set S divided into m classes can be expressed as:

\[
E(S) = - \sum_{i=1}^{m} P_i \cdot \log_2(P_i)
\]
Among the many test attributes of the sample, we chose one of the test attributes $A$. Assuming that $A$ has $k$ different values, we can also use the test attribute $A$ to divide the sample set $S$ into $k$ classes, namely $S_j$ ($j=1, 2, ..., K$). At this time, $A_{ij}$ ($i=1, 2, ..., M$) ($j=1, 2, ..., K$) is expressed as the number of samples belonging to both the $S_j$ class and the $C_i$ class. Then the expected entropy caused by attribute $A$ partitioning sample set $S$ can be expressed as:

$$E(S) = - \sum_{i=1}^{m} P_{ij} \cdot \log_2(P_{ij})$$

The information entropy of the sample subset $S_j$ can be expressed as:

$$E(S, A) = \sum_{j=1}^{k} \frac{A_{ij} + A_{i2} + \cdots + A_{in}}{n} \cdot E(S_j)$$

In this way, we can calculate the information gain of attribute $A$ based on the above information, which can be expressed by the following formula:

$$\text{Gain}(S, A) = E(S) - E(S, A)$$

The information gain rate is actually equal to the information gain/entropy. The reason why the information gain rate index is used is that if the information gain is used to select the test attribute, the result will make the decision tree tend to tend to those attributes with a variety of attribute values, so we To avoid this phenomenon, the information gain rate is calculated as follows:

among them:  
$$\text{GainRatio}(A) = \frac{\text{Gain}(S, A)}{\text{SplitInfo}(S, A)}$$  
$$\text{SplitInfo}(S, A) = \sum_{j=1}^{k} P_{j} \cdot \log_2(P_{j})$$

C4.5 Decision Tree Algorithm Modeling Process

This study is based on WEKA data mining platform C4.5 algorithm data modeling and data analysis, in the WEKA platform C4.5 algorithm implementation process shown in Figure 1.

![Diagram of C4.5 algorithm implementation process in WEKA platform.](image)

Figure 1. Implementation of C4.5 algorithm in WEKA platform.

Problems with Decision Tree Algorithms

The decision tree algorithm classifies a set of chaotic and disordered data sets. Information entropy is usually used to represent the degree of confusion in the sample set. The larger the information entropy, the greater the degree of confusion in the sample set, and the more difficult it is to completely classify the sample set. When the branch of the decision tree has a large information entropy and has fewer instances, it will make the branch too large and there are not enough instances to support and produce overfitting phenomena. The method of solving the fitting usually adopts the pruning of the decision tree. The commonly used methods include: pre-pruning and post-pruning. The C4.5 algorithm adopts post-pruning, that is, the pruning is after the decision tree is fully grown.
Because the C4.5 algorithm uses a post-pruning strategy, the following problems occur when modeling with the C4.5 algorithm:

(1) Using the C4.5 algorithm to model, the decision tree is pruned after full growth, thus increasing the time spent in the modeling process.

(2) The pruned branch also contains part of the feature description of the decision tree, but it has not been used after pruning. As a result, resources are wasted.

### C4.5 Decision Tree Algorithm Improvement

#### The Core Idea of Improvement

Aiming at the problems existing in the C4.5 algorithm in 1.3, an improved C4.5 algorithm with variable parameters is proposed to make full use of the data set and incorporate the former pruning idea.

The core idea of the improved algorithm is: When a certain node of the decision tree needs to be split, the information entropy of each instance subset related to that node is calculated in advance. If it will contain fewer instances (less than $\beta$) and the information entropy is greater than the average information entropy of the instance subset. The subset of $(1+\alpha)$ is marked and treated as a hashed subset, and the splitting process of this subset is stopped early. In order to avoid being caught in the problem of "visual constraints", these disorganized subsets are merged as a new instance subset of the node for the subsequent splitting process. Where $\alpha$ represents the magnitude of the increase in the average information entropy, ranging from 0 to 1, where $\beta$ represents the minimum number of instances, and the introduction of $\alpha$ mainly leads to a decrease in the accuracy of the decision tree for some branches. The introduction of $\beta$ is used to control where the decision tree merges so that it is neither close to the leaf nor to the root of the decision tree. So that the combined subset of instances can have more data support, so that the sample data can be effectively used, and the consolidation of the instance subsets can also avoid the process of redundant branches continuing to split downward, thereby reducing the modeling. Time and time spent after pruning.

#### Improved Algorithm Flow

The improved algorithm is mainly to determine in advance which branch is easy to produce the overfitting phenomenon in the process of creating the decision tree, and merge these branches. The implementation process of the improved algorithm is shown in Figure 2, where in the dotted line in the figure represents the place the modified part based on the original algorithm. The specific implementation process of the improvement part is shown in Figure 3:
Treatment of Special Problems

There are problems in modeling with improved algorithms. We have adopted the following methods to deal with these problems:

1. The test attribute will still appear in the subsequent splitting process after the sub-collection of the instance sub-collection. This increases the depth of the tree to a certain extent. Here, it is necessary to unify the value of the merged data set on the test attribute. Change to the same value.

2. If the merged branch is still a disorderly branch, the decision tree is pruned after the growth of the decision tree using the post-pruning strategy.

Experimental Comparison and Analysis

This paper uses Java language to rewrite the original C4.5 algorithm, and integrates the algorithm into the WEKA data mining platform, selects the pre-processed data in the engineering quality supervision system for testing. Using 10 cross-validation methods, the C4.5 algorithm and the improved algorithm were compared and analyzed in terms of modeling speed, model generation scale, and model accuracy. The experimental environment is CPU: 2.6 GHz, memory: 4 GB, and Windows 7 operating system.

Two kinds of algorithms were used to model the quality evaluation data of the barrage, and the comparison between the modeling time and the scale of the nodes was shown in Figure 4 and Figure 5.

From the perspective of modeling speed, although the improved algorithm requires time-consuming merger of the redundant branches during the modeling process (the time complexity is linear), but because the algorithm removes the modeling process of the redundant branches, it is greatly improved the modeling speed. From the experimental results, the improved algorithm modeling time is less than the original C4.5 algorithm.

From the node scale, the original C4.5 algorithm has 70 nodes in the decision tree model generated by the two algorithms. The improved algorithm has 62 nodes, the number of nodes is reduced by 11.42%, the leaf node is also reduced from 47 to 41. Because of the decrease of the leaf node, the prediction rules are also relatively reduced. Therefore, the model of the algorithm is improved. The pattern becomes clearer and easier to understand, and the rules generated are more succinct and clearer, which provides convenience for the application of subsequent models.

Then use the two algorithms to model the data of the sluice project (11 test properties) and the step-up transformer project (9 test properties) in the project quality supervision data. The
comparison results of the accuracy of the three groups of decision tree models are shown in the figure. 6.

![Figure 6. Comparison of model accuracy.](image)

It can be seen from Fig. 6 that the improved algorithm has higher accuracy in the three sets of experimental data, and the improved algorithm has high accuracy in the decision tree models built on the boosting substation and barrage projects. In the original C4.5 algorithm, although the accuracy of the model in the sluice project is slightly lower than that of the original C4.5 algorithm, the average accuracy improvement of the three models is 0.65% higher than that of the original C4.5 algorithm. Looking at the overall situation of the data, the model accuracy of the improved algorithm is slightly better than the original C4.5 algorithm.

Table 1 shows the overall performance comparison between the two algorithms. It can be seen from Table 1 that the improved algorithm is superior to the C4.5 algorithm in the number of nodes in the decision tree, the modeling speed, and the average accuracy of the model. Therefore, the improvement is improved. The algorithm is more adaptable to engineering quality supervision decisions.

<table>
<thead>
<tr>
<th>Description</th>
<th>C4.5 algorithm</th>
<th>Improve algorithm</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modeling speed</td>
<td>0.22s</td>
<td>0.15s</td>
<td>Improved Algorithm Accelerates 0.07s</td>
</tr>
<tr>
<td>Summary point quantity</td>
<td>70</td>
<td>62</td>
<td>Improved Algorithm reduced by 8</td>
</tr>
<tr>
<td>Number of leaf nodes</td>
<td>47</td>
<td>41</td>
<td>Improved Algorithm reduces six</td>
</tr>
<tr>
<td>Forecast rule quantity</td>
<td>12</td>
<td>9</td>
<td>Improved Algorithm reduces 3</td>
</tr>
<tr>
<td>Model Accuracy (Barrage Project)</td>
<td>90.80%</td>
<td>91.91%</td>
<td>Improved Algorithm improved by 1.11%</td>
</tr>
<tr>
<td>Model Accuracy (Boost Substation Project)</td>
<td>89.42%</td>
<td>90.77%</td>
<td>Improved Algorithm improved by 1.35%</td>
</tr>
<tr>
<td>Model Accuracy (Sluice Project)</td>
<td>88.46%</td>
<td>87.94%</td>
<td>Improved Algorithm is reduced by 0.52%</td>
</tr>
<tr>
<td>Model average accuracy</td>
<td>89.56%</td>
<td>90.21%</td>
<td>Improved Algorithm improves by 0.65%</td>
</tr>
</tbody>
</table>

**Improved C4.5 Algorithm Application**

The application of data mining classification technology to the engineering quality supervision process is of great help to the supervisory authorities. Based on the large amount of historical data in the engineering construction supervision process, an improved algorithm is used in the WEKA data mining platform to model the “Assessment of project quality is shown in Figure 7 as an example. After data preprocessing, the data set to be analyzed is shown in the following figure.
The decision tree model generated using the improved algorithm for data analysis and modeling is shown in Figure 8.

![Decision Tree Model](image)

**Figure 7. Data set to be analyzed.**

**Figure 8. Decision tree model generated by improved algorithm modeling.**

In the model of the decision tree, the priority of the parent node is higher than that of the child node. That is, the more subprojects closer to the root of the decision tree, the more important it is. Through the decision tree model, the importance of each subproject in project construction is clearly known. Therefore, the supervisory work of the supervisory authority is highly targeted. Through the decision tree model, a more streamlined forecasting rule can be obtained than the C4.5 algorithm, and this rule can be applied to the engineering quality supervision and decision-making. Through forecasting the future project, it can help the supervisory and
management department to discover the impact on the project quality early. The factors that enable them to find problems in advance, solve problems in time, and achieve the purpose of rainy days.

**Conclusion**

The above experimental analysis and application can see that the improved algorithm is less in the modeling time than the original C4.5 algorithm, and is also smaller than the original C4.5 algorithm in the model scale, and the improvement in the overall accuracy of the model is also considered to be better than the original C4.5. The algorithm is high, which fully shows that the improved algorithm improves the efficiency of decision tree construction. This algorithm is applied to engineering quality supervision and decision-making, predicts future projects through decision tree rules, and helps regulatory authorities to discover potential problems in engineering projects in advance. To reduce the occurrence of engineering accidents.

**Acknowledgements**

Guizhou Province Big Data Application Engineering Research Center (Qian Jiao KY[2017]022)

**References**