Research on the Application of Decision Tree Algorithm in the Selection of Variables in Cost Estimation Model

Ping-Hao ZHANG\textsuperscript{a,*} and Kang FU\textsuperscript{b}

Naval Research Academy, Beijing, China
\textsuperscript{a}85106639@qq.com, \textsuperscript{b}84557328@qq.com

*Corresponding author

Keywords: Decision Tree, Cost Estimation, Model Prediction, Variable Selection.

Abstract. In order to solve the problem that there may be too many input variables, which affect the convergence speed of the model and the easy formation of over fitting, this paper proposes to use the decision tree algorithm to search for the best grouping variables when constructing the decision tree. Explore a new method to filter model input variables. Through the case analysis, the precision of the forecast model is satisfied by using the decision tree algorithm. This method will be an important knowledge method reserve in the processing of specific engineering problems. Especially for complex problems such as low-dimensional analysis, rapid modeling will play an important role.

Introduction

Decision tree algorithm is an important algorithm for classification prediction. The goal is to establish classification model or regression model. The key to establishing the decision tree is how to find the best grouping variable from the many input variables and how to find the best segmentation point from the grouping variables. It is not difficult to see that the grouping variable must be the most important variable in the classification prediction model, and this variable must be the key variable that can most affect the accuracy of model prediction. To this end, this paper proposes to use the decision tree algorithm to analyze how to select the key variables among the many variables with similar degree of correlation with the target variables. The model dimension is reduced without significantly reducing the model precision. Finally, by comparing the model prediction values established after using the decision tree algorithm to screen the variables, and other models such as time series models and neural network algorithms, using all the relevant variables to model the prediction results, analyze the error range, and summarize the application laws.

Feasibility Analysis of Decision Tree Algorithm Screening Variables

According to the different types of output variables, the decision tree has the difference between the classification prediction model and the regression prediction model, which corresponds to the sub-type variable and the continuous numerical variable respectively. The difference of variable types also determines the difference between the two forecasting models in determining the grouping variable algorithm. The classification prediction takes the C5.0 algorithm as an example. Based on information entropy, the best grouping variables and segmentation points are determined by calculating the information increase rate. The regression prediction based on CART is based on Gini coefficient and variance.

According to the classification variables, the best grouping variables of the forecasting model are selected from the information theory. Information entropy is the mathematical expectation of the amount of information, and it is the average uncertainty before the source sends the information. Its mathematical definition is:

\[
\text{Ent}(U) = \sum_i P(u_i) \log_2 \frac{1}{P(u_i)} = -\sum_i P(u_i) \log_2 P(u_i)
\]

(1)
P(u_i) is the probability distribution of the transmitted information. If the information entropy \( \text{Ent}(U) = 0 \), it means that only the only possibility of sending information is \( P(u_i) = 1 \). Therefore, the smaller the difference in \( P(u_i) \), the greater the information entropy, the smaller the average uncertainty. When receiving the signal \( V = v_i \), the probability distribution of the sending signal is expressed as the conditional probability under the \( v_i \) signal, \( \text{IE } P(U | v_i) \). The mathematical definition of the average uncertainty of the source is based on the formula:

\[
\text{Ent}(U | v_i) = \sum_i P(u_i | v_i) \log_2 \frac{1}{P(u_i | v_i)} = -\sum_i P(u_i | v_i) \log_2 P(u_i | v_i)
\]

Here \( \text{Ent}(U | v_i) \) is called post-entropy, reflecting the information measurement of the signal \( U \) obtained after receiving \( v_i \). Usually \( \text{Ent}(U | V) \) \( \text{Ent}(U) \), so there are:

\[
\text{Gains}(U, V) = \text{Ent}(U) - \text{Ent}(U | V)
\]

Called information gain, it reflects the degree to which information eliminates random uncertainty. According to this, during the establishment of the decision tree, the \( \text{Ent}(U) \) of the output variable that is the source of the information \( U \) is calculated first, that is, the average uncertainty; Secondly, calculate the post-entropy \( \text{Ent}(U | v_i) \) under the input variable that receives a series of information \( V \) as a letter; Then, the information gain of the input variable is calculated by equation 3; Finally, the maximum of the information gain from all input variables is selected, that is, the best grouping variable. From the point of view of information theory, because the signal(input variable) has the strongest ability to eliminate the average uncertainty of the trust to the source, the resulting sample packet output variable has the highest degree of convergence within the two groups. There are large differences in \( P(u_i) \) within each group.

In order to avoid how much the number of category values affects the situation where the input variable is reasonably used as the best grouping variable, the information increase rate is introduced:

\[
\text{GainsR}(U, V) = \frac{\text{Gains}(U, V)}{\text{Ent}(V)}
\]

If the input variable \( V \) has more classification values, its own information entropy will be too large, and the information increase rate will be reduced, thus eliminating the impact of the number of categories. Similarly, for continuous numerical variables, regression model prediction is used, Gini coefficients and variance are used as the basis for selection, and overfitting problems are avoided through pruning of test sample sets. The specific application will be implemented through SPSS Model in subsequent case analysis.

**Case Analysis**

With the continuous development and enhancement of our economy and national strength, the scale and technology of our port are accelerating and optimizing. The pier is the main component of the port hydraulic building. It is a hydraulic building that provides ships with functions such as docking, loading and unloading cargo, and passengers. Investment estimation in the feasibility study stage is an important basis for project decision-making and scheme selection, and also an important basis for studying and analyzing the economic effects of project investment. This chapter will use the cost estimate of the dock construction project in the feasibility study stage as a case study to analyze how to use the decision tree algorithm to select key variables from many influencing factors and simplify the model without significantly reducing the accuracy of the model. In turn, it helps to understand the physical meaning of the model.

**Resource Estimation Model Impact Factors**

According to the research experience of domestic and foreign experts and scholars, the influence factors summarized will be integrated and classified. The initial formation of 17 factors that affect the
cost of the dock project: water flow, wind, waves, water level, water depth, Geological conditions, surrounding environment, Foundation bed thickness, Foundation bed dredging, Foundation bed throwing, caisson volume, caisson prefabricated, caisson installation, caisson backfill, behind the wall Backfill materials, upper facilities, gravel prices. Then qualitative screening was conducted through expert consultation methods, and 11 variables were finally selected as the independent variables of the estimation model: pier length (meters, X1), pier width (meters, X2), and design depth (meters, X3). Water distance (km, X4), land distance (km, X5), dredging volume (Square, X6), base bed throwing volume (Square, X7), caisson volume (Square, X8), steel composite index (X9), National Cement Index(X10), Regional(X11), etc.. Y is the terminal project cost (10,000 yuan). By literature[ 3] Fifteen sample data were obtained and normalized as shown in table 1.

<table>
<thead>
<tr>
<th>project</th>
<th>Y</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
<th>X11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.38</td>
<td>0.13</td>
<td>0.00</td>
<td>0.51</td>
<td>1.00</td>
<td>0.67</td>
<td>0.08</td>
<td>0.00</td>
<td>0.63</td>
<td>1.00</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.09</td>
<td>0.38</td>
<td>0.00</td>
<td>0.41</td>
<td>0.72</td>
<td>0.44</td>
<td>0.67</td>
<td>0.16</td>
<td>0.20</td>
<td>1.00</td>
<td>0.35</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.24</td>
<td>0.38</td>
<td>0.42</td>
<td>0.30</td>
<td>1.00</td>
<td>1.00</td>
<td>0.33</td>
<td>0.49</td>
<td>0.04</td>
<td>0.65</td>
<td>0.94</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.38</td>
<td>0.43</td>
<td>0.13</td>
<td>0.85</td>
<td>0.41</td>
<td>1.00</td>
<td>0.33</td>
<td>0.39</td>
<td>0.45</td>
<td>0.47</td>
<td>0.42</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.65</td>
<td>0.83</td>
<td>0.37</td>
<td>0.70</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.30</td>
<td>1.00</td>
<td>0.54</td>
<td>0.54</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.91</td>
<td>0.84</td>
<td>0.61</td>
<td>0.79</td>
<td>0.05</td>
<td>0.22</td>
<td>0.33</td>
<td>0.36</td>
<td>0.84</td>
<td>0.47</td>
<td>0.31</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
<td>0.88</td>
<td>1.00</td>
<td>0.70</td>
<td>0.65</td>
<td>0.11</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.61</td>
<td>0.40</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0.40</td>
<td>0.28</td>
<td>0.67</td>
<td>0.83</td>
<td>0.00</td>
<td>0.22</td>
<td>0.67</td>
<td>0.01</td>
<td>0.42</td>
<td>0.69</td>
<td>0.69</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.49</td>
<td>0.77</td>
<td>0.25</td>
<td>0.73</td>
<td>0.07</td>
<td>0.22</td>
<td>0.33</td>
<td>0.00</td>
<td>0.23</td>
<td>0.62</td>
<td>0.85</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0.69</td>
<td>1.00</td>
<td>0.25</td>
<td>1.00</td>
<td>0.69</td>
<td>0.78</td>
<td>0.33</td>
<td>0.10</td>
<td>0.82</td>
<td>0.00</td>
<td>0.07</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0.00</td>
<td>0.00</td>
<td>0.47</td>
<td>0.85</td>
<td>0.54</td>
<td>0.78</td>
<td>0.33</td>
<td>0.04</td>
<td>0.10</td>
<td>0.09</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0.14</td>
<td>0.32</td>
<td>0.23</td>
<td>0.53</td>
<td>0.26</td>
<td>0.56</td>
<td>0.33</td>
<td>0.06</td>
<td>0.18</td>
<td>0.36</td>
<td>0.34</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0.03</td>
<td>0.31</td>
<td>0.23</td>
<td>0.41</td>
<td>0.02</td>
<td>0.22</td>
<td>0.33</td>
<td>0.05</td>
<td>0.00</td>
<td>0.62</td>
<td>0.76</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0.35</td>
<td>0.38</td>
<td>0.47</td>
<td>0.50</td>
<td>0.04</td>
<td>0.44</td>
<td>0.67</td>
<td>0.17</td>
<td>0.08</td>
<td>0.42</td>
<td>0.40</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0.18</td>
<td>0.30</td>
<td>0.29</td>
<td>0.55</td>
<td>0.03</td>
<td>0.33</td>
<td>0.33</td>
<td>0.05</td>
<td>0.03</td>
<td>0.64</td>
<td>0.75</td>
<td>0</td>
</tr>
</tbody>
</table>

### Variable Screening of Decision Tree Algorithm

For table 1 data, SPSS Model 14.2 software is used to establish the decision tree prediction algorithm model workflow. Among them, the cost Y of the dock project is used as the target, and other variables are used as input; 80 % of the data is used as a training sample, 20 % of the data is used as a test sample, and other settings follow the default option to obtain the importance distribution of the predicted variables.

The variable of pier length is the most important and belongs to the best grouping variable in the first layer. The binary tree is then generated according to whether the length of the pier is greater than a certain critical value. Then the decision tree continues to be constructed by the design depth of the variable importance secondary level. The two branches formed by the corresponding second dock length did not build a follow-up decision tree. It was due to the default setting of the pruning standard that cut off the subsequent bifurcation of the branch to avoid overfitting. The subsequent bifurcation of the decision tree's second layer design depth introduces the pier width as the best grouping variable at this stage, and forms the fourth and final decision tree. The same follow-up variable is discarded due to the default fitting pruning setting. Since this article focuses on the feasibility of using the decision tree search for the best packet variable as the key modeling variable of the cost prediction model, the specific pruning strategy of SPSS Model software is not repeated here. This example adopts the software default algorithm setting.

Through the above analysis, it can be seen that the establishment of the dock engineering estimation model can be established by the best grouping variables in the three stages of the decision tree establishment process, namely the pier length X1, the design depth X3, and the pier width X2.
Budget Model Based on Screening Variables Established

According to the above three variables, the model is established by SPSS Model 14.2. The modeling node selects regression. In addition to the above three variables, the other variable roles are defined as none in the type node. The model that generates the yellow diamond shape through the software runs internally.

The error of the two new sets of data is 5.41% and 3.74% respectively, which satisfies the forecast precision of the project cost in the feasibility study stage.

Comparative Analysis of Predictive Results of Non-Decision Tree Algorithm

In this chapter, the 11 variables summarized in the previous article are input. Two methods, time series model and neural network algorithm, are used to establish the prediction model.

Time Series Mode

The establishment of the time series model uses SPSS Model 14.2, the model type is ARIMA (0, 0, 0), and the fitting statistical indicator R is selected. MAPE evaluates the prediction results. In general, the closer the R party is to 1, the closer the MAPE is to 0, indicating that the better the fitting degree of the model is; R party 0.998 is close to 1, and MAPE 1.669 is closer to 0 than other project analysis results.

Based on the time series model, and the fitting state of the observed and quasi-values is better. The relative errors are 7.12% and 6.14%.

Neural Network Algorithm

The MLP neural network analysis was conducted using the SPSS Modeler 14.2 software "multi-layer receptor". Because the variable selected the terminal project cost, the factor was the region, and the other ten input variables were used as covariables; 80% of the sample is used as a training sample, and 20% of the sample is used as a test sample; Other settings use software default values to generate the structure diagram, including 1 output layer, 1 hidden layer, and 1 output layer. The number of neurons in the input layer is 12, and the number of neurons in the hidden layer is 8, using hyperbolic tangent. Activate function. The largest relative error of the model is that the No. 13 sample is 4.89%, and the smallest error is 0.06% of the No. 10 sample, and the fitting state is good. From this, the relative error of the sample results of No. 16 and No. 17 is predicted to be 4.85% and 3.89%, respectively.

Comparative Analysis

A summary of the relative errors of the above three models for the predicted results of samples 16 and 17 is shown in table 2.

<table>
<thead>
<tr>
<th>Name</th>
<th>Three Variable Model</th>
<th>Time Series Model</th>
<th>Neural Network Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 16</td>
<td>5.41%</td>
<td>7.12%</td>
<td>4.85%</td>
</tr>
<tr>
<td>Sample 17</td>
<td>3.74%</td>
<td>6.14%</td>
<td>3.89%</td>
</tr>
</tbody>
</table>

It can be seen from the comparison that the relative errors of samples 16 and 17 obtained by using the three-variable prediction model generated by the decision tree algorithm are in the same order as the errors of the other two algorithms, and they are all within a reasonable range of 10% of the error in the feasibility study stage of the project. Under the premise of similar precision requirement, the decision tree method can be used to select a small number of variables to quickly establish a prediction model.
Summary

Decision tree algorithm is an important algorithm for classification prediction, and it has a more complete system prediction process. In this paper, the process of determining the optimal grouping variables in the decision tree algorithm is selected to realize the variable selection of the project. By comparison with other prediction results of all-variable modeling, it is shown that this method can be used for other data analysis, data mining and other applications. At the same time, with the increasing complexity of the research object and the increasing number of case samples, the significance of the variable selection method and the improvement of modeling efficiency will become more and more significant. Therefore, as a new variable selection method, it will be an important knowledge method reserve in the treatment of specific engineering problems. In particular, the low-dimensional analysis and rapid modeling of complex problems will play an important role.

References