Research on Construction Engineering Project Management Optimization Based on C4.5 Improved Algorithm

Yawen Zhong
School of Civil Engineering & Architecture, Southwest Petroleum University, Sichuan, 637000, China
*Corresponding author’s e-mail: arwenzyw@163.com

Abstract. In view of the problem that enterprises are not quick enough to respond to market changes in current construction engineering project management, this paper proposes to use the improved C4.5 decision tree to optimize construction project management. Firstly, the optimization attributes of construction project management are constructed. Relevant experts are hired to score the attributes and project risks. Then, the evaluation is discretized by rough set theory. Finally, the C4.5 decision tree algorithm is applied to utilize the project information and generate the decision tree of project management optimization, find out the classification rules affecting the optimal management of engineering projects, analyze the implementation of construction projects, and give personalized tips, so as to achieve effective control and optimization of enterprise engineering projects.

1. Introduction
The rapid development of social economy has brought tremendous development opportunities to construction companies. In order to make construction enterprises adapt to economic development and meet social needs, it is necessary to do a good job in construction project management [1-2].

Because there are certain differences in the management methods and competition levels, the management effects of engineering projects of different enterprises are obviously different. Therefore, in order to better meet the needs of society and market, different enterprises should optimize the engineering project management according to their own resources advantages and environment of competition [3].

In this paper, the C4.5 decision tree algorithm is introduced into the optimization process of enterprise project management. Decision tree algorithm is mainly used to predict and classify research data. It is currently the most widely used data utilizing algorithm decision tree. Due to its high efficiency and small error, it has been widely used in classification problems [4]. Reference [5] proposes a project management optimization research based on ID3 decision tree, but ID3 directly uses the gain of information entropy, which will lead to a relatively large information entropy due to the fact that a certain attribute has more category values.

2. C4.5 Decision Tree Algorithms

2.1. Decision Tree Algorithms
The Decision Tree [6] algorithm is a supervised classification algorithm for data utilization. By learning the sample set to construct a classification model or function, the classification model or function can map data records to a specific category one by one, so that the corresponding relationship can be applied
to data classification. The decision tree usually consists of decision points, branches, and leaf nodes, and the final decision model is represented by a binary tree or a multi-fork tree.

2.2. Construction of C4.5 Decision Tree Algorithm

C4.5 uses information gain rate, and is the deformation of information entropy. Considering the "cost" of splitting information, the information gain rate can partially offset the influence of the number of attribute values, so it is one of the important improvements of the ID3 algorithm by C4.5. The C4.5 algorithm increases the information gain rate on the basis of computing information entropy, and takes the information gain rate as the basis of selecting conditional attributes, which greatly improves the accuracy of decision tree construction [7].

The main process of C4.5 algorithm can be divided into the following steps:

Information Entropy:

\[
\text{Entropy}(D) = -\sum_{i=1}^{m} p_i \log_2 p_i
\]  

(1)

Where \(D\) is the training set for creating the decision tree, \(p_i\) (i=1, 2, ..., m) is the frequency of occurrence of each value of the attribute \(A\) in the total training set;

(1) Category information entropy:

If the training set \(D\) is divided by attribute \(A\), the entropy of the data set \(\text{Entropy}_A(D)\) divided by attribute \(A\) is calculated. Assuming that \(A\) has \(k\) values, then attribute \(A\) classifies \(D\) into \(k\) subsets \(\{D_1, D_2, D_3, ..., D_k\}\) according to \(K\) values. Attribute \(A\) divides the information entropy of \(d\) as [8-9]:

\[
\text{Entropy}_A(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \text{Entropy}(D_i)
\]  

(2)

In the formula, \(|D_i|\) and \(|D|\) are the number of samples included in \(D_i\) and \(D\), respectively; The difference between the information entropy of training set \(D\) and the information entropy divided by attribute \(A\) is the information gain (Equation 3). Generally, the larger the information gain value is, the higher the purity of the divided data set will be, and the better the partitioning effect will be:

\[
\text{Gain}(D, A) = \text{Entropy}(D) - \text{Entropy}_A(D)
\]  

(3)

The calculation formulas of split information and information gain rate are shown in formula (4) and formula (5), respectively. It can be seen from equation (4) that the value of the split information increases as the value of the attribute increases, and it can be seen from equation (5) that when calculating the information gain rate, the split information is used as the denominator, thus, to some extent, it can eliminate the influence caused by a great number of attribute values.

Split information:

\[
\text{Split}(A) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \log_2 \left(\frac{|D_i|}{|D|}\right)
\]  

(4)

Information gain ratio:

\[
\text{GainRatio}(A) = \frac{\text{Gain}(D, A)}{\text{Split}(A)}
\]  

(5)

The C4.5 algorithm overcomes the problems of the ID3 algorithm, and at the same time develops strategies for improvement. Its algorithm principle is simple, the classification accuracy is high, and the speed is fast. It has been widely used and studied in both scientific research and practical work. Especially in the single-machine decision tree algorithm, the C4.5 algorithm not only has high classification accuracy but also has fast classification speed [10-11].

2.3. C4.5 decision tree improvement

The C4.5 algorithm is improved based on the mean value theorem. When calculating the entropy and information gain in the C4.5 algorithm, it involves the operation of the logarithm. The system function
is called every time, and the amount of calculation is large. By simplifying the calculation of entropy and information gain through Taylor's Mean Value Theorem, the C4.5 algorithm is improved, and the performance of the C4.5 algorithm is improved in terms of efficiency.

In order to solve this problem, an improved method for calculating formula of information gain rate is proposed, which simplifies the calculation of information gain rate of C4.5 algorithm by using Taylor formula and McLaughlin formula and greatly reduces the computational complexity of the algorithm. The improved C4.5 algorithm is named TAM-C4.5 algorithm (Taylor and McLaughlin-C4.5).

Since the derivative of $\ln(x)$ at $x = 0$ is meaningless, and the probability range commonly used in the formula of information gain rate is between $[0, 1]$, the McLaughlin formula of $\ln(x)$ selected in this paper improves the calculation formula of information gain ratio in traditional C4.5 algorithm, such as:

$$\ln(x + 1) \approx x - \frac{1}{2} x^2 + \frac{1}{3} x^3 \ldots + (-1)^{n-1} \frac{1}{n} x^n$$  \hspace{1cm} (6)

$$\ln(x - 1) \approx \frac{1}{2} (x - 1)^2 + \frac{1}{3} (x - 1)^3 \ldots + (-1)^{n-1} \frac{1}{n} (x - 1)^n$$  \hspace{1cm} (7)

When $x \in (0, 1)$,

$$\ln(x) \approx (x - 1) - \frac{1}{2} (x - 1)^2 + \frac{1}{3} (x - 1)^3$$  \hspace{1cm} (8)

Through the above approximate simplification, the logarithmic operation can be converted into a non-logarithmic operation, and the complex logarithm operation in the formula of the information gain rate can be eliminated by using the above-mentioned conversion feature, thereby simplifying the calculation formula and improving the efficiency of building trees. And the formula (8) is more accurate than the equivalent infinitesimal $\ln(x + 1) = x$.

The conversion formula of information entropy is shown in equation (9):

$$\text{Entropy} (D) = -\sum_{i=1}^{m} p_i \log_2(p_i) = -\sum_{i=1}^{m} \frac{D_i}{D} \log_2 \left( \frac{D_i}{D} \right) = -\sum_{i=1}^{m} \frac{D_i}{D} \ln \left( \frac{D_i}{D} \right)$$

$$= -\frac{1}{\ln 2} \times \sum_{i=1}^{m} D_i \times \ln \left( \frac{D_i}{D} \right) = -\frac{1}{\ln 2} \times \sum_{i=1}^{m} D_i \times \left[ \frac{(D_i - D)(11D^2_i + 2D^3_i - 7D_i D)}{6D^2} \right]$$  \hspace{1cm} (9)

Similarly, the conversion formula of category information entropy and split information entropy is shown as equation (10) and equation (11):

$$\text{Entropy}_A (D) = -\frac{1}{\ln 2} \times \sum_{i=1}^{k} \sum_{j=1}^{m} \left[ \frac{D_{ij}(D_{ij} - D)(11D^2_{ij} + 2D^3_{ij} - 7D_{ij} D)}{6D^2_{ij}} \right]$$  \hspace{1cm} (10)

$$\text{SplitE} (A) = -\frac{1}{\ln 2} \times \sum_{j=1}^{k} \left[ \frac{D_j(D_j - s)(11D^2_j + 2D^3_j - 7D_j D)}{6D^2_j} \right]$$  \hspace{1cm} (11)

Therefore, the calculation formula of the converted information gain ratio is shown as equation (12):

$$\text{GainRatio} (A) = \frac{\text{SplitE} (D) - \text{SplitE}_A (D)}{\text{SplitE} (A)}$$  \hspace{1cm} (12)

Analysis of the improved calculation formula shows that the category information entropy is the same when calculating the value of information gain ratio of the condition attribute. Since $\frac{1}{\ln 2 \times s}$ of all parts are omitted in the simplification formula, in order to ensure the classification accuracy of the
algorithm, in this paper, the improved formula is adopted when calculating the category conditional entropy to ensure that the order of information gain ratio of each condition attribute is unchanged, and the classification accuracy is not affected. The traditional C4.5 algorithm needs to call the function to perform a large number of logarithmic function operations. The improved algorithm proposed in this paper only needs a simple four-mixed operation, which eliminates the logarithm operation frequently called in the information gain ratio calculation formula, and the system operation speed is greatly increased.

3. Improved project management optimization algorithm for C4.5 decision tree algorithm
1) Select a certain construction project of the enterprise, construct the relevant attributes of the project management optimization, hire relevant experts to score the attribute values, collect the scores given by the experts, and construct a training set optimized by the project management;
2) Use the rough set theory to discretize the training set;
3) Use the improved C4.5 decision tree algorithm, select the node attribute with the highest information gain as the root node;
4) Then the second highest node is found as a branch of the decision tree, and the decision tree is generated recursively by analogy.
5) The C4.5 decision tree algorithm is executed once per loop, and the generated decision tree is updated once.

3.1. Three Elements Analysis of Project Management
Project management aims at balancing resources at the lowest cost and controlling project quality. According to relevant research results, project management attributes can be classified into four basic attributes: resources, technology, cost and duration.

In the process of project management, the plan is to ensure the smooth implementation and completion of the project. In advance, it is necessary to ensure the quality of the project and predict the risks of the project in implementation; the technology applied is the technical guarantee to ensure the design and implementation of the project; cost is the financial budget to ensure the project in the implementation process; progress is as follows: the estimated and allocated operating time for project design, construction and installation.

From the perspective of economics, the mutual constraint relationship between resources, technology, cost and construction duration in construction project management has produced multi-objective synergy problems, namely multi-objective optimization.

3.2. Engineering project management optimization based on C4.5 decision tree algorithm
The software of MATLAB platform is used to simulate the experiment. Taking the optimization of project management of a state-owned construction enterprise as the objective attribute, four attributes of project management are selected to optimize the project management, and the useful classification rules are extracted by using C4.5 decision tree algorithm to guide the optimal operation of enterprise project. The operation of a certain engineering management project in a state-owned enterprise is analyzed. Relevant experts are invited to score the four indicators according to the actual situation of the sample, and the comprehensive score is obtained according to certain rules. Finally, the risk of the project is evaluated according to the comprehensive score of each attribute. Table 1 shows the discretization results of 10 experts' evaluation on the construction project.

3.3. Information Gain of Traditional C4.5 Decision Tree Algorithms
In order to reduce operational errors, the expert scoring table of the project is discretized by rough set algorithm, and then the expected information and gain rate of resources \(a_1\), technology \(a_2\), cost \(a_3\), duration \(a_4\) and project risk \(a_5\) of the project are calculated by using formula (1) ~ formula (3).
In order to improve the efficiency of project management optimization, the C4.5 decision tree algorithm is improved, and the calculations are carried out by using equations (9) to (12):

The information entropy of each item category is:

\[ \text{Entropy}_{\text{a}}(a_s) = 3.1273 \]
\[ \text{Entropy}_{\text{a}}(a_s) = 35.5956 \]
\[ \text{Entropy}_{\text{a}}(a_s) = 36.8136 \]
\[ \text{Entropy}_{\text{a}}(a_s) = 35.5956 \]

The information entropy: \( \text{Entropy}(a_i) = 25.6918 \)

Split information:

\[ \text{SplitE}(a_i) = 8.4709 \]
\[ \text{SplitE}(a_i) = 5.3689 \]
\[ \text{SplitE}(a_i) = 5.3520 \]
\[ \text{SplitE}(a_i) = 6.5378 \]

Improved information gain ratio: \( \text{GainRatio}(a_i) = 0.8854 \)
\( \text{GainRatio}(a_i) = 1.1765 \)
\( \text{GainRatio}(a_i) = 1.2246 \)
\( \text{GainRatio}(a_i) = 1.1310 \)

In order to intuitively reflect the size of each project in the C4.5 decision tree, the calculation of the above data has not been simplified \( \frac{1}{\ln 2 \times s} \) and \( \frac{1}{\ln 2 \times s} \) of all parts in practical engineering application calculation have been omitted.

### Table 1. Discretization results of construction project management indicators.

| Project | a₁ | a₂ | a₃ | a₄ | a₅ |
|---------|----|----|----|----|----|
| 1       | 0  | 1  | 3  | 2  | 1  |
| 2       | 2  | 0  | 2  | 3  | 2  |
| 3       | 1  | 2  | 2  | 2  | 1  |
| 4       | 0  | 2  | 1  | 2  | 2  |
| 5       | 2  | 1  | 0  | 0  | 1  |
| 6       | 1  | 1  | 3  | 1  | 0  |
| 7       | 2  | 2  | 1  | 1  | 1  |
| 8       | 2  | 1  | 2  | 2  | 1  |
| 9       | 1  | 0  | 3  | 3  | 1  |
| 10      | 1  | 3  | 1  | 1  | 2  |
The GainRatio ($a_i$) is relatively small, that is, the progress indicator should be used as the test attribute to create a decision tree, and then each branch is recursively calculated.

To compare the performance changes before and after the improvement of C4.5 decision tree algorithm, for engineering project management with different numbers, the comparison of time required to process the same data is shown in Figure 1. It can be seen from Figure 1 that the improved C4.5 algorithm can effectively reduce the classification time. In other words, the improvement of classification efficiency is not at the expense of the accuracy rate, and the improved C4.5 decision tree algorithm is much faster than the traditional C4.5 decision tree algorithm. Besides, the work efficiency is greatly enhanced.

![Figure 1. C4.5 Comparison of running time before and after improvement of decision tree algorithm](image)

4. Conclusion

Traditional management methods have been difficult to adapt to changes in the environment. In order to better adapt to the needs of society and the market, targeted management of enterprise engineering projects has become an inevitable trend. In this paper, by analyzing the shortcomings of the traditional C4.5 algorithm, the calculation method of attribute selection is improved. The improved algorithm uses the Taylor formula and the McLaughlin expansion formula to obtain an approximate formula to complete the approximate conversion of the information gain ratio formula. It not only ensures that the simplified algorithm retains high precision, but also eliminates complex logarithmic operations in the function, which ensures the accuracy of classification while speeding up the classification speed.

The C4.5 decision tree algorithm constructs the enterprise engineering project management decision tree, obtains the classification rules, utilizes the project management information, and realizes the personalized prompt and targeted management of the project management.

Acknowledgments

This work was supported by the Nanchong key laboratory of robot engineering and intelligent manufacturing under Grant NCKL201706

References

[1] Chen Yongqiang. (2006) Gu Wei. Summary of Research on Risk Management of Engineering Projects. Science & Technology Progress and Policy., 29(18):157-160
[2] Zhang Jian. (2018) Talking about the Improvement of the Ability of Integrated Management of International Engineering Projects. Yangtze River., 49(2):223-225
[3] Wei Wenping. (2015) Influencing factors and countermeasures of construction engineering management. Research on Financial and Economic Issues.,2015,6 ,(379):67-70
[4] Fan Yanyan. (2016) Application of ID3 decision tree mining in teaching assistant system. Journal of Qiqihar University (natural science edition)., 32(1):85-89
[5] Pu Tiantian. (2018) Study on engineering project management optimization based on decision tree. Modern Electronics Technique., 41(1):169-172
[6] Lu Wei, Wang Cong. (2007) Optimization and Comparison of Decision Tree Algorithm. Computer Engineering., 33(16):189-190

[7] Zhou Tao, Ji Weixia, Song Chengxuan. (2018) Manufacturing Process Quality Management Based on Decision Tree C4.5 Algorithm. Modular Machine Tool & Automatic Manufacturing Technique., 12:134-141

[8] An Weipeng, Shang Jiaze. (2019) Improvement and Analysis of C4.5 Decision Tree Algorithm. Computer Engineering and Applications., 55(12):169-173

[9] Wei Hao, Ding Yaojun. (2014) An Improved Algorithm of C4.5 Decision Tree Based on Attributes Correlation., 35(4):402-406

[10] Yao Yatao, Xing Liutao. (2011) Improvement of C4.5 decision tree continuous attributes segmentation threshold algorithm and its application. Journal of Central South University (Science and Technology), 42(12):3372-3376

[11] Wang Miao, Chai Ruimin. (2010) Improved classification attribute selection scheme for decision tree. Computer Engineering and Applications., 46(8):127-129

[12] Liu Peng, Yao Zheng, Yin Junjie. (2006) Improved decision tree of C4.5. J Tsinghua Univ (Sci &Tech), 46(S1):997-1001