Research on Civil Engineering Cost Prediction Based on Decision Tree Algorithm

Yin Bai

Liaoning Institute of Science and Technology, Benxi, 117004, China

Abstract: *Civil engineering includes not only all engineering construction on the ground, but also the maintenance, exploration and design of equipment and materials used in the whole construction process. During the construction of civil engineering, there is no scientific and reasonable supervision system for the supervision of construction, which greatly affects the quality of engineering and the control and management of engineering cost. In this paper, through the application of DT (Decision tree) algorithm, the research of civil engineering cost prediction is carried out. Aiming at the application of the algorithm in civil engineering cost management, this paper tries to improve the C4.5 algorithm. DT civil engineering life-cycle cost analysis and prediction model is trained by training set, and the optimal result of civil engineering life-cycle cost analysis is obtained by inputting sample data for model prediction. The application results show that the above algorithm has lower computational complexity and higher prediction efficiency in civil engineering cost prediction, which can better meet the needs of actual civil engineering cost analysis and prediction.*

Keywords: *Decision tree; Civil engineering; Cost prediction*

1. Introduction

In order to improve the economic benefits and meet the actual needs of economic and social development, civil engineering construction units must adopt practical cost management strategies, effectively control economic elements, straighten out the relationship among various cost input elements, and cater to the actual needs of market and policy dual regulation [1]. To do a good job in a civil engineering project, we must have a detailed and effective management of the whole project cost. The control and management of civil engineering cost is mainly in the design stage, decision-making stage, implementation stage and completion stage. Its main purpose is to reduce the difference between actual expenditure and planned expenditure in time and realize reasonable financial expenditure.

The main function of civil engineering cost control management is to minimize the gap between actual expenditure and planned expenditure, implement the cost accounting of civil engineering, and realize the principle of balance cost and reasonable expenditure [2-3]. Through the application of DT (Decision tree) algorithm, this paper explores the law and connotation between civil engineering investment and various influencing factors, provides a method to quickly predict civil engineering cost under certain preconditions, and also provides reference opinions for informatization of cost management.

2. Overview of DT algorithm

Although DT is also a tree structure, it is different from a binary tree. DT can have two or two child nodes, and each child node represents an attribute. As shown in Figure 1. DT algorithm uses certain methods to classify and summarize a large number of disordered and complicated data, so that a tree-like classification rule can be obtained. The advantage of DT algorithm is that data analysts can make good use of this algorithm to analyze data without having a very rich knowledge background when using DT.

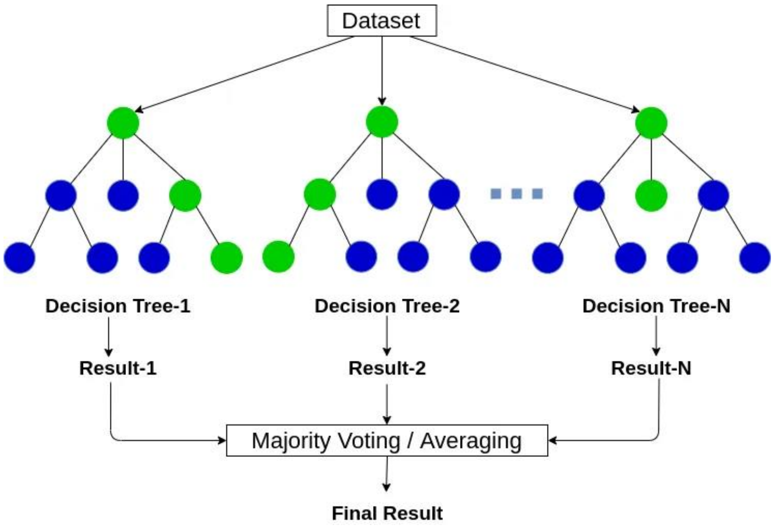

Figure 1: DT structure diagram

The reason why DT technology is so popular is that it has many advantages: First, DT's classification model uses a tree structure, and the classification is easy to understand and impressive; Second, users often can understand and realize the application of DT without having a profound and professional knowledge background; Third, DT's data preparation is relatively easy, and DT can handle discrete and continuous data. The C4.5 algorithm is a DT generation algorithm developed on the basis of ID3 algorithm, which overcomes the shortcomings of ID3 algorithm, such as its inability to process continuous data and its biased information gain [4-5].

The general rules of DT generation are: firstly, find out the most discriminating features in the training samples, then divide the data set into several subsets, then select the most discriminating features for each subset to divide, and finally divide them until all subsets only contain the same type of data, thus obtaining a DT[6]. DT pruning is a basic technique to overcome data noise, and it can also simplify DT [7-8]. DT pruning has two pruning methods, one is front pruning and the other is back pruning. Start statistical analysis from the leaf nodes of the tree to decide whether to keep branches, and then make decisions one by one in the direction of the root node.

3. Research method

3.1. Cost control management in civil engineer

All construction activities in real life are called civil engineering. Civil engineering includes not only all engineering construction on the ground, but also the maintenance, exploration and design of equipment and materials used in the whole construction process. Reasonable cost control of civil engineering can limit all the expenses of the whole project to the previous budget and maximize profits. In some civil engineering projects, in order to finish the construction period ahead of schedule and improve the performance level of the whole project, there are some problems that violate the relevant engineering construction rules and regulations. In some civil engineering projects, in order to finish the construction period ahead of schedule and improve the performance level of the whole project, there are some problems that violate the relevant engineering construction rules and regulations.

During the construction of civil engineering, there is no scientific and reasonable supervision system for the supervision of construction, which greatly affects the quality of engineering and the control and management of engineering cost. The engineering design must be strictly examined and verified, so that the engineering cost can be preliminarily estimated, and then reasonably controlled. The estimated cost should be regarded as the highest standard of the total engineering cost, and then reasonably controlled. This means the rational use of funds for civil engineering [9]. In the final stage of final accounts, attention should be paid to bidding, visa, negotiation, change and other links, so as to avoid repeated operations in the settlement process. Bill of quantities pricing is a pricing mode that is in line with international standards, which can truly reflect the open, fair and just competition principle of bidding. Bill of quantities valuation not only reflects the level of construction enterprises, but also effectively controls the investment of the project in the implementation stage.

The construction stage is that the whole process has entered the practical operation stage, which is materialized in accordance with the established schemes. For example, the influence of many uncontrollable factors, such as the human factors of construction personnel and the influence of objective environment [10]. The quota design should be adhered to in the whole design process. The so-called quota design is to design a specific construction scheme when the funds are determined. The liquidation of the project is to compare and audit all kinds of expenses that have already occurred with the project budget in the design stage, and count out the discrepancy and gap between the final actual construction expenses and the expenses scheduled in the design stage. The actual control or compression of the project cost in this stage can no longer be carried out, which is actually the summative stage of the whole project cost management and control.

3.2. Research on civil engineering cost prediction based on DT algorithm

As the cost control management of civil engineering runs through the whole construction project, it is dynamic for the whole project, and many factors will affect it. In the whole civil engineering, the design of civil engineering is the core content and the prerequisite for the smooth implementation of the project. When necessary, cost personnel, managers and designers can demonstrate together, so that the project cost in the design stage can be effectively controlled. In addition, factors such as possible claims in the construction process should be fully considered and predicted, so as to avoid the occurrence of claims. Therefore, the management of engineering change visa should be strengthened, and the construction unit should try its best to take effective measures in the construction process to avoid engineering change. Enterprises should strengthen the supervision of projects, strictly examine the changes of projects, and ensure that there will be no unreasonable engineering changes.

There are some problems in civil engineering, such as unreasonable resource allocation and excessive expenditure. Therefore, it is necessary to analyze and predict the whole life cycle of civil engineering, realize the optimal allocation of resources, and provide a reliable foundation and basis for civil engineering bidding and civil engineering management, which has important practical significance for promoting the scientific development of China's economy. This paper studies the DT-based civil engineering cost prediction algorithm to improve the accuracy of civil engineering cost prediction and meet the relevant requirements. DT civil engineering life-cycle cost analysis and prediction model is trained by training set, and the optimal result of civil engineering life-cycle cost analysis is obtained by inputting sample data for model prediction.

The author uses the method of deviation square sum to cluster the whole life cycle cost data of civil engineering, and divides the set n samples into k classes represented by $\{B_1, B_2, \cdots, B_k\}$, the number of samples in B_t is represented by n_t, the i th sample in B_t is represented by x_{it}, the center of gravity of B_t is represented by $\overline{x}_t$, and the calculation formula of deviation square sum of B_t is as follows:

$$S_t = \sum_{i=1}^{n_t} \left(x_{it} - \overline{x}_t\right)' \left(x_{it} - \overline{x}_t\right) \tag{1}$$

p_k is the proportion $p_k\left(k = 1,2,\cdots,|y|\right)$ of the k -th sample in the current sample set D, and the information entropy can be used as the purity index of the sample set. The lower the entropy, the higher the information purity.

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k \tag{2}$$

$Ent(D) = 0$, which means that all data sets are of the same category, so the information purity is the highest.

C4.5 when dealing with continuous value attributes, the algorithm first inserts a segmentation point among all values of continuous value attributes, then calculates the information gain rate based on these segmentation points, and finally selects the segmentation point with the largest information gain rate as the segmentation threshold. This poses a severe challenge to the efficiency of the algorithm. However, the data set of bill of quantities in the project cost forecast contains a large number of continuous attributes, and each continuous attribute has a large range of values, which is the data sample that this paper needs to study.

Based on the analysis and discussion of C4.5 algorithm, this paper tries to improve C4.5 algorithm for its application in civil engineering cost management. Discretize continuous attributes, select two or three segmentation points as thresholds for classification based on historical experience, and divide attribute values into three or four data sets. According to the pruning strategy that DT obeys most principles, when DT builds trees, it is set that when more than 85% of the samples in a given node belong to the same class, DT will stop recursion.

Through the above ideas and according to the actual civil engineering situation, set reasonable DT algorithm parameters to realize the civil engineering cost forecast. See Figure 2 for the specific process.

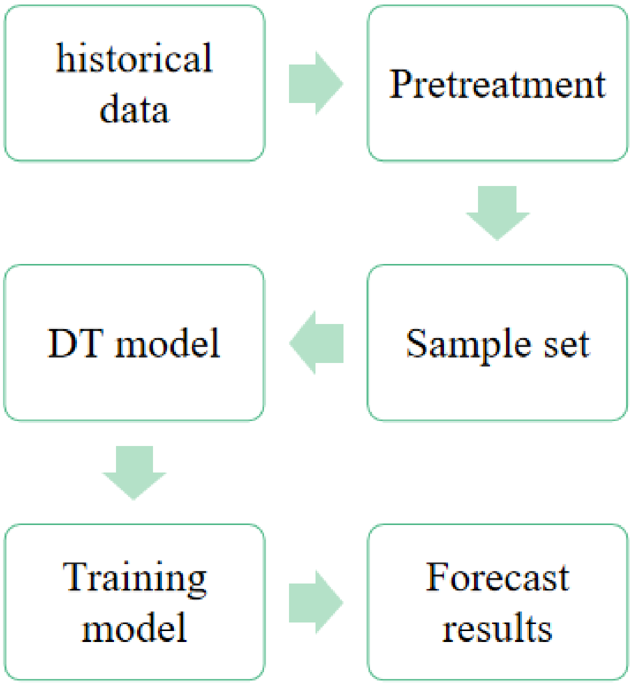

Figure 2: Forecast process

4. Applied analysis

In order to verify the universality and applicability of the classification rules extracted by the improved C4.5 algorithm DT in the application of project cost prediction, this paper randomly downloads a tender announcement project construction general contracting in the tender announcement column of public resources trading network as the verification research object. The collected engineering data is the training sample set.

According to statistics, there are 25 items that conform to the DT classification rules of the improved C4.5 algorithm of civil engineering, accounting for 83.3%, and 5 items that do not conform to the classification rules, accounting for 16.7% (as shown in Table 1). If the abnormal part of the training set data is excluded, this example verifies that the success rate is 100% in accordance with the extraction and classification rules. It proves the universal applicability of this research method.

Table 1: Distribution statistical results conforming to classification rules

	Quantity	Proportion
Conform to the rules	25	83.3%
Do not conform to the rules	5	16.7%

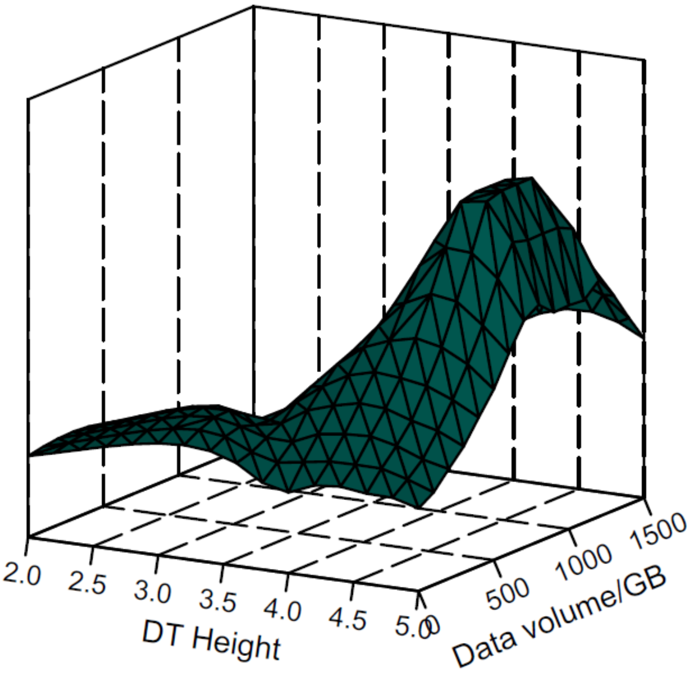

Figure 3: DT height map with different data volume

As can be seen from the above figure 3, with the increase of data volume, the DT height increases gradually at first and then remains stable, but the DT height never exceeds 5. It shows that DT obtained by using the above algorithm for analysis and prediction has a low height, and the above algorithm has low computational complexity and higher prediction efficiency in civil engineering cost prediction, which can better meet the needs of actual civil engineering cost analysis and prediction.

5. Conclusions

The main function of civil engineering cost control management is to minimize the gap between actual expenditure and planned expenditure, implement the cost accounting of civil engineering, and realize the principle of balance cost and reasonable expenditure. Through the application of DT algorithm, this paper explores the law and connotation between civil engineering investment and various influencing factors, and provides a method to quickly predict the cost of civil engineering under certain preconditions. The application results show that the improved C4.5 algorithm has lower computational complexity and higher prediction efficiency in civil engineering cost prediction, which can better meet the needs of actual civil engineering cost analysis and prediction.

References

[1] Zhang Xiaobo. (2022). Research on the Application of Engineering Cost in Civil Engineering. Engineering Seismic Resistance and Reinforcement, 2022(003), 044.
[2] Zhang Yongcheng, Guo Shuai, &Ye Yanbing. (2020). Engineering cost data information service system from the perspective of big data. Journal of Civil Engineering and Management, 37(1), 6.
[3] Hu Danping, &Tao Xueming. (2018). Improved design of cost model of post-earthquake reconstruction project based on improved genetic algorithm. Journal of Earthquake Engineering, 40(4), 6.
[4] Wang Xinyue, Zeng Hui, &Liu Tongfei. (2021). Simulation Research on Dispute Resolution Factors of Construction Cost Based on netlogo. Construction Economy, 042(005), 113-116.
[5] Jiang Hongyan, &Bai Yuqing. (2019). Cost estimation of high-rise housing based on grey correlation pso-bp neural network. Journal of Engineering Management, 33(1), 5.
[6] Xu Bing,&Yao Junyi. (2019). Design and Implementation of Price Adjustment Formula Platform for Construction Engineering. Construction Economy, 40(2), 5.
[7] Wang, Y., Xia, S. T., & Wu, J. (2017). A less-greedy two-term tsallis entropy information metric approach for decision tree classification. Knowledge-Based Systems, 120(15), 34-42.
[8] Xu, H., Wang, L., & Gan, W. (2016). Application of improved decision tree method based on rough

set in building smart medical analysis crm system. International Journal of Smart Home, 10(1), 251-266.

[9] Chen Dachuan, Yu Yi,&Liu Yuelong. (2021). Construction of construction project cost standard system based on system engineering. Journal of Civil Engineering and Management, 38(6), 6.

[10] Julia, Chen Fei, & Jing Liu. (2016). Engineering cost prediction model based on normalized network and generalized network. Practice and understanding of mathematics, 2016(7), 6.