



Research on the Application of Data Mining in the Construction Project Cost Estimation Model

Rui Wang*, Jing He, Jinrui Pan, Cancan Liao

China Coal Technology & Technology (Chongqing) Co., LTD. Chongqing, 400000, China

*ruiw_77@163.com

Abstract. In construction engineering management, cost estimation is crucial, but traditional models often lead to inaccurate estimation due to the failure to accurately capture the dynamic cost changes. To this end, we developed a data mining-based model for construction engineering cost estimation, using least squares support vector machine (LS-SVM) to establish a mathematical framework for SVM, and refined by an improved particle swarm optimization algorithm (PSO). The results confirm the estimation of construction cost. The results show that the data mining technology significantly improves the accuracy of construction engineering cost estimation, and the model performs better than other existing estimation models.

Keywords: Construction Engineering; Quality Management; Cost Estimation; Data Mining.

1 Introduction

With the economic progress and the improvement of people's living standards, the housing demand has promoted the rapid development of the construction industry in various cities [1]. Construction project costs increase with rising materials, labor, and management costs. Accurate estimation of these costs is essential for project management [2-3]. Traditional statistical methods are complex and prone to error. Although the multiple linear regression model provides more reliable estimates, its linear assumption is inconsistent with the actual cost change [4]. Data mining techniques, such as SVMs, can estimate the relationship between cost and impact factors more accurately, but are inefficient when processing large amounts of data [5].

Least squares support vector machine (LSSVM) is a fast learning and generalization data mining method [6]. In view of the deficiency of the existing model in reflecting the cost trend of construction projects, we propose an LSSVM model (PSO-LSSVM) based on particle swarm optimization. This model uses LSSVM to effectively capture the changing trend of construction cost, significantly improve the estimation accuracy and reduce the estimation error [7].

2 Model of Project Cost Estimation

2.1 Data Standards for the Cost Management System

Only by taking advantage of the opportunities brought by the era of big data, actively learning from foreign mature ways and methods, establishing management modes and measures in line with China's national conditions, and strengthening data analysis and management, can China's construction enterprises remain in an invincible position.

In big data research, ensuring data unification and standardization is crucial for subsequent analysis and mining. Although China construction enterprises have unified engineering measurement standards, but most enterprises still evaluate according to their own standards, which affects the data versatility and analysis efficiency. Therefore, the establishment of unified data standards is the key to big data processing, which helps to facilitate data exchange, improve processing efficiency, save time and reduce costs.

2.2 Cost Management System Data Mining Method

Construction cost is related to many factors, such as labor cost, materials, etc. There are n factors, which can be expressed as $x_i, i = 1, 2, \dots, n$, construction cost is y [8], so the change between the two is shown in formula 1.

$$y = f(x_i), i = 1, 2, \dots, n \quad (1)$$

The formula represents the estimated function of construction cost and cost, which directly affects the estimation of construction cost.

2.3 The Process of Data Mining in the Cost Management System

Information management system, based on the management concept, support enterprise strategic decision, create management atmosphere. After mastering the cost information, we will use the data mining technology to deeply understand the project cost, and fully support the cost control and decision-making of construction enterprises. Data mining of cost management information system includes three main parts: preprocessing, mining and knowledge conversion, each containing multiple sub-functions. Human-computer interaction is realized through a visual user interface. Enterprises use data warehouse and mining technology, combined with history and basic information for project cost accounting, prediction, decision-making, analysis and control, to achieve timely project cost management.

3 Construction Project Cost Estimation Model of Data Mining

3.1 Least-Squares Support Vector Machine

Set the historical sample of construction project cost for $(x_i, y_i), x_i \in R^n$ is the influence of construction cost, the input vector, $y_i \in R$ represents the corresponding construction cost value, using nonlinear mapping $\phi(\cdot)$ to map x_i [9]. When establishing

the accurate construction project cost performance estimation model, the optimal weight vector ω and deviation b are determined. According to the principle of structural risk minimization, the relaxation variable ζ_i is introduced to change it into the optimization problem of equation constraint, as shown in formula 2.

$$\begin{aligned} \min_{\omega, b, \zeta} \quad & \frac{1}{2} \omega^T \omega + \frac{\gamma}{2} \sum_{i=1}^N \zeta_i^2 \\ \text{s. t. } \quad & y_i = \omega^T \phi(x_i) + b + \zeta_i \end{aligned} \tag{2}$$

The Lagrangian is used to solve the formula 2 and establish the Lagrangian as follows as shown in formula 3.

$$L(\omega, b, \zeta, \alpha) = \frac{1}{2} \omega^T \omega + \frac{\gamma}{2} \sum_{i=1}^N \zeta_i^2 - \sum_{i=1}^N \alpha_i [\omega \phi(x_i) + b + \zeta_i - y_i] \tag{3}$$

Formula 4 is obtained from the Karush-Kuhn-Tucher condition.

$$\begin{bmatrix} 0 & S^T \\ S & K + 1/c \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{4}$$

The number in formula 4 can be represented by formula 5.

$$\begin{aligned} S &= [1, 1, \dots, 1]^T \\ \alpha &= [\alpha_1, \alpha_2, \dots, \alpha_N] \\ Y &= [y_1, y_2, \dots, y_N] \\ K &= \phi^T(x_k) \phi(x_N) = K(x_k, x_N) \end{aligned} \tag{5}$$

For this purpose, the kernel functions selected in this paper are specific as shown in formula 6.

$$K(x, x_i) = \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) \tag{6}$$

Finally, the construction cost estimation model based on least squares support vector machine is shown in formula 7.

$$f(x) = \sum_{i=1}^N \alpha_i K(x, x_i) + b \tag{7}$$

Each intermediate node of the least-squares SVM structure corresponds to one support vector. To model the construction cost performance of the least square SVM, the parameters γ and σ_2 , in order to improve the estimation effect of the construction cost and affect the construction cost, the particle group algorithm is used to determine the optimal value of parameters γ and σ_2 .

3.2 Particle Population Algorithm

The particle swarm algorithm simulates bird foraging behavior, using particles to represent the potential solution, sets the fitness function according to the target, and evaluates the particle position quality [10]. The particle state updates are performed as described in formula 8.

$$\begin{cases} v_{id}^{t+1} = wv_{id}^t + c_1r_1(p_{id}^t - x_{id}^t) + c_2r_2(p_{gd}^t - x_{gd}^t) \\ s_{id}^{t+1} = s_{id}^t + v_{id}^{t+1} \end{cases} \quad (8)$$

Inertial weight w affects local and global search performance, in order to improve the search capability of the algorithm and ensure the diversity of particles, automatically adjusted w , and the particle diversity function is defined as shown in formula 9.

$$F_{diversity}(t) = \frac{f_{min}(\alpha(t))}{f_{min}(\alpha(t)) + f_{max}(\alpha(t))} \quad (9)$$

$F_{diversity}(t)$ Represents the motion properties of the particle and defines the nonlinear function $\delta(t)$ as shown in formula 10.

$$\delta(t) = e^{(F_{diversity}(t)-L)^{-t}} \quad (10)$$

3.3 Construction Estimation Step of Data Mining

The estimation steps of the construction project cost of data mining are as follows:

Step1: For specific construction projects, the historical data of the cost is collected through relevant data.

Step2: initialize the position and velocity of the particles, and determine their value range, and map the parameters γ and σ_2 of the least squares SVM into the particle positions.

Step3: input the training sample of the construction cost estimation to the least squares support vector machine, and inverse code the particle position vector to obtain the values of parameters γ and σ_2 , and obtain the fitness value of each particle through training.

Step4: determine the optimal position p_i^t and the population optimal position p_g^t for each particle.

Step5: Update the inertial weight w .

Step6: Update of the particle state to create new particle populations.

Step7: The fitness value of the new particle population is calculated by training samples and least squares support vector machines.

Step8: See whether the end conditions are met. If not, proceed to **Step4**.

Step9: According to the global optimal combination of the values of parameters γ and σ_2 obtained by p_g^t , learn the construction cost training sample again, and establish the construction cost estimation model.

4 Simulation Experiment

4.1 Data Sources

The cost data of 200 construction projects in a certain city are studied, as shown in Figure 1, and the analysis shows that these data have a significant non-linear change trend. Based on the delay time and embedding size of cost data, a mathematical model of building engineering cost estimation is constructed, see formula 11.

$$y(n + 1) = [x(n), x(n - \tau), \dots, x(n - (m - 1)\tau)]^T \tag{11}$$

According to the difference entropy, the delay time of the construction cost data in Figure 1 is $\tau = 12$, and the embedded dimension is $m = 7$. The last 50 data are used as the test samples for the construction cost, and the others are the training samples for the estimation of the construction cost. Before using the least squares support vector mechanism to build the construction project cost estimation model, the particle swarm algorithm is used to determine the parameters γ and σ_2 , which are $\gamma = 100$, $\sigma_2 = 6.725$, the number of particles is 20, $c_1 = c_2 = 2$, and the maximum number of iterations is 500.

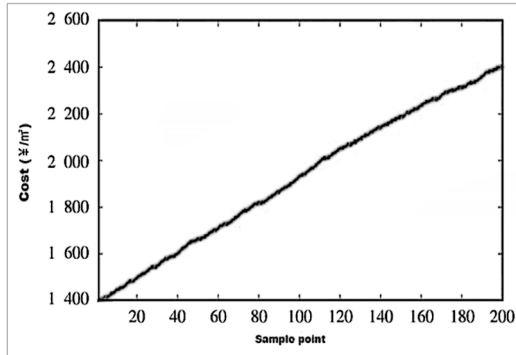


Fig. 1. Sample of construction project cost estimation.

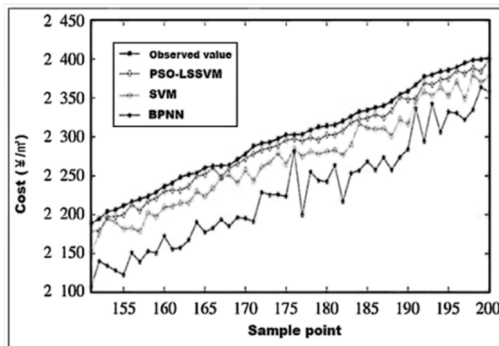


Fig. 2. Construction project cost estimation result.

4.2 Results and Analysis

The performance of BPNN and SVM is compared on the construction cost dataset. The results are shown in Figure 2, BPNN had the worst estimates, while SVM and LSSV had similar and ideal estimates. Table 1 presents the training time and testing results for BPNN, SVM, and PSOLSSVM. SVM is long and the modeling efficiency is low. Although BPNN is short, the estimation error is large and does not meet the actual management requirements. In contrast, PSOLSSVM showed a clear advantage in both speed and effect.

Table 1. Temporal comparison of training and testing.

Model	Training time	Testing time
BPNN	5.43	1.78
SVM	100.72	3.54
PSO-LSSVM	10.28	1.80

5 Conclusions

In order to accurately predict the construction cost, data mining technology is modeling, and the following conclusions are obtained through simulation experiments:

- The traditional method cannot fit the changing trend of construction cost with high precision, resulting in large error of construction cost estimation.
- The least squares support vector machine learns the connection between construction cost and influencing factors, which can effectively excavate its fitting trend and improve the accuracy of construction cost estimation.
- The particle swarm algorithm is used to optimize the parameters of the construction project cost estimation model, improve the ability of the model to describe the cost change, and the estimation results are better than other models.
- The model is versatile and can be applied to study prediction problems in other non-linear systems, with broad applications.

References

1. Wang Li. (2021). Analysis of construction project cost control under engineering cost big data based on BIM. *Old brand brand marketing* (11), 51-52.
2. Kurasova, O., Marcinkevičius, V., Medvedev, V., & Mikulskienė, B. (2021). Early cost estimation in customized furniture manufacturing using machine learning. *International journal of machine learning and computing.*, 11(1), 28-33.
3. Yang, X., Yu, M., & Liu, F. (2022, January). Construction of power network operation and maintenance cost prediction model based on data information mining. In *2022 International Conference on Big Data, Information and Computer Network (BDICN)* (pp. 124-127). IEEE.
4. Diallo, M. A. (2022, June). Prediction and Early Warning Model of Substation Project Cost Based on Data Mining. In *Application of Intelligent Systems in Multi-modal Information Analytics: The 4th International Conference on Multi-modal Information Analytics (ICMMIA 2022)*, Volume 2 (Vol. 138, p. 400). Springer Nature.
5. Zhang Hong. (2023). Research on big data application of project cost control platform. *Fujian Computer* (01), 39-44.
6. Hou Hong & Meng Hui. (2021). Research on engineering cost data Mining under the background of big data. *Adhesive* (01), 151-155 + 162.
7. Patel, T., & Patel, V. (2020). Data privacy in construction industry by privacy-preserving data mining (PPDM) approach. *Asian Journal of Civil Engineering*, 21(3), 505-515.

8. Qi-rong, X. (2020). Evaluation Model of Case Teaching Effect of Engineering Cost Based on Data Mining. In *e-Learning, e-Education, and Online Training: 6th EAI International Conference, eLEOT 2020*, Changsha, China, June 20-21, 2020, Proceedings, Part I 6 (pp. 16-28). Springer International Publishing.
9. Tayefeh Hashemi, S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Applied Sciences*, 2(10), 1703.
10. Wang Demei, Chen Hui, Xiao Zhihong, Xia Songlin, Fan Shuqian, Cui Changhui & Zhang Qinghua. (2021). Forecdiction of residential engineering cost based on data mining. *Journal of Civil Engineering and Management* (01), 175-182.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

