WARSAW UNIVERSITY OF TECHNOLOGY Index 351733
FACULTY OF CIVIL ENGINEERING
COMMITTEE FOR CIVIL AND WATER ENGINEERING
POLISH ACADEMY OF SCIENCES ISSN 1230-2945

DOI: 10.24425/ace.2023.147667

ARCHIVES OF CIVIL ENGINEERING

Vol. LXIX ISSUE 4 2023

pp. 413 –429

Research paper

Research on cost prediction for construction project based on Boruta-SO-BP model

Hao Cui[1], Junjie Xia[2]

Abstract: Cost prediction for construction projects provides important information for project feasibility studies and design scheme selection. To improve the accuracy of early-stage cost estimation for construction projects, an improved neural network prediction model was proposed based on BP (back propagation) neural network and Snake Optimizer algorithm (SO). SO algorithm is adopted to optimize the initial weights and thresholds of the BP neural network. Cost data for 50 construction projects undertaken by Shandong Tianqi Real Estate Group in China was collected, and the data samples were clustered into three categories using cluster analysis. 18 engineering feature indicators were determined through a literature review and 10 feature indicators were selected using Boruta algorithm for the input set. Compared to BP neural network and PSO–BP neural network, the results show that the improved SO–BP model has higher prediction accuracy, stability, better generalization ability and applicability. Therefore, based on reasonable feature indicators, the method proposed in this paper has certain guiding significance for predicting engineering costs.

Keywords: Boruta, BP neural network, intelligent prediction,project cost, Snake Optimizer algorithm

[1]PhD., College of Civil Engineering, Jiangxi Science and Technology Normal University, No. 605 Fenglin Avenue, 330013, Nanchang, China, e-mail: cuihaoch@qq.com, ORCID: 0000-0002-8292-9555

[2]B. Eng., College of Civil Engineering, Jiangxi Science and Technology Normal University, No. 605 Fenglin Avenue, 330013, Nanchang, China, e-mail: 2897152269@qq.com, ORCID: 0000-0002-5993-6474

1. Introduction

For a proposed construction project, project cost control is integral to the project's entire life cycle, including the project proposal stage, feasibility study stage, blueprint design stage, tender stage and construction stage. Research has shown that the investment decision-making stage in the early stages of the project development has the greatest impact on project costs, accounting for as much as 75% to 95% of the total cost, followed by the design stage, which accounts for 35% to 75% of the cost [1]. It shows that the accuracy of project cost forecasting plays a decisive role in early investment and the preparation of feasibility study reports. In addition, project costs tend to be determined by historical experience in construction practice, which can cause serious problems for owners and contractors. Accurate prediction of actual project cost is one way to mitigate these problems. It can also better assist owners in making scientific investment decisions and help contractors with cost control [2].

To explore how to predict project cost more accurately, cost prediction models using various methods have been developed in recent years. These methods can be broadly classified into two categories: traditional prediction methods and machine learning ones.

Traditional methods require a clear understanding of the relationship between the dependent and independent variables [3]. The most typical method is regression analysis, which is widely used in the prediction field due to its simplicity, speed and ability to visually reflect mathematical relationships between variables. For instance, a multiple regression model was used by Prasetyono [4] to predict the cost of a project with insufficient information during the design phase. The mean absolute percentage error (MAPE) was 19.3%. A standardized cost model for a water treatment plant project using regression analysis was proposed by Ma [5]. To further improve the accuracy of cost prediction, some researchers have integrated case-based reasoning (CBR) into multiple regression models based on numerous typical cases of building construction projects [6, 7]. The method was used to predict cost by adjusting historical similar cases. Although to a certain extent, this approach has improved the prediction accuracy compared to the multiple linear regression model, it faces challenges in computing the similarity of categorical variables.

With the continuous development of big data and artificial intelligence technologies, an information integration platform has been provided for project cost prediction, fully excavating the hidden rules behind complex data and providing convenience for accurate and fast engineering cost prediction. Among machine learning methods, neural networks are the most widely used in the field of cost prediction. A neural network with bootstrapping prediction intervals for estimating the range of engineering costs was proposed by Hwang [8]. A new short-term cost prediction model was presented by Xie [9]. However, in practical applications, neural networks also have some drawbacks, such as difficulty in determining the number of hidden neurons, susceptibility to local optima and poor generalization ability, which limit the accuracy of neural network predictions. To improve the prediction accuracy of neural networks, some researchers have optimized the parameters of neural networks. A genetic algorithm was adopted by Lu [10] to optimize the neural network model, simulating the generation of construction waste with the progress of the project.

The results showed that the prediction error, using the improved model, was lower than that without using GA. The combination of the AdaBoost algorithm with neural networks was done by Sun and Gao [11] to overcome the instability of a single neural network. This approach provided more accurate and stable predictions for new datasets. Cost prediction was performed by Ye [12] using a BP neural network based on particle swarm optimization. The results showed that the BP neural network had a faster convergence speed and better generalization ability.

However, in real engineering projects, there are inevitably differences. If the samples are directly trained and learned, it may result in long model training time, low prediction accuracy, and even serious deviation. Based on big data, the prediction of construction project cost can reduce redundant information, the dimension of the original feature set and the training time of the model, and improve the prediction accuracy of the model by selecting the main features. The contribution analysis algorithm of a neural network was employed by Wang [13] to screen out 12 main feature factors. It was discovered that utilizing these main feature factors as input variables can significantly enhance the model's prediction accuracy. The intuitive fuzzy analysis method was utilized by Liang [14] to identify the 5 major influencing engineering characteristics. These characteristics were subsequently used as inputs for the BP neural network, resulting in a significant improvement in the prediction accuracy of construction project costs. The dimensionality of sample data was reduced and index correlation was eliminated by Qin [15] through principal component analysis. The resulting data was then inputted into SVM and LS-SVM models for further analysis. The results showed that the relative error of the prediction model was controlled within $\pm 7\%$, and the results were stable. In addition, the reliability of this method has been further demonstrated by other scholars who have applied feature selection methods to identify remodeling risk factors of project schedules.

Therefore, in this research context, the Boruta feature selection method [16] is introduced in this paper. Based on cluster analysis, the sample data is selected and the SO algorithm [17] is used to optimize the weights and thresholds of the BP neural network, which is then used to construct the SO–BP neural network model. It can both improve the generalization ability of BP neural network and ensure the safety and scientificity of cost prediction for construction project.

2. Data collection

The selection of engineering characteristic indicators is a crucial issue that affects the efficiency and accuracy of cost prediction models. Relevant literature on factors influencing construction project costs was collected from both domestic and international databases to perform a preliminary selection of indicators. After considering factors such as the quality and publication year of the papers, a comprehensive analysis of the internal attributes and external influencing factors of construction projects was conducted to statistically analyze the factors influencing construction project costs. A total of 18 common characteristic indicators were selected as the main influencing factors for construction cost prediction, as shown in Table 1.

Table 1. Literature statistics on influencing factors of engineering cost

No.	Author	Factors influencing the cost of construction projects	Year
1	Shen [18]	fundamental data types, pile foundation type, building structure, number of stories, building area, door and window type, interior and exterior wall decoration, degree of installation completion	2018
2	Dursun [19]	building area, number of stories, number of above-ground and underground floors or stories, average floor height, interior and exterior wall area, soil conditions	2016
3	Wang [20]	building area, number of above-ground and underground floors or stories, seismic design level, structural type, foundation type, seismic intensity, decoration category	2021
4	Xu [23]	building area, number of above-ground and underground floors or stories, average floor height, pile foundation type, structural type, seismic design level, foundation type, project management level	2021
5	Liang [14]	building area, standard floor area, building floor-to-floor height, number of stories, structural type, plan shape, seismic design level, foundation type and depth	2017
6	Dimitrijevic [21]	building area, number of stories, floor-to-floor height, excavation depth, structural type, resource unit price, construction period, door and window type	2019
7	Ji [22]	building area, structural type, number of stories, number of units, floor-to-floor height, construction period, roof type, foundation type, decoration	2019
8	Sheikh [24]	contractor's professional level or Contractor's expertise, construction period, number of stories, total building area, materials, foundation type	2019

For instance, Ground floor area (V1),Basement area (V2), Above-ground stories (V3), Below-ground stories (V4), Average height of above-ground floors (V5), Average height of below-ground floors (V6), Pile foundation type (V7), Foundation type (V8), Building structure type (V9), Seismic resistance level (V10), Ground material (V11), Decoration material (V12), Door and window type (V13), Fire protection system (V14), Degree of equipment installation (V15), Project management level (V16), Construction environment (V17) and Project duration (V18).

The feature indicators selected can be divided into quantitative and qualitative ones. Quantitative indicators, such as building area, can be directly inputted with actual engineering data. Qualitative indicators, such as seismic grade, require quantification before inputting into the prediction model. Based on the equal interval partition method, representative category indicators such as building structure type are determined, and are expressed on a scale of 1–5 to represent the structure type, thereby discretizing and quantifying the qualitative indicators. The specific quantification method is shown in the Table 2.

Table 2. Quantitative processing of qualitative indicators

Characteristic Indicator	Quantitative Value				
	1	2	3	4	5
Seismic Design Category	level 6	level 7	level 8	–	–
Building Structural Type	brick and concrete structure	frame structure	shear wall structure	frame-shear wall structure	steel structure
Foundation Type	independent foundation	strip foundation	pile foundation with a pile cap	raft foundation	box foundation
Pile Foundation Type	bored pile	drilled pile	precast pile	–	–
Project Management Level	level 1	level 2	level 3	level 4	–
Construction Environment	poor	qualified	medium	good	excellent
Completeness of Installation	simple	general	basically intact	intact	very intact
Fire protection system	automatic alarm system	automatic sprinkler system	foam fire suppression system	gas fire suppression system	–
Internal and external wall decoration	coatings	ceramic tiles	natural stone	curtain wall	–
Flooring materials	rough surface	cement mortar	floor tiles	terrazzo	wooden floor
types of doors and windows	sliding	sliding and folding	hinged	folding	–

Features in projects may have different scales and units, which can result in significant differences. Normalizing the data can eliminate the impact of large scale differences and improve the speed of model training. In this paper, we use the commonly used maximum-minimum normalization method. The specific formula is as follows:

$$i = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2.1}$$

where: i – the normalized input data, x – the actual value, $x_{\min}$ – the minimum value of the measured data, $x_{\max}$ – the maximum value of the measured data.

3. Methodology

3.1. Clustering analysis

Based on extensive literature research, it is known that in the process of establishing a project cost prediction model for construction engineering, relevant engineering information from the historical database is often directly used without distinction. Further prediction for the cost of the construction project is made. However, there are inevitably differences between actual engineering projects, and in the case of multiple project cost indicators, it is difficult to objectively judge the similarity of the proposed construction project. If similar projects are directly used for sample learning and training, the prediction accuracy may be biased. Systematic cluster analysis is an effective statistical analysis method that can objectively determine the similarity of projects. There are two types of cluster analysis based on different classification objects, sample cluster analysis (Q-type cluster analysis) and variable cluster analysis (R-type cluster analysis). Since the purpose of this work is to classify engineering samples, the Q-type cluster analysis method is adopted. Based on the analysis of various indicators, similar cases are grouped together. Generally, if there are too many isolated points in the clustering results, it indicates that the clustering method is not effective. From the perspective of reducing isolated points, the best clustering effect is obtained using the sum of squares of differences (also known as Ward's method). Therefore, Ward's method is adopted in this paper, and the specific formula is as follows:

$$\boldsymbol{D}_t = \sum_{i=1}^{n_t} (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_t) \cdot (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_t) \tag{3.1}$$

where: $\boldsymbol{x}_{it}$ – the i-th sample in a cluster, $\bar{\boldsymbol{x}}_t$ – the center of the cluster, n_t – the number of samples in the cluster.

3.2. Feature selection

In the process of predicting project costs in actual construction projects, the characteristic attributes of a project can affect the cost, but there may be instances where individual attributes are not highly correlated, resulting in redundant feature indicators. To minimize workload and increase the computational speed of the prediction model, the Boruta algorithm is used to perform feature selection on cost prediction indicators of the input model, reducing the input variables and data to improve the learning efficiency of the prediction model. Additionally, the main purpose of feature selection is to identify the most important features in a given dataset, maintaining their importance even in the presence of noise and highly correlated features, improving the prediction accuracy of the model. The main steps for using the Boruta algorithm for feature selection in predicting project costs are as follows:

1. The original feature matrix $\boldsymbol{R}$ is randomly shuffled and connected to a shadow matrix $\boldsymbol{S}$ that has the same features, creating a new feature matrix $\boldsymbol{N} = [\boldsymbol{R}, \boldsymbol{S}]$.

2. A random forest algorithm is trained on the ***N*** matrix, and the importance scores for ***R*** and ***S*** are obtained.
3. The importance score for each feature in ***R*** is compared with the maximum score in ***S***.

If the former is greater than the latter, the feature is marked as "confirmed". Otherwise, it is marked as "to be further validated".

4. For all features marked as "to be further validated", steps (1)–(3) are repeated until all features are either confirmed or invalidated.
5. Based on the final scores, the most important features are selected for retention.

3.3. BP neural network

Back Propagation neural network (BPNN) was proposed by Rumelhart and McClelland in 1986, and it is a multi-layer feedforward neural network. It uses the backpropagation algorithm to train the network by propagating the error between the predicted and actual values. The network uses a perceptron layer to handle nonlinear mapping problems and has strong adaptive learning and data processing capabilities. In recent years, the BP neural network has been widely used in the field of engineering cost prediction. In this paper, a typical BPNN was employed to establish a prediction model. The network topology has three layers, namely, the input layer, the hidden layer and the output layer, as shown in Fig. 1.

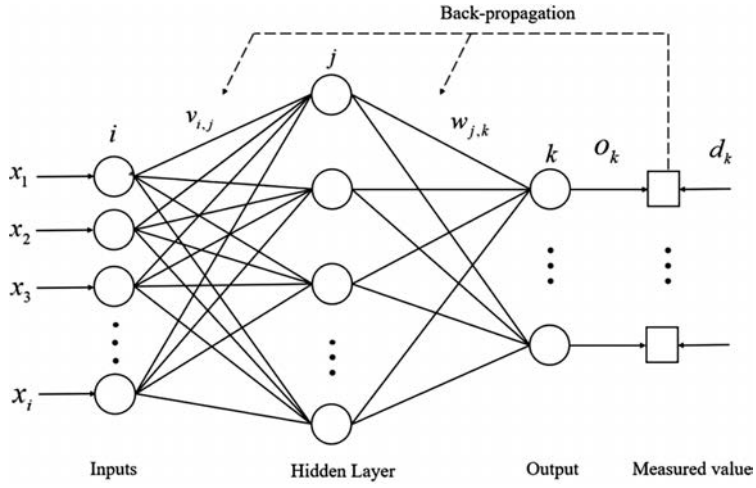

Fig. 1. 3-layer BP neural network topology

The mathematical relationships between the input layer, output layer and hidden layer in a three-layer perceptron are expressed as follows.

First, output value of hidden layer neurons can be described as follows:

$$\begin{cases} y_j = f\left(\text{net}_j\right) \\ \text{net}_j = \sum_{i=0}^{n} v_{i,j} x_i + b_j \end{cases} \tag{3.2}$$

where: y_j – the output value of the j-th hidden layer neuron ($j = 1, 2, \ldots, m$), x_i – the i-th input signal ($i = 1, 2, \ldots, n$), $v_{i,j}$ – the weight value from the i-th input neuron to the j-th hidden layer neuron, b_j – the threshold of the j-th hidden layer neuron.

Second, output value of output layer neurons can be expressed as follows:

$$\begin{cases} o_k = f(\text{net}_k) \\ \text{net}_k = \sum_{j=0}^{n} w_{j,k} y_j + b_k \end{cases} \tag{3.3}$$

where: o_k – the output value of the k-th output layer neuron ($k = 1, 2, \ldots, l$), $w_{j,k}$ – the weight value from the j-th hidden layer neuron to the k-th output layer neuron, b_k – the threshold of the k-th output layer neuron.

In Eq. (3.2) and Eq. (3.3), the function typically used is the Sigmoid function, which is chosen for its continuity and differentiability. The formula for the Sigmoid function is as follows:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.4}$$

where $\exp(-x)$ represents the exponential function with a negative exponent. The Sigmoid function is commonly employed as the activation function in neural networks due to its desirable properties.

In the error backpropagation process, the difference between the output value and the expected value is compared and then the error for each layer is calculated in a backward, step-by-step manner. The formula for calculating the error is as follows:

$$E = \frac{1}{2} \sum_{k=1}^{l} (d_k - o_k)^2 \tag{3.5}$$

By substituting Eq. (3.3) into Eq. (3.5), the error of the output layer is obtained:

$$E = \frac{1}{2} \sum_{k=1}^{l} \left(d_k - f\left(\sum_{j=0}^{m} w_{j,k} y_j + b_k \right) \right)^2 \tag{3.6}$$

By substituting Eq. (3.2) into Eq. (3.6), the error is further extended to the input layer:

$$E = \frac{1}{2} \sum_{k=1}^{l} \left(d_k - f\left(\sum_{j=0}^{m} \left(w_{j,k} f\left(\sum_{i=0}^{n} v_{i,j} x_i + b_j \right) \right) + b_k \right) \right)^2 \tag{3.7}$$

According to Eq. (3.7), the output error of the neural network is a function of the weights and thresholds. Therefore, by adjusting the weights and thresholds, the error can be reduced, making the training data closer to the expected value.

When using a BPNN for prediction, the number of hidden layers and their nodes can have a significant impact on the network topology and model performance. Based on the Hecht-Nielson theory that a single hidden layer BPNN can approximate any continuous function in any interval, this paper adopts a single hidden layer structure. The number of nodes in the hidden layer is determined using an empirical Eq. (3.8) that minimizes the root mean square error value of the training set:

$$h = \sqrt{n+m} + c \tag{3.8}$$

where: h – the number of nodes in the hidden layer, n – the number of nodes in the input layer, m – the number of nodes in the output layer, c – a constant in the interval $[1, 10]$.

3.4. Basic principle of snake optimizer algorithm

The Snake Algorithm is a heuristic optimization algorithm proposed by Hashim and Hussien, inspired by the foraging and reproduction behaviors of snakes [17]. The algorithm first initializes a population of snakes in the feasible solution space, with each individual representing a potential solution to the optimization problem. The snake population is randomly divided into two groups, male and female, with equal numbers. In each iteration of the search process, individuals update their positions based on the amount of food available in each dimension. When food is scarce, individuals continue to search for food while when food is abundant, they select a survival mode based on temperature, either fighting or mating. Then, natural selection takes place, replacing the worst individuals in the old snake population with better individuals or producing new individuals. The determination of food and temperature is shown in Eq. (3.9):

$$\begin{cases} Q = c_1 \cdot \exp\left((t-T)/T\right) \\ T_p = \exp\left(-t/T\right) \end{cases} \tag{3.9}$$

where: T_p – temperature, Q – the amount of food, t – the current number of iterations, T – the maximum number of iterations, c_1 – a constant with a value of 0.5.

3.5. The proposed model

BP neural networks have strong nonlinear mapping capabilities, but two important parameters need to be determined during the error backpropagation process: weights and thresholds. The common method to adjust the connection weights and thresholds of the network is the gradient descent method. However, when the step size is too large, it may skip the global optimal solution. When the step size is too small, it may get stuck in a local optimum, which affects the prediction error of the model. The SO–BP neural network model utilizes the global optimization ability of the SO algorithm to search for the network's weights and thresholds in its solution space. The algorithm flow chart is showed as Fig. 2.

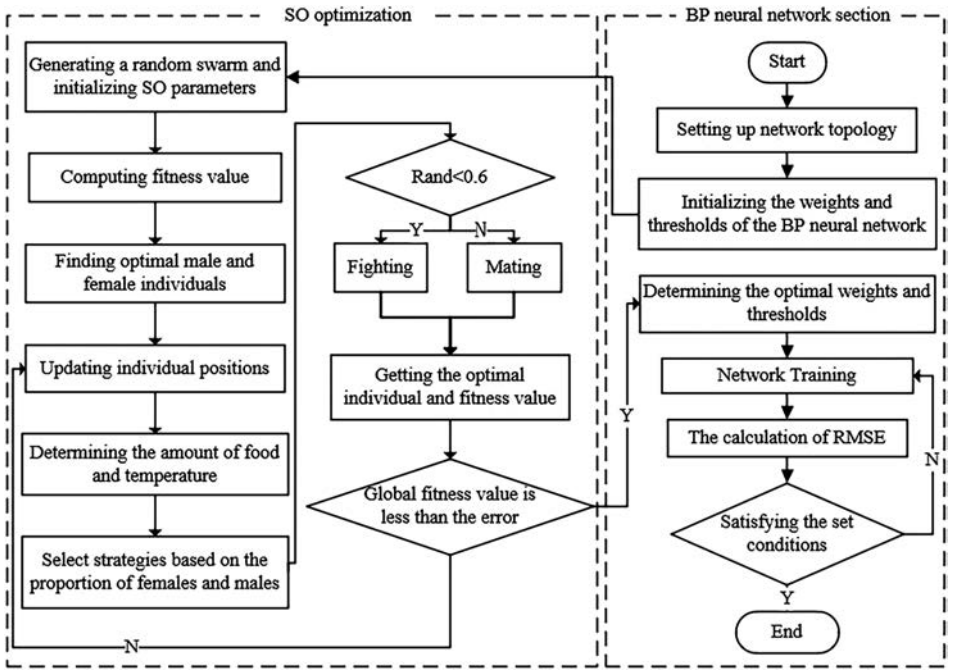

Fig. 2. Algorithm flow chart

The position of the individual on each dimension is the solution sought, and a mapping relationship between the dimension of the individual and the weights and thresholds is established. The dimension of a BP neural network with one hidden layer is as follows:

$$D = i \times h + h + h \times k + k \tag{3.10}$$

where: i – the nodes in the input layer, h – the nodes in the hidden layer, k – the nodes in the output layer.

The quality of an individual's position is determined by the fitness function, which is the mean squared error of the training and testing sets in this paper. The smaller the value of the fitness function, the better the training and testing results, and the higher the accuracy of the model. The specific formula is as follows:

$$f = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{Z} \left(o_{ij} - d_{ij}\right)^2 \tag{3.11}$$

where: N – the total number of input learning samples, Z – the number of output nodes, o_{ij} – the actual output value of the corresponding parameter, d_{ij} – the expected output value of the corresponding parameter.

3.6. Accuracy evaluation

From a statistical perspective, it is not sufficient to use only one performance metric for evaluation [25]. Therefore, three performance metrics, namely, the coefficient of determination (R^2), root-mean-square error (e_{RMSE}) and mean absolute percentage error (e_{MAPE}), are used to comprehensively evaluate the model. R^2 represents the linear correlation between the measured values and the predicted values. e_{RMSE} is used to indicate the dispersion of the results and e_{MAPE} represents the accuracy of the results:

$$R^2 = 1 - \sum_{i=1}^{N} (o - \hat{o})^2 \Big/ \sum_{i=1}^{N} (o - \overline{o})^2 \tag{3.12}$$

$$e_{\mathrm{RMSE}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (o - \overline{o})^2} \tag{3.13}$$

$$e_{\mathrm{MAPE}} = \frac{100\%}{N} \sum_{i=1}^{N} |(o - \hat{o}) / o| \tag{3.14}$$

where: o – the expected value, $\hat{o}$ – the predicted value, $\overline{o}$ – the mean value of the outputs, N – the size of the dataset.

4. Case analysis

4.1. Select samples

Cost data for 50 construction projects undertaken by Shandong Tianqi Real Estate Group in China was collected. The projects were categorized based on their building area, with a division into small-scale projects (less than 3000 square meters) and medium-scale projects (ranging from 3000 to 100,000 square meters). The selected samples included 2 small-scale projects and 48 medium-scale projects. Furthermore, the data was standardized using Eq. (2.1) and then divided into three categories using Eq. (3.1), as shown in Table 3.

Table 3. Sample System Cluster Category

Category	Sample ID		
Class 1	7, 37, 50, 23, 45, 30, 38, 11, 39, 2, 14, 34	13, 1, 42, 46, 25, 29, 3, 17, 16, 32, 21, 49	40, 36, 41, 12, 18, 35, 19, 5
Class 2	6, 22, 27, 26, 43, 9	33, 44, 10, 47, 28, 20	31, 4
Class 3	15, 24, 8, 48	–	–

To ensure that the model achieves better results and overcomes the errors caused by small samples, 32 similar samples from the first category were selected as input set. The Boruta feature selection method was used to obtain the feature selection results, as shown in Fig. 3.

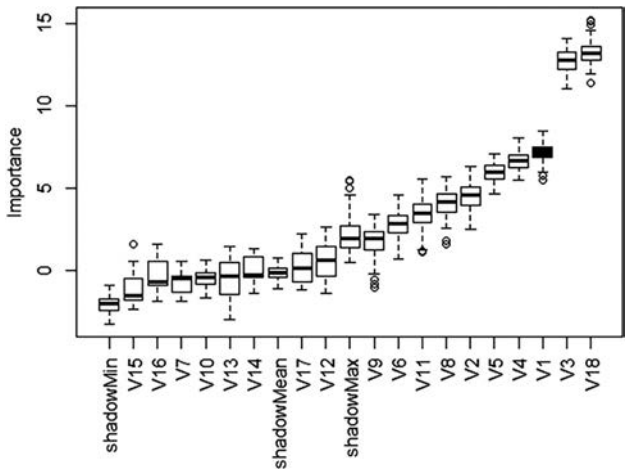

Fig. 3. Boruta feature selection result

From the figure, on the one hand, it can be observed that the Z-scores of V1–V6, V8, V9, V11 and V18 are greater than the maximum Z-score of the shadow attribute. Therefore, these indicators are considered important. On the other hand, the Z-scores of V7, V10 and V12–V17 are lower than the maximum Z-score of the shadow attribute, indicating that they should be excluded. Among them, the Z-scores of V7, V10 and V13–V16 fall between the minimum and average Z-scores of the shadow attribute, suggesting that their importance is relatively low. The Z-scores of V12 and V17, however, fall between the maximum and average Z-scores of the shadow attribute. Consequently, after applying the Boruta algorithm for feature selection in cost prediction indicators, a total of 10 highly important indicators have been identified as input variables for the prediction model. These indicators are above-ground building area V1, underground building area V2, above-ground stories V3, underground stories V4, above-ground average height V5, underground average height V6, foundation category V8, building structure type V9, ground material V11 and construction period V18.

4.2. Determine model input indexes

The preprocessed samples described above were used, with the first 27 sets of data used as the training samples for the network and the last 5 sets of data used for prediction. The specific parameter settings of the SO–BP model are shown in Table 4.

According to Eq. (3.10), 10 different numbers of nodes were obtained. Network topology models with different numbers of neurons were established, and the network's performance was evaluated using Eq. (3.13). The structure model with the minimum mean square error (MSE) in the training set was selected, as shown in Table 5.

It is evident that 8 hidden layer nodes are optimal, and an 8-8-1 model topology structure was constructed.

Table 4. SO–BP algorithm parameter settings

Parameter name	Parameter setting	Parameter name	Parameter setting
Training function	trainlm	Training iterations	1000
Transfer function of the hidden layer	tansig	Initial population	50
Output layer transfer function	purelin	Initial boundary	[–5, 5]
Learning rate	0.01	Food threshold	0.25
Target error	1.00e-06	Temperature threshold	0.6

Table 5. Selection of hidden layer nodes

No.	4	5	6	7	8	9	10	11	12	13
MSE	127.59	85.86	135.10	83.89	61.54	70.57	115.63	81.28	324.44	126.88

4.3. Determine model input indexes

The SO–BP model was trained and the established network topology structure was used to predict. The fitting effect of the training set sample values and predicted values is shown in Fig. 4.

Based on the results shown in Fig. 4 and Fig. 5, it can be seen that the SO–BP model performs well in fitting the training set, with a regression fit of $R^2 = 0.982$. This indicates that the trained neural network has good linear fitting ability. Additionally, the model performs well in predicting the test set, suggesting that it is suitable for predicting construction project costs.

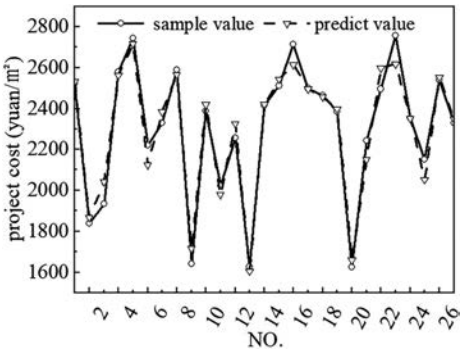

Fig. 4. Fitting graph of SO–BP model training set

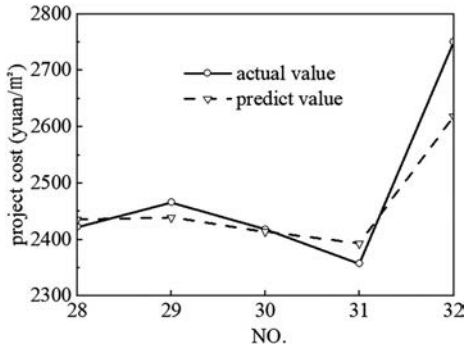

Fig. 5. Comparison of prediction results on the testing set for the SO–BP model

The comparison between the predicted and actual values of the test set for case verification is shown in Fig. 5.

In this study, the performance of the proposed SO–BP model was compared with the standard BPNN and the PSO–BP model. The input and output variables of each model were identical to those used in the SO–BP model. The PSO parameters were set as follows: learning factors c_1 and c_2 were both set to 2; the maximum and minimum velocities of the particles were set to $V_{max} = 0.5$ and $V_{min} = -0.5$, respectively; the maximum and minimum inertia weights were set to $popmax = 10$ and $popmin = -10$, respectively. All other parameters were kept the same.

Performance evaluations of each model were conducted, and the results are presented in Table 6, Fig. 6 and Fig. 7.

Table 6. Comparison and analysis of test results

Index	Actual value (¥/m^2)	BP prediction value (¥/m^2)	Error (%)	PSO–BP prediction value (¥/m^2)	Error (%)	SO–BP prediction value (¥/m^2)	Error (%)
28	2422.00	2556.57	5.56%	2439.97	0.74%	2435.90	0.57%
29	2465.90	2608.90	5.80%	2523.79	2.35%	2439.36	1.08%
30	2418.12	2435.05	0.70%	2403.06	0.62%	2413.67	0.18%
31	2357.06	2304.80	2.22%	2406.91	2.11%	2392.99	1.52%
32	2750.88	2273.02	17.37%	2927.53	6.42%	2618.72	4.80%
e_{MAPE}	–	–	6.33%	–	2.45%	–	1.63%
e_{RMSE}	–	232.35	–	86.71	–	62.73	–
R^2	–	–0.407	–	0.989	–	0.994	–

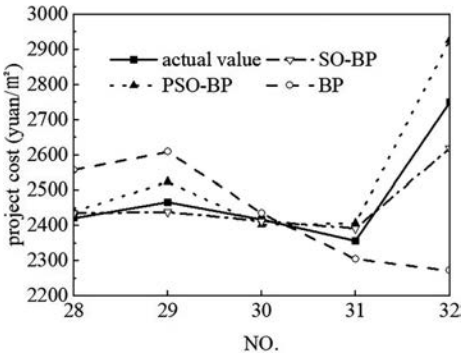

Fig. 6. Comparison chart of test results

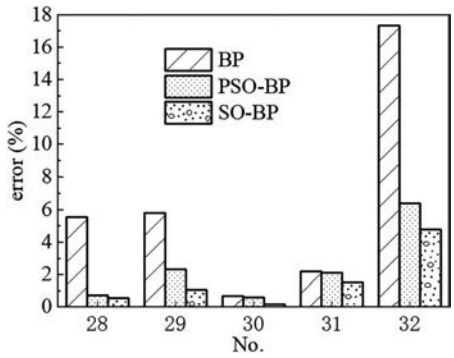

Fig. 7. Error comparison chart of cost prediction of different models

It can be concluded that all three models have good prediction performance. The accuracy of the results can be controlled within ±20%, which meets the requirements for

prediction accuracy in the investment decision-making stage. Moreover, it can be found that both PSO–BP and SO–BP neural network prediction models significantly reduce the root mean square error compared to the BPNN. Among them, the SO–BP model has a maximum error of 4.80%, an e_{MAPE} of 1.63% and an R^2 of 0.994 while the PSO–BP model has a maximum error of 6.42%, an e_{MAPE} of 2.45%, and an R^2 of 0.989. It is obvious that the SO–BP model has better prediction performance than the BPNN and the PSO–BP model. The prediction results of the SO–BP model are more accurate, closer to the actual values, and have better generalization ability.

5. Conclusions

1. In light of the numerous shortcomings of traditional engineering cost prediction methods, this paper proposes the SO–BP neural network prediction model. Through a literature review and analysis, 18 factors affecting the cost of building engineering were selected. The collected sample data was clustered, and the class with the most data was selected to ensure the effectiveness of the data. Boruta was used to reduce the 18 factors and eliminate indicator correlations, resulting in the selection of 10 feature indicators as input parameters, ensuring the scientificity and completeness of the constructed cost indicator system. The feasibility and effectiveness of the SO–BP neural network prediction model were demonstrated.
2. Comparing the prediction data of the three models with the actual values, it can be observed that all models exhibit a certain level of predictive performance. However, the SO–BP neural network model has the smallest prediction error, highest precision and stability, and exhibits better generalization ability. Thus, this model can be applied to practical project cost prediction.
3. The present study is limited by the small sample size of 50 collected data points. In practice, it is necessary to collect a large amount of project cost data for construction projects to provide the SO–BP neural network model with more data support and improve its predictive accuracy.

Acknowledgements

The research reported in this paper was financially supported by the Research Project of Humanities and Social Sciences (No.GL21118), Doctor Start-up Fund of Jiangxi Science & Technology Normal University, China (No.2020BSQD018) and Graduate Innovation Special Fund Funding Project of Jiangxi Science & Technology Normal University (No.YC2022-x08).

References

[1] S. Demirkesen and B. Ozorhon, “Impact of integration management on construction project management performance”, *International Journal of Project Management*, vol. 35, no. 8, pp. 1639–1654, 2017, doi: 10.1016/j.ijproman.2017.09.008.

[2] W. Hu, Y. Chang, and X. He, "Impact factors and prediction models of building construction duration", *China Civil Engineering Journal*, vol. 51, no.2, pp. 103–112, 2018, doi: 10.15951/j.tmgcxb.2018.02.012.

[3] J. Xu and S. Moon, "Stochastic forecast of construction cost index using a cointegrated vector autoregression model", *Journal of Management in Engineering*, vol. 29, no. 1, pp. 10–18, 2013, doi: 10.1061/(ASCE)ME.1943-5479.0000112.

[4] P. N. Prasetyono, H. M. Suryanto, and H. Dani, "Predicting construction cost using regression techniques for residential building", *Journal of Physics: Conference Series*, vol. 1899, no. 1, art. no. 012120, 2021, doi: 10.1088/1742-6596/1899/1/012120.

[5] K. Ma, Q. Wang, L. Zhou, et al., "Standardized construction cost estimation models for drinking water treatment plant", *China South-to-North Water Transfers and Water Science &Technolog*y, vol. 19, no. 1, pp. 191–197, 2021, doi: 10.13476/j.cnki.nsbdqk.2021.0019.

[6] R. Jin, K. Cho, C. Hyun, and M. Son, "MRA-based revised CBR model for cost prediction in the early stage of construction projects", *Expert Systems with Applications*, vol. 39, no. 5, pp. 5214–5222, 2012, doi: 10.1016/j.eswa.2011.11.018.

[7] J. Ahn, S. Ji, S. J. Ahn, et al., "Performance evaluation of normalization-based CBR models for improving construction cost estimation", *Automation in Construction*, vol. 119, art. no. 103329, 2020, doi: 10.1016/j.autcon.2020.103329.

[8] S. Hwang, "Dynamic regression models for prediction of construction costs", *Journal of Construction Engineering and Management*, vol. 135, no. 5, pp. 360–367, 2009, doi: 10.1061/(ASCE)CO.1943-7862.0000006.

[9] Y. Xie, W. Huang, and L. Gao, "A novel short-term forecasting model of construction cost based on intelligent information processing technology", *China Mathematics in Practice and Theory*, vol. 37, no. 6, pp. 24–31, 2007.

[10] W. Lu, Y. Peng, X. Chen, et al., "The S-curve for forecasting waste generation in construction projects", *Waste Management*, vol. 56, pp. 23–34, 2016, doi: 10.1016/j.wasman.2016.07.039.

[11] W. Sun and Q. Gao, "Exploration of energy saving potential in China power industry based on Adaboost back propagation neural network", *Journal of Cleaner Production*, vol. 217, pp. 257–266, 2019, doi: 10.1016/j.jclepro.2019.01.205.

[12] D. Ye, "An algorithm for construction project cost forecast based on particle swarm optimization-guided BP Neural Network", *Scientific Programming*, vol. 2021, art. no. 4309495, 2019, doi: 10.1155/2021/4309495.

[13] J. Wang and Y. Lu, "Prediction model of subway station civil engineering cost based on ANN contribution analysis and GEP algorithm", *China Journal of Railway Science and Engineering*, vol. 17, no. 8, pp. 2152–2162, 2020.

[14] X. Liang and Y. Liu, "Predicting model for construction engineering cost based on fuzzy neural network", *China Technology Economics*, vol. 36, no. 3, pp. 109–113, 2017, doi: 10.3969/j.issn.1002-980X.2017.03.014.

[15] Z. Qin, X. Lei, D. Zhai, et al., "Forecasting the costs of residential construction based on support vector machine and least squares-support vector machine", *China Journal of Zhejiang University (Science Edition)*, vol. 43, no. 3, pp. 357–363, 2016, doi: 10.3785/j.issn.1008-9497.2016.03.017.

[16] H. Guo, B. Gao and H. Lu, "Research on stock yield based on Boruta-PSO-SVM", *China Transducer and Microsystem Technologies*, vol. 37, no. 3, pp. 51–53+57, 2018, doi: 10.13873/J.1000-9787(2018)03-0051-03.

[17] F. Hashim and A. Hussien, "Snake Optimizer: A novel meta-heuristic optimization algorithm", *Knowledge-Based Systems*, vol. 242, art. no. 108388, 2022, doi: 10.1016/j.knosys.2022.108320.

[18] J. Shen, S. Wang, and X. Sun, "Research on cost prediction of construction engineering in design stage based on LS-SVM", *China Architecture Technology*, vol. 49, no. 2, pp. 209–212, 2018, doi: 10.3969/j.issn.1000-4726.2018.02.027.

[19] O. Dursun and C. Stoy, "Conceptual estimation of construction costs using the multistep ahead approach", *Journal of Construction Engineering and Management*, vol. 142, no. 9, art. no. 04016038, 2016, doi: 10.1061/(ASCE)CO.1943-7862.0001150.

[20] D. Wang, H. Chen, Z. Xiao, et al., "Prediction of housing project cost based on data mining", *China Journal of Civil Engineering and Management*, vol. 38, no. 1, pp. 175–182, 2021, doi: 10.13579/j.cnki.2095-0985.20201027.001.

[21] B. Dimitrijevic, Z. Stojadinovic, D. Marinkovic, et al., "Influence of structural system on the construction time and cost of residential projects", *Gradevinar*, vol. 71, pp. 681–693, 2019, doi: 10.14256/JCE.2315.2018.

[22] S. Ji, J. Ahn, H. Lee, and K. Han, "Cost estimation model using modified parameters for construction projects", *Advances in Civil Engineering*, vol. 2019, art. no. 8290935, 2019, doi: 10.1155/2019/8290935.

[23] X. Xu, L. Peng, and Z. Ji, "Research on substation project cost prediction based on sparrow search algorithm optimized BP Neural Network", *Sustainability*, vol. 13, no. 24, art. no. 13746, 2021, doi: 10.3390/su132413746.

[24] A. Sheikh, M. Ikram, R. Ahmad, et al., "Evaluation of key factors influencing process quality during construction projects in Pakistan", *Grey Systems: Theory and Application*, vol. 9, no. 3, pp. 3 21–335, 2019, doi: 10.1108/GS-01-2019-0002.

[25] M. F. Hasan, O. Hammody, and K. S. Albayati, "Estimate final cost of roads using support vector machine", *Archives of Civil Engineering*, vol. 68, no. 4, pp. 669–682, 2022, doi: 10.24425/ace.2022.143061