Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **1**, February **2025**

A Hybrid Machine Learning Model-Based Approach to Cost Prediction for Construction Projects

Yan-Hua Du*

School of Civil Engineering and Environment
Zhengzhou University of Aeronautics, Zhengzhou 450000, P. R. China
duyanhua9@126.com

Benson Liu

College of Information Science
Mapúa University, Intramuros, 1002 Metro Manila, Philippines
dk0253@163.com

*Corresponding author: Yan-Hua Du
Received February 27, 2024, revised May 29, 2024, accepted July 21, 2024.

Abstract. *Estimates in the design phase of construction projects have a crucial impact on the whole life cycle cost of the project, and it is necessary to study and analyze them separately. However, there is a large amount of features in construction project cost samples. In addition, the traditional construction project cost prediction model based on Support Vector Machine (SVM) has relatively limited feature extraction and transformation capabilities for data, and cannot automatically extract complex features. Therefore, this work proposes a hybrid machine learning model-based construction cost prediction method. Firstly, a reasonable construction project cost prediction index system is constructed on the basis of expert scoring, combined with the composition of construction project cost and its influencing factors. At the same time, principal component analysis is used to reduce the dimensionality of the attribute indexes to get the unrelated composite indexes, which reduces the sample complexity and improves the learning efficiency of the model. Then, the probability mean is obtained from the front and back neighboring values, so as to get the a priori model of the corresponding features. Next, a prediction model based on BP neural network and Twin Support Vector Machine (TWSVM) is constructed based on the 17 extracted feature indicators as the input set. Aiming at the defects of TWSVM model that the parameter settings depend on empirical values, with the advantage of particle swarm algorithm in the field of parameter optimization, a prediction model based on particle swarm optimization parameters is proposed, enhancing the predictive accuracy and robustness. Finally, three methods, BP neural network, LS-TWSVM and BP-PSO-TWSVM, are selected for the simulation and prediction of engineering cost, respectively. The experimental results show that the BP-PSO-TWSVM model is the optimal model, and the gap between its cost prediction value and the real value is small, and the prediction results are stable and highly accurate.*
Keywords: Construction cost; Machine learning; Predictive modeling; TWSVM; BP neural network; PSO

1. **Introduction.** How to efficiently and rapidly carry out construction cost prediction has become an important part of engineering construction. Construction cost prediction has a significant impact on project investment, and accurate construction cost prediction is crucial for realizing correct investment decisions [1, 2]. Therefore, it is very necessary to conduct an in-depth study on the project cost prediction in the pre-project stage.

As a necessary item in the pre-project preparation, construction cost prediction refers to the estimation of the cost required for a certain project construction, including all steps of construction [3]. The construction unit needs to take the prediction results as an important basis for construction decision-making in order to minimize the construction budget and avoid waste. However, due to the construction process, there are too many uncertainties in the costing of a large number of items and a large number of items, the difficulty of the prediction is also increased accordingly. In addition, due to the relative complexity of the engineering structure, efficient prediction based on the current engineering environment is still the difficulty of the current research [4, 5].

Construction project cost prediction is realized by certain mathematical models and relevant historical engineering data. Previous cost estimation models usually focus on regression theory [6], fuzzy mathematics [7], gray system theory [8] and artificial neural networks [9]. However, the traditional construction project cost prediction model mainly suffers from the defects of low computational accuracy, poor stability, and insufficient generalization ability, which makes it difficult to give reasonable and effective guidance to cost management work in the early stage of the construction project. The traditional construction cost prediction model based on support vector machine (SVM) has limited ability to model complex nonlinear relationships [10]. BP neural networks, on the other hand, are good at learning and capturing complex nonlinear relationships in data, which can help to supplement the shortcomings of SVM in this regard, thus making the model more accurate in predicting construction project cost. Therefore, the research objective of this work is to combine BP neural network and SVM to construct a hybrid machine learning model, which can further improve the performance of construction cost prediction.

1.1. **Related Work.** Many scholars have used linear regression [11], GM(1,1) gray prediction model [12], time series analysis [13], moving average method and other methods for project cost prediction, but these traditional statistical learning methods have large defects in the process of application, and the applicability is not strong. With the development of machine learning algorithms, BP neural network [14], SVM, genetic algorithm (GA), particle swarm algorithm (PSO), and others are applied to the cost prediction model. Compared to traditional algorithms, these algorithms have more improvements in prediction accuracy and sample requirements.

An et al. [15] proposed to establish an SVM model by using the dimensionality reduction of least squares regression and the original feature indicators to predict the cost of subprojects and one-sided costs and found that the prediction efficiency of the SVM model based on dimensionality reduction indicators has been improved. Yi et al. [16] established a construction cost prediction model using the least squares support vector machine, and optimized the model using the improved PSO algorithm, verifying the significance of the model through examples. Yi et al. [16] used the PSO algorithm to optimize the model and verified the significance of the model through examples. Lin et al. [17] used Principal Component Analysis (PCA) technology to process the residential engineering data and used the support vector machine and least squares support vector machine for training, prediction, and comparative analysis to select a more reasonable prediction model.

Jafarzadeh et al. [18] proposed an Artificial Neural Network (ANN) as the basis for extracting the main feature parameters of the project and determining the internal storage weights after sufficient learning of the training samples, so as to carry out the cost estimation at the design stage without the complete information of the project. Hong et al. [19] further improved the BP neural network algorithm using PSO to optimize the initial weights and thresholds of the model. neural network algorithm to optimize the

initial weights and thresholds of the model. However, such neural network algorithms require a large number of training samples, and the training speed is slow, which fails to take into account the limited number of historical reference cases that can be collected during the actual prediction process and the insufficient resources of the sample base. Car-Pusic et al. [20] used neural networks to establish a project cost prediction model. By collecting field data and historical data with certain similarity of the project to be predicted, the linear dynamic values of the cost parameters over time are extracted, and the laws are converted into comparison factors to realize the prediction. However, this prediction method has more limitations, and the prediction error is larger for projects with a large number of items. Dang-Trinh et al. [21] used a machine learning-based project cost prediction method. The center value of each step of building construction is extracted by principal component analysis method. Then the minimum cost required for each step is calculated in steps. The machine learning method is used to filter and solve the prediction samples to obtain the cost trend and then realize the prediction. However, this method does not take into account the influence of uncertainty and unexpected factors in building construction, and the prediction results are affected by the large error in the a priori numerical calculation.

1.2. **Motivation and Contribution.** SVM-based prediction models have relatively weak fitting ability for complex datasets and nonlinear relationships. By combining with BP neural networks, the BP-SVM model can make full use of the powerful generalization ability of neural networks and the convex optimization property of SVM, so as to improve the adaptability of the model and achieve better prediction results for different types of datasets and scenarios. In addition, Twin Support Vector Machine (TWSVM) is an improved algorithm based on SVM [22, 23], which learns two optimal partition hyperplanes at the same time to improve the classification performance and generalization ability. Compared with SVM, TWSVM has advantages in both learning efficiency and calculation efficiency. It learns two independent maximum boundaries instead of one, which makes the calculation more efficient. At the same time, the models learned by TWSVM are usually more sparse and easier to explain. In this paper, we try to combine TWSVM with BP neural networks to realize the cost prediction model. The main innovations and contributions of this work include:

(1) On the basis of expert scoring, combined with the composition of construction project cost and its influencing factors, a sample of 28 cases of engineering data with 17 characteristic indicators was initially constructed. At the same time, principal component analysis was used to downscale the attribute indicators to obtain 9 comprehensive indicators with higher contribution rates, thus reducing the sample complexity and improving the learning efficiency of the model.

(2) In order to further improve the accuracy of the prediction model, the a priori information and probabilities of different characteristic data in the historical data samples of construction project cost are collected. The solution results are used as the initial comparison conditions of the subsequent prediction model to minimize the determination error.

(3) In construction cost prediction, it is proposed to combine BP neural networks and TWSVM to obtain a more powerful and accurate hybrid prediction model. Aiming at the defects of the TWSVM model that the parameter settings depend on empirical values, the PSO-TWSVM model is proposed with the advantage of PSO in the field of parameter optimization, which enhances the hybrid model's stability even further.

2. **Determination of model indicators and pre-processing.**

2.1. **Construction of cost prediction index system.** To carry out the prediction of construction project cost, it is necessary to provide a set of construction project characteristic vectors as input variables for the prediction model, so as to output the predicted value of project cost after intelligent learning of the collected project data in the model. Engineering characteristic indicators are key parameters that represent the unique features of a building project and provide the breakdown of construction costs. The engineering cost characteristic index is essential for predicting various engineering costs accurately. The prediction index system condenses numerous small factors of a construction project into a comprehensive overview, reflecting the project's overall situation. Thus, it is essential to adhere rigorously to the fundamental construction principles and procedures while systematically and thoroughly developing the building project cost prediction index system to facilitate the model's input.

In order to identify the characteristic indicators of construction project cost more objectively, this study, based on the basic composition of construction project cost, divides the prediction indicator system into different levels, and carries out the refinement of the characteristic indicators respectively, so as to strengthen the systematicity and completeness of the indicator system. To summarize, the five indicators related to architectural features, structural foundation-related indicators, decoration-related indicators, installation engineering-related indicators, and project features-related indicators are taken as the first-level indicators. Also, in order to reflect the commonalities of construction projects comprehensively and objectively, the first-level indicators are further divided according to the main features of individual engineering cases and the indicator selection principle. This allows for the initial acquisition of 20 second-level indicators.

The preliminary identification results of the prediction indicators will inevitably have the phenomenon of repetition, cross-definition, and unclear level. For this reason, combined with the basic principles of the construction project cost prediction indicator system construction, the expert scoring approach is selected to eliminate duplicate indications with little influence on the project cost projection. The numerical values used in the five-level assessment approach are 1, 2, 3, 4, and 5. 20 feature indicators were evaluated for their value based on specialists' practical experience. To quantitatively examine the experts' viewpoints, the concentration and dispersion of their scores are used.

$$E_i = \frac{1}{p} \sum_{j=1}^{5} E_j n_{ij} \tag{1}$$

$$\sigma_i = \sqrt{\frac{1}{p-1} \sum_{j=1}^{5} n_{ij} \left(E_j - E_i\right)^2} \tag{2}$$

where E_j denotes the value of the j-th level of importance, n_{ij} denotes the proportion of experts agreeing on the importance level of the indicator, p denotes the total amount of experts, and σ_i denotes the variation in expert ratings for the indicator's importance.

Generally speaking, when $E_i \leq 3$ and $\sigma_i \leq 1$, it is considered that the index item reaches the degree of importance and above. The results of expert scoring are shown in Table 1 (the numbers corresponding to each level of importance in the table are the number of experts who chose the item).

As can be seen from the above table, the concentration of experts' opinions for "5", "8" and "12" are all greater than 3, and the dispersion of experts' opinions is greater than 1. After detailed communication with experts, it is determined that the above three indicators have less influence on the project cost forecast, so these three redundant indicators are deleted. Finally, the remaining types of feature indicators are 17.

Table 1. Statistical Tables for Evaluating the Importance of Construction Cost Forecasting Indicators

No.	Extremely important	Important	Critical	General	Unimportant	Concentration E_i	Dispersion σ_i
1	23	8	2	1	1	1.12	0.32
2	23	7	3	1	2	1.15	0.43
3	21	8	4	3	2	1.39	0.59
4	22	6	5	2	2	1.42	0.66
5	4	3	6	9	3	3.62	1.15
6	23	9	2	1	1	1.12	0.33
7	21	11	1	2	2	1.15	0.31
8	3	5	5	10	12	3.59	1.17
9	17	9	3	2	1	1.52	0.69
10	25	3	3	1	1	1.12	0.47
11	19	11	3	2	1	1.29	0.44
12	3	6	4	8	4	3.62	1.15
13	16	13	3	2	1	1.45	0.59
14	11	2	4	2	2	1.75	0.76
15	2	3	3	9	14	3.82	0.94
16	15	11	2	3	1	1.62	0.74
17	17	13	2	1	2	1.42	0.67
18	18	10	4	2	1	1.42	0.63
19	22	7	3	1	2	1.25	0.59
20	31	3	2	2	2	1.14	0.11

2.2. **Indicator dimensionality reduction based on PCA.** For construction projects, there are many factors affecting their project cost, and although this paper has initially screened out 17 factors that are representative of them, there are still quantitative direct or indirect relationships between the indicators. Some of the factors can be represented by some existing factors, resulting in a large amount of overlap of sample information. Theoretically, it seems that any one of the characteristic indexes may have a certain impact on the construction project cost; however, in the actual cost prediction process, the staff does not need to consider all the characteristic indexes one by one. On one hand, if all the feature indicators are included in the input set during the modeling process, it will cause the workload of the SVM model to rise sharply and affect the learning efficiency. On the other hand, if some important influencing factors are neglected in the data input process, it will lead to the lack of accuracy of the prediction model and poor prediction results, so it is necessary to reasonably analyze the feature indicators to obtain the comprehensive main influencing factors.

In summary, in order to make the construction cost prediction model based on SVM achieve a better generalization effect, this paper proposes to use PCA combined with the prediction model, and the comprehensive variables obtained by PCA are used as input vectors of SVM, eliminating the high correlation between selected feature indicators, thus overcoming the dimensionality catastrophe to a certain extent, and improving the prediction accuracy of the model. PCA can prevent the correlation of different influencing factors from affecting the model. PCA can prevent the correlation of different influencing factors from affecting the actual prediction, thus avoiding the over-fitting problem caused by excessive input [24]. The 28 engineering data samples obtained from screening were analyzed by PCA, and the equivalent transformation of a small number of principal components was carried out by dimensionality reduction.

Based on the standardization, the corresponding matrix is constructed for a given indicator value, as shown below:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,17} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,17} \\ \vdots & \vdots & \ddots & \vdots \\ x_{28,1} & x_{28,2} & \cdots & x_{28,17} \end{bmatrix} \tag{3}$$

where the 29×17 matrix represents the 28 sample engineering data samples with 17 types of samples.

Combined with the correlation theory of matrices in SPSS, the correlation of variables was tested using Bartlett's spherical test and the corresponding KMO test [25], and the results are shown in Table 2.

Table 2. Correlation Test Results

Test methods	**Norm**	**Numerical value**
KMO test	KMO	0.658
Bartlett's test of sphericity	Approximate chi-square	457.3
	Degrees of freedom	164
	p	0.0002

In PCA analysis, the total variance, i.e., the cumulative contribution, is usually required to exceed 80% in order to comply with the requirements for the end of principal components [26]. The design of this paper sets this value to 85%. Among the 17 feature components, the first 9 features with high contribution rates are extracted, and their vote responsibility rate is 85.530, which meets the overall explanation requirements. The factor analysis of industrial building construction cost of crushed stone diagram is shown in Figure 1, where it can be seen that the 9th factor turn is more obvious, so the first 9 components can be extracted.

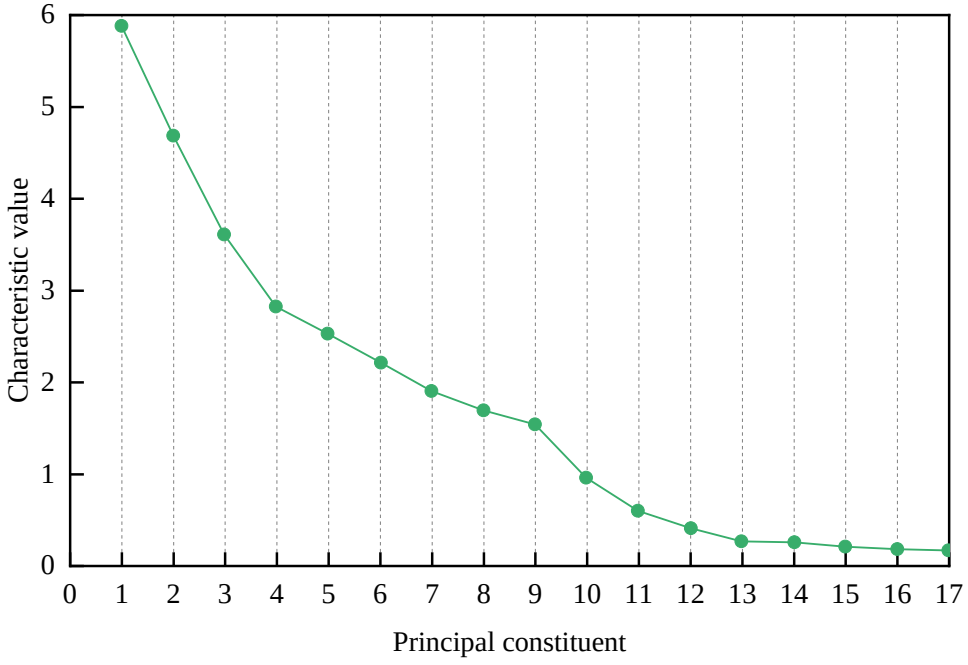

Figure 1. Factorization of gravel plots

3. **Hybrid machine learning model-based construction cost prediction.**

3.1. **Historical construction cost a priori model.** In order to improve the accuracy of construction cost prediction, the a priori information, as the first step of the prediction algorithm, needs to be ensured in order to improve the quality of the subsequent prediction.

Initialization is done with the initial historical data.

$$x_1^{(0)}(k) = \sum_{m=1}^{k} x_1^{(0)}(m) = \sum_{m=1}^{5} x_1^{(0)}(m) \tag{4}$$

where m is the data vector.

Create a data matrix and find the pooled mean based on the front and back neighboring values of the sample data.

$$a^{(1)}\left[x_1^{(1)}(k+1)\right] = -az_1^{(1)}(k+1) + \sum_{i=2}^{N} b_i z_i^{(1)}(k+1) \tag{5}$$

where $a^{(1)}$ is the parameter column; b_i is the parameter column of neighboring points; and $z_i^{(1)}$ is the set mean.

Let $a^{(1)} = [a, b_2, b_3, \ldots, b_N]$, according to the cost of construction project cost of the previous years of the use of cost (labor, machinery, materials, and the total cost of the total cost of construction) substitution into Equation (4), to get the total cost of the total cost of the sample data $x_N^{(0)}$ of the a priori model for the following.

$$x_N^{(0)}(k) = \sum_{m=1}^{k} b_i x_i^{(1)}(k) - az_1^{(1)}(k) \tag{6}$$

The a priori models corresponding to labor, machinery use, and material costs are then calculated separately.

(1) A priori model of labor cost $x_1^{(0)}$.

$$x_1^{(0)}(k) = -0.52457z_3^{(1)}(k) + 1.868196x_3^{(1)}(k) - 0.16192x_4^{(1)}(k) - 0.99745z_1^{(1)}(k) \tag{7}$$

(2) A priori model of machinery usage cost $x_2^{(0)}$.

$$x_2^{(0)}(k) = 1.445301x_2^{(1)} - 1.377999z_2^{(1)}(k) \tag{8}$$

(3) A priori model of material costs $x_3^{(0)}$.

$$x_3^{(0)}(k) = 1.060936 + 0.01682z_3^{(1)}(k) \tag{9}$$

The a priori probability calculation model for the characteristics of labor, machinery use, and material costs in construction cost can be obtained after the above process calculation. The a priori information obtained can be used as the condition reference of the subsequent prediction model. The comparison of the information greatly improves the accuracy of the cost data prediction, and can effectively avoid the judgment error to ensure the quality of prediction.

3.2. **PSO-TWSVM algorithm design.** TWSVM is an enhanced iteration of classical machine learning, namely the SVM branch. It seeks for a set of non-parallel hyperplanes, resulting in superior classification capabilities, making it ideal for addressing approximation sample classification challenges. TWSVM is more computationally efficient than regular SVMs because it solves two SVM-type problems. The time complexity of SVM is $O(n^3)$, and with the increase of the number of samples n, the amount of calculation will

increase sharply. TWSVM learns two independent maximum boundary problems, and the time complexity of each sub-problem is about $O(n^2)$, and the overall time complexity is $2O(n^2)$. This means that the computation of TWSVM is linear, and with the increase of sample number, the computation will grow more slowly. In addition, TWSVM learns two sparse models, which are not only easier to explain, but also more efficient in prediction. Generally speaking, TWSVM is significantly superior to standard SVM in learning efficiency and prediction efficiency in big data scenarios, which is also an important reason for its wide adoption.

Assume that the set of training samples in the n-dimensional real number space R^n is (x^j, y_j), $i = 1, 2$, $j = 1, 2, \ldots, m$. The total number of samples is $m = m_1 + m_2$, where m_1 is the number of sample points in the positive category and m_2 is the number of sample points in the negative category. Then the method for seeking a nonlinear TWSVM hyperplane is:

$$K(x^T, C^T)y_1 + b_1 = 0, \quad K(x^T, C^T)y_2 + b_2 = 0 \tag{10}$$

Similarly, the plane dividing the positive and negative classes is obtained by solving the following two quadratic programs.

$$\min_{u_1,b_1,\xi} \frac{1}{2}\left\|K(A, C^T)u_1 + e\rho_1\right\|^2 + c_1 e^T\xi \quad s.t. -(K(B, C^T)u_1 + e_2 b_1) + \xi \geq e_2, \xi \geq 0 \tag{11}$$

$$\min_{u_2,b_2,\xi} \frac{1}{2}\left\|K(B, C^T)u_2 + e_2 b_2\right\|^2 + c_2 e_1^T\xi \quad s.t.(K(A, C^T)u_2 + e_2 b_2) + \xi \geq e_1, \xi \geq 0 \tag{12}$$

where $e_1 = (1, \ldots, 1)^T \in R^n$, $e_2 = (1, \ldots, 1)^T \in R^h$.

To further simplify Equation (11) and Equation (12), they are pairwise transformed.

$$\max_{\alpha} e_2^T\alpha - \frac{1}{2}\alpha^T R(S^T S)^{-1}R^T\alpha \quad s.t. \quad 0 \leq \alpha \leq c_1 e_2 \tag{13}$$

$$\max_{\gamma} e_1^T\gamma - \frac{1}{2}\gamma^T S(R^T R)^{-1}S^T\gamma \quad s.t. \quad 0 \leq \gamma \leq c_2 e_1 \tag{14}$$

$$R = [K(B, C^T)e_2], \quad S = [K(A, C^T)e_1] \tag{15}$$

Solving Equation (13) and Equation (14) gives:

$$(u_1^T, b_1)^T = -(S^T S)^{-1}R^T\alpha \tag{16}$$

$$(u_2^T, b_2)^T = -(R^T R)^{-1}S^T\gamma \tag{17}$$

It can be seen that according to u_1, u_2, b_1 and b_2, the hyperplane for classification can be obtained, then its classification decision function is:

$$classlabel = \arg\min_{k=+1,-1} |K(x^T, C^T)u_k + b_k| \tag{18}$$

PSO is a heuristic algorithm by simulating the collective behavior of pigeons or fish schools. It is often used to solve optimization and machine learning problems. Let $W = (W_1, W_2, ..., W_n)$ denotes a population of n particles in S-dimensional space, $W_i = (W_{i1}, W_{i2}, ..., W_{in})^T$ denotes an S-dimensional vector of the i th particle with velocity denoted as $V_i = (V_{i1}, V_{i2}, ..., V_{is})^T$. In the iterative computation step, two particles are searched for, the previous particle for finding its own optimal solution. The previous individual extremum is denoted as $P_i = (P_{i1}, P_{i2}, ..., P_{is})^T$; the latter individual is the

optimal solution for the race at this stage, i.e., the global optimal solution, denoted as $P_g = (P_{g1}, P_{g2}, ..., P_{gs})^T$.

Using the race-optimal solution and itself, the particle updates its own position and velocity as follows:

$$P_g^k = \{P_1^k, P_2^k, ..., P_n^k\} \quad s.t. \quad f(P_g^k) = \min\{f(P_1^k), f(P_2^k), ..., f(P_n^k)\} \tag{19}$$

$$P_i = \frac{\varphi_1 P_{id} + \varphi_2 P_{gd}}{\varphi_1 + \varphi_2} \tag{20}$$

$$m_{best} = m\left(\prod_{i=1}^{m} P_i\right) \tag{21}$$

$$W_{id}^{k+1} = \tau\left(P_{id} - P_g\right) \pm \gamma\left|m_{best} - W_{id}^k\right| \times \ln\left(\frac{1}{u}\right) \tag{22}$$

where $d = 1, 2, ..., S$, $i = 1, 2, ..., n$; k denotes the number of iterations; $\varphi_1 = c_1 r_1$, $\varphi_2 = c_2 r_2$; c_1 and c_2 denote the non-negative constants, i.e., the acceleration factors; r_1 and r_2 denote two random numbers with distribution range $[0, 1]$; m_{best} denotes the average optimal value in the population; γ denotes the dilation adjustment factor, which is able to regulate the rate of convergence; u and τ denote two random numbers with value range $(0, 1)$.

It is found that TWSVM parameter setting is more difficult in the process of classification and identification, so this paper uses PSO algorithm for parameter optimization of TWSVM to further improve the accuracy of construction cost prediction. The PSO-based TWSVM algorithm needs to determine three core parameters of SVM when solving nonlinear problems, including the penalty factors (C_1, C_2) and the kernel parameter σ in the Gaussian kernel function. Initialize a particle swarm, the position of its i-th particle is represented as a 3D vector $X_i = (x_{i1}, x_{i2}, x_{i3})$, x_{i1} and x_{i2} denote the two penalization factors, and x_{i3} denotes the kernel parameter σ.

3.3. **Construction cost prediction based on BP-PSO-TWSVM.** In construction cost prediction, this paper combines BP neural nets and TWSVM in order to utilize the advantages of each of them. This combination aims to create a more powerful and accurate prediction model. Traditional SVM-based construction cost prediction models have limited ability to model complex nonlinear relationships. BP neural networks, which are good at learning and capturing complex nonlinear relationships in the data, can help to complement SVM in this regard, thus making the model more accurate in predicting the construction cost.

Multi-layer feed-forward neural networks trained via error back-propagation are known as BP neural networks, which is good at capturing and learning complex nonlinear relationships in data, but may suffer from overfitting problem and local optimization problem. Compared with SVM, TWSVM is a more advanced supervised learning algorithm, which searches for the optimal segmentation hyperplane in the data feature space to distinguish between different data classes and shows good generalization ability for small samples and nonlinear problems. In this way, the BPNN acts as a feature converter to capture complex patterns in the data. In this way, the high fitting ability of BP neural network and the strong generalization ability of TWSVM can complement each other to improve the performance of the whole construction cost prediction model. The main steps of the hybrid model proposed in this paper are as follows:

(1) Historical data is preprocessed by PCA technique for dimensionality reduction to form sample matrix. The dataset is divided into training and testing sets using cross-validation (k-fold cross-validation).

(2) Set up a three-layer BP neural network with one Sigmoid function hidden layer. Randomly initialize the weights and biases. Nonlinear feature learning of PCA processed input sample engineering data using BP neural network;

(3) Set the value range of parameters C and σ, and initialize a population $X = \{X_1, X_2, \cdots, X_m\}$ of m particles;

(4) Use the trained BP neural network to forward propagate the training and test sets to obtain the activation values of the last hidden layer before the output layer as a new feature set for TWSVM. The new feature set obtained from the BP neural network is utilized to train the TWSVM model;

(5) Calculate particle fitness in conjunction with the training dataset and compare. Iterate repeatedly until the end condition is satisfied (adaptation value is minimized);

(6) Output the optimization hyperparameters C and σ and assign them to the TWSVM prediction model;

(7) Train the TWSVM model with the test set and input the predicted sample data for model prediction to obtain the optimal results.

4. **Experimental results and analysis.**

4.1. **Test environment.** In order to verify the effectiveness of the construction cost prediction method of Kijen BP-PSO-TWSVM, the test selects the first-line high-rise residence as the prediction object. This kind of high-rise residential project has a long cycle and large input capital, and its cost index change has a close relationship with the macro environment. Usually, the cost of construction project contains a lot of detailed features, and a complete set of construction cost data is composed of multiple sub-cost features. The proportion of sub-features is related to the role of the building, the shape and the basic data. Twenty of the thirty-eight projects served as training samples for the learning and training process, and the eight remaining data sets were evaluated with the trained model, according to the sample matrix that was collected earlier. We chose three different approaches, LS-TWSVM [29] and BP-PSO-TWSVM, for project cost prediction and simulation so that we could more thoroughly test the suggested hybrid model's efficacy.

4.2. **Analysis of the accuracy of the prediction model.** The prediction results of selecting the construction project cost within the year 2013-2023 are shown in Figure 2.

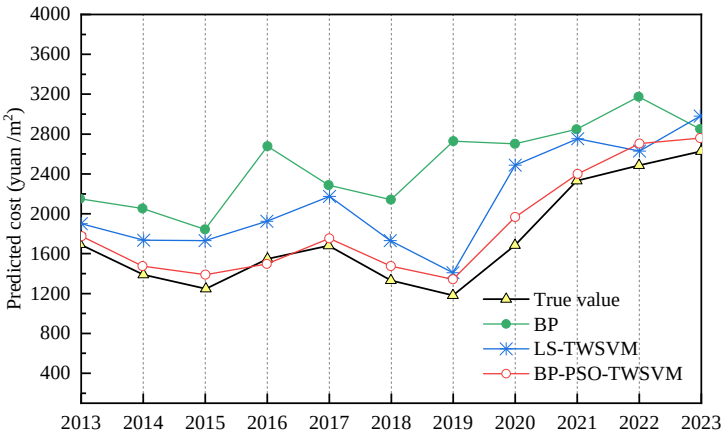

Figure 2. Stability curves for different clustering methods

It can be seen that among all the methods, only the prediction results of the mentioned BP-PSO-TWSVM model have the highest degree of agreement and the closest change amplitude between the prediction results and the real values, which indicates that the accuracy of the prediction results is higher. On the other hand, there is an obvious gap between the prediction results of the other two prediction models and the real values, with a lower similarity of the curves and a larger degree of deviation. Comprehensive comparison of the results shows that the proposed BP-PSO-TWSVM model is more practical and has a higher accuracy in predicting the cost of construction projects in practical applications.

4.3. **Stability Analysis of Prediction Models.** The impact of various algorithms on the predictive performance of the model is determined by the relative error δ and the mean absolute percentage error ($MAPE$).

$$\delta = \frac{y_i - \hat{y}_i}{y_i} \tag{23}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} |\delta| \times 100\% \tag{24}$$

where y_i and $\hat{y}_i$ denote the actual and predicted cost values of the i-th sample, and n is the number of test samples.

Figure 3 displays the discrepancies between the anticipated and actual results of the test samples for the three distinct prediction models. It can be seen that the distribution interval of the prediction relative errors of BP neural network model is [-1.37%, 5.65%]; the distribution interval of the prediction relative errors of LS-TWSVM model is [-2.05%, 3.58%]; and the distribution interval of the prediction relative errors of BP-PSO-TWSVM model is [-4.65%, 2.11%]. Therefore, the proposed hybrid machine learning model outperforms the BP and LS-TWSVM models in terms of the stability of construction cost prediction and obtains high robustness. In summary, the BP-PSO-TWSVM prediction model has better guidance for construction project cost and is more applicable to the prediction of pre-construction project cost.

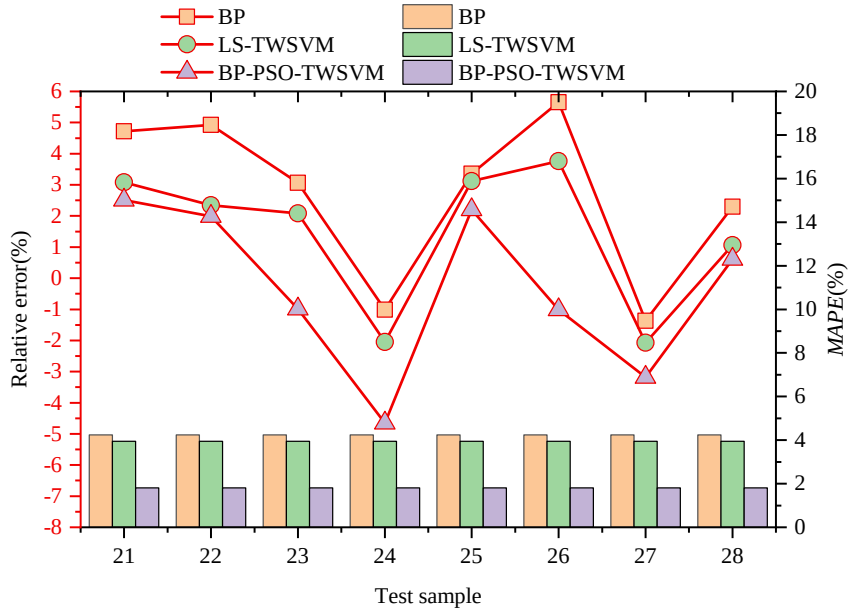

Figure 3. The discrepancies between the anticipated and actual results

5. **Conclusion.** This work proposes to build a hybrid machine learning model by combining TWSVM and BP neural network for construction work cost prediction. On the basis of expert scoring, combined with the composition of construction work cost and its influencing factors, 28 examples of engineering data samples with 17 attribute indicators were initially constructed. At the same time, principal component analysis was used to downscale the attribute indicators to obtain 9 comprehensive indicators with higher contribution rate, thus reducing the sample complexity and improving the learning efficiency of the model. In order to further improve the accuracy of the prediction model, the a priori information and probability of different characteristics of data in the historical data samples of construction cost are collected. In construction cost prediction, it is proposed to combine BP neural net and TWSVM to obtain a more powerful and accurate hybrid prediction model. The experiment selects a first-line high-rise residence as the prediction object. The results show that the proposed hybrid machine learning model outperforms the BP and LS-TWSVM models in terms of stability of construction cost prediction and obtains high robustness. Therefore, the BP-PSO-TWSVM prediction model is more suitable for the prediction of pre-construction project cost.

REFERENCES

[1] S. M. Shahandashti, and B. Ashuri, "Highway construction cost forecasting using vector error correction models," Journal of Management in Engineering, vol. 32, no. 2, 04015040, 2016.
[2] S. Hwang, M. Park, H.-S. Lee, and H. Kim, "Automated time-series cost forecasting system for construction materials," Journal of Construction Engineering and Management, vol. 138, no. 11, pp. 1259-1269, 2012.
[3] S. Kim, B. Abediniangerabi, and M. Shahandashti, "Pipeline construction cost forecasting using multivariate time series methods," Journal of Pipeline Systems Engineering and Practice, vol. 12, no. 3, 04021026, 2021.
[4] R. M. Skitmore, and S. T. Ng, "Forecast models for actual construction time and cost," Building and Environment, vol. 38, no. 8, pp. 1075-1083, 2003.
[5] B. Aslam, A. Maqsoom, H. Inam, M. u. Basharat, and F. Ullah, "Forecasting Construction Cost Index through Artificial Intelligence," Societies, vol. 13, no. 10, 219, 2023.
[6] S. M. Shahandashti, and B. Ashuri, "Forecasting engineering news-record construction cost index using multivariate time series models," Journal of Construction Engineering and Management, vol. 139, no. 9, pp. 1237-1243, 2013.
[7] J.-w. 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.
[8] F. M. S. Al-Zwainy, and N. T. Hadal, "Application artificial forecasting techniques in cost management," Journal of Engineering, vol. 22, no. 8, pp. 1-15, 2016.
[9] R. Skitmore, S. G. Stradling, and A. P. Tuohy, "Human effects in early stage construction contract price forecasting," IEEE Transactions on Engineering Management, vol. 41, no. 1, pp. 29-40, 1994.
[10] K. Pujitha, and K. Venkatesh, "Forecasting the construction cost by using unit based estimation model," Materials Today: Proceedings, vol. 33, pp. 613-619, 2020.
[11] H. K. Park, S. H. Han, and J. S. Russell, "Cash flow forecasting model for general contractors using moving weights of cost categories," Journal of Management in Engineering, vol. 21, no. 4, pp. 164-172, 2005.
[12] D. Ye, "An algorithm for construction project cost forecast based on particle swarm optimization-guided BP neural network," Scientific Programming, vol. 2021, pp. 1-8, 2021.
[13] M. Juszczyk, K. Zima, and W. Lelek, "Forecasting of sports fields construction costs aided by ensembles of neural networks," Journal of Civil Engineering and Management, vol. 25, no. 7, pp. 715-729, 2019.
[14] A. Mahmoodzadeh, M. Mohammadi, A. Daraei, H. Farid Hama Ali, A. Ismail Abdullah, and N. Kameran Al-Salihi, "Forecasting tunnel geology, construction time and costs using machine learning methods," Neural Computing and Applications, vol. 33, pp. 321-348, 2021.
[15] S.-H. An, U.-Y. Park, K.-I. Kang, M.-Y. Cho, and H.-H. Cho, "Application of support vector machines in assessing conceptual cost estimates," Journal of Computing in Civil Engineering, vol. 21, no. 4, pp. 259-264, 2007.

[16] T. Yi, H. Zheng, Y. Tian, and J.-p. Liu, "Intelligent prediction of transmission line project cost based on least squares support vector machine optimized by particle swarm optimization," Mathematical Problems in Engineering, vol. 2018, pp. 1-11, 2018.
[17] T. Lin, T. Yi, C. Zhang, and J. Liu, "Intelligent prediction of the construction cost of substation projects using support vector machine optimized by particle swarm optimization," Mathematical Problems in Engineering, vol. 2019, pp. 1-10, 2019.
[18] R. Jafarzadeh, J. M. Ingham, S. Wilkinson, V. González, and A. A. Aghakouchak, "Application of artificial neural network methodology for predicting seismic retrofit construction costs," Journal of Construction Engineering and Management, vol. 140, no. 2, 04013044, 2014.
[19] Y. Hong, H. Liao, and Y. Jiang, "Construction Engineering Cost Evaluation Model and Application Based on RS-IPSO-BP Neural Network," Journal of Computers., vol. 9, no. 4, pp. 1020-1025, 2014.
[20] D. Car-Pusic, S. Petruseva, V. Zileska Pancovska, and Z. Zafirovski, "Neural network-based model for predicting preliminary construction cost as part of cost predicting system," Advances in Civil Engineering, vol. 2020, pp. 1-13, 2020.
[21] N. Dang-Trinh, P. Duc-Thang, T. Nguyen-Ngoc Cuong, and T. Duc-Hoc, "Machine learning models for estimating preliminary factory construction cost: case study in Southern Vietnam," International Journal of Construction Management, vol. 23, no. 16, pp. 2879-2887, 2023.
[22] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," Computers, Materials & Continua, vol. 79, no. 1, pp. 19-46, 2024.
[23] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on SVM in VR art media interaction environment," Human-centric Computing and Information Sciences, vol. 9, 40, 2019.
[24] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," Journal of Intelligent & Fuzzy Systems, vol. 43, no. 2, pp. 2121-2133, 2022.
[25] B. M. S. Hasan, and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," Journal of Soft Computing and Data Mining, vol. 2, no. 1, pp. 20-30, 2021.
[26] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," Nature Reviews Methods Primers, vol. 2, no. 1, 100, 2022.
[27] R. S. Rao, A. R. Pais, and P. Anand, "A heuristic technique to detect phishing websites using TWSVM classifier," Neural Computing and Applications, vol. 33, pp. 5733-5752, 2021.
[28] M. Tanveer, T. Rajani, R. Rastogi, Y.-H. Shao, and M. Ganaie, "Comprehensive review on twin support vector machines," Annals of Operations Research, pp. 1-46, 2022.
[29] M.-x. Chu, A.-n. Wang, R.-f. Gong, and M. Sha, "Multi-class classification methods of enhanced LS-TWSVM for strip steel surface defects," Journal of Iron and Steel Research International, vol. 21, no. 2, pp. 174-180, 2014.