Cost Prediction and Control Measures for Road and Bridge Projects Using Combined PCA, MLRM, and SVM

Lei Zhai*, Xiangqin Yan, Guangbiao Liu
Road & Bridge Engineering College, Xinjiang Vocational & Technical College of Communications, Urumqi 831401, China
E-mail: jelly202301@163.com
*Corresponding author

Keywords: engineering cost, multiple linear regression, support vector machine, principal component analysis

Received: June 5, 2024

This study proposes a cost prediction model for road and bridge projects that combines principal component regression (PCR) and support vector machine (SVM). Principal component analysis (PCA) identifies the primary cost-influencing factors, while PCR and SVM predict costs. In this model, PCR is used to analyze and reduce the dimensionality of assumed independent variable data, and handle linear factors. SVM is used to handle nonlinear factors and predict costs. In order to determine the performance of the proposed model, the R-squared index is used to evaluate it. The weight proportions of each factor were analyzed. The model is compared with PCR and SVM-RBF models, showing superior performance with convergence after 70 iterations and an R-squared value of 0.87. Key cost factors include materials (63%), labor (13%), and equipment (12%). The above results show that PCR-SVM model can accurately predict the cost and influence factors of road and bridge engineering.

Povzetek: Inovativni model za napoved stroškov projektov cest in mostov z združitvijo PCA, MLRM in SVM učinkovito identificira ključne stroškovne dejavnike.

1 Introduction

With population growth and rapid economic development, infrastructure is under increasing pressure. At the same time, road and bridge facilities are under particularly high pressure, because car ownership is growing rapidly. If the pressure on the road and bridge network is to be relieved, it is necessary to expand or build additional road and bridge facilities. Road and bridge construction is often characterized by long construction periods and large capital investments. To ensure that the road and bridge construction process is adequately funded, it is necessary to accurately predict the cost of road and bridge projects to reasonably control the construction cost. Currently, the commonly used project cost prediction methods are least squares regression method, index measurement method, engineering coefficient estimation method, etc. These methods often have low accuracy and require constant adjustment, making it difficult to meet the requirements of cost prediction for road and bridge engineering. They lack good decision-making for cost control in the later stages of construction. The multiple linear regression model (MLRM) is more accurate, because it predicts the dependent variable by multiple independent variables. However, factor selection and expression in regression analysis are only guesses, which affects the diversity of independent variables and the unpredictability of some factors. At the same time, it is difficult to consider all the factors that affect the project cost prediction results, so it is necessary to identify the main influencing factors. Principal component analysis (PCA) can analyze and simplify many factors to get the main influencing factors and reduce the influence between evaluation indicators and engineering budget calculations. However, the interpretation meaning of PCA is often vague and not as clear as the original data [1-2]. Among the influencing factors of engineering cost, there are both linear and nonlinear influencing factors. This requires the model to be able to deal with both linear and nonlinear problems. Support vector machine (SVM) is widely used because it can solve both linear and nonlinear problems, as well as classification and regression problems at the same time [3]. Therefore, a road and bridge project cost prediction model based on principal component regression and SVM (PCR-SVM) is proposed by combining the above three algorithms. It accurately predicts project costs while reflecting the impact of various influencing factors on the prediction results.

The main contents are as follows. Chapter 1 gives a brief introduction to the current research status of engineering cost prediction models and the application of MLRM. Chapter 2 investigates the algorithm of PCR-SVM model. Chapter 3 analyzes the project cost prediction results of PCR-SVM model and the degree of main influencing factors. Chapter 4 summarizes the research results of the full text.

2 Literature review

The cost of roads and bridges is influenced by numerous

factors, causing the composition and calculation to be complex. Wang et al. proposed a cost prediction model based on PCA, particle swarm optimization algorithm, and least squares SVM. The model used PCA to process the data and particle swarm algorithm for the optimal parameters and regularization parameters of the least squares SVM. The mean square error and average relative error of the prediction model results were tested to be 10.01% and 0.79%, respectively, which accurately predicted the cost of highway engineering [4]. Sharma et al. proposed an assisted optimization scheme based on machine learning for the problem of how to optimize the construction cost. The method used gradient augmentation tree to predict the construction cost and then used Bayesian optimization to optimize the construction cost. The test results showed that the method calculated a suitable construction cost optimization scheme [5]. Fan and Sharma proposed a cost prediction model based on SVM and LSSVM for the problem of how to accurately predict cost. The test results indicated that the relative error of the model was less than 7% and the accuracy met the requirements [6]. Priti and Salunkhe proposed an engineering cost prediction model based on artificial neural network for construction cost prediction. It used material cost as input data for project cost prediction. The prediction results of construction cost were accurate [7]. Ashour et al. proposed a cost control method based on earned value correction for the problem of how to achieve construction cost control. The method compared the difference between the predicted cost and the actual cost by correcting the earned value. The test results indicated that the final cost obtained by this method was lower than the predicted cost and lower than the actual cost [8].

MLRM is widely used in many fields such as biology, medicine, and air quality because they can calculate multiple independent variables to produce a unique result and the result is realistic. Croteau et al. proposed a mixed toxicity evaluation model based on MLRM and stepwise regression to evaluate the toxicity of nickel to aquatic organisms. Test results showed that the model was more accurate in evaluating the magnitude of elemental nickel toxicity to aquatic organisms than the Pooledll yuck MLR model [9]. Zi et al. proposed a prediction model based on multiple linear regression and SVM for the problem of how to predict the electrical conductivity of imidazole-based ionic liquids at different temperatures. The model was used to calculate and correlate the electrical conductivity of different imidazole ionic liquids by quantitative structure-property relationship method. The R-square of the model was tested to be about 0.99, and the mean absolute relative deviation was about 7.5% [10]. Attanayake et al. proposed a MLRM based morbidity prediction model for the dengue morbidity prediction. The model used interval value data analysis method to process the data, and then used temperature, rainfall and other data to predict the interval value of dengue incidence rate [11]. Mansor et al. proposed a PM10 prediction model based on MLRM for the problem of how to achieve PM10 prediction within three hours. The test results showed that the R-square of PM10 prediction results for 1, 2 and 3 hours were about 0.61, 0.42 and 0.35, respectively, which showed that the MLRM had the most accurate prediction for PM10 within 1 hour [12]. Hashemi et al. proposed a linear regression method for the problem of how to analyze the factors affecting the shear performance of reinforced concrete deep beams. Test results showed that this model fitted well with the actual results and fully reflected the degree of influence of factors [13].

In summary, the current research on construction cost prediction models has achieved certain results, but most of the project cost prediction models are predicted based on historical data, which are weak in interpretability and can hardly reflect the importance of influencing factors. Therefore, the research proposes a road and bridge project cost prediction model based on MLRM, PCA and SVM to accurately predict cost and reflect the importance of each influencing factor. The relevant research is shown in Table 1.

Table 1: Summary of the literature

Author	Method	Key results
Wang et al [4]	Cost prediction model based on PCA, particle swarm algorithm and least squares SVM	The mean relative error of the prediction results was significantly reduced by only 0.79%
Sharma et al [5]	Machine learning optimization scheme based on auxiliary optimization	Accurately calculated the project cost optimization scheme
Fan and Sharma [6]	Cost prediction model based on SVM and LSSVM	The prediction relative error was less than 7%
Priti and Salunkhe [7]	Engineering cost prediction model based on an artificial neural network	High accuracy of the cost prediction results
Ashour et al [8]	Cost control method based on cost value correction	The final cost was lower than the predicted cost and actual cost
Croteau et al [9]	Mixed toxicity evaluation model based on MLRM and stepwise regression	The evaluation accuracy of element nickel toxicity was better than the soil MLR model

Zi et al [10]	Conductivity prediction model based on multiple linear regression and SVM	The R-squared test of the model (0.99), mean absolute relative deviation (7.5%)
Attanayake et al [11]	Prediction model of dengue incidence, based on a MLRM	The prediction accuracy was better than the other models
Mansor et al [12]	A PM10 prediction model based on the MLRM	Realized the accurate prediction of PM10 within 1 hour
Hashemi et al [13]	Analysis model of influencing factors on shear performance of reinforced concrete deep beams based on linear regression	In good agreement with the actual results

3 Cost prediction model for road and bridge project based on PCA and MLRM

Roads and bridges, as essential basic transportation facilities, provide the foundation for the normal operation of traffic activities. At the same time, with the rapid increase in the number of vehicles, existing roads and bridges are under great pressure, so the scale of road and bridge construction is expanding by leaps and bounds. However, due to the high construction cost, how to control cost has become an important issue. This chapter explores the factors and control measures that affect the cost of road and bridge construction through PCR modeling.

3.1 PCR model based on PCA and MLRM

In the process of cost prediction, factors such as materials, design, construction methods, and quality of work can affect the prediction results. As a statistical method to reflect whether there is a linear relationship between the independent and dependent variables, MLRM can make more accurate predictions for problems influenced by multiple factors. Therefore, it is widely used in various fields. The process of MLRM is shown in Figure 1.

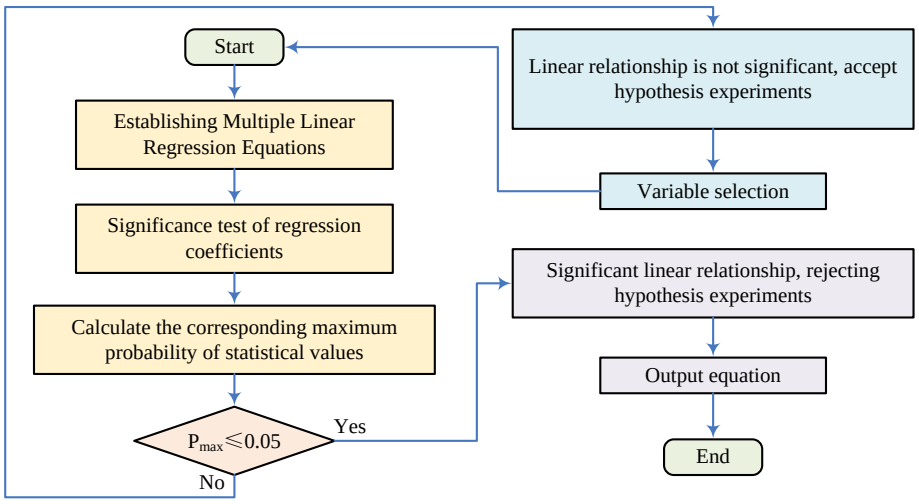


Figure 1: Process of multiple linear regression model

As can be seen from Figure 1, after establishing the multiple linear regression equation, the significance of the regression coefficient needs to be tested and the maximum probability value is calculated. When the maximum probability is less than or equal to 0.05, the linear relationship is significant. The hypothesis experiment is rejected and the equation is output. Otherwise, the linear relationship is not significant, the hypothesis experiment is accepted, and the regression model is re-established after screening the variables [14-15]. The MLRM general form is shown in equation (1).

$$y_i = b_0 + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} + \varepsilon_j \quad (1)$$

In equation (1), b_0 denotes the intercept of y. $b_1, b_2, \mathrm{L}, b_k$ denote the slope change of y on the independent variable x. ε_j denotes the random error of the first j observation for y. Since the relationship between the independent variable and the dependent variable is often not obvious in practical problems, it needs to be tested for significance. The significance testing is shown in equation (2).

$$F = \frac{SSR/k}{SSE/(n-k-1)} \sim F(k, n-k-1) \quad (2)$$

In equation (2), SSE is the squares regression sum. n is the number of rows of the variable matrix. k denotes the number of independent variables. Since there are many influencing factors, the study introduces PCA to solve the problem. Principal component analysis can transform multiple variables into a small number of comprehensive indicators, effectively compressing the number of variables and reducing the computational effort. The schematic diagram of PCA is shown in Figure 2.

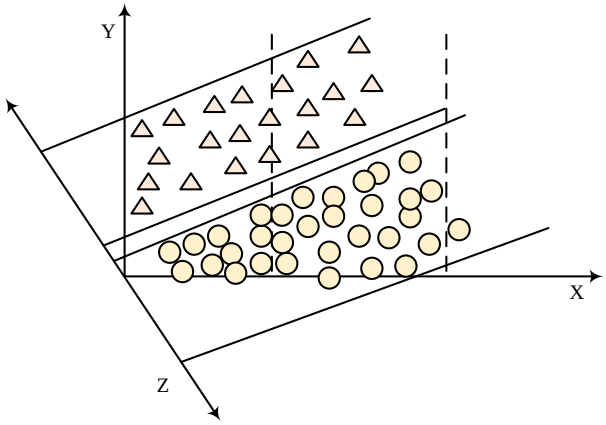


Figure 2: PCA schematic diagram

In Figure 2, the PCA transforms the component-related random vectors into component-unrelated new vectors by orthogonal transformation on the basis of the minimum loss of the original data. The transformed vectors are the principal components. Geometrically, it is represented as the original coordinate system being transformed into a new one [16-17]. The PCA is shown in equation (3).

$$\begin{cases} F_1 = l_{11}zx_1 + l_{12}zx_2 + \cdots + l_{1p}zx_p \\ F_2 = l_{21}zx_1 + l_{22}zx_2 + \cdots + l_{2p}zx_p \\ \vdots \\ F_m = l_{m1}zx_1 + l_{m2}zx_2 + \cdots + l_{mp}zx_p \end{cases} \quad (m \le p) \quad (3)$$

In equation (3), F_1, F_2 and F_m denote the reduced dimensional principal components. l_{ij} denotes the loading of the original vector mapped to the principal components. zx_{ij} denotes the normalized feature factors. The feature factor is shown in equation (4).

$$\begin{cases} zx_{ij} = \dfrac{x_{ij} - \mu_i}{\sigma_i} \\ \sigma_i = \sqrt{\dfrac{1}{n}\sum_{j=1}^{i}(x_{ij} - \mu_i)^2} \end{cases} \quad (4)$$

In equation (4), x_{ij} denotes the original variables. μ_i is the mean of the variables in column i. σ_i is the variance of that. The principal component loading is shown in equation (5).

$$l_{ij} = p(F_i, zx_j) = \sqrt{\lambda_i} e_{ij} \cdots (i, j = 1,2,\ldots, p) \quad (5)$$

In equation (5), F_i denotes the i th principal component. λ_i denotes the eigenvalue. e_{ij} is the j th component of the i th eigenvector. Since the correlation information between different variables has differences, the correlation test of the eigenfactors is necessary, as shown in equation (6).

$$KMO = \frac{\sum\sum_{i \ne j} r_{ij}^{\ 2}}{\sum\sum_{i \ne j} r_{ij}^2 + \sum\sum_{i \ne j} p_{ij}^2 +} \quad KMO \in [0,1] \quad (6)$$

In equation (6), r_{ij} indicates the correlation coefficient between the variables. P_{ij} indicates the partial correlation coefficient between the variables. When $KOM > 0.5$, the variables are suitable for factor analysis. The correlation coefficient and partial correlation coefficient are shown in equation (7).

$$\begin{cases} r = \dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}} \\ P_{ij,k} = \dfrac{r_{ij} - r_{ih}r_{jh}}{\sqrt{\left(1 - r_{ih}^2\right)\left(1 - r_{jh}^2\right)}} \end{cases} \quad (7)$$

In equation (7), X_i and Y_i denote the observed values of different random variables at the point i. $\bar{X}$ and $\bar{Y}$ denote the mean of different samples, respectively. The above algorithm constitutes the PCR model. The PCR model flow is shown in Figure 3.

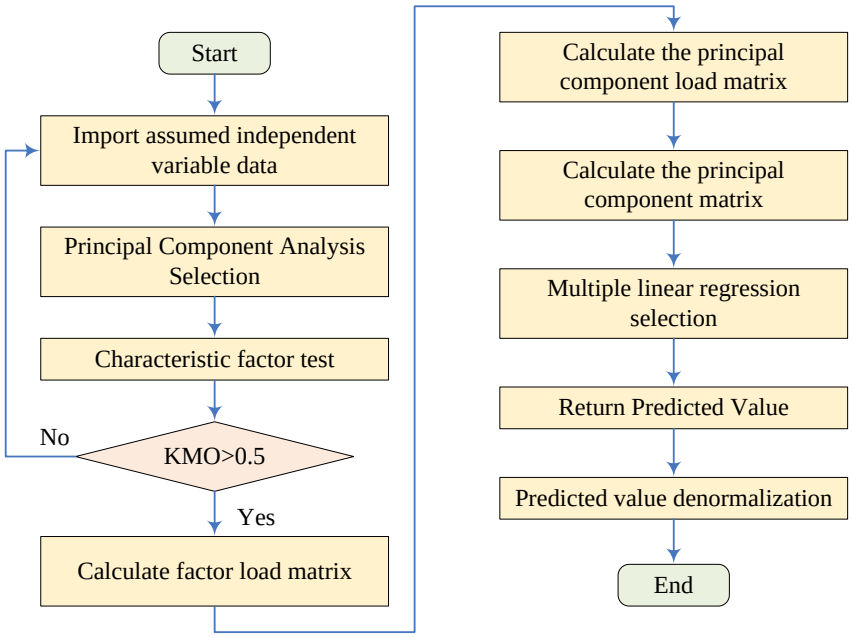
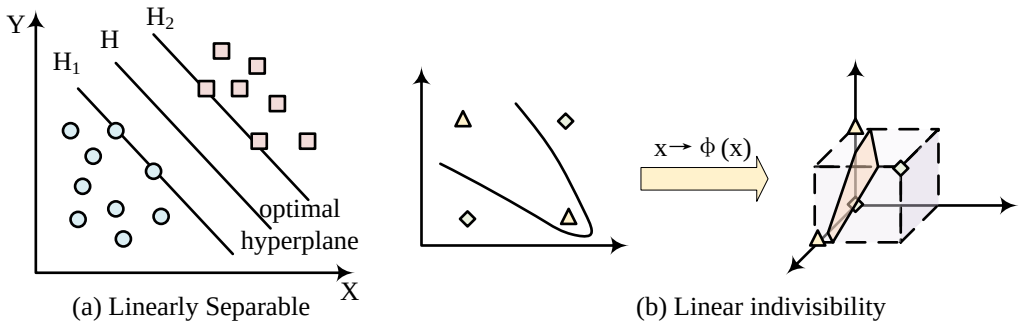


Figure 3: PCR model process

In Figure 3, PCR first analyzes and reduces the dimensionality of the hypothetical independent variable data. Then, it tests the significance and independence of the characteristic factors, and outputs the correlation matrix if it meets the requirements. Otherwise, the variable structure is changed until it meets the requirements. After obtaining the factor loading matrix, the principal component loading matrix is calculated. The principal component matrix is calculated. The dependent variable and the independent variable are selected to calculate the predicted dependent variable. Finally, the predicted values are counter-standardized and the results are output.

3.2 PCR-SVM-based cost prediction model for road and bridge projects

In addition to linear factors, there are also nonlinear factors when predicting project costs. Therefore, it is difficult to handle nonlinear factors through the simple PCR model. The study introduces SVM to enhance the processing capability of nonlinear problems. The SVM schematic diagram is shown in Figure 4.



(a) Linearly Separable (b) Linear indivisibility

Figure 4: Schematic diagram of SVM principle

In Figure 4(a), in the linearly divisible case, the SVM can find the correct partition line that is not unique. The further the distance between the two nearest partition lines from the partition interface, the smaller the confidence range of the generalization capability boundary. When this distance reaches the maximum, the optimal classification surface, i.e., the optimal hyperplane, is obtained. In Figure 4(b), in the case of indistinguishable linearity, the SVM classifies based on soft spacing, i.e., a small number of samples are allowed to often appear in the spacing band [18-20]. The hyperplane is shown in equation (8).

$$w^T x + b = 0 \tag{8}$$

In equation (8), w denotes the normal hyperplane vector. b denotes the displacement term. x denotes the sample. The distance from an arbitrary sample to the hyperplane is given in equation (9).

$$r = \frac{w^T x + b}{\|w\|} \tag{9}$$

In equation (9), r denotes the distance from any sample to the hyperplane. In order to find the divided hyperplane with the maximum distance, the constraints shown in equation (10) need to be satisfied.

$$\begin{cases} w^T x_i + b \geq 1, y_i = 1 \\ w^T x_i + b \leq -1, y_i = -1 \end{cases} \tag{10}$$

In equation (10), x_i and y_i denote different samples, respectively. w^T denotes the transpose vector of w. The SVM algorithm flow is shown in Figure 5.

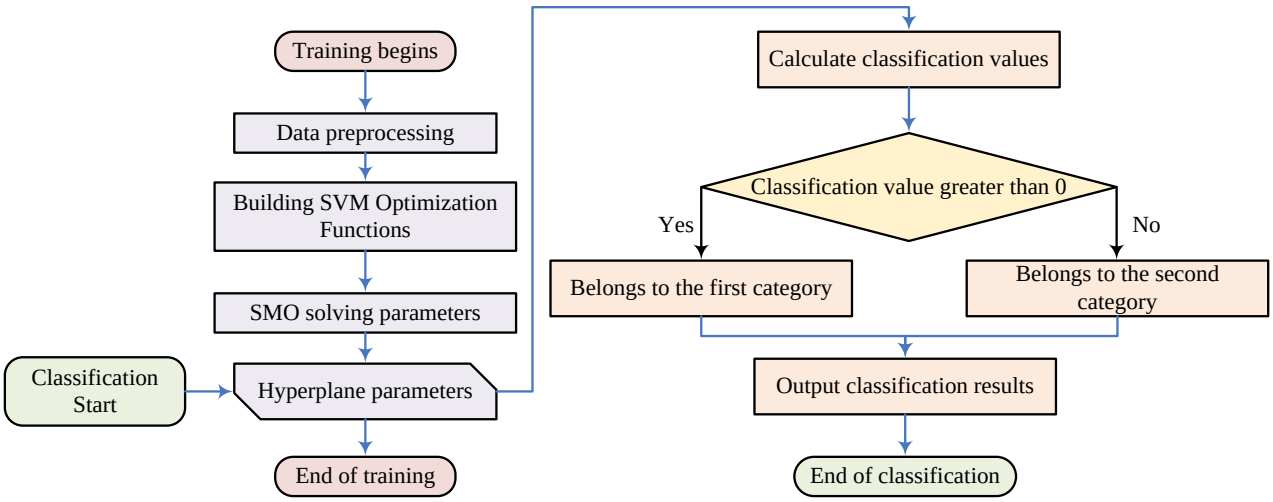


Figure 5: SVM algorithm process

In Figure 5, the first step is data pre-processing, and then the optimization function is constructed. After the optimization function is constructed, the parameters are solved by SMO to obtain the hyperplane parameters. In the linear regression model, the distance between the optimal hyperplane and all samples is minimized by

satisfying the condition shown in equation (11).

$$\begin{cases} \min\limits_{w,b} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{n}\left(\xi_i + \hat{\xi}_i\right) \\ s.tf\left(x_i\right) - y_i \le \varepsilon + \xi_i \\ y_i - f\left(x_i\right) \le \varepsilon + \hat{\xi}_i \\ \xi_i \ge 0, \hat{\xi}_i \ge 0, i = 1,2,\ldots,n \end{cases} \tag{11}$$

In equation (11), both ξ_i and $\hat{\xi}_i$ denote the relaxation variables. C denotes the penalty coefficients. $f\left(x_i\right)$ denotes the linear regression function. To solve the optimization problem of SVM, Lagrangian duality is introduced, as shown in equation (12).

$$L\left(w,b,\alpha,\hat{\alpha},\xi_i,\hat{\xi}_i,\mu,\hat{\mu}\right) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i - \hat{\xi}_i\right) - \sum_{i=1}^{n}\mu_i\xi_i - \sum_{i=1}^{n}\hat{\mu}_i\hat{\xi}_i + \sum_{i=1}^{n}\alpha_i\left(f\left(x_i\right) - y_i - \varepsilon - \xi_i\right) + \sum_{i=1}^{n}\hat{\alpha}_i\left(y_i - f\left(x_i\right) - \varepsilon - \hat{\xi}_i\right) \tag{12}$$

In equation (12), n denotes the number of variables. α denotes the scale factor, which is greater than 0. The SVM after Lagrangian duality transformation is shown in equation (13).

$$\begin{cases} \max\limits_{\alpha,\hat{\alpha}} \sum\limits_{i=1}^{n} y_i\left(\hat{\alpha}_i - \alpha_i\right) - \varepsilon\left(\hat{\alpha}_i + \alpha_i\right) - \dfrac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{i=1}^{n}\left(\hat{\alpha}_i - \alpha_i\right)\left(\hat{\alpha}_j - \alpha_j\right)k\left(x_i, x_j\right) \\ s.t.\sum\limits_{i=1}^{n}\left(\hat{\alpha}_i - \alpha_i\right) = 0 \\ 0 \le \alpha_i \le C, 0 \le \hat{\alpha}_i \le C \end{cases} \tag{13}$$

In equation (13), ε denotes the random error. $k\left(x_i, x_j\right)$ denotes the kernel function. The normal vector and bias term of the hyperplane are given in equation (14).

$$\begin{cases} w = \sum\limits_{i=1}^{n}\left(\hat{\alpha}_i - \alpha_i\right)k\left(x_i, x_j\right) \\ b = \sum\limits_{i=1}^{n}\left(\hat{\alpha}_i - \alpha_i\right)k\left(x_i, x_j\right) \end{cases} \tag{14}$$

In the SVM algorithm, the kernel function is significant in the prediction results of the model. To ensure the superior prediction performance, Gaussian function is chosen for the study, as shown in equation (15).

$$k\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|}{2\delta^2}\right), \delta \ge 0 \tag{15}$$

In equation (15), $\delta \ge 0$. The function has a good anti-interference ability for the noise in the data. The flow of the engineering cost prediction model combining SVM and PCR algorithm is shown in Figure 6.

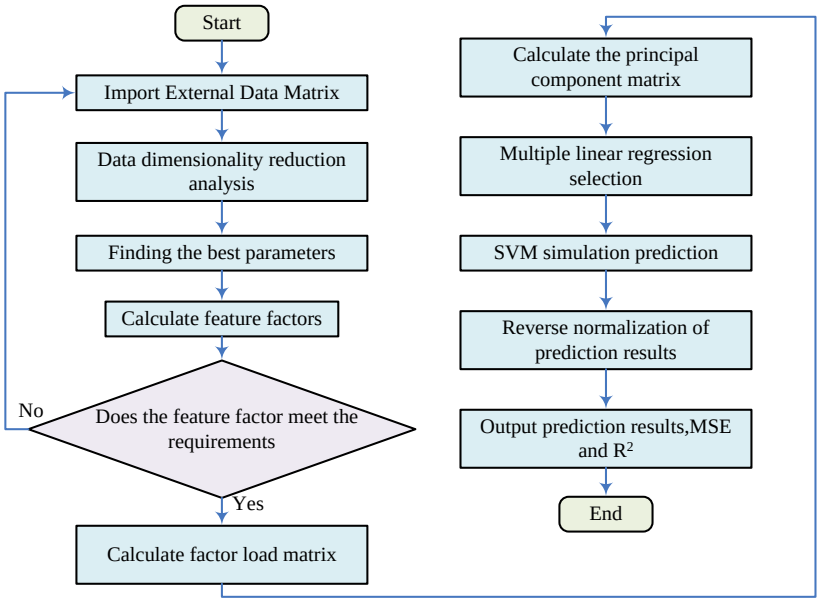


Figure 6: PCR-SVM model process

In Figure 6, PCR-SVM first normalizes the data, then performs dimensionality reduction and analysis. The optimal parameters are determined through the kernel function. Then the feature factors are calculated and evaluated, and the correlation matrix is output if it meets the requirements. Otherwise, the variable structure is changed until it meets the requirements. Next, the principal component loading matrix and principal component matrix are calculated, and the independent and dependent variables are selected. Finally, the prediction is simulated by SVM, and the prediction results are back-normalized to output the results.

4 Road and bridge project cost prediction results analysis and influence factors control

To verify the prediction accuracy of PCR-SVM and the degree of influence of different influencing factors on the prediction results, the study selects a bridge as a research sample and evaluates the reliability by calculating the error level between the predicted and actual values. The PCR-SVM is compared with PCR and SVR-RBF models. The magnitude of each influencing factor is also reflected by the importance of the variable coefficients. Before making prediction, the influencing factors is screened and the principal components is determined. The characteristic values and sum of squared loads of various influencing factors on engineering cost are shown in Table 2.

Table 2: Eigenvalues and sum of load squares of various influencing factors on engineering cost

Assembly	Initial eigenvalue			Extract the sum of squares of the load		
	Total	Variance percentage	Accumulate (%)	Total	Variance percentage	Accumulate (%)
1	8.73	45.9	45.9	8.73	45.9	45.9
2	3.56	18.7	64.6	3.56	18.7	64.6
3	1.85	10.8	75.4	1.85	10.8	75.4
4	1.75	10.3	85.7	1.75	10.3	85.7
5	1.37	8.1	93.8	1.37	8.1	93.8
6	0.31	1.8	95.6			
7	0.25	1.5	97.1			
8	0.19	1.2	98.3			
9	0.17	1.1	99.4			
10	0.09	0.4	99.8			
11	0.04	0.2	100			

As shown in Table 2, the eigenvalues numbered 1-11 were all greater than 0, satisfying PCA requirements. Among the 11 principal components, principal component 1 carried the largest amount of original

information, about 45.9%. The original information carried by the subsequent individual principal components gradually decreases. The cumulative contribution rate of the information carried by the first 5 principal components was about 93.8, which was much greater than 85% and satisfied the set reliability interval, so the first 5 principal components were selected as input indicators for the study. The convergence of PCR-SVM, PCR and SVM-PBF is shown in Figure 7.

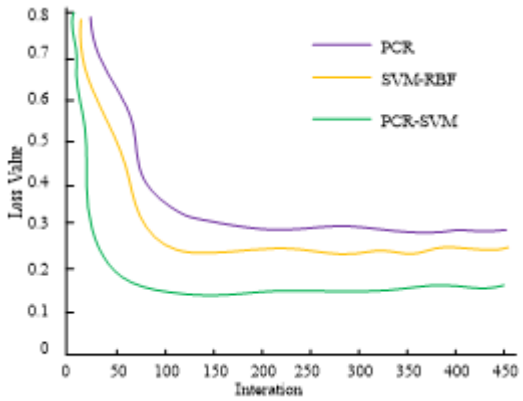
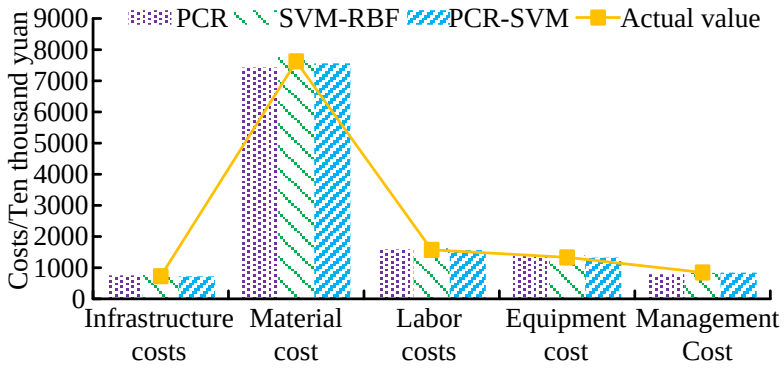
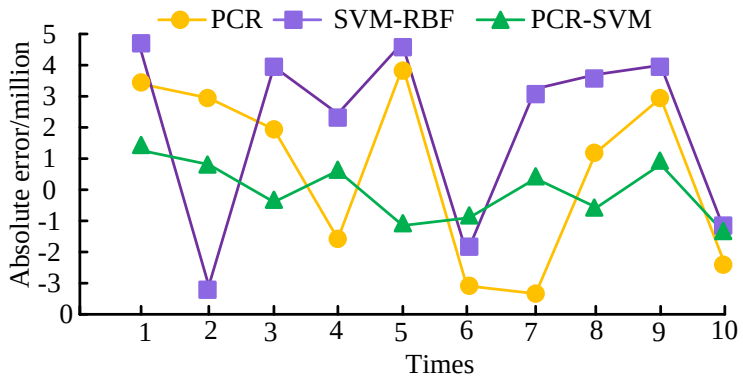


Figure 7: Convergence of PCR-SVM, PCR, and SVM-PBF

In Figure 7, the PCR converged after about 120 iterations, and the loss value was about 0.3. The SVM-RBF model converged after about 100 iterations, and the loss value was about 0.25. The PCR-SVM model converged after about 70 iterations, and the loss value at this time was about 0.16. The convergence of PCR-SVM was better. The prediction results of the three models for the bridge construction cost and the absolute error with the actual total cost are shown in Figure 8.



(a) Prediction Results of Bridge Construction Cost



(b) Absolute error between predicted result and actual total cost

Figure 8: Prediction results and errors of bridge construction cost

From Figure 8(a), the prediction results of PCR for the infrastructure cost, material cost, labor cost, equipment cost and management cost of the bridge were 7.54 million, 74.28 million, 15.89 million, 14.06 million and 7.92 million, respectively, with absolute errors of 280,000, 1.95 million, 160,000, 750,000 and 550,000 from the actual values. The prediction results of SVM-RBF for the five types of costs were 7.67 million, 74.28 million, 15.89 million, 14.06 million and 7.92 million, respectively. The prediction results of SVM-RBF for the five types of costs were 7.67 million, 7.853 million, 16.25 million, 13.97 million, and 8.33 million, with absolute errors of 0.41 million, 2.30 million, 0.52 million, 0.66 million, and 0.14 million from the actual values. PCR-SVM predicted 7.23 million, 75.63 million, 15.64 million, 13.21 million, and 8.47 million for the five costs, with absolute errors of 0.03 million, 0.6 million, 0.09 million, 0.10 million, and 0.07 million from the actual values. In Figure 8(b), the average absolute errors between the prediction results of the three models and the actual total cost were about 2.56 million, 3.32 million and 0.78 million. The prediction results of PCR-SVM had smaller errors with the actual results. The error rates and RMSEs of the three models are shown in Figure 9.

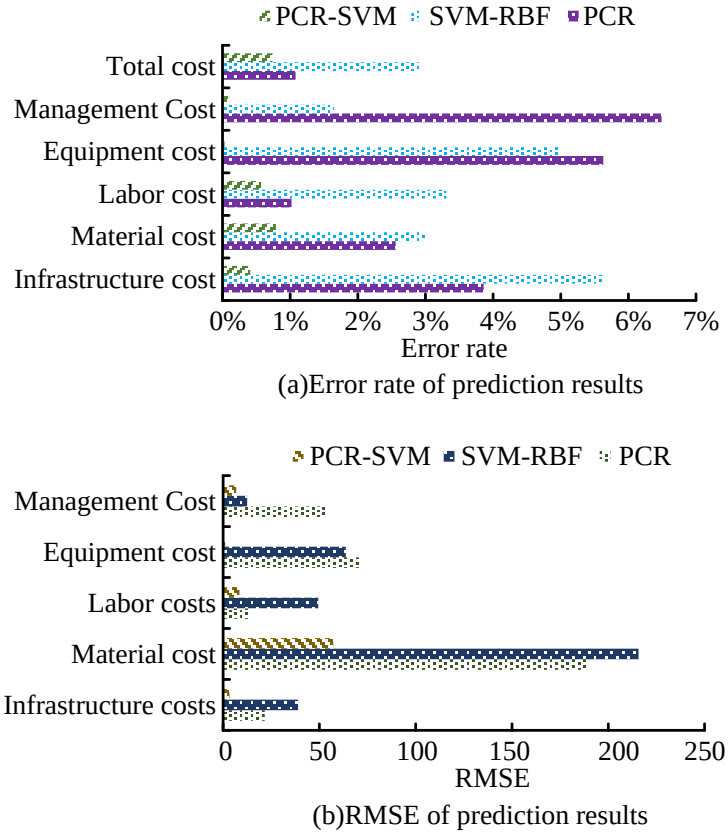
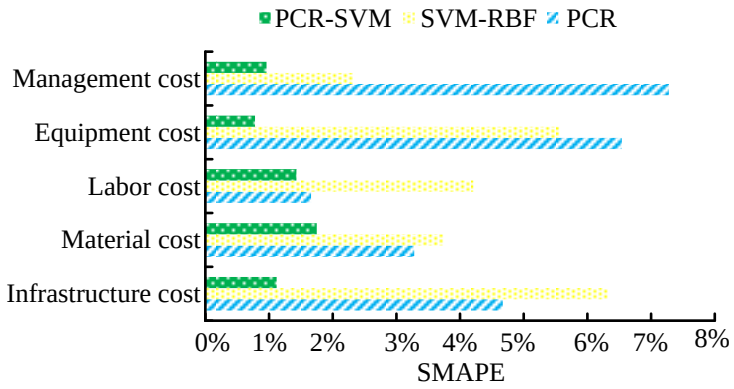
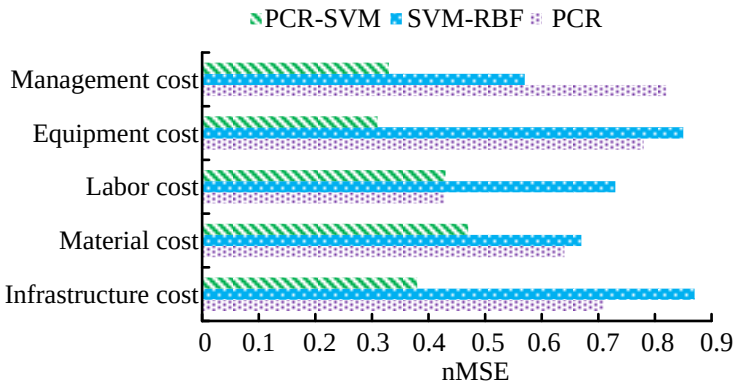


Figure 9: Prediction result error rate and RMSE

From Figure 9(a), the error rates of PCR for the prediction results of infrastructure cost, material cost, labor cost, equipment cost, management cost and total cost were about 3.86%, 2.56%, 1.02%, 5.63%, 6.49% and 1.08%, respectively. The error rates of SVM-RBF for the prediction results of the six costs were about 5.65%, 3.02%, 3.31%, 4.96%, 1.65% and about 2.95%, respectively. The error rates of the prediction results of PCR-SVM were about 0.41%, 0.79%, 0.57%, 0.01%, 0.08% and 0.74%, respectively. The PCR-SVM had the lowest error rate. From Figure 9(b), the RMSE values of the PCR prediction results for the five costs were about 21.6, 188.5, 14.2, 71.3, and 54.4, respectively. The RMSE values of the SVM-RBF were about 38.9, 215.7, 49.4, 63.7, and 12.5, respectively. The RMSE values of the PCR-SVM were about 3.3, 57.2, 8.5, 1.1, and 6.8. The prediction results of PCR-SVM had the least dispersion. The SMAPE and nMSE of the prediction results are shown in Figure 10.



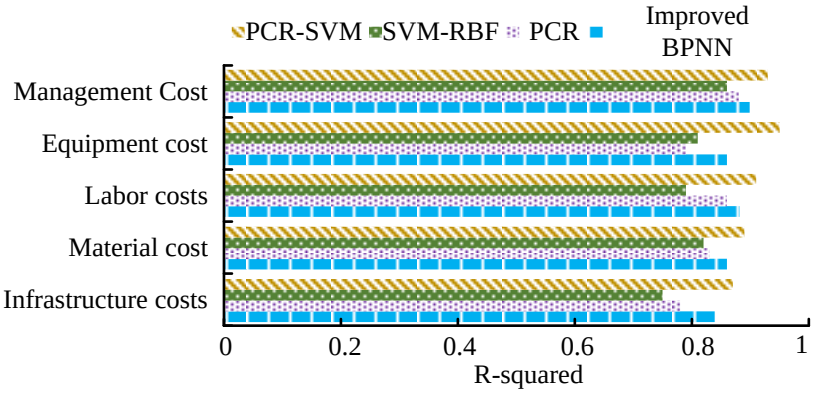
(a) SMAPE of prediction results



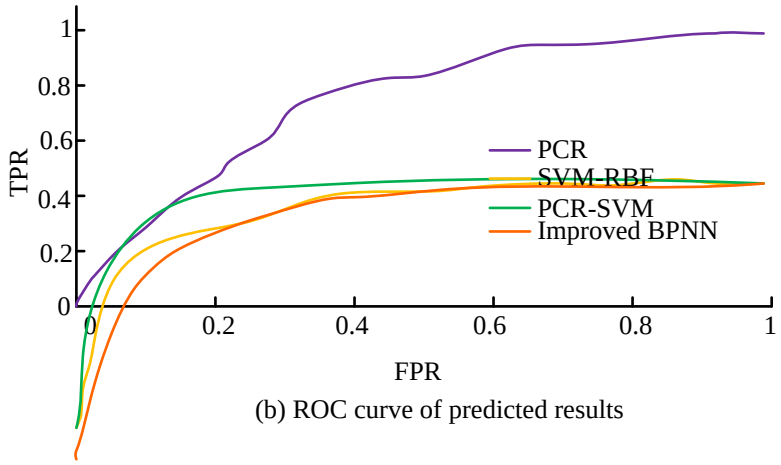
(b) nMSE of prediction results

Figure 10: SMAPE and nMSE of predicted results

In Figure 10(a), SMAPE values of PCR for infrastructure cost, material cost, labor cost, equipment cost and management cost were about 4.67%, 3.28%, 1.66%, 6.54% and 7.28%, respectively. The SMAPE values of the SVM-RBF predicted results for the five costs were about 6.32%, 3.74%, 4.21%, 5.56% and 2.33% respectively. The SMAPE values of the PCR-SVM prediction results were about 1.12%, 1.75%, 1.43%, 0.78% and 0.96%, respectively. The SMAPE values of the PCR-SVM were the smallest. From Figure 10(b), the nMSE values of the PCR for the five costs were approximately 0.71, 0.64, 0.43, 0.78, and 0.82, respectively. The nMSE values of the SVM-RBF were approximately 0.87, 0.67, 0.73, 0.85, and 0.57, respectively. The nMSE values of the prediction results of PCR-SVM were approximately 0.38, 0.47, 0.43, 0.31 and 0.33, respectively. The PCR-SVM had the smallest nMSE value. The R-squared, and ROC curves of the three models are shown in Figure 11.



(a) R-squared of predicted results



(b) ROC curve of predicted results

Figure 11: R-squared and ROC curves of predicted results

From Figure 11(a), the R-squared of PCR predictions for the five costs was about 0.78, 0.83, 0.86, 0.79 and 0.88, respectively. The R-squared of the SVM-RBF for the five costs was approximately 0.75, 0.82, 0.79, 0.81 and 0.86, respectively. The R-squared of the improved BPNN prediction results for the five costs was 0.82,0.84,0.87,0.85 and 0.88, respectively. The R-squared of the PCR-SVM was approximately 0.87, 0.89, 0.91, 0.95 and 0.93. The prediction results of PCR-SVM have a better fit with the actual results. From Figure 11(b), the area under the RCO curve of PCR was approximately 0.78, and the SVM-RBF was about 0.83. The ROC curve of improved BPNN was close to SVM-RBF, about 0.81, and the PCR-SVM was approximately 0.92. The prediction performance of PCR-SVM is better than that of PCR and SVM-RBF models. The percentages of the five costs in the total cost are shown in Figure 12.

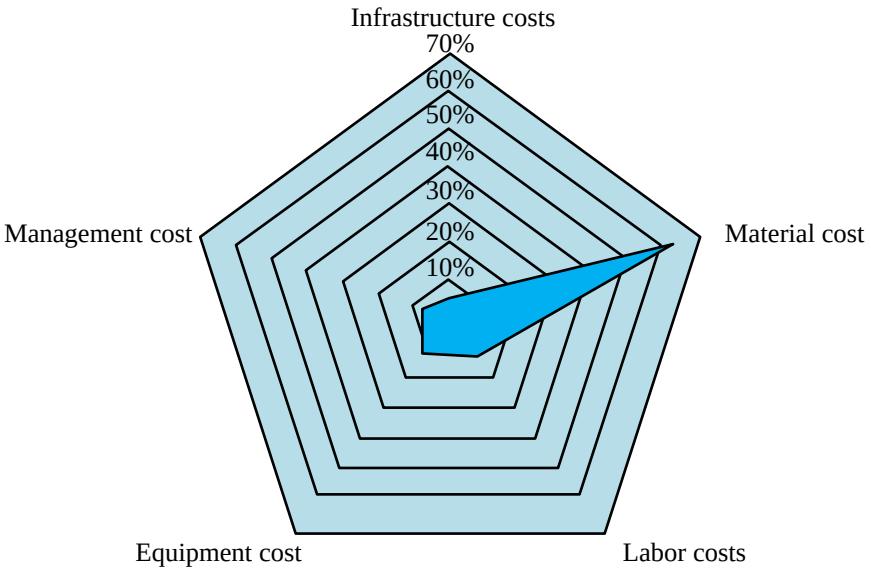


Figure 12: Proportion of various costs in total cost

From Figure 12, the total cost, infrastructure cost, material cost, labor cost, equipment cost, and management cost accounted for 5%, 63%, 13%, 12%, and 7%, respectively. In the total cost prediction, material has the greatest influence on the prediction result, followed by labor cost and equipment. The least influence is the infrastructure cost. From this, to control the project cost budget, it is necessary to start from the above five aspects. In terms of material cost, it is difficult to compress material cost, because it is often determined by the market price of materials. It is possible to minimize some material costs by thoroughly investigating market prices. In terms of labor cost, they are often determined by the supply and demand in the employment market. It is very difficult to compress labor cost. Equipment is often recycled, it only needs to consider the depreciation cost. It can be compressed part of the cost. Infrastructure cost generally includes the cost of geological exploration before construction, infrastructure construction, and other work. Geological exploration has a significant impact on the formulation of later construction plans, while infrastructure construction has a significant impact on the safety and lifespan of construction sites. Therefore, these costs are difficult to compress. Management costs usually consist of on-site management costs and salaries of management personnel. The proportion of management personnel salaries is relatively high, as construction sites often have low management efficiency and redundant management personnel. Therefore, streamlining the management team and improving management efficiency can reduce management costs. To further validate the performance of the PCR-SVM model, the study is subjected to cross testing, and the experimental results are shown in Table 3.

Table 3: Cross experimental results

Nth experiment	Training Set 1/Billion	Training Set 2/Billion	Training Set 3/Billion	Training Set 4/Billion	Test set/Billion	Mean of training set/Billion	True value/Billion
1	1.121	1.124	1.183	1.120	1.183	1.188	1.196
2	1.147	1.152	1.167	1.159	1.164	1.161	1.153
3	1.099	1.112	1.109	1.111	1.105	1.119	1.124
4	1.183	1.188	1.179	1.174	1.182	1.176	1.171
5	1.124	1.131	1.128	1.122	1.129	1.127	1.122
6	1.195	1.212	1.206	1.213	1.199	1.208	1.211
7	1.218	1.211	1.221	1.217	1.215	1.222	1.234
8	1.176	1.169	1.164	1.171	1.172	1.673	1.159
9	1.249	1.244	1.251	1.249	1.256	1.248	1.243
10	1.176	1.177	1.182	1.179	1.177	1.181	1.186
11	1.173	1.175	1.167	1.162	1.168	1.166	1.172
12	1.215	1.209	1.214	1.213	1.211	1.215	1.201

From Table 3, the relative error of the PCR-SVM prediction result did not exceed 7%, so the calculated result was as high as 0.97. The above results showed that the PCR-SVM model had good robustness. In addition, in order to explore the influence of each input factor on the cost prediction, the study conducted a sensitivity analysis. The analysis results are shown in Table 4.

Table 4: Results of the sensitivity analysis

Investment factors	Weight ratio	Predicting sensitivity
Infrastructure cost	0.06	0.25
Material cost	0.63	2.41
Labor costs	0.13	1.02
Equipment cost	0.11	0.77
Management cost	0.07	0.21

From Table 4, among the input factors, the management cost and infrastructure cost had little impact on the total cost, while materials and labor had a significant impact on the cost prediction.

5 Discussion

In order to achieve accurate cost prediction of roads and bridges, a cost prediction model based on PCR-SVM was proposed. The prediction error rate of PCR-SVM for total cost was only 0.74%, which was lower than other models. Compared with the research by Wang X et al. [4], the prediction accuracy of the PCR-SVM model was higher. This is because the significance and independence of the feature factors are tested to meet the requirements, and the principal component loading matrix is used to calculate the principal component matrix, allowing the algorithm to select appropriate dependent and independent variables. In addition, PCR-SVM can handle both linear and nonlinear factors simultaneously. Compared with the research of Fan M and Sharma A [6], the computational efficiency of the PCR-SVM model was higher. This is because the research combines PCA and multiple linear regression to construct the PCR model. PCA can convert multiple variables into a small number of comprehensive indicators, effectively reducing the number of variables and decreasing computational complexity. At the same time, the PCR-SVM model can effectively manage both linear and nonlinear factors. For KOMs greater than 0.5, they are considered linear factors and analyzed using PCR. Otherwise, they are considered nonlinear factors and analyzed using SVM. According to the weight analysis of influencing factors, the proportions of infrastructure cost, material cost, labor cost, equipment cost, and management cost were 5%, 63%, 13%, 12%, and 7%, respectively. The material factor had the greatest impact on the total cost. This is because personnel and equipment are managed by construction companies, and their costs and quantities are generally stable. However, the required amount and loss of materials are relatively large, and they need to be purchased in the market with significant price fluctuations. Therefore, the prediction results of road and bridge construction costs are greatly affected.

6 Conclusion

Before the road and bridge construction, the project cost forecast is needed to ensure reasonable budget use. However, during the calculation, many factors can affect the prediction results. Therefore, a project cost prediction model-PCR-SVM model by PCA, multiple linear regression, and SVM was established. The PCR-SVM model showed that infrastructure cost, material cost, labor cost, equipment cost and management cost accounted for 5%, 63%, 13%, 12% and 7%, respectively, with material having the greatest impact. In addition, to test the prediction performance of PCR-SVM, the study selected PCR and SVM-RBF models as control groups. The test results showed that the PCR and SVM-RBF models converged after about 120 and 100 iterations, respectively, and the loss values were about 0.3 and 0.25 at this time. PCR-SVM converged after about 70 iterations, and the loss value was about 0.16 at this time. It is evident that the convergence of PCR-SVM is better than that of PCR and SVM-RBF models. The lowest error rates of the three models for the prediction results of infrastructure cost, material cost, labor cost, equipment cost, management cost and total cost were about 1.02%, 1.65% and 0.01%, respectively, among which the PCR-SVM model had the smallest error rate. The lowest SMAPE values of the prediction results of the three models were about 1.66%, 2.33% and 0.78%, respectively. The lowest nMSE values were about 0.43, 0.57 and 0.31, respectively. The SMAPE and nMSE values of the PCR-SVM were the smallest. The minimum RMSE values of the prediction results of the three models were about 14.2, 12.5 and 1.1. The prediction results of PCR-SVM had the least dispersion. The area under the RCO curve of PCR, SVM-RBF and PCR-SVM models were about 0.78, 0.83 and 0.92. The prediction performance of PCR-SVM was better than that of PCR and SVM-RBF models. The above results indicate that the PCR-SVM model can accurately predict engineering costs and accurately reflect the impact of major influencing factors on the prediction results.

References

[1] N. Akiba, A. Nakamura, T. Sota, K. Hibino, H. Kakuda, and M. Aalders, "Separation of overlapping fingerprints by principal component analysis and multivariate curve resolution–alternating least squares analysis of hyperspectral imaging data," Journal of Forensic Sciences, vol. 67, no. 3, pp. 1208-1214, 2022. https://doi.org/10.1111/1556-4029.14969

[2] Y. Dai, Y. Yu, X. Wang, Z. Jiang, Y. Chen, K. Chu, and Z. J. Smith, "Hybrid principal component analysis denoising enables rapid, label-free morpho-chemical quantification of individual nanoliposomes," Analytical Chemistry, vol. 94, no. 41, pp. 14232-14241, 2022. https://doi.org/10.1021/acs.analchem.2c02518

[3] A. Raj, J. P. Misra, and D. Khanduja, "Modeling of wire electro-spark machining of inconel 690 superalloy using support vector machine and random forest regression approaches," Journal of Advanced Manufacturing Systems, vol. 21, no. 3, pp. 557-571, 2022. https://doi.org/10.1142/S0219686722500196

[4] X. Wang, S. Liu, and L. Zhang, "Highway cost prediction based on LSSVM optimized by intial parameters," Computer Systems Science and Engineering, vol. 36, no. 1, pp. 259-269, 2021.

[5] V. Sharma, M. Zaki, K. N. Jha, and N. Krishnan, "Machine learning-aided cost prediction and optimization in construction operations," Engineering Construction and Architectural Management, vol. 29, no. 3, pp. 1241-1257, 2022. https://doi.org/10.1108/ECAM-10-2020-0778

[6] M. Fan, and A. Sharma, "Design and implementation of construction cost prediction model based on SVM and LSSVM in industries 4.0," International Journal of Intelligent Computing and Cybernetics, vol. 14, no. 2, pp. 145-157, 2021. https://doi.org/10.1108/IJICC-10-2020-0142

[7] M. Priti, and A. A. Salunkhe, "Comparative analysis of construction cost estimation using artificial neural networks," Xi'an Dianzi Keji Daxue Xuebao/Journal of Xidian University, vol. 14, no. 7, pp. 1287-1305, 2020. https://doi.org/10.37896/jxu14.7/146

[8] M. Ashour, A. A. Ahmed, and E. A. Aldahhan, "Using modified earned value for cost control in construction projects," Periodicals of Engineering and Natural Sciences (PEN), vol. 8, no. 1, pp. 156-168, 2020. https://doi.org/10.21533/pen.v8i1.1103

[9] K. Croteau, A. C. Ryan, R. Santore, D. Deforest, C. Schlekat, E. Middleton, and E. Garman, "Comparison of multiple linear regression and biotic ligand models to predict the toxicity of nickel to aquatic freshwater organisms," Environmental Toxicology and Chemistry, vol. 40, no. 8, pp. 21189-2205, 2021. https://doi.org/10.1002/etc.5063

[10] K. K. Zi, Z. Wan, and K. A. Kurnia, "Prediction of ionic conductivity of imidazolium-based ionic liquids at different temperatures using multiple linear regression and support vector machine algorithms," New Journal of Chemistry, vol. 45, no. 39, pp. 18584-18597, 2021. https://doi.org/10.1039/D1NJ01831K

[11] A. M. C. H. Attanayake, S. S. N. Perera, and U. P. Liyanage, "Multiple linear regression models on interval-valued dengue data with interval-valued climatic variables," International Journal of Applied Mathematics and Statistics, vol. 59, no. 3, pp. 49-60, 2020.

[12] A. Mansor, S. Abdullah, N. C. Dom, N. Napi, A. N. Ahmed, M. Ismail, and M. F. R. Zulkifli, "Three-hour-ahead of multiple linear regression (MLR) models for particulate matter (PM_(10)) forecasting," International Journal of Design and Nature and Ecodynamics, vol. 16, no. 1, pp. 53-59, 2021. https://doi.org/10.18280/ijdne.160107

[13] S. S. Hashemi, K. Sadeghi, S. Javidi, and M. Malakooti, "A parametric shear constitutive law for reinforced concrete deep beams based on multiple linear regression model," Advances in Structural Engineering, vol. 8, no. 4. pp. 285-294, 2019. https://doi.org/10.12989/acc.2019.8.4.285

[14] T. T. Cai, H. Li, and S. Li, "Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 84, no. 1, pp. 149-173, 2022. https://doi.org/10.48550/arXiv.2006.10593

[15] S. K. Damarla, X. Sun, and F. Xu, "Practical linear regression-based method for detection and quantification of stiction in control valves," Industrial and Engineering Chemistry Research, vol. 61, no. 1, pp. 502-514, 2022. https://doi.org/10.1021/acs.iecr.1c02723

[16] K. Ghoulem, T. Kormi, and N. Ali, "Damage detection in nonlinear civil structures using kernel principal component analysis," Advances in Structural Engineering, vol. 23, no. 11, pp. 2414-2430, 2020. https://doi.org/10.1177/1369433220913207

[17] X. Dong, T. Zhong, and Y. Li, "New suppression technology for low-frequency noise in desert region: the improved robust principal component analysis based on prediction of neural network," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 7, pp. 4680-4690, 2020. https://doi.org/10.1109/TGRS.2020.2966054

[18] R. Rangayya, V. Virupakshappa, and N. Patil, "An enhanced segmentation technique and improved support vector machine classifier for facial image recognition," International Journal of Intelligent Computing and Cybernetics, vol. 15, no. 2, pp. 302-317, 2022.

https://doi.org/10.1108/IJICC-08-2021-0172
[19] Y. Guo, Z. Mustafaoglu, and D. Koundal, “Spam detection using bidirectional transformers and machine learning classifier algorithms,” Journal of Computational and Cognitive Engineering, vol. 2, no. 1, pp. 5-9, 2022. https://doi.org/10.47852/bonviewJCCE2202192
[20] W. Zhang, Z. Wu, and D. W. Bunn, “Optimal hybrid framework for carbon price forecasting using time series analysis and least squares support vector machine,” Journal of Forecasting, vol. 41, no. 3, pp. 615-632, 2022. https://doi.org/10.1002/for.2831