Research on overhead line Engineering Cost prediction based on PCA-LSSVM model

Shen Yabo¹, Ma Ning¹, Men Yan¹, Zhao Xin¹, Jia Zhenghao^{1*}

¹ Economy and Technology Research Institute, State grid Xin Jiang Electric Power Corporation, Urumqi, Xinjiang, 830011, China

Abstract. In recent years, the investment projects of overhead line engineering increase year by year. Establishing scientific cost prediction concept and optimizing cost prediction method can improve the investment utilization efficiency. Based on the actual cost data of 110kV overhead line project, this paper extracts the principal component factor through principal component analysis and eliminates the correlation between the original indicators. Then, the training sample is input into the least-squares support vector machine model to build a learning network. Finally, the predicted value of the model is compared with the actual cost level for analysis. The prediction results show that the average error rate is less than 5%, indicating that the PA-LSSVM model constructed in this paper can effectively predict the overhead line engineering cost.

1 Introduction

The cost level prediction of overhead line engineering is a complex process with multiple variables and nonlinearity, and its cost influencing factors mainly include line length, wire quantity, single wire area, tower base, tower material quantity, tower material price, altitude, earth and stone quantity, base steel quantity, base steel price, etc. [1]. Because there are many influencing factors of project cost, quantitative prediction method is more applied to overhead project cost prediction. Based on the principle of structural risk minimization, SUPPORT vector machine has better generalization ability. Moreover, its optimal solution is based on limited sample information, which is more suitable for cost prediction in the case of small samples. Moreover, the training time of support vector machine model is shorter, which reduces the empirical prediction component [2]. Therefore, based on the main points analysis (PCA) and improve the support vector machine (SVM) - least squares support vector machine (LSSVM) hybrid algorithm, this paper establishes the overhead line construction cost prediction model through feature extraction and small sample research, so as to achieve the goal of faster prediction of overhead line construction cost and guidance of construction project cost management.

2 Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction method of data. Through feature extraction, fewer comprehensive independent indicators are obtained, and the original information is not lost to the greatest extent. The advantage of principal component analysis is that multiple variables originally linearly related are transformed into several independent variables, which not only achieves the purpose of dimension reduction but also retains most of the information of the original variables [3].

3 Least squares support vector machines

Support vector machine (SVM) is a new machine learning method proposed by Vapnik et al. in 1995 based on statistical learning theory. It replaces the principle of minimizing empirical risk in traditional statistical learning by seeking structural risk Minimize to improve the generalization ability of the learning machine, so that in the case of small samples, it can also achieve better prediction accuracy [4]. In 1999, Suykens J.A.K introduced the least squares linear system into the support vector machine, forming the theory of least squares support vector machine. LSSVM uses the quadratic loss function to replace the insensitive loss function in SVM, and changes the inequality constraint conditions into equality constraints, thereby further improving the learning accuracy [5].

*Corresponding author's e-mail: 2651308061@qq.com

Assuming training sample set $S = \{x_i, y_i\}, i = 1, 2, \dots, n$, $x_i, y_i \in R^n$, The optimization problem of LSSVM can be expressed as:

$$\begin{cases} \min_{w, b, e} J_p(\omega, e) = \frac{1}{2} \omega^T \omega + \gamma \frac{1}{2} n \sum_{i=1}^n e_k^2 \\ s.t. \quad y_i = \omega^T \varphi(x_i) + b + e_i, i = 1, 2, \dots, n \end{cases} \quad (1)$$

In this formula, ω is the weight vector; γ is the regularization parameter; b is the deviation value; $\varphi(\cdot)$ is the kernel function; $e_i (i = 1, 2, \dots, n)$ is the n sample point error variable[6].

Because it is difficult to solve the original problem, the Lagrange function is used to transform the original problem into a dual problem, so as to optimize the dual problem. And by introducing a kernel function, this paper maps the sample in the original space to a vector in the high-dimensional feature space to solve the linear inseparable problem in the original space. There are many settings for kernel functions, and radial basis kernel functions are used in this article[7].

4 Overhead line project cost prediction model

4.1 Principal component analysis

This article takes the data of 46 110kV overhead lines completed in 2019 in an area as an example, and collects 16 indicators in the project construction to verify the established cost prediction model. In the model, the unit length cost index is used as the output dependent variable, and the total length of the line (folded order), the number of loops, the amount of tower material, the price of tower material, the ratio of tensile strength, the area of a single wire, the amount of wire, the price of wire, the amount of steel, A total of 15 indicators, such as steel price, topography distribution, geological conditions, earthwork volume, total amount of foundation concrete and altitude are used as input independent variables of the model.

4.2 Cost forecast indicator data

Combined with the actual situation of overhead line engineering cost, this paper established a cost prediction model based on principal component analysis and least squares support vector machine. After pre-processing the collected data samples of 46 impact line projects on 15 impact factors according to 4.1, this paper uses SPSS software to extract the principal components of many impact factors to achieve the purpose of dimensionality reduction and index correlation reduction, Table 1 Interpretation table of total variance obtained by principal component analysis.

Table 1. Total variance interpretation.

ingredient	Initial eigenvalue			extracting the sum of squared loads		
	total	percentage of variance	cumulative %	total	percentage of variance	cumulative %
1	5.657	45.738	45.738	5.657	45.738	45.738
2	1.654	16.146	61.884	1.654	16.146	61.884
3	1.522	13.54	75.424	1.522	13.54	75.424
4	1.131	7.134	82.558	1.131	7.134	82.558
5	1.076	4.075	86.633	1.076	4.075	86.633
6	0.924	3.518	90.151			
7	0.807	2.582	92.733			
8	0.754	2.001	94.734			
9	0.513	1.905	96.639			
10	0.36	1.102	97.741			
11	0.267	0.843	98.584			
12	0.145	0.645	99.229			
13	0.121	0.474	99.703			
14	0.068	0.256	99.959			
15	0.001	0.041	100			

In actual engineering applications, when the cumulative contribution rate is generally set to be greater than or equal to 85%, it can be considered that the extracted principal component factor can already replace the original index data. According to Table 1, the cumulative contribution rate of the first five factors has

reached 86.633%, and the eigenvalues are all greater than 1, which has reached the requirement of principal component extraction. Therefore, the first five factors are extracted as principal component factors. After the main component factors are extracted, the impact factor data matrix and the component score coefficient matrix are

synthesized to obtain the scores of each factor, which are used as the final training set and test set of the least square support vector machine model in the next stage to

predict the cost of the overhead line project. As shown in table 2.

Table 2. Component score coefficient.

	Component				
	1	2	3	4	5
line folding length	.168	.035	.029	-.076	-.048
number of loops	.048	-.190	.397	.394	.008
tower material	.136	-.080	-.088	-.012	.030
tower price	.022	.270	.381	.412	-.118
tensile ratio	-.128	-.100	-.015	-.110	.222
single conductor area	-.013	-.257	-.063	.354	.677
wire amount	.168	.027	.027	-.068	-.030
altitude	-.034	.219	.334	-.395	.256
total amount of foundation concrete	.158	-.077	-.146	-.039	.168
amount of earthwork	.154	.009	-.038	-.209	-.024
geological conditions	-.020	.284	-.388	.309	-.015
terrain distribution	.066	.351	-.067	.272	.035
steel price	.051	.249	.157	-.186	.472
amount of steel	.157	-.130	-.009	.072	.079
wire price	.017	-.272	.165	-.018	-.281

4.3 LSSVM cost prediction model

In this paper, the LSSVM toolkit is added to the MATLAB software, and the training program is written to predict the cost. According to 4.2, the first five principal component factors are taken as input variables, and the cost per unit length is used as output variable, and then 36 projects are randomly selected as training sets from 46 overhead line projects and input into the

least squares support vector machine model. After training and learning, the remaining 10 overhead line projects are used as the test set, and the predicted results are tested according to the trained model, and compared with the actual unit length cost to calculate the model error rate. The fitting situation of the training samples is shown in Figure 1, and the prediction results and error rates of the test samples are shown in Table 3.

Figure 1. Training sample fitting diagram.

Table 3. Test error table.

sample	1	2	3	4	5	6	7	8	9	10
The actual value	74.48	88.85	34.53	72.47	52.89	92.5	97.54	38.98	74.41	91.95
Predictive value	73.15	85.52	35.64	69.21	53.21	96.17	93.51	40.85	75.05	89.29
Error rate (%)	1.79	3.75	3.21	4.50	0.61	3.97	4.13	4.80	0.86	2.89

According to Table 3, the transmission and transformation project cost predicted by the model is basically consistent with the actual project cost, and the error rate of the 10 test samples is all below 5%, which meets the 5%~10% error rate allowed in the project budget estimate And the requirement of 3%~5% error rate allowed in the project budget. Therefore, the cost prediction model based on PCA-LSSVM established in this paper has high accuracy, which can accurately predict the cost of overhead line engineering.

5 Conclusion

Aiming at the problems that there are many factors influencing the overhead line project cost and the difficulty of prediction, this paper builds a PCA-LSSVM-based overhead line project cost prediction model, and uses 46 overhead line project cost index data in a certain area to test and verify the constructed model. The results show that the error rate between the predicted value and the actual value of the 10 test samples is below 5%, and the prediction accuracy is high. Therefore, the cost prediction model based on PCA-LSSVM established in this paper is scientific and effective, and can accurately predict the cost in the case of a small sample, which has certain reference significance for the actual budget work of overhead line engineering and cost control.

References

1. Geng P Y, AN L, WANG X. Establishment and Implementation of Transmission Engineering Cost Prediction Model based on Data Mining Technology [J]. Modern Electronics Technology, 2008,41(4):157-160.
2. Dong zhen, wang yanqin, wang yong. Prediction of power grid technical reconstruction cost based on DEA and genetic BP neural network [J]. Hebei electric power technology,2020,39(1):25-29.
3. Li wang, Wang mianbin, Rao rao. Research on cost prediction index system and model of 220kv substation project in settlement stage [J]. Journal of electric power,2014,29(2):145-148.
4. He yongping, Gu xin, Yong hao, Gou ruixin, Wei dongni. Prediction model establishment and analysis of power transmission and transformation cost weight based on pso-svr [J]. Automation technology and application,2020,39(03):98-102.
5. Wang Xin, Geng Pengyun, YUAN Jingzhong. Research on overhead Power Line Engineering Cost Prediction Model [J]. China Power Enterprise Management,2018(24):76-77.
6. Yu, B., Xiao, Y., Liu, S., Liu, X. Cost Estimation Analysis of Transmission and Distribution Engineering Based on PSO-ELM Algorithm[J].Information Technology,2019,43(04):148-151+156.
7. Xu, L. Research on engineering cost prediction model of uhv transmission lines -based on factor analysis and BP neural network [J]. Industrial technology economics,2017,36(07):18-26.