Transmission line engineering cost prediction based on principal component analysis and least square support vector machine

Bo Yu^{1*}, Zheng Wang¹, Xiaomin Liu¹, Tong Liu² and Xinyi Liu¹

¹Economic and Technological Research Institute, State Grid Ningxia Electric Power Co., Ltd., Huyingkangchen No. 1 office building, Baohu Middle Road, Yinchuan City, Ningxia, 750000, China

²Power Construction Technology and Economic Consultation Center, China Electric Power Enterprise Association, No. 13 Baiguang Road, Xicheng District, Beijing, 100053, China

Abstract. Due to the many factors affecting the cost of transmission line engineering and the lack of mutual independence, it is difficult to predict the cost. Firstly, the principal component analysis is used to process the original indicator data, eliminating the correlation between the original indicators and extracting the potential comprehensive independent indicators. Then, the new indicator is used as the input set to construct the predictive learning model based on the least squares support vector machine, and the predicted output and the actual value are compared and analyzed. The results show that the model can achieve the desired prediction effect in the case of small samples.

1 Introduction

The cost level of transmission line engineering is a multi-variable and highly nonlinear problem, which is manifested in the fact that there are many factors affecting the cost of transmission and transformation engineering, and they are not independent of each other, but form a complex relationship. It has made it more difficult to predict the cost of power transmission and transformation projects. Traditionally used cost prediction methods include fuzzy mathematics, analog engineering, etc., but these methods are cumbersome and the prediction accuracy can not meet the demand of reasonable cost prediction[1]. With the application of some mathematical methods and intelligent algorithms in forecasting, many new prediction methods have emerged, including classical prediction methods represented by trend analysis and regression analysis, and intelligent prediction methods represented by neural networks and support vector machines[2]. The classical prediction method is easy to understand and easy to operate, but it has great limitations in dealing with nonlinear problems. The intelligent prediction method has great advantages in dealing with nonlinear problems, and the prediction accuracy is greatly improved compared with the classical prediction method, but it is also prone to over-fitting, falling into local optimum and so on[3]. In this paper, principal component analysis is firstly adopted for factor dimensionality reduction, and then the least square support vector machine is used for cost prediction.

2 Principal Component Analysis

Principal Component Analysis (PCA) is a dimension reduction algorithm commonly used in data mining. It was first proposed in Pearson's research on non-random variables in 1901, and later Hotelling extended this method from the case of non-random variables to the case of random vectors in 1933[4]. Principal Component Analysis is a multivariate statistical method that uses the idea of dimensionality reduction to convert multiple indicators into several comprehensive indicators under the premise of losing little information. Usually, the comprehensive index generated by transformation is called the principal component, and each of the main components are linear combinations of the original variables, and the main components are not related to each other, so that the principal components have some superior performance than the original variables.

The basic principle of principal component analysis is to regroup a set of variables $X_1, X_2, \dots, X_p$ with certain correlation into a set of new variables $U_1, U_2, \dots, U_m$ that are independent of each other. Among them, $m < p$, $Cov(U_r, U_t) = 0$ $r, t \in (1, 2, \dots, m)$. Moreover, m new variables $U_j (j=1, 2, \dots, m)$ are required to represent most of the variation information of p original variables $X_i (i=1, 2, \dots, p)$. Usually the mathematical process is to linearly combine the original p variables, that is to find a_{ij} which can make

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

$U_j = a_{1j}X_1 + a_{2j}X_2 + \dots + a_{pj}X_p (j=1,2,\dots,m)$ meet the above requirements.

3 Least Squares Support Vector Machine

Support Vector Machine (SVM) is a new algorithm based on statistical machine learning theory proposed by Vapnik et al. in 1995. It has a wide range of applications in feature extraction, pattern recognition and regression analysis. The support vector machine improves the generalization ability of the learning machine by seeking the minimum structural risk rather than the empirical risk in the traditional statistical learning, and finally achieves the minimum empirical risk and confidence range[5]. The significant advantage of the support vector machine is that it can better adapt to the small sample situation, and has a good treatment for the over-learning problem. Suykens J.A.K proposed the Least Squares Support Vector Machine Theory (LSSVM) based on SVM in 1999. The significant difference between it and SVM is that LSSVM uses another optimization objective function to replace the SVM inequality constraint by least squares method. The equality constraint further improves the accuracy of learning[6].

Assuming training sample set $S = \{x_i, y_i\}, i=1,2,\dots,n$, $x_i, y_i \in R^n$, among them, x_i is the input of training samples, y_i is the output of training samples, and n is the number of training samples. The optimization problem of LSSVM is as follows:

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

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

The original problem can be transformed into a dual problem through Lagrange duality, which makes the problem easy to solve. At the same time, the Lagrange function is constructed for the purpose of extending to nonlinear classification problems and then introducing kernel functions[7]:

$$L(\omega, b, e, \alpha) = J_p(\omega, e) - \sum_{i=1}^n \alpha_i \{y_i [\omega^T \varphi(x_i) + b] - 1 + e_i\} \quad (2)$$

The optimal solution of equation (2) is as follows:

$$\begin{cases} \frac{\partial L}{\partial \omega} = 0 \rightarrow \omega = \sum_{i=1}^n \alpha_i y_i \varphi(x_i) \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^n \alpha_i y_i = 0 \\ \frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i, i=1,2,\dots,n \\ \frac{\partial L}{\partial \alpha_i} = 0 \rightarrow y_i [\omega^T \varphi(x_i) + b] - 1 + e_i = 0 \end{cases} \quad (3)$$

Eliminate ω , e and then get a set of linear equations about α , b :

$$\begin{bmatrix} 0 & y^T \\ y & \Omega + I/\gamma \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (4)$$

In this formula, $\Omega_{ij} = y_i y_j \times \varphi(x_i)^T \varphi(x_j)$, $i, j=1,2,\dots,n$; I is an n -order unit matrix. Then the regression function of the least squares support vector machine can be expressed as:

$$y_k = \sum_{k=1}^n \alpha_k \varphi(x_i)^T \varphi(x_j) + b \quad (5)$$

4 Transmission line engineering cost prediction model

Combined with the actual situation of transmission line engineering cost, this paper establishes a cost prediction model based on principal component analysis and least squares support vector machine. Firstly, the raw data index is extracted into the principal component to achieve the purpose of dimension reduction and correlation reduction. Then the principal component is used as a new input indicator set, and the sample data is divided into two parts, one is used as training sample and the other is used as test sample. First use the training samples to let LSSVM learn autonomously, then input the test samples into the training network, get the predicted output, and compare with the actual output to verify the validity of the model.

4.1 Sample indicator data selection

This paper selects the cost data of 110kV overhead lines whose number is 46 in a certain area in 2018 as sample data to verify the cost prediction model. Because of the large number of index data, it is necessary to firstly screen and simplify the index data. 16 index data such as line folding length, tower material quantity and unit length cost were selected, in which the unit length cost was used as the dependent variable, and the other 15 indicators were the influencing factors, as the independent variables, as shown in Table 1. Among the 15 influencing factors obtained above, the number of line loops, topographical distribution and geological conditions are descriptive variables, so they need to be quantified. The principle is to classify them by numbers 1, 2, 3, etc., the weighted averaging process is then performed on an overall scale.

Table 1. Overhead line engineering cost index.

serial number	indicators	unit	serial number	indicators	unit
1	line folding length	m	9	amount of steel	t
2	number of loops	/	10	steel price	yuan / t
3	tower material	t	11	terrain distribution	/
4	tower price	ten thousand yuan / t	12	geological conditions	/
5	tensile ratio	/	13	amount of earthwork	m ³
6	single conductor area	mm ²	14	total amount of foundation concrete	m ³
7	wire amount	t	15	altitude	m
8	wire price	yuan / t	16	unit length cost	ten thousand yuan / km

4.2 Principal component extraction

After obtaining data samples of 15 influencing factors in 46 projects, principal component analysis of the influencing factors can be performed using SPSS

software. When the correlation test is passed, the principal component extraction can be performed, and the number of factors is generally determined by a combination of the feature value and the stone diagram.

Table 2. 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	48.738	48.738	5.657	48.738	48.738
2	1.654	14.147	62.885	1.654	14.147	62.885
3	1.522	10.54	73.425	1.522	10.54	73.425
4	1.131	7.143	80.568	1.131	7.143	80.568
5	1.076	4.575	85.143	1.076	4.575	85.143
6	0.924	3.818	88.961			
7	0.807	2.682	91.643			
8	0.754	2.426	94.069			
9	0.513	1.918	95.987			
10	0.36	1.301	97.288			
11	0.267	0.983	98.271			
12	0.145	0.764	99.035			
13	0.121	0.604	99.639			
14	0.068	0.456	99.995			
15	0.001	0.005	100

Figure 1. Gravel map.

From the total variance interpretation table, the eigenvalues of the first five factors are greater than 1,

and when the five factors are extracted, the cumulative contribution rate has reached 85.143%, which explains the variance content of more than 85%, so the required extraction can be determined. The number of principal factors is five, and the same conclusion can be obtained intuitively through Figure 1. After determining the principal component factor, the original influencing factor data matrix and the component score coefficient matrix (Table 3) are combined to calculate the factor score value, which is used as the final training and test data of the next stage overhead line engineering cost prediction model.

Table 3. Component score coefficient.

component	1	2	3	4	5
line folding length	0.168	0.035	0.029	-0.076	-0.048
number of loops	0.048	-0.190	0.397	0.394	0.008
tower material	0.136	-0.080	-0.088	-0.012	0.030
tower price	0.022	0.270	0.381	0.412	-0.118
tensile ratio	-0.128	-0.100	-0.015	-0.110	0.222
single conductor area	-0.013	-0.257	-0.063	0.354	0.677
wire amount	0.168	0.027	0.027	-0.068	-0.030
wire price	-0.034	0.219	0.334	-0.395	0.256
amount of steel	0.158	-0.077	-0.146	-0.039	0.168
steel price	0.154	0.009	-0.038	-0.209	-0.024
terrain distribution	-0.020	0.284	-0.388	0.309	-0.015
geological conditions	0.066	0.351	-0.067	0.272	0.035
amount of earthwork	0.051	0.249	0.157	-0.186	0.472
total amount of foundation concrete	0.157	-0.130	-0.009	0.072	0.079
altitude	0.017	-0.272	0.165	-0.018	-0.281

4.3 LSSVM prediction

This article uses the LSSVM toolkit added to the MATLAB software to write a program for cost prediction. First, the sample data obtained in 3.2 is divided into two groups. The first 36 sample projects are grouped as a training set, and the last 10 sample projects are grouped as a test set. Then, the training sample data is input into the least squares support vector machine model for training learning, and the training sample fitting situation is shown in the figure 2. Finally, the test sample data is used as an input to test and obtain the predicted result. Compared with the actual value, the

predicted error rate is calculated, as shown in the table 4 and figure 3.

Figure 2. Training sample fitting diagram.**Table 4.** Test error table.

sample	1	2	3	4	5	6	7	8	9	10
The actual value	38.53	85.47	54.89	74.41	91.95	99.50	88.85	97.54	38.98	74.48
Predictive value	36.64	88.21	53.21	75.05	89.29	96.17	85.52	93.51	40.85	73.15
Error rate	4.90%	3.21%	3.06%	0.85%	2.89%	3.35%	3.75%	4.13%	4.81%	1.79%

Figure 3. Test error figure.

As can be seen from table 4, the deviation rate between the predicted cost and the actual cost of the 10 projects in this cost prediction is less than 5%, and the relative average absolute error is 4.03%. The deviation

between the estimated cost and the actual project settlement is allowed to be 5%~10%, and the deviation between the project budget and the actual project settlement is allowed to be 3%~5%. Therefore, the prediction results show that the project cost prediction model based on principal component analysis and least square support vector machine has high accuracy in the project cost prediction of overhead lines.

5 Conclusion

In this paper, the principal component analysis and least squares support vector machine are combined to construct the transmission line engineering cost prediction model, and the model is trained and tested by using the 110kV overhead line engineering cost index data in a certain area. The training results are better and

the prediction results are obtained. The difference between the actual and the actual is only 4.03%, which indicates that the cost prediction model established in this paper is scientific and effective, and the prediction accuracy is high. In the case of small samples, the cost of the power transmission and transformation project can be accurately predicted, which has certain reference to the project budget. So this model has a good application prospect.

References

1. Yu, B., Xiao, Y., Liu, S. Cost Estimation Analysis of Transmission and Distribution Engineering Based on PSO-ELM Algorithm[J]. Information Technology, 2019, 43(04): 148-151+156.
2. Wang, X. Research on Cost Estimation Model of Power Network Overhead Line Engineering[J]. China Power Enterprise Management, 2018(24): 76-77.
3. 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.
4. Geng, P., Liu, C., Yang, G. Establishment and Implementation of Transmission Engineering Cost Prediction Model Based on Data Mining Technology[J]. Modern Electronic Technique, 2018, 41(04): 157-160.
5. Kong, J., Cao, X. Research on cost prediction model of transmission and distribution engineering based on Support Vector Machine[J]. Modern Electronic Technology, 2018, 41(04): 127-130.
6. Hao, K. Study on the Fast Prediction Method of Construction Engineering Cost Based on Fuzzy Least Squares Support Vector Machine[J]. Railway Engineering Cost Management, 2012, 27(03): 1-4+18.
7. Liu, Y. Optimization of wind power time series prediction model based on two-parameter least squares support vector machine (TPA-LSSVM)[J]. Journal of Beijing University of Chemical Technology(Natural Science Edition), 2019, 46(02): 97-102.