Research on the Forecast Model of Transmission and Distribution Project Cost Based on Support Vector Regression Algorithm

Ming Zhou, Shengwei Lu, Li Ma, Yue Ming, Li-ping Sun, Qiupeng Zhou, Sicong Wang and Jiong Yan

1State Grid HuBei Electric Power Company Limited Economic, Hubei, China
2State Grid HuBei Electric Power Company Limited Economic Research Institute, Hubei, China

*Corresponding author: zhoum50@hb.sgcc.com.cn

Abstract. Based on historical power transmission and transformation project data, this paper constructs a support vector regression forecasting model by screening the key influencing factors of power transmission and transformation projects, and then predicts the cost of power transmission and transformation projects above 110kV in a certain province. The results show that the support vector regression forecasting model can quickly predict the project cost of power transmission and transformation projects by inputting key factors of the project. The establishment of this prediction model provides a more scientific basis for rationally controlling the cost of power transmission and transformation projects.

Keywords: Transformer Substation, Neural Network, Investment Plan, Investment Surplus Rate.

1. Introduction

At present, there are many domestic researches on the influencing factors, cost control and management of substation engineering costs, but relatively few researches on the establishment of cost prediction models. In actual substation engineering, the cost level of the project may be affected by many factors such as topography, substation capacity, the number of outgoing circuits, and the layout of the station area, etc. These factors often show a non-linear relationship to the cost, and it is difficult to use traditional prediction methods to perform linear regression and prediction.

Support vector regression has high self-organization, self-adaptation, nonlinear mapping ability and high fault tolerance. Using the support vector regression forecasting method to study the sample can grasp the complex linear relationship between the project cost and its influencing factors. This method is more practical than the traditional linear prediction method. Therefore, this article attempts to use the support vector regression algorithm to establish a model to predict the cost of substation projects.
2. Construction method of cost prediction model

2.1. Literature review of researches on cost forecasting methods
Support vector machine is developing on the base of Vladimir Vapnik statistical theory. It has good pattern recognition and regression prediction effects especially in small sample learning. Compared with traditional methods, the advantages of support vector machines are reflect in the training convergence time, training speed and accuracy. The support vector machine model is simple but the theory is solid. It has made great progress in regression prediction, pattern recognition, and function evaluation. Therefore, it has been widely used in the fields of text recognition, sign language recognition, facial image recognition, gene classification, and engineering cost prediction [1]. As one of the main intelligent algorithms at present, genetic algorithm fully draws on the natural law of survival of the fittest in the theory of biochemical evolution; high search ability, strong robustness, and wide application range are its biggest advantages. The algorithm is mainly based on the exchange of group optimization and individual optimization information, and finds the global optimization by changing the way of gene configuration. It is mainly used in function optimization, combination optimization, power production, power dispatch, image processing, etc. [2].

Scholars at home and abroad have done a lot of research on power engineering cost prediction. Zhao Xiaofang et al. [3] combined the random forest algorithm, particle swarm algorithm and PSO-SVM algorithm in data mining to establish a smart transmission line project cost forecasting model, and verified the effectiveness of the model through actual sample data in Sichuan, and made power grid investment and the decision is more reasonable. Liu Shili et al. [4] used analytic hierarchy process to analyze the key indicators of substation project cost; Hu Jinlan et al. [5] comprehensively used principal component analysis and multiple regressions models to systematical study the investment influencing factors of power engineering, and construct a forecasting model based on BP neural network. This model is optimize by particle swarm optimization and has been applied in related examples. Principal component analysis is a widely used method in factor analysis. Many scholars apply the principal component analysis method to the analysis of factors affecting the cost of power engineering [6-8].

2.2. Support vector regression algorithm
Based on the support vector regression method, combined with the project characteristics of power transmission and transformation projects, this paper constructs an intelligent prediction model for the cost of power transmission and transformation projects, and compares and analyzes the prediction results to improve the prediction accuracy and efficiency.

For the sample \((x, y)\), suppose that the support vector regression algorithm can tolerate a deviation of at most \(\epsilon\) between \(f(x)\) and \(y\), that is, the loss is calculated when the absolute value of the difference between \(f(x)\) and \(y\) is greater than \(\epsilon\). As shown in Figure 1, this is equivalent to taking \(f(x)\) as the center and constructing an interval band with a width of \(2\epsilon\) (red area in the figure). If the training sample falls into this interval band, it is considered correct.

![Figure 1. Typical diagram of support vector regression.](image)
Unlike other regression methods, SVR tolerates a maximum deviation of $\varepsilon$ between $f(x)$ and $y$, that is, the loss is calculated only when the absolute value of the difference between $f(x)$ and $y$ is greater than $\varepsilon$. This is equivalent to taking $f(x)$ as the center and constructing an interval band with a width of $2\varepsilon$. If the training sample falls into this interval band, no loss is calculated. So the SVR problem can be formalized as:

Determine the parameters $w, b$;

$$
\min \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{m} l_{\varepsilon}(f((x_i) - y))
$$

Where $C$ is the regularization constant, and $l_{\varepsilon}$ is the insensitive function of $\varepsilon$. By introducing slack variables $\varepsilon_i$ and $\tilde{\varepsilon}_i$, equation 1 can be rewritten as:

Determine the parameters $w, b, \varepsilon_i$ and $\tilde{\varepsilon}_i$;

$$
\min \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{m} (\varepsilon_i + \tilde{\varepsilon}_i)
$$

By introducing Lagrangian multipliers $u_i \geq 0$, $(u_i) \geq 0$, $(a_i) \geq 0$, $a_i \geq 0$, the Lagrangian function of formula 3 is obtained, and the duality of SVR is obtained by seeking partial derivative problem, and get the solution of SVR through KKT condition:

$$
f(x) = \sum_{i=1}^{m} (a_i - \tilde{a}_i) x^T
$$

The samples with $(\tilde{a}_i - a_i) \neq 0$ in formula 3 are the support vectors of SVR, and they must fall outside the $\varepsilon$-interval band. The support vector of SVR is only a part of the training sample, and its solution is sparse.

If the feature mapping form is considered, formula 3 can be expressed as:

$$
f(x) = \sum_{i=1}^{m} (a_i - \tilde{a}_i) k(x, x_i) + b
$$

Where $k(x, x_i) = \Phi(x_i)^T\Phi(x_i)$ is the kernel formula. Support vector regression is a better machine learning algorithm when the data samples of substations and overhead lines are relatively small at this stage. Support vector regression uses part of the support vectors to make hyperplane decisions, instead of relying on all the data. The learned model can always be expressed as a linear combination of the kernel function $k(x, x_i)$. According to the pattern recognition theory, the linearly inseparable patterns in the low-dimensional space may be linearly separable by non-linear mapping to the high-dimensional feature space. However, if this technique is directly used for classification or regression in the high-dimensional space, there is a need to determine the nonlinearity issues such as the form and parameters of the mapping function and the dimension of the feature space. The biggest obstacle is the "curse of dimensionality" when computing high-dimensional feature spaces. The kernel function transforms the inner product operation of the high-dimensional space into the kernel function calculation of the low-dimensional input space, so as to subtly solve the problem of "dimensionality disaster" calculated in the high-dimensional feature space.
3. Construction of intelligent forecast model of engineering cost

3.1. The case of training samples
The typical engineering samples selected for modeling training are the real power transmission and transformation engineering data of a province from 2016 to 2018 of 110kV-500kV. Since the modeling is in the verification and testing stage, this stage only performs data preprocessing on samples of substation projects that affect the cost of transmission and transformation projects. There are 119 110kV substation-engineering samples in the training samples.

3.2. Extraction of key factors
The factors selected after data preprocessing do not necessarily have a significant impact on the cost results, and changes in some factors within a certain range will not have a major impact on the overall cost level. Therefore, the key factors can be screened out by principal component analysis.

Screening steps:

By comparing the correlation coefficients of the variables in each of the main components, the main variables with larger correlation coefficients are selected, and all key factors are finally screened. The key factor screening results are shown in the following table:

Table 1. Key factors of substation project.

| X1                  | X2                           | X3                           | X4                           | X5                           | X6                           |
|---------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| Construction nature | Total capacity               | High voltage side outlets    | Medium voltage side outlets  | Low voltage side outlets     | Low voltage capacitors       |
| X7                  | X8                           | X9                           | X10                          | X11                          | X12                          |
| Land acquisition area | Area of main control building | Construction site requisition and cleaning fees | Main transformer price | High voltage side power distribution unit type | High voltage side circuit breakers unit price |
| X13                 | X14                          | X15                          | X16                          | X17                          | X18                          |
| Medium voltage side circuit breakers unit price | Cost of inbound road | Foundation treatment cost | Field leveling cost | Retaining wall and slope protection | Power cable |

3.3. Model training
There are 119 data samples of 110kV substation, 10 data are randomly selected as the verification sample set, and the remaining 109 data are used as the training sample set. Test the training effects of linear kernels, polynomial kernels and Gaussian kernels respectively. After comparative analysis, the effect of linear kernels is better than that of polynomial kernels and Gaussian kernels. Therefore, linear kernels are selected as the final kernel function of the support vector regression model.

3.4. Model validation
The 10 pieces of data are verified, and the verification result table is as follows:
Table 2. Substation engineering verification result.

| Project No. | Project nature | Substation capacity (MVA) | Actual static investment (10K Yuan) | Forecast static investment (10K Yuan) | Absolute deviation rate (%) |
|-------------|----------------|---------------------------|-----------------------------------|---------------------------------------|----------------------------|
| 1           | New transformer | 50                        | 3127.07                           | 2943.79                               | 5.86                       |
| 2           | New transformer | 50                        | 2756.41                           | 2631.00                               | 4.55                       |
| 3           | New transformer | 50                        | 2804.49                           | 2839.53                               | 1.25                       |
| 4           | New transformer | 50                        | 2724.43                           | 2772.40                               | 1.76                       |
| 5           | New transformer | 50                        | 2296.86                           | 2319.10                               | 0.97                       |
| 6           | New transformer | 100                       | 3054.68                           | 2944.73                               | 3.60                       |
| 7           | New transformer | 50                        | 2276.30                           | 2373.49                               | 4.27                       |
| 8           | New transformer | 50                        | 2613.55                           | 2537.09                               | 2.93                       |
| 9           | New transformer | 50                        | 2420.39                           | 2412.91                               | 0.31                       |
| 10          | Expansion transformer | 50                  | 827.83                             | 934.13                                | 12.84                      |

According to the chart, there are 9 samples with predicted static investment and real static investment within 10%, and 1 sample with more than 10%. Overall, the average deviation rate is 3.83%, and the model performs well on the validation set.

Support vector regression prediction has a certain accuracy rate, but some samples have large deviations. The reason for this situation is that the sample size is too small or the particularity of the sample itself, which leads to large prediction deviations.

4. Application of cost prediction model

According to the support vector regression model, the scale and technical parameters of 9 substation projects to be predicted are input into the model, and the output forecast investment and balance rate results are as follows.

Table 3. Forecast results of substation projects.

| Project No. | Voltage level | Project nature | Feasibility study estimate (10K Yuan) | Forecast static investment (10K Yuan) | Forecast balance rate (%) |
|-------------|---------------|----------------|--------------------------------------|---------------------------------------|---------------------------|
| 1           | 110           | New transformer | 3015.15                             | 2447.41                               | 18.82                     |
| 2           | 110           | New transformer | 3274                                 | 2488.42                               | 23.99                     |
| 3           | 110           | New transformer | 3117.05                             | 2744.63                               | 11.94                     |
| 4           | 110           | New transformer | 4486                                 | 4690.32                               | -4.55                     |
| 5           | 110           | New transformer | 3580.25                             | 3234.18                               | 9.67                      |
| 6           | 110           | Expansion transformer | 954.52                          | 863.51                                | 9.53                      |
| 7           | 110           | Expansion transformer | 1279.46                        | 916.23                                | 28.39                     |
| 8           | 110           | New transformer | 4465.54                             | 3526.24                               | 21.03                     |
| 9           | 110           | New transformer | 2870.27                             | 2022.38                               | 29.54                     |
6

Figure 2. Prediction model actual value and prediction value comparison.

Figure 3. The distribution of the error rate between the actual value and the predicted value of the prediction model.

The error of the verification sample is required to be within 10%, and it is reasonable to control the predicted investment balance rate at about 20%. As can be seen from the above figure and table, there are 2 samples with predicted balance rate within 10%; 2 samples with predicted balance rate within 10%-20%, and 4 with predicted balance rate within 20%-30%, and 1 unreasonable result.

The unreasonable project is a 110kV new transformer substation project, and the new transformer has a single capacity of 50MVA. There are 2 sets in this period, belonging to outdoor stations, with 2 times on the high-voltage side of the outlet scale, 0 times on the medium-voltage side and 24 times on the low-voltage side. It was found that the construction site and clean-up costs were relatively high, about 7 million, which was higher than the average level of conventional similar projects; the reported construction area of the main control building and the entire station were relatively high. Therefore, there is a certain deviation in the model prediction, which leads to unreasonable prediction results.

5. Conclusion
This paper establishes a substation project cost prediction model based on the principle of support vector regression algorithm, which can establish corresponding prediction models for different voltage levels, and then predict the project cost. From the analysis of the model's balance rate, the model can accurately and quickly predict the cost level of the substation project, and the cost balance rate calculated by the model is very close to the actual situation. The predictive model can be used for cost
control of substation projects in the initial design, feasibility study and other stages. Because the support vector regression prediction model is highly dependent on the quality and quantity of the database, in future work, researchers will continue to accumulate actual project cost data, further improve the model, and apply the prediction model to all aspects of cost management, provide a more scientific basis for controlling the cost.

References
[1] Jiang Lina. Research on construction cost prediction based on support vector machine [D]. Hebei University of Engineering.
[2] Jing-ming, YANG, et al. Genetic Algorithm-Based Optimization Used in Rolling Schedule [J]. Journal of Iron & Steel Research International, 2008
[3] Zhao Xiaofang, Shi Yu, Zhao Kuiyun, et al. Research on Forecasting Model of Transmission Line Engineering Cost [J]. China Power Enterprise Management, 2018, 525 (12): 72-74
[4] Liu Shili, Zhu Xiaohu, Chen Fulei, etc. Critical Factors for Cost Management of Power Grid Projects Based on Analytic Hierarchy Process [J]. Journal of Engineering Management, 2016, 30 (6): 129-133.
[5] Hu Jinlan, Gao Xiaobin, Liu Chongming, et al. Empirical Study on the Influencing Factors of Power Grid Project Investment Based on BP Neural Network [J]. Engineering Economy, 2017, 27 (10): 15-19.
[6] Ji Yongmei, Wu Dongping, Peng Ying, et al. Analysis of influencing factors of substation project cost—based on SPSS software principal component analysis method [J]. Operation and Management, 2014 (2): 125-130.
[7] Hu Jinlan, Gao Xiaobin. Identification and Data Modeling of Transmission and Transformation Projects’ Factors [J]. Journal of North China Electric Power University (Social Sciences), 2016 (5): 55-59.
[8] Xiao Junyi, Hao Xiaoling, Tang Xiaoqian, etc. Influencing Factors of Electric Power Project Cost Based On Principal Component Analysis [J]. Management and Administration, 2014 (3): 128-133.