

Research on Forecasting the Cost of Residential Construction Based on PCA and LS-SVM

Zhongfu Qin ^{1,a*}, Xiaolong Lei ^{1,b}, Liqing Meng ^{1,c}

¹College of civil Engineering and Architecture, Zhejiang University, China.

^axl5358949@126.com; ^b1056581442@qq.com; ^c1575840551@qq.com

Keywords: Residential construction; Indicators; Costs forecasting; Principal component analysis; Least squares support vector machine

Abstract. To forecast the costs of a residential construction rapidly and accurately in the initial stage of construction and lack of relevant information. Based on the strengths and weaknesses of previous studies about it, a new model to forecast the costs of a residential construction which is based on Principal Component Analysis (PCA) and Least Squares Support Vector Machine (LS-SVM) is proposed. First based on the factors analysis in the costs of a residential construction, chooses the indicators and samples of residential construction for the prediction of the Construction Costs, after which submits the selected indicators data to the Principal Component Analysis(PCA) to eliminate the Correlation in it; then, the new indicators data is imported into the Least Squares Support Vector Machine (LS-SVM) and training in it to build a new model to forecast the costs of a residential construction. Finally, selects 5 projects in conjunction with the new model for simulation analysis, the relative error are controlled within $\pm 7\%$.

Introduction

Construction project cost forecasting method was initially implemented by the unit index [1]. Since the estimated indicators are the industry or local uniform, without considering the quality of the projects, schedule, construction of enterprise technology and management level and other factors, it is not well adapted to the market economy. The traditional statistical analysis method, the linear regression method was established by regression analysis model to predict the project cost, when the ways to deal with the uncertain factors and limited data samples showed poor, even invalidation method [2]. The time series prediction can only be applied to short-term factors and stable construction cost forecast [3]. It can be known that the traditional statistical learning methods during the construction project cost prediction, defect are obvious.

Statistical learning theory (SLT) [4] is a machine learning theory developed by Vapnik. It is a training samples that is studied under the condition of limited machine learning rules of discipline. Support Vector Machine [5] (SVM) is based on the development of the theory, and introduces the structural risk minimization principle and the kernel function. Compared with the traditional machine learning, support vector machine is good at the small sample learning and nonlinear problems [6], Support vector machine theory and neural network forecasting technology can be compared to more accurate predictions and forecast further enhance the speed [7]. Construction Cost prediction model, however, based solely on SVM theory building, its efficiency can still be further optimized [8].and in many scientific research fields have shown a good generalization ability. Therefore, the study of support vector machine has become a hot spot in the world. Many of them are for the improvement of SVM algorithm itself, such as block processing algorithm, fixed

working sample set algorithm, SMO algorithm [9]. Least square support vector machine (LS-SVM) [10] is an improved algorithm of SVM, which is used to solve the dimension disaster of support vector machine. And it can further improve the training speed and accuracy of learning machine.

Through the above analysis, this study selected LS-SVM algorithm to construct the building construction cost forecasting model; at the same time, in order to further optimize the model prediction effect, this study by means of principal component analysis (PCA) to remove the correlation between the selected predictor, and the process variable dimension [11].

Support Vector Machine

Support vector machine (SVM) theory is developed from the optimal classification line [5]. The optimal classification line [12] can not only separate the two types of samples, but also get the maximum classification interval. The optimal classification line is increased to a multi dimension space. Finding the optimal classification surface is a two time planning problem. The support vector machine is used to search the optimal classifier, and the linear map of the original space can be classified by using the kernel function, and the fault tolerance is introduced by using the loss function.

LS-SVM is an improved algorithm of SVM. And the support vector machine is improved in the process of optimizing problem. The different loss function is selected, the inequality constraints are changed into equality constraints, so the optimization problem of LS-SVM is:

$$\begin{aligned} \min \quad & \phi(\omega, \varepsilon) = \frac{1}{2} \|\omega\|^2 + \frac{1}{2} \gamma \sum_{i=1}^l \varepsilon_i^2 \\ \text{subject to} \quad & y_i((\omega * x_i) + b) = 1 - \varepsilon_i, i = 1, \dots, l; \varepsilon_i \geq 0 \end{aligned} \quad (1)$$

Index Selection

The influence factors of construction are complicated, which are influenced by different technical conditions, natural conditions and economic conditions, which make the cost analysis face difficulties. This study selects the unit cost of construction (yuan /m²) as the output set. Through the above analysis, this study selects 15 indicators as the input set. { Construction area on the ground and below the ground, Building floors on the ground and below the ground , Layer high on the ground and below the ground, Base type, Structure type, Pile Type, Concrete price growth, Steel price growth, Seismic rating, Project management level, Earthwork processing difficulty, Facades }

Simulation Analysis of Construction Cost Prediction Model

Case Description. This study collected 45 construction projects in Hangzhou area. Due to the construction project cost that the construction area is too small is influenced by accidental factors easily. And then affect the forecast effect from construction project cost prediction model. Therefore, the combination of expert opinion, eliminated the five construction project which's Built-up area is less than 3 million square meters, ultimately retained 40 projects as project case. Then the construction cost prediction model is simulated and analyzed.

Simulation Analysis.

Principal Component Analysis. By using MATLAB, the original input set X was standardized by the standard score (Z-SCORE) Then call the principal component analysis toolbox in MATLAB to analysis. The principal component analysis get the characteristic value z_i 、the contribution

rate q_i and the cumulative contribution rate Q_i which are shown in Table 1:

Table 1 Characteristic value and contribution rate of principal components

z_i	λ_i	q_i (%)	Q_i (%)
1	3.6898	24.5986	24.5986
2	2.3428	15.6183	40.2169
3	1.8631	12.421	52.6379
4	1.6626	11.084	63.7219
5	1.1954	7.9691	71.691
6	0.9897	6.5982	78.2892
7	0.9157	6.1045	84.3937
8	0.8756	5.8377	90.2314
9	0.4747	3.1647	93.3961
10	0.3175	2.1169	95.513
11	0.281	1.8735	97.3865
12	0.2442	1.6279	99.0144
13	0.1083	0.7217	99.7361
14	0.0273	0.1821	99.9182
15	0.0123	0.0819	100

From the above Table 1 we can see that the eigenvalues of the 15 principal components are more than 0. The characteristic values and the contribution rate of each principal component are decreased from z_1 to z_{15} in turn. The general cumulative contribution rate is $Q_i \geq 85\%$, which is considered that the main components of the former i can be used to instead of the original index information^[13]. In this study select the first 8 principal components ($Q=90.2314\%$), the first 9 principal components ($Q=93.3961\%$) and the first 10 principal components ($Q=95.513\%$).

Model training and prediction of LS-SVM. First of all, separate the 40 samples of retention, the former 35 samples as the prediction model of training samples, the remaining 5 samples as the test samples; secondly, in the preinstalled LS-SVM toolbox in MATLAB environment, the training sample data processing by PCA import the "LS-SVM model" (in view of the parameters in LS-SVM model can be adjusted, starting parameters were: the penalty coefficient In view of the adjustable parameters of LS-SVM model, we select initial parameters: the penalty coefficient $\gamma=100$, the kernel function of the width coefficient $\sigma^2=0.01$ in the MATLAB for training and prediction). Finally, the test sample data is introduced into the construction project cost forecasting model, and the results of the model are analyzed in order to test the effect of the model. The following Fig. 1 is showed the first 8 principal components, the first 9 principal components and the first 10 principal components of the construction project cost prediction model comparison.

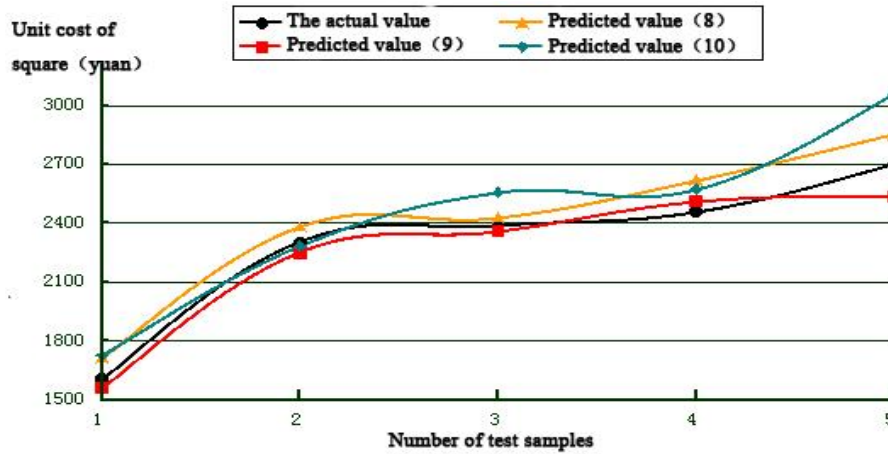


Figure 1. The prediction effect of LS-SVM model based on the first 8, 9 and 10 principal components respectively

Through the Fig. 1 shows that the first 9 principal components, based on the PCA and LS-SVM of the construction project cost prediction model predict the better results.

Evaluation Model. The absolute error and relative error between the predicted value and the actual value of the "construction project cost prediction model" built by this study are shown in the following Table 2.

Table 2 The error between the predicted value and the actual value

Test sample	Actual value(yuan/m2)	Forecast value(yuan /m2)	Absolute error(yuan /m2)	Relative error (%)
1	1607	1563	-44	-2.738
2	2306	2249	-57	-2.472
3	2391	2360	-31	-1.297
4	2460	2513	53	2.154
5	2701	2538	-163	-6.035

From the Table 2, the relative error of the model prediction is -6.035%, the minimum is -1.297%, and the relative error is less than 7%. Due to the project cost forecast generally occurs in the earlier stage of project construction, scheme comparison and selection phase, therefore, the relative error of project cost prediction is controlled less than 7% [14]. The construction of the construction project cost forecast model prediction model performs well which can be basis for scheme selection.

Summary

In this paper, the combination of principal component analysis and LS-SVM is applied to the prediction of construction cost. The construction cost prediction model based on PCA and LS-SVM is constructed. Combined with the selected sample project, a case study is carried out. Then, the error of the model prediction and the stability of the prediction results are analyzed. Through the analysis we can know that the new model is good at solving the small sample and nonlinear problem of the construction project cost prediction. And the relative error control is reasonable and stable. The forecast effect is better. Therefore, the construction project cost prediction model

constructed by this study can bring the help to the construction project plan comparison and the investment estimate, and has the very high popularization value.

References

- [1] Liu Zhongying, Xu Hong. Construction cost and tender offer [M]. Southeast University press, 2002.
- [2] Xue Xiangyang. Improved Linear Regression Forecast Model[J].Science Technology and Engineering,2010,10(12):2970-2973.
- [3] Box G. Box and Jenkins: time series analysis, forecasting and control [J]. A Very British Affair: Six Britons and the Development of Time Series Analysis During the 20th Century, 2012, 161.
- [4] Vapnik V. The nature of statistical learning theory[M]. Springer Science & Business Media, 2013.
- [5] Harman G, Kulkarni S. Statistical Learning Theory and Induction [M]//Encyclopedia of the Sciences of Learning. Springer US, 2012: 3186-3188.
- [6] Zhu Jiayuan, Yang Yun, Zhang Hengxi. Study on small sample prediction based on optimized least square support vector machine [J]. Journal of Aeronautical Science, 2004,25 (6): 565-568.
- [7] Xie Ying, Gao Linan, Shi Zhenwu. Highway engineering cost prediction model based on least squares support vector machine [J]. Chinese and foreign highway, 2007,03:242-245.
- [8] Jiang Lina. Research on construction engineering cost prediction based on support vector machine [D]. Hebei University of Engineering, 2009.
- [9] Liu Huichun, Research status of support vector machine [J]. Chinese Journal of image and graphics, 2002,7 (6): 618-623.
- [10]Liu Jian. Study on the online learning algorithm based on support vector machine [D]. Zhejiang University, 2013.
- [11]Boolchandani D, Sahula V. Exploring efficient kernel functions for support vector machine based feasibility models for analog circuits[J].International Journal on Design, Analysis and Tools for Circuits and Systems, 2011:1-8.
- [12]Aldrich C, Auret L. Statistical Learning Theory and Kernel-Based Methods[M].Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods. Springer London, 2013: 117-181.
- [13]Xiao Chi, Li Wenjuan. Personal credit evaluation based on principal component analysis and support vector machine [J]. Technology economy, 2010,03:69-72.
- [14]Shen Ruina, Cao Chang, fan Zhongjun. Based on principal component analysis support vector machine model for Shanghai housing prices prediction [J]. Mathematical practice and understanding, 2013,23:11-16.