Construction Project Cost Prediction Based on Genetic Algorithm and Least Squares Support Vector Machine

Ming Xu[1, a], Bingfeng Xu[1,b*], Lanjiang Zhou[2,c] and Lin Wu[2,d]

[1]School of Civil Engineering and Architecture, Kunming University of Science and Technology, Kunming 650500, China

[2] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

[a]mingxukm@sina.com, [b]xbf321@sina.com,[c]915090822@qq.com,[d]52038994@qq.com

Keywords: construction project cost; forecast model; genetic algorithm; least squares support vectormachines; small-sample learning

Abstract. For the small sample data and complex nonlinear characteristic of construction project cost, a new hybrid prediction model combing genetic algorithm and small sample learning model based on least squares support vector machines is proposed. First, all the candidate features are ranked by correlation with the dependent variable, the front ranking features are used to initialize part of population for the genetic algorithm to get better feature subset, and then the construction cost prediction model of the least square support vector machine is constructed. Experiments on Jiangsu Province housing project data show an improved performance over other models in prediction accuracy, it is an effective method of project cost forecasting.

Introduction

Construction industry plays an important role due to its large share in economy of China. The core of construction industry is construction project, thus construction project management has a high practical significance. Construction project management includes investment estimation, design estimates, project budget, project settlement and project accounts, the most important part of construction project management is on investment estimation for construction cost, which directly determines the profitability of a project. Construction cost estimation plays an important role in construction projects investment, and provides an important basis for project feasibility study, comparison and selection of design. The accuracy of construction cost estimation directly affects the investment decision of project, thus has a high research value.

There has been a lot of research for the construction project cost forecast. Currently the project cost prediction methods mainly focus on probability analysis method [1], artificial neural network analysis method [2-6], fuzzy analogy method [7-8], regression analysis [9], time series forecasting method [10], case-based derivation method [11-13].Calculation process of the method based on probability analysis is complex and it requires a lot of previously project dataset in order to obtain priori probabilities. Fuzzy analogy method describes the complex issue of project cost estimation relatively simple, and it is difficult to determine the fuzzy membership function , thus its prediction accuracy is not good enough. Regression analysis establishes a relationship between the input attributes and predicted cost, and does not take uncertain factors into account. Time series forecasting method uses parametric model to analyze and process the ordered random data, whose prediction reliability is not high[14]. The method of artificial neural network is difficult to design the network structure, and needs a large number of data samples, it is prone to over-fitting, and has a poor generalization. Also case-based reasoning method requires large sample of cases. The actual project cost is affected by the comprehensive regional economic level of construction in addition to the building structure, building materials prices, etc. During the same period, the collected comparable data is not much more. Investors and construction unit urgently need to find an ideal forecasting method, using the limited historical engineering data to estimate the new project cost quickly, and make the investment plan reasonably. Least squares support vector machine(LSSVM) is a machine learning method based on the principle of structural risk

minimization, it is specifically aimed at the problem of small sample, nonlinear, high dimension, and has a good generalization capability. The change in project cost is a multivariable, nonlinear complex system, which is difficult to describe with mathematical models, LSSVM does not require any assumption on the data distribution, it can effectively deal with the small sample data with multiple variables, thus is suitable for the estimation of the cost of the project. As genetic algorithm has better global optimization ability and characteristics of maintaining population diversity , in order to obtain a good performance and prediction results, we propose a hybrid algorithm based on genetic algorithm and the least square support vector machine ,which constructs the project cost forecasting model by the feature extraction and the small sample learning, and experiments are conducted to verify the prediction performance of the proposed model.

Least square support vector machine regression algorithm

On the basis of standard SVM, Suykens proposed LSSVM[15] on the basis of standard SVM, which sets the function of the support vector machine as the sum of squared error, and replaces inequality constraint with equality constraint. LSSVM greatly facilitates the solution of Lagrange multiplier, improves the adaptability and accuracy of the model, and broadens the application space of support vector machine.

For the training sample set {(Xi, Yi)} (i=1,2,... , n) xi $\in$ ^{n}, yi $\in$, where Xi denotes input and Yi denotes output, the input sample is mapped to a high dimensional feature space by nonlinear mapping function j (.) , and the linear regression is done in Hilbert space.

$$f(x) = w^T j(x) + b. \tag{1}$$

where w is the weight vector, b is the bias. Formula (1) can be converted to:

$$\begin{aligned} &\underset{w,b,x}{Min}\ J(w,x) = \frac{1}{2} w^T w + g \sum_{k=1}^{n} x_k^{\,2}. \\ &s.t.\ y_k - [w^T j(x_k) + b] = 1 - x_k. \end{aligned} \tag{2}$$

γis a adjustable parameter or penalty factor; $x_k \in$ is an error variable. Similar to SVM, Lagrange multiplier is introduced:

$$L(w,b,x,a) = \frac{1}{2} w^T w + g \sum_{k=1}^{n} x_k^{\,2} - \sum_{k=1}^{m} a_k \{ y_k - [w^T j(x_k) + b] + x_k - 1 \}. \tag{3}$$

where Lagrange multiplier $a_i \in$.

According to the optimized conditions, $\frac{\partial L}{\partial w} = 0$ $\frac{\partial L}{\partial b} = 0$ $\frac{\partial L}{\partial x_k} = 0$ $\frac{\partial L}{\partial a_k} = 0$

Then

$$\begin{aligned} &w = \sum_{i=1}^{n} a_i j(x_i),\ \sum_{i=1}^{n} a_i = 0,\ a_i = c x_i. \\ &w j(x_i) + b + x_i - y_i = 0. \end{aligned} \tag{4}$$

According to the Mece condition [16], there must be a mapping function $\Psi(\cdot,\cdot)$, such that formula(5) holds:

$$\Psi(x_k, x_l) = j(x_k)^T j(x_l). \tag{5}$$

Then the LSSVM regression model is:

$$f(x) = \sum_{i=1}^{n} a_i \Psi(x_i, x_j) + b. \tag{6}$$

In general, the performance of the radial basis kernel function is good. In this paper, radial basis function is selected, which is defined as

$$\Psi(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2s^2}\right). \tag{7}$$

Where σ is the width of the radial basis kernel function, LSSVM regression model is given by formula (8).

$$f(x) = \sum_{i=1}^{n} a_i \exp\left(-\frac{\| x_i - x_j \|^2}{2s^2}\right) + b. \tag{8}$$

Feature selection based on genetic algorithm.

Feature selection removes irrelevant or redundant features, so as to achieve the purpose of reducing the number of features, improving the accuracy of the model and saving the time of operation. Genetic algorithm is a random search strategy, which achieves improved adaptability of each individual, and then solves complex problems via evolutionary operations of selection, heredity and variation[17]. Genetic algorithm can evaluate the feature subsets, it evaluates the contribution of a combination of multiple features to the prediction of target variable to ensure the combination optimization of the selected subset, also eliminates the need to consider the correlation between all features.

Feature representation. We utilize a simple and efficient binary encoding scheme. Individual of genetic algorithm is a binary string of length M, each bit of the string represents a feature , where M is the number of candidate features. Bit i of the string is 1 indicates that feature i is selected, else feature i is not selected. Each individual represents a feature selection scheme.

Initial population. As a kind of uncertain search algorithm, the performance of Genetic algorithm is greatly influenced by the initial population. Pearson correlation between candidate features and target variable is computed and sorted in descending order as a priori knowledge to genetic algorithm. Meanwhile, in order to make up the deficiency of only considering the correlation between single feature and target variable, and then add some of the features randomly generated, in order to avoid filtering useful combination of features.

Let the initial population size is P, M is the length of the individual; C, I are user-defined parameters, $1 \le C \le M, 1 \le I \le P$. C represents the number of high-ranking feature extracted from the correlation sort list, I is the number of initial individuals associated with C, more clearly, for I individuals, set the corresponding genes associated with the top-C features extracted from the correlation sorted list in the first phase to 1, the rest genes are randomly set to 0 or 1. The remaining N individuals are randomly initialized.

Fitness function. In this paper, the objective function is designed to minimize the mean absolute percent error(MAPE), by support vector regression machine learning, use formula (9) to evaluate fitness of the i-th individual:

$$f(x_i) = \frac{1}{l_1 * MAPE(i) + l_2 * \frac{n * feature(i)}{M}}. \quad (9)$$

where MAPE is defined as

$$MAPE = \frac{\sum_{i=1}^{m} | y_i - \hat{y}_i | / y_i}{m} * 100 \quad (10)$$

m is the number of test samples, y_i is the actual cost value, $\hat{y}_i$ is the model output value , l_1, l_2 are user-defined parameters, which are used to balance the two items of the fitness function , that is, the prediction accuracy and the number of features to the fitness function.

Genetic Operators. In this paper, we use roulette proportional selection method and single-point crossover algorithm where the cross point position is randomly selected. Termination condition is reaching the maximum number of iterations or the optimal solution unchanged during five consecutive times[17].

Application of GA-LSSVM in project cost forecast

Experimental data. We select residential construction projects as the object of Construction Cost Forecasting. The Data is residential cost information of Jiangsu Province ,China, published from 2014 to 2015, where 28 records published in 2014 and 14 records published in 2015. We randomly select 38 records as learning sample, the rest as test set. By consulting the relevant civil engineer in the field of construction to preprocess all of the sample index, we can get the following original set of inputs and outputs set: input set ={construction area, foundation type, foundation bottom elevation, structure type, seismic intensity, roof waterproof, roof insulation, façade, wall insulation, windows and doors, ceiling, number of stories, ground floor, wall insulation, number of floors, floor height, top of wall level, Concrete price, steel price},output set ={cost per square meter}.First, turn the qualitative data into quantitative data, the value of construction area, foundation bottom elevation, number of stories, number of floors, floor height, top of wall level are unchanged, building materials prices are set to the median price announced by the government website according to date of the project. To avoid the impact on prediction caused by the large data dimension difference between different dimension, all data of p indexes is normalized by formula (11):

$$x_i' = \frac{x_i - E(x_i)}{\sqrt{D(x_i)}}. \quad (11)$$

where i = 1,2, ..., p; $E(x_i)$ denotes the expectation of variable ; $D(x_i) = \frac{1}{n}\sum_{i=1}^{p}(x_i - x)^2$ is the variance of the variable . There are a total of 18 input indexes in the original data set, whose data are distributed between 0.5-5219, data values with different dimension are quite different, especially values between foundation bottom elevation and the steel price. After normalizing, data value difference between different indexes is not big, the maximum value is 2.882307, minimum is -2.09783, normalizing relieves the impact on regression analysis caused by the large magnitude difference between different indexes. The correlation coefficient between each index and the dependent variable is shown in Table 1.

Table 1 correlation coefficient between the input index and the dependent variable

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18
0.378	0.226	0.456	0.242	0.099	0.464	0.068	0.310	0.216	0.371	0.071	0.082	0.221	0.188	0.186	0.187	0.368	0.309

where p1-p18 are as follows: construction area, foundation type, foundation bottom elevation, structure type, seismic intensity, roof waterproof, roof insulation, façade, wall insulation, windows and doors, ceiling, number of stories, ground floor, wall insulation, number of floors, floor height, top of wall level, concrete price, steel price.

Experiments and results. In order to verify the effectiveness of the proposed feature selection method and the forcasting model, two sets of experiments are designed. Experiment 1 selects all the indexes whose correlation coefficient with the dependent variable is greater than 0.15, the project cost forecasting model is established by LSSVM and back propagation neural network (BPNN) respectively, According to experience, the LSSVM kernel function width coefficient is set to 0.2, the penalty coefficient 50[18]. The results are shown in Table 2.

Table 2 Prediction results of two models

model	MSE	MAPE
BPNN	29331.63	10.09%
LSVSM	17415.52	7.76%

From table 2 we can see that the prediction result of BPNN is worse than that of LSSVM, for its structure design, parameter determination is more difficult and it requires a lot of training samples, is prone to overfitting , while LSSVM is suitable for the prediction of small sample data, and it can describe the complex nonlinear variation of the engineering cost.

In order to verify the effectiveness of the feature selection method, experiment 2 uses different features to predict the cost of the project and compared the results. Three sets of feature are used. The first set of feature F1 consists of all the index whose correlation coefficient with the dependent variable is greater than 0.15, The second set of feature F2 is obtained by the genetic algorithm which uses all of indexes rather than correlation coefficient filtered and selects the highest individual fitness as a result. The third set of feature F3 is obtained by the genetic algorithm utilizing the correlation sorted result to construct some individuals of initial population ,we set the initial population size P=20, the maximum number of running I=30, crossover and mutation probability are 0.6 and 0.02,set the number of high-ranking feature extracted from the correlation sort list C as 30% of total features, set the number of initial individuals associated with C , I as 25% of the population size. Prediction results are shown in table 3.

Table 3 Prediction results of different feature selection methods

Feature set	MSE	MAPE
F1	17415.52	7.76%
F2	14750.10	6.94%
F3	12097.81	6.47%

As can be seen from Table 3, models which use feature set F2 or F3 achieve better performance than that uses feature set F1,which is obtained by correlation filtering only, the feature selection is an effective method to improve the accuracy of prediction, it proves that the feature selection algorithm is effective. Compared to feature set F1,F2, since feature set F3 is obtaineded by utilizing the correlation sorted result to construct some individuals of initial population so that GA can get good prior knowledge for better feature subset. Moreover, under the premise of ensuring the prediction performance, the number of feature is reduced, experiment results show the effectiveness in terms of dimensionality reduction.

Conclusions

In this paper, we propose a hybrid forecasting method, which combines feature extraction and small sample learning, and apply it to the construction cost prediction. For construction project cost is affected by many complex factors and the sample data is limited, the prediction model based on genetic algorithm and LSSVM is constructed. LSSVM can effectively deal with the small sample data with multiple variables, and does not require any assumption on the data distribution, thus the GA-LSSVM is introduced into the construction project cost forecast. The results show that GA-LSSVM is very

suitable for modeling and prediction for construction project cost with small sample, the performance is better than BP neural network, and it can provide help for the construction project investment. In future work, we plan to optimize the parameters of LSSVM.

Acknowledgements

This work was supported by National Natural Science Foundation of China(Grant No. 61163004), National Nature Science Foundation of China (Grant No.61462055).

References

[1] Nassar, K. M., Gunnarsson, H. G. and Hegab, M. Y., Journal of Construction Engineering and Management. 131(2005) 1257–1262.
[2] Attalla, M., Hegazy, T., Journal of Construction Engineering and Management.129(2003) 405–411.
[3] Denwen Zhang , Hongyan Jiang and Ziyuan Zhang, Journal of Water Resources and Architectural Engineering. 8(2010)61-73.(In Chinese)
[4] Chengjun Wang, Xinhui Zuo, Journal of Shanxi Finance and Economic University, 32(2010)327-330. (In Chinese)
[5] Tao Yin, Jihui Yu, Journal of Chongqin Universiy(Natural Science Edition),30(2007)36-41. (In Chinese)
[6] Hong Ren , Qiming Zhou, China Civil Engineering Journal.38(2005)139-142. (In Chinese)
[7] Yifei Chen , Fangcheng Li, Engineering Journal of Wuhan University. 44(2011)93-73,101. (In Chinese)
[8] Xiaochen Duan, Xiaoping Zhang, Lijun Li and Jianlong Zhang, Journal of Shijiazhuang railway institute (social sciences). 1(2007)38-43. (In Chinese)
[9] Xiaolong Chen, Yingguang Wang, Journal of Tongji University(Natural Science). 37(2009)1115-1112. (In Chinese)
[10] Liuxing Hu, Journal of Taiyuan university of technology. 43(2012) 706-714. (In Chinese)
[11] Doğan Sevgi Zeynep, David Arditi, and H. Murat Günaydin, Journal of Construction Engineering and Management.132(2008) 1092–1098.
[12] Doğan Sevgi Zeynep, David Arditi, and H. Murat Günaydin, Journal of Construction Engineering and Management. 134(2008)146–152.
[13] RunZhi Jin , KyuMan Cho , ChangTaek Hyun and MyungJin Son, J. Expert Systems with Applications. 39 (2012) 5214–5222.
[14] Yan Wang, Application of Time Series Analysis, China Renmin University Press , Beijing,2008.
[15] Weiwu Yan, Huihe Shao, J.Control and Decision,18(2003)358-360. (In Chinese)
[16] Cristianinin Shawe Taylor J., An introduction to support vector machines and other kernel-based learning methods , Cambridge University Press , Cambridge,2000,pp30-34.
[17] Xioaping Wang, Liming Cao, Application and software implementation of genetic algorithm, Xi'an Jiao Tong University press, Xian,2002. (In Chinese)
[18] Guangjin Peng, Jihui Yu, Juntao Wei and Guang Yang, Journal of Chongqing University. 32(2009)1104-1110. (In Chinese)