Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Application of support vector regression in prediction model using genetic algorithm optimized

To cite this article: Wenke Du *et al* 2021 *J. Phys.: Conf. Ser.* **1982** 012048

View the article online for updates and enhancements.

Application of support vector regression in prediction model using genetic algorithm optimized

Wenke Du[1, *], Ruihan Chen[2], Zhenglong Cong[3]

[1]Renmin University of China, Beijing, 100872
[2]Zhejiang Normal University, Jinhua, Zhejiang, 321000
[3]University of Chinese Academy of Social Sciences, Beijing, 102488

*Corresponding author: DuWenke@ruc.edu.cn

Abstract. No matter in any period, housing is the most basic demand of people's life, and it is closely related to people's daily life. Although many domestic and foreign experts have done research and prediction on housing prices, the causes of housing prices are complex, and the results of housing price research in different regions and different periods are very different, and most forecasting models have limitations in the use of them. Therefore, this article is based on the performance of support vector regression depends on the characteristics of key parameter selection, and uses genetic algorithm to optimize the penalty parameters, kernel function parameters and insensitive loss function of the support vector regression model. The optimized parameters are used to establish a support vector regression prediction model, and the housing price is predicted through the prediction model. The simulation results show that the convergence speed and prediction accuracy of the support vector regression prediction model optimized by genetic algorithm have been greatly improved, and the prediction results verify the feasibility and effectiveness of the model..

Keywords: Genetic algorithm; support vector regression; housing price prediction.

1. Introduction

Regardless of whether it is the Chinese market or the international market, in the development of the national economy, the real estate industry is a pillar industry, and the price of real estate has always been the focus of attention. Studying the trend of housing prices can not only provide the basis for the government to strengthen the macro-control of the real estate industry, but also can fit the changes in real estate prices and predict real estate prices well through the forecasting research on housing prices. This can not only provide a reference for buyers and sellers in the real estate industry, but also help both parties to make correct investment and purchase decisions. It can also provide the government with a basis for formulating policies to ensure the stable development of the real estate market and avoid big fluctuations in real estate prices.

Sun Shanshan [1] uses the correlation vector machine method with the kernel function as the radial basis to analyze the monthly price data. According to experiments, the fitting effect of this method is better than that of the traditional data mining method. Liu Feng et al [2]. considered the differential impact of time on housing prices in Chongqing, used variable coefficient regression models to fit housing price data, and compared and analyzed the fitting effects of linear regression models， it is concluded that the variable coefficient model has a better fitting effect. Gao Yuming [3] and others used genetic algorithm (GA) optimized BP neural network to predict housing prices in Guiyang. The results show that the BP neural network prediction model optimized by GA can increase the network training speed and increase the accuracy of predicting housing prices. Chen Shipeng [4] established a random forest model based on Xiangyang housing loan data, and compared it with the prediction results of the ARIMA model and the multiple linear regression model. The experiment proved that the random forest model has a better prediction effect. Wang Zhihai [5] used Beijing's second-hand housing transaction price as the research object, and used online housing sales agencies as the data source to construct a GA-BP forecasting model to predict the Beijing second-hand housing transaction price. He Ling [6] analyzed the influencing factors of Harbin's commodity housing prices from the three levels of supply, demand and comprehensiveness, and used a 5-degree polynomial model to predict the trend of Harbin's housing prices.

Through the above analysis, this paper uses genetic algorithm to optimize the penalty parameters, kernel function parameters and insensitive loss function of the support vector regression model, and uses the optimized parameters to establish a support vector regression prediction model to predict the housing price. The prediction results show that this method can accurately predict the changes in my country's housing prices, which has very important reference value.

2. Basic Principles of Support Vector Regression

Support vector regression is a method of fitting regression to data using the idea of support vector and Lagrangian multiplier. Support vector regression is the promotion of support vector machines in regression problems. Support vector regression maps the sample set from the original feature space to the high-dimensional feature space, and then performs regression analysis on the sample set in the high-dimensional space.

The basic model corresponding to the hyperplane of support vector regression is

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

Where: *w* is the weight vector, *x* is the feature vector of the sample in the original space, and *b* is the bias.

On this basis, introducing loss deviation $\in$, formula (1) can be transformed into

$$\min_{w,b}\frac{1}{2}\left\|w\right\|^{2}+C\sum_{i=1}^{m}l_{\in}\left[f\left(x_{i}\right)-y_{i}\right] \tag{2}$$

In formula (2): C is the regularization constant; is the insensitive loss function of $\in$:

$$l_{\in}=\begin{cases}0, & \left|z\right|<\in\\ \left|z\right|-\in, & \left|z\right|>\in\end{cases} \tag{3}$$

In formula (3): $\left|z\right|=\left|f\left(x_{i}\right)-y_{i}\right|$, introduce slack variables ξ and ζ_{i}, and formula (2) can be rewritten as

$$\min_{w,b,\xi_{i},\zeta_{i}}\frac{1}{2}\left\|w\right\|^{2}+C\sum_{i=1}^{m}l_{\in}\left(\xi_{i}+\zeta_{i}\right)$$

$$\begin{aligned}\text{s.t.}\quad & f\left(x_i\right)-y_i \le \in +\zeta_i \\ & y_i - f\left(x_i\right) \le \in +\xi_i \\ & \xi_i \ge 0, \zeta_i \ge 0, \quad i = 1,2,\cdots,m\end{aligned} \tag{4}$$

By introducing Lagrangian multipliers $\mu_i \ge 0$, $\hat{\mu}_i \ge 0$, $\alpha_i \ge 0$, $\hat{\alpha}_i \ge 0$, Lagrangian function:

$$\begin{aligned}& L\left(w,b,\alpha,\hat{\alpha},\xi,\hat{\xi},\mu,\hat{\mu}\right)=\frac{1}{2}\left\|w\right\|^2 + C\sum_{i=1}^{m} l_{\in}\left(\xi_i+\zeta_i\right)- \\ & \sum_{i=1}^{m}\left(\mu_i\zeta_i\right)-\sum_{i=1}^{m}\mu_i\xi_i+\sum_{i=1}^{m}\alpha_i\left[f\left(i\right)-y_i-\in-\xi_i\right]+ \\ & \sum_{i=1}^{m}\hat{\alpha}_i\left[f\left(i\right)-y_i-\in-\xi_i\right]\end{aligned} \tag{5}$$

Let the partial derivative of $L\left(w,b,\alpha,\hat{\alpha},\xi,\hat{\xi},\mu,\hat{\mu}\right)$ with respect to *w*, *b*, ξ_i , ζ_i be 0, we get:

$$w=\sum_{i=1}^{m}\left(\hat{\alpha}_i+\alpha_i\right)x_i \tag{6}$$

$$\begin{cases} 0=\sum_{i=1}^{m}\left(\hat{\alpha}_i+\alpha_i\right) \\ C=\sum_{i=1}^{m}\left(\hat{\alpha}_i+\alpha_i\right)+\mu_i \\ C=\hat{\alpha}_i+\hat{\mu}_i \end{cases} \tag{7}$$

Substituting formula (6) and formula (7) into formula (5), the dual form of support vector regression can be obtained:

$$\begin{aligned}& \max_{\alpha,\hat{\alpha}}\sum_{i=1}^{m} y_i\left(\hat{\alpha}_i-\alpha_i\right)-\in\left(\hat{\alpha}_i+\alpha_i\right)- \\ & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right)\left(\hat{\alpha}_j-\alpha_j\right)x_i^T x_j \\ \text{s.t.}\quad & \sum_{i=1}^{m}\left(\hat{\alpha}_i-\alpha_i\right)=0, 0\le\alpha_i,\hat{\alpha}_i\le C\end{aligned} \tag{8}$$

The above process satisfies the Karush Kuhn-Tucker (KKT) conditions, that is, the requirements:

$$\begin{cases} \alpha_i\left[f(i)-y_i-\in-\xi_i\right]=0 \\ \hat{\alpha}_i\left[y_i-f(i)-\in-\xi_i\right]=0 \\ \alpha_i\hat{\alpha}_i=0 \\ \xi_i\hat{\xi}_i=0 \\ (C-\alpha_i)\xi_i=0 \\ (C-\hat{\alpha}_i)\xi_i=0 \end{cases} \tag{9}$$

In the case of meeting KKT conditions, there are:

$$b=y_i+\in-\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^T x \tag{10}$$

$$f(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^T x+b \tag{11}$$

Introducing the kernel function, formula (6), formula (10) and formula (11) become formula (12) ~ formula (14) respectively:

$$w=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)\varphi(x_i) \tag{12}$$

$$b=y_i+\in-\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^T x \tag{13}$$

$$f(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^T x+b \tag{14}$$

In the formula: $k(x_i,x_j)=\varphi(x_i)^T\varphi(x_j)$, the selected kernel function is the radial basis kernel function:

$$k(x_i,x_j)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{\sigma^2}\right)=\exp\left(-g\left\|x_i-x_j\right\|^2\right) \tag{15}$$

In formula (15); $g=\frac{1}{\sigma^2}$.

Support vector regression has many kernel functions to choose from, such as Marton kernel function, linear kernel function, tensor product kernel function and so on. The Gaussian radial basis kernel function (RBF kernel) is the most widely used kernel function. It has good support for large samples and small samples. It should be preferred in the absence of prior knowledge. Therefore, the Gaussian radial basis kernel function is selected in this study, as shown in the above formula (15).

2.1. Basic principles of genetic algorithm

Genetic algorithm is a heuristic optimization algorithm based on biological evolution process, which has the characteristics of high efficiency, parallelism and global optimization. For a problem O(t) to be optimized, the genetic algorithm regards each of its solutions t as a gene code in the chromosome, and generates a new gene code through repeated gene recombination and mutation operations, finally, the optimal solution t* of O(t) is obtained. The specific content of the genetic algorithm is as follows:

(1) Population initialization. Initialize a population S={s_1， s_2， …， s_n} randomly, where n represents the number of chromosomes in the population, and s_i represents the i-th chromosome in the population. Commonly used encoding methods include binary encoding and floating-point number encoding. The encoding vector is expressed as b_i, which corresponds to the solutiont_i of O(t).

(2) Fitness calculation. The fitness function f(t) is used to evaluate the fitness of each chromosome in the population S, expressed as f_1， f_2， …， f_n in turn, where f_i represents the fitness of s_i in the population.

(3) Parental choice. According to the fitness of the chromosome, the probability of each chromosome being selected as the parental chromosome is calculated as follows.

$$p_i = f_i / \sum_{k=1}^{n} f_k \tag{16}$$

Where: p_i represents the probability of chromosome s_i being selected as the parental chromosome. According to the probability corresponding to each chromosome, two chromosomes are randomly selected from the population S as the parental chromosomes, denoted as s_f and s_m respectively.

(4) Gene recombination. The parental chromosomes s_f and s_m are genetically cross-recombined, and the respective gene codes are updated as follows

$$\begin{cases} b'_f = \alpha b_f + (1-\alpha) b_m \\ b'_m = (1-\alpha) b_f + \alpha b_m \end{cases} \tag{17}$$

Where: b_f andb'_f are the codes before and after chromosome s_f gene recombination, b_m and b'_m are codes before and after chromosome s_m gene recombination, α are random recombination factors, and $\alpha \in$(0,1).

(5) Gene mutation. The newly generated gene codes b'_f and b'_m have a certain mutation probability p_v for a certain position of code x_j to mutate

$$\begin{cases} b'_f = \alpha b_f + (1-\alpha) b_m \\ b'_m = (1-\alpha) b_f + \alpha b_m \end{cases}$$
$$x_j = U(x_1, x_2) \tag{18}$$

Where: $U(x_1, x_2)$ represents a random number in the interval (x_1, x_2).

Repeat (2) ~ (5) until the termination rule is met, and finally select the optimal solution t^* of O(t) corresponding to the genetic code of the chromosome with the largest fitness in the population. In this article, set the maximum number of iterations T as the termination rule of the genetic algorithm, and$\alpha = 0.8$, $p_v = 0.05$.

2.2. Genetic algorithm optimization support vector machine algorithm flow

The prediction accuracy of support vector regression largely depends on the kernel function and the selection of its parameters. The introduction of the kernel function avoids complicated calculations. The change of its form and parameters will implicitly change the mapping relationship of the slave function, and then change the complexity of the mapping feature space, thereby affecting the performance of the support vector regression machine. The key parameter of the RBF kernel function is σ, the width of the kernel, which controls the radial range of the function. The penalty parameter C acts as a balance between model complexity and training error. The larger the value of C, the greater the penalty for data exceeding the loss function, which affects the promotion ability of the model. The loss function ε represents the approximation accuracy of the training data and the real data. If the value is too small, the regression accuracy is higher, but it may cause the model to be too complex and poor promotion ability; if it is too large, the model is simple, resulting in insufficient learning accuracy. At present, there is no definite theory to guide the parameter selection of the support regression machine. For this problem, the parameter selection is regarded as a multivariate combination optimization problem.

The basic idea of genetic algorithm for support vector machine regression optimization is: introducing the principle of biological evolution into the coded concatenation group formed by the optimization parameters (C, σ, ε), and according to the selected fitness function, through the selection, crossover and mutation in heredity, the individual is iterated until the termination condition is met, and the purpose of intelligent optimization is achieved.

The basic flow of the algorithm optimization support vector machine regression prediction model is as follows:

(1) Obtain housing price data and preprocess it to determine the training sample set and the test sample set.

(2) Set genetic algorithm parameters and encode support vector regression parameters.

(3) Calculate the fitness of each individual.

(4) Determine whether the iteration conditions are met. If it is not satisfied, perform selection, crossover and mutation operations.

(5) Construct a support vector regression prediction model, and substitute the parameters finally obtained by genetic algorithm optimization into the support vector machine prediction model for simulation prediction.

3. Establishment of housing price prediction model of support vector machine based on genetic algorithm optimization

3.1. Establishment of housing price prediction model

In order to be able to accurately and accurately reflect the changing trend of housing prices, this article takes full advantage of the various factors that affect housing prices, including the average sales price of residential commercial housing, the investment in real estate development, the area of completed housing, the per capita disposable income of urban residents, and the town at the end of the year. Variables such as population, commodity retail price index, average sales price of residential commercial houses, per capita disposable income of urban residents, urban population at the end of the year, housing investment in real estate development, and completed residential housing area are used as housing prices. This article uses Harbin City’s urban residential and commercial housing data from 2000 to 2017. The data comes from the "China Statistical Yearbook" and "Harbin Statistical Yearbook".

This paper uses the RBF kernel function to establish a support vector regression prediction model. The parameter settings of the model are as follows: select ε-SVR, the value range of the penalty factor C is [0,100], the value range of the RBF kernel function parameter σ is [0,1000], and the maximum evolutionary algebra of the genetic algorithm is 120， the maximum number of the population is 30, the crossover probability is 0.4, the mutation probability is 0.01, and the cross validation parameter is set to 5.

3.2. Simulation results

In this paper, a total of 13 sets of normalized data from 2000 to 2012 are used as the sample training set data, and a total of 5 sets of data from 2013 to 2017 are used to predict housing prices through the trained model. Support vector regression and genetic algorithm optimization support vector regression to predict the value of housing prices are shown in Table 1.

Tab 1 Housing price forecast results

years	Actual value /hundred yuan	GA-SVR	error/ %	SVR	error/ %
2013	58.84	57.97	1.5	56.45	4.1
2014	57.51	56.01	2.6	60.01	4.3
2015	61.24	63.05	3.0	65.13	6.4
2016	63.38	65.18	2.8	60.5	4.5
2017	78.61	76.52	2.7	74.39	5.4

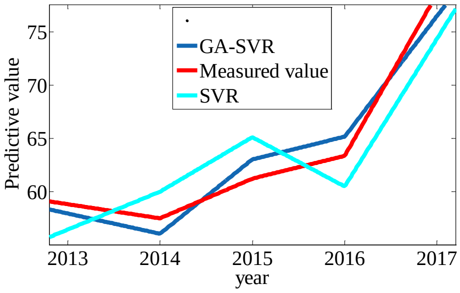

Fig 1 Graph of house price prediction results

According to the prediction results of the model in this paper, it can be clearly seen from Table 1 that the highest error of the support vector regression prediction through genetic algorithm optimization is 3.9%, the lowest is 1.5%, the average error is 2.52%, and the highest error of the support vector regression model is 6.4%, , the lowest is 4.1%, and the average error is 4.94%. Use genetic algorithm to optimize the penalty parameters, kernel function parameters and insensitive loss function of the support vector regression model, and use the optimized parameters to establish the support vector regression prediction model. the prediction accuracy of the optimized model is higher than that of a single support vector regression prediction model.

4. Conclusion

1) By combining genetic algorithm and support vector regression, using optimized parameters to establish a support vector regression prediction model, and verifying housing prices with examples, the results show that the prediction model has good prediction accuracy. The prediction model in this paper is suitable for the prediction of housing prices.

Compared with the support vector regression model optimized by genetic algorithm, its convergence speed and prediction accuracy have been greatly improved, the genetic algorithm optimized support vector regression model can predict housing prices more accurately, providing a new reference method for housing price prediction.

References

[1] Sun Shanshan. Real estate price prediction based on data mining [J]. Modern Electronic

Technology, 2017(05):134-137.

[2] Liu Feng, Zhang Xing, Zhang Guangfeng. Modeling and analysis of variable coefficient regression model of housing prices in Chongqing［J］. Journal of Chongqing University of Technology: Natural Science Edition, 2014, 28(4): 150 -154.

[3] Gao Yuming, Zhang Renjin. Housing price prediction analysis based on genetic algorithm and BP neural network [J]. Computer Engineering, 2014, 40(4):187-191.

[4] Chen Shipeng. Housing price prediction based on random forest model [J]. Science & Technology Innovation and Application, 2016(4):52-52.

[5] Wang Zhihai. Prediction of second-hand housing transaction price in Beijing based on GA-BP model [D]. Hebei University of Technology, 2019.

[6] He Ling. Research on the price of commercial housing in Harbin [D]. Harbin: Harbin Institute of Technology, 2010.