Transaction price prediction of second-hand houses in Wuhan based on GA-BP model

Can Peng[1, #, *], Heng Xiao[2, #], Kaiman Ou[3, #]

[1]School of Finance and Taxation Southwestern University of Finance and Economics Chengdu, China

[2]School of Finance Guangdong University of Foreign Studies Guangdong, China

[3]School of data science Guangzhou Huashang College Guangzhou, China

*Corresponding author. Email: pengcan220328@163.com

[#] Equally contributed to this works

Abstract. The significance of the used housing market to the stability of people's lives and social development is increasing, and the prediction of its prices is becoming a key concern for society. In this paper, BP neural network model is used to predict the price of second-hand properties. In view of the disadvantages of slow convergence and the tendency to obtain local optimal solutions, the BP neural network model's input layer is optimized by a genetic algorithm to speed up the convergence speed and accuracy of the BP neural network model. The experimental results show that the improved GA-BP neural network has high prediction accuracy, with RMSE and MAE of 398.72 and 170.18, respectively. The difference between the actual and predicted values is small, which can provide a more stable tool for predicting house prices in the second-hand property market and broaden a new channel for the practical application of GA-BP neural network.

Keywords: GA-BP neural network; Second-hand house prices; Predictive models.

1. Introduction

Real estate is an important part of China's economy and has long occupied a dominant position in the market economy. With the real implementation of the "no speculation in housing" and "three red lines" policies, the trend of property prices has received widespread attention from society. However, high property prices are still daunting and the supply of real estate is limited due to limited land resources. As a result, second homes are gradually becoming the object of trade and the importance of the second home market in stabilizing the lives of residents and social development is increasing, making effective forecasting of second home prices essential.

The three traditional valuation methods used in financial estimation of second-hand property prices, namely the market comparison method, the cost method and the income method, rely on the appraiser's experience and are subjective and require specific analysis of a property's information, and have no universal predictive significance. Using some mathematical models, deep learning and machine learning methods to obtain prediction levels by processing the overall market data is becoming the mainstream method of house price prediction.

Sun Bo, Luo Zhikun[1] achieved the prediction of the house price trend in Harbin based on the GM (1, 1) model. However, since the discrete fitting equation legally established by the gray differential fitting is a rough distinction equation, it is hard to ensure that the fitted equation is rigorously similar to the fitted system's differential equation. Moreover, it is also impossible to ensure that the inherent error of the grey model established is an infinitesimal amount, resulting in prediction errors. Yang Mu Hei [2] used the random forest method as a regression prediction analysis of the characteristic price model to establish a second-hand house appraisal model, and used it for a certain area of Guangzhou, with high accuracy. However, for data with attributes with distinct values, attributes with more value departments will have a more significant effect on the random forest, so the attribute weights produced by the random forest on such data are unreliable. Tang, Xiaobin, Zhang, Rui and Liu, Lixin [3] used the bat algorithm to optimize the parameters in the SVR model to improve the generalization ability of the SVR model, and achieved better results in the prediction of second-hand house prices in Beijing. However, this model algorithm is difficult to implement for large-scale training samples.

In recent years, with the gradual development of support vector machine models and artificial neural network models, these models have shown excellent performance on large data. In addition, more and more scholars have started to try to improve them and apply them to the prediction of housing prices. Shanshan Wu[4] established a BP neural network model with the real estate price of Nanjing city as a reference and used the model established from 2001 to 2013 to predict the house price from 2014 to 2016 compared with the real data from 2014 to 2016. The model prediction error was smaller and better in the short term, but the long-term prediction error was larger.

As BP neural networks have the defects of slow convergence and tendency to fall into local minima, many scholars have tried to combine other models to optimize BP neural networks to achieve more accurate estimation.

Gao Wen [5] collected 20 years of data from 1997 to 2016 on the impact of seven on house prices in Xi'an city, firstly pre-processed the data, then used the PCA method to reduce the dimensionality of the data, and then did prediction analysis of Xi'an house prices based on BP neural network, the shortcoming is that the data is old and has little reference value for predicting the current house prices. Ding Feiji [6] improved the house price prediction by improving the lion swarm algorithm and BP neural network model.

In summary, predicting house prices is a typical non-linear problem, and BP neural networks are particularly suitable for predicting non-linear variables such as house prices because of their strong mapping ability. At the same time, this paper adopts an optimization algorithm combining genetic algorithm and BP neural network. In the learning and training process of BP neural network, the genetic network algorithm is used to optimize its weights and thresholds, and then the prediction range is narrowed, trying to solve the shortcomings of slow convergence speed and easy falling into local optimal solution of BP neural network. At present, the second-hand housing market is an important part of China's real estate market, but there are few studies on house price prediction related to the forecast of second-hand housing prices. Therefore, this paper applies the GA-BP model to the prediction of second-hand house prices and optimizes second-hand house prices, adding a new path for house price prediction, which has very great practical significance in the second-hand house market.

2. Data pre-processing

In this paper, a total of 5,417 pieces of data on second-hand properties in Wuhan from 2019 to 2021 were collected through the official website of Chain Home, and 5,352 pieces of valid data were obtained through data cleaning, missing value processing and outlier processing.

The factors affecting the sales price of houses are complex, ranging from macro-level factors such as policies and economic conditions to micro-level factors such as house size and geographical location, which all impact the sales price of houses. This paper does not cover the impact of macro-level factors on house prices, but only examines second-hand house transaction prices at the micro-level. Three major characteristic variables are selected, namely location characteristics, architectural characteristics and transaction indicators [7].

The building characteristics include floor area, internal area, building type, house type, house orientation, decoration, floor level, year of construction, proportion of staircase, lift, house structure and building structure. The transaction indicators include transaction ownership, house use, house age, mortgage information and ownership.

The number of bedrooms, living rooms, kitchens and bathrooms was also quantified and standardized, and the number of variables used in the construction of the model was 122.

3. Model building

Because of the random nature of the generation of both the weights and thresholds of the BP neural network, there is no way to intervene in the process. If the final prediction result is satisfactory, the program can only be repeated, but there is no guarantee that the prediction result will be in a better direction. So we improved the model using a genetic algorithm, a kind of evolutionary merit-seeking algorithm. This algorithm takes the trained BP neural network prediction results as an individual fitness function, and finds the optimal value of the function through selection, crossover and variation operations. On this basis, a GA-BP model is constructed, and the threshold and weights of BP neural network are optimized by using genetic algorithms, so as to optimize the performance of BP neural network model..

3.1. BP neural network building

The implicit layer activation function f is set to the Relu function [8].

$$\mathrm{Re}lu(x) = \begin{cases} x & x > 0 \\ 0 & x \le 0 \end{cases} \tag{1}$$

Set the figure of nodes in the input layer l, the figure of nodes in the concealed layer m, the number of nodes in the output layer n, initialize the connection weight ω_{ij} of the input layer and the hidden layer and the threshold a of the hidden layer, and set parameters such as the number of iterations and the learning rate [9]. The sample data is input and processed layer by layer by the implicit layer, and the implicit layer is output.

$$H_j = f(\sum_{i=1}^{l} \omega_{ij} x_i - a_j) \quad j = 1, 2, \cdots, m \tag{2}$$

Initialize the weights ω_{ij} between the output layer and the implied layer and the output layer threshold b, and calculate the predicted output.

$$O_k = \sum_{j=1}^{m} H_j \omega_{jk} - b_k \quad k = 1, 2, \cdots, n \tag{3}$$

The measurement of the model test error can be performed according to Eq.

$$MAPE = \frac{1}{n} \sum_{k=1}^{n} \left| \frac{Y_k - O_k}{Y_k} \right| \times 100\% \tag{4}$$

Where Y_k is the true value of the d5ependent variable on the test set and O_k is the model's predicted value on the test set.

3.2. Genetic algorithm settings

Genetic algorithm problem solutions were set between [T,1], encoded by the floating-point method, and the gene values were obtained in the maximum-minimum interval of the training features and set [10].

The test error is calculated as the fitness function value, and the fitness value is used to judge the solution's superiority.

$$F = k(\sum_{t=1}^{n} abs(y_t o_t)) \tag{5}$$

Where y_t is the true value of the ith node of the network and o_t is the predicted output.

In the evolutionary process, roulette selection is used as the selection operator, the crossover operator is set to two-point crossover, the crossover probability, and the variation probability are set, the crossover operator is used cyclically, and finally, the optimal solution with initial weights and thresholds is obtained.

After training, the neural network is able to memorize the weights and thresholds to minimize the test error and thus achieve optimal prediction.

3.3. GA improved BP model building

In this paper, the genetic algorithm is built based on custom development in Python language, and the artificial intelligence algorithm framework PyTorch, which supports the python language, is used to build BP neural networks. Comparing the GA-BP model with the traditional BP neural network model, it can be seen that the former network weights and thresholds obtained by genetic algorithm optimization can make the estimated performance of the BP neural network better to some extent. GA-BP model The process framework is shown in Figure 1.

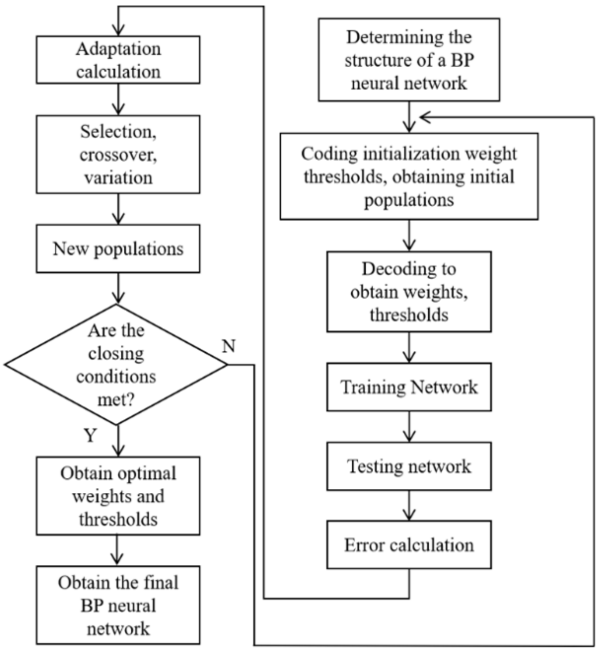

Figure 1 Structure of the GA-BP model

Figure 1 illustrates the basic execution process of the GA-BP model, which can usually be divided into the following steps.

(1) Determine the structure of the BP neural network. This process mainly determines the number of layers of the hidden layers of the BP neural network and the number of nodes in each hidden layer.

(2) Determine the fitness function. The primary purpose of this process is to calculate the fitness value of the chromosomes in the population and provide a basis for the subsequent genetic algorithm to perform genetic operations.

(3) Initialisation of the population. The genetic algorithm randomly generates a population based on the neural network's structure. Each individual in the population is encoded as a binary string, which contains all the information about the weights and thresholds of the neural network. These binary strings are the first individuals to participate in the training of the BP neural network model, and the error between the predicted and desired data is used as a fitness function.

(4) The fitness value is calculated for the individuals in the initial population based on the test error, then the genetic operation is applied to the individuals based on the fitness value to produce a new population, and the new population is decoded and substituted into the BP neural network model again for the fitness value calculation, cycling through the selection, crossover, variation and fitness calculation operations until the termination condition is met or the maximum number of iterations is reached, and the genetic algorithm stops.

(5) Output the final optimal individuals and decode them; decoding the most will get the best initial weights and thresholds to build the BP neural network.

4. Empirical Analysis

This section provides a detailed description of the model settings involved in the BP neural network and the genetic algorithm of the GA-BP model.

4.1. Model settings

Setting of BP neural network

(1)Input and output indicators for the network

After quantifying the data, the number of feature variables used to build the model in this paper is 122, and these variables will be used as the input indicators of the BP network, so the number of neuron nodes in the input layer of the BP network is set to 122 nodes; secondly, the research object of this paper is the transaction price per square meter of second-hand properties in Wuhan, i.e., the transaction price is used as the output indicator of the BP network, so the output layer of the BP network The number of neurons is set to 1 node.

(2)Implicit layer setting

The Relu activation function used in this paper has good predictability for deep neural networks, and the number of hidden layers is set to 3.

(3)Number of implied layer nodes set

In this paper, the number of nodes in the hidden layer is determined empirically: assuming that the number of nodes in the input layer is n, then the number of the hidden layer nodes is $2n+1$. The input layer in this paper has 113 input nodes. Based on this empirical formula, the final number of nodes in the three hidden layers is set to 227, 227, and 113 nodes, respectively.

(4)Other parameter settings for BP neural networks

In this paper, we set 10,000 iterations of the BP neural network, and the learning rate is 0.001.

Setting of the genetic algorithm

This paper sets up the genetic algorithm according to Table 1.

Table 1: Genetic algorithm setting table

Settings	Set values
Solution sets	[T,1]
Coding method	Floating-point encoding
Adaptation function*	Test error
Number of iterations	100 times
Population size	300
Number of populations	1pc

* The final choice of test error as the value of the fitness function is based on two considerations: firstly, the test error in this paper can be calculated quantitatively based on the division of the dataset; secondly, if the training error is chosen as the value of the fitness function, it will lead to overfitting of the final model, thus making the model not have excellent generalization ability.

In addition, the gene values in the algorithm are taken within the limits of the given interval, and the maximum and minimum values of the interval training feature set $[Max,Min]$.

(5)Setting of genetic manipulation

①Selection operator: The selection operator is set to roulette selection. The probability of selection is divided into individual and cumulative probabilities, and a random number is generated between [0,1], and the individual in the cumulative probability interval of is selected.

②Crossover operator: The crossover operator is set to a two-point crossover with a crossover probability of 0.8.

③Variation operator: variation probability of 0.01.

The above genetic operations start by determining the domain of feasible solutions and the encoding method according to the specific problem, representing each feasible solution in the domain in the form of a numerical string or string; Construct a fitness function to measure each solution, which is a non-negative function; Determine the size of the population, the mode of selection, crossover and variation, the probability of crossover and variation, and determine the termination conditions.

4.2. Model training

In order to make the model more generalizable, the paper divides the 5351 data records into two parts; the training set and the test set, corresponding to 4828 and 523 pieces of data respectively. In this paper, the training set is used to train the BP neural network part of the model, and the test error value of the BP neural network is calculated according to the test set data, which is also the value of the fitness function of the genetic algorithm in the improved model. Figure 2 shows the training results of the improved model.

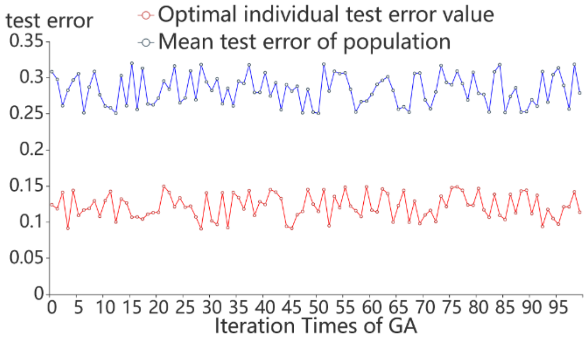

Figure 2 Training results of GA-BP model

As can be seen from Figure 2, the optimal individual test error and the mean value of the test error of the population tend to decrease as the number of evolutionary generations increases, indicating that the genetic algorithm does optimize the performance of the BP neural network. Further, the optimal individual test error value reaches a minimum value of 0.0908 when the genetic algorithm is iterated for the 28th time. This means that we have reached the optimal solution, i.e., the BP neural network weights and thresholds have reached the optimal state, and this value reflects the good fit of the model on the training set.

4.3. Model comparison

Differences in forecast results

In order to compare the differences in the effectiveness of different models, we compared the accuracy of three methods, GA-BP model, multiple linear regression model and BP neural network model, by testing the same data, and the final accuracy and error of the three methods obtained are shown in Table 1.

Table 1 Comparison table of accuracy rates

	R^2	RMSE	MAE	MAPE
MR	0.9259%	1945.90	1200.38	7.0069%
BP	0.9970%	1801.56	1573.53	9.3865%
GA-BP	0.9970%	398.72	170.18	0.9762%

$$R^2 = \frac{ESS}{TSS} \tag{6}$$

R^2 shows the fitting degree of the predicting model.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2} \tag{7}$$

RMSE is used to measure the direct deviation between the actual and predicted values.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i' - x_i| \tag{8}$$

MAE is the sum of the absolute value of the difference between the predicting value and the actual value.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i' - x_i}{x_i}\right| * 100\% \tag{9}$$

MAPE is a relative value rather than absolute value, so it is meaningless to look at the size of MAPE alone.

The formula of each index is shown above. As can be seen from the table, among the three methods, the GA-BP model clearly has a higher accuracy rate for predicting second-hand house prices, so the GA-BP model outperforms the other two methods on this dataset.

Conclusions of the model comparison

In order to improve the accuracy of prediction for second-hand house prices, this paper proposes a prediction method based on a modified BP neural network model with a genetic algorithm. Through analysis and comparison, the following conclusions are drawn.

(1) There are quite a number of factors influencing the price of second-hand properties, and there are many missing values for the relevant variables, making it difficult to make accurate predictions. Therefore, this paper first uses interpolation and logistic regression to supplement the missing values of the data, and then uses genetic algorithms to optimize the original data. The results show that the improved algorithm can better accelerate the convergence of the model.

(2) Compared to the multiple regression model and the BP neural network model, the improved GA-BP neural network-based model has a higher prediction accuracy, with less deviation of actual values from predicted values. Its RMSE, MAE and MAPE are 398.72, 170.18 and 0.9762%, respectively, all of which are the lowest values of the corresponding errors of the three models, with the multiple regression model having the worst results were obtained from the multiple regression model.

(3) The improved GA-BP model still has some errors in the case of outliers in the second-hand property prices. However, regarding prediction accuracy and model applicability, the improved GA-BP model has shown better results in predicting second-hand property prices, and can be used as a reference model for relevant practitioners.

5. Conclusion

In this paper, 122 indicators were selected as factors affecting the transaction prices of second-hand houses in Wuhan from three levels: location characteristics, architectural characteristics and transaction indicators from a micro perspective.

A GA-BP model is constructed in this paper, and the GA algorithm optimizes the weights and thresholds of the BP network. The test error on the final prediction set is 0.0098, which means that the GA-BP model achieves a more accurate prediction for the transaction price of second-hand properties in Wuhan. It futher indicates that the improved BP model has a better performance than the traditional BP network model and other conventional models in prediction.

Reference

[1] Sun B. and Luo Z., " Forward-looking house price trend in Harbin based on GM(1,1) model ", Journal of Harbin University of Commerce (Social Science Edition),vol.01,Jan.2017,pp.108-113.

[2] Yang M., " A Study of Second-House Price Evaluation Based on Random Forest Model ", unpublished.

[3] Tang X., Zhang R. and Liu L., " A Study on Second-House Price Forecasting in Beijing Based on Bat Algorithm SVR Model ", Statistical Research,vol.11,Nov.2018,pp.71-81, DOI:10.19343/j.cnki.11-1302/c.2018.11.006.

[4] Wu S., " BP Neural Network Based House Price Forecasting in Nanjing ", Market Week, vol.02,Feb.2019,pp. 58-60.

[5] Gao W., " A Study of BP Neural Networks for House Price Prediction Based on Genetic Algorithm Optimization ",unpublished.

[6] Ding F. and Jiang M. " House Price Prediction Based on Improved Lion Swarm Algorithm and BP Neural Network Model ", Journal of Shandong University (Engineering Edition), vol.04, May 2021, pp.8-16.

[7] Jie C., " A Study of Property Valuation and Property Tax Fairness ", Fiscal Studies, vol.08, Aug.2018, pp. 76-92, DOI:10.19477/j.cnki.11-1077/f.2018.08.007.

[8] Dong Z., Xu W., Zhuang H. and Qiu G., " Deep Learning BP Neural Networks for GNSS Level Fitting ", Marine Mapping, vol.05, Sep.2019, pp.26-29.

[9] Bai B., Zhu H. and Fan Q., " Application of BP Neural Networks in Dairy Product Quality and Safety Risk Warning " ,vol.07, Jul.2020, pp. 42-45+57, DOI:10.19827/j.issn1001-2230.2020.07.009.

[10] Song J., Ren L. and Li Z., " GA-BP Neural Network Based Urban Residential Price Forecasting Method " , Modern Business, vol.06, Feb.2020, pp.13-16, DOI:10.14097/j.cnki.5392/2020.06.005.