Automation and Machine Learning (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2023.040103
ISSN 2516-5003 Vol. 4 Num. 1

House Price Forecasting in Ames Based on Bayesian Regularized BP Neural Network

Haiqing Bai[1,a], Xiaoyong Chen[2,b]

[1]School of Computer Science & Engineering Artificial Intelligence, Wuhan Institute of Technology, Wuhan, 430205, China
[2]School of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing, 210037, China
[a]g.raser0001@gmail.com, [b]2531037370@qq.com

Keywords: Bayesian regularization, BP neural network, house price prediction

Abstract: Housing has always been an important issue related to the national economy and people's livelihood. House prices not only affect people's welfare, but also have a significant impact on the national economy and social stability. Therefore, the prediction of house prices is also a necessary means. This paper forecasts the housing price data of Ames City in the United States, and establishes a Bayesian regularized BP neural network model to solve the nonlinear mapping relationship between housing prices and indicators. Through curve fitting analysis of samples, the overall R value is 0.9667, the overall result is better, and the error histogram also conforms to normal distribution. The experimental results show that the BP neural network model based on Bayesian regularization is very effective in dealing with the problem of house price prediction, and can better analyze and predict the trend of house price. This study provides a basis for real estate developers to develop and position products, and also provides model support for buyers to better judge the real price of houses.

1. Introduction

Housing satisfaction is an important factor for people to obtain happiness. In recent years, with the increase of people's demand for housing, the building area of housing in China also shows an increasing trend, and more people choose to buy houses in working cities for different reasons. Based on the results of big data analysis, house price forecasting research plays a great role in the evaluation of house selection value. So far, scholars from all over the world have studied the problem. For example, Chinese scholars Wu Xiuli and Zhang Feng used the time series analysis method to establish a prediction model based on the housing price data of several representative administrative districts in Guangzhou, and tested the error through residual analysis. The predicted value basically coincides with the actual observation value, achieving the purpose of prediction[1]; Wang Dongxue and Guo Xiujuan used XGBoost algorithm to model and train the data set in 2021 for the characteristic housing price data[2]; Internationally, Quigley uses parallel data regression analysis to explain the trend of housing prices in various cities with relevant indicators of economic fundamentals; Potepan adopts the two-stage least squares method to predict the prices of different cities.

Due to the highly nonlinear mapping relationship between house prices and the impact indicators of house prices, and the large data volume of house prices and their indicators in a certain area, compared with the traditional multiple linear regression prediction model, the nonlinear mapping relationship of neural network can deal with the nonlinear problem of house price prediction. In order to improve the accuracy and fitting degree of BP neural network prediction, this paper studies the improved BP neural network based on Bayesian regularization. The application of this model will help practitioners predict the house price, and also provide model support for buyers and potential buyers to effectively purchase the ideal residence, so that buyers can truly understand the real value of the house and judge whether the current house price is reasonable.

2. Data preprocessing

2.1 Data source and characteristics of data

Realizing the BP neural network model to predict house prices requires a large number of data training models. Because it is difficult to collect house price data in China and there are some non-quantifiable factors, the data in this study comes from the authoritative data mining website Kaggle, and uses the data of Ames City, central Iowa, USA (quoted data) to verify the effectiveness of this method. The data set used in this study contains a total of more than 1400 samples, including residential type, block frontage, block area, street, overall quality, overall conditions and construction year and other evaluation indicators with great reference value. Some key data samples of the dataset are shown in Table 1.

Table 1: Data samples

MSSubClass	LotFrontage	LotArea	Street	OverallQual	OverallCo	Yearbuit
0.06732	-0.18444	-0.2178	0.646073	-0.50719	1.046078	0.896679
-0.8734	0.458096	-0.0720	-0.06317	2.187904	0.154737	-0.39553
0.06732	-0.05593	0.137173	0.646073	-0.50719	0.980053	0.848819

2.2 Data normalization

Due to the different dimensions of each indicator, the features with higher values in the training process may have a greater impact on the network weight, weakening the impact of the features with smaller values on the results. Therefore, at the beginning of the study, it is necessary to normalize the data of the indicators to avoid causing large errors. Specifically, normalization is to map the data to the range of 0 to 1 for processing, which plays the role of transforming the dimensional expression into the dimensionless expression, so that the indicators of different units can be operated and weighted, and the overall operation can be simplified[3]. The normalization formula used in this study is shown:

$$y = \frac{x - E[x]}{Var[x] + \in} * \gamma + \beta \tag{1}$$

2.3 Visualize the correlation between numerical variables

Due to the possible correlation between variables, which may cause the weight of variables to change, the correlation test must be carried out before the formal experiment, and the correlation thermodynamic diagram between variables can be obtained through python operation, as shown in Figure 1 above. The more the color tends to dark blue, the stronger the correlation, and vice versa. It can be seen from the depth of the blue area in the figure that most variables have weak correlation and do not need data cleaning.

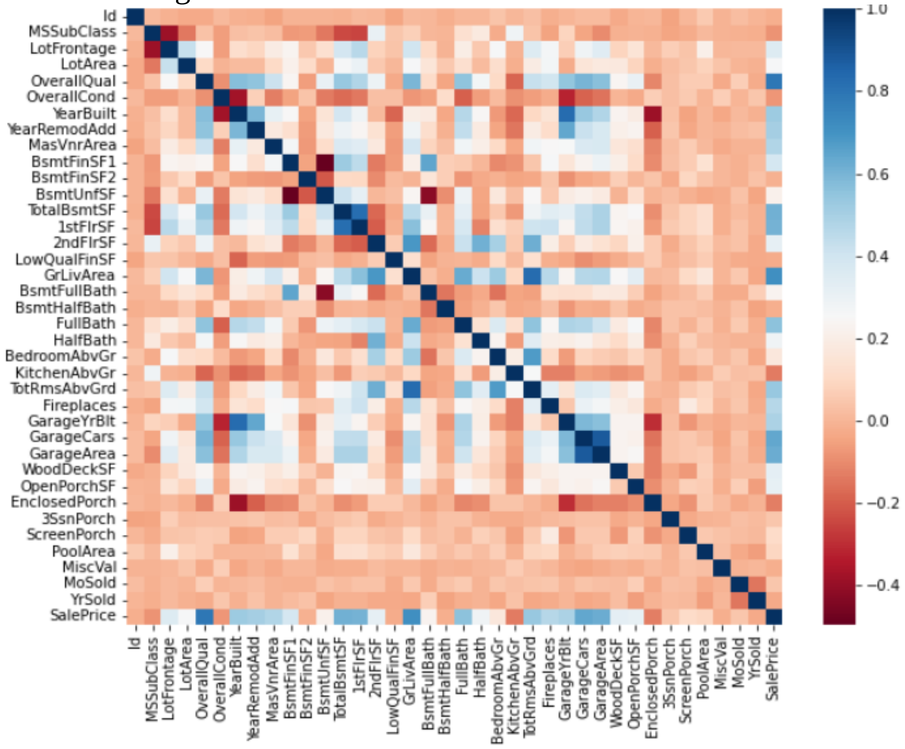


Figure 1: Variable heat map

3. Model building

3.1 Predictive model structure building

A BP neural network is an artificial neural network that was proposed in 1986 by Rumelhart and McClellend[4] and consists of multiple neuron nodes, each with a weight that can be adjusted by a back-propagation algorithm to achieve the best results.BP neural networks can be used for a variety of applications such as speech recognition, image recognition, natural language processing, etc. It has the advantage that it can learn automatically to solve complex problems and the weights can be adjusted to make the network more flexible[5]. The BP neural network-based prediction model constructed in this paper uses a non-linear logarithmic Sigmoid function for the activation function of the nodes in the hidden layer, while a linear purelin function is chosen for the activation function of the nodes in the output layer.

In this paper, a classical three-layer BP forward network structure with one hidden layer is used, in which the neurons within the layers of the neural network are not connected to each other, while the neurons between adjacent layers are all connected. Based on the characteristics of the data samples, the number of neurons in the input layer is determined to be 331 and the number of neurons in the output layer is 1, i.e. the output variable is the house price corresponding to the input features, as there are 331 valid features.

The choice of the number of nodes in the hidden layer has an important impact on the accuracy and speed of data training: if the number of nodes is too small, it will lead to an undertrained neural network and the training results will not reach the expected accuracy; while if the number of nodes is too large it will result in low overall efficiency and overfitting. From the number of implied layers refer to the empirical formula (2):

$$l = \sqrt{n+m} + q \tag{2}$$

where l, n, m and q represent the number of nodes in the hidden layer, the number of nodes in the input layer, the number of nodes in the output layer and a constant greater than 1 and less than 10, respectively. As the number of input layer nodes is 331 and the number of output layer nodes is 1, the number of nodes in the hidden layer is recorded as 21 in consideration of accuracy and efficiency.

3.2 BR-BP neural network structure

The neural network was constructed by matlab, the maximum number of iterations of the neural network was set to 1000 and the maximum number of non-decreasing steps was 6. The data set was divided by random method, the neural network was trained by Bayesian regularization algorithm and the model performance was measured by MSE. The flow chart of the bp neural network algorithm for Bayesian neural network optimisation is shown in Fig. 2.

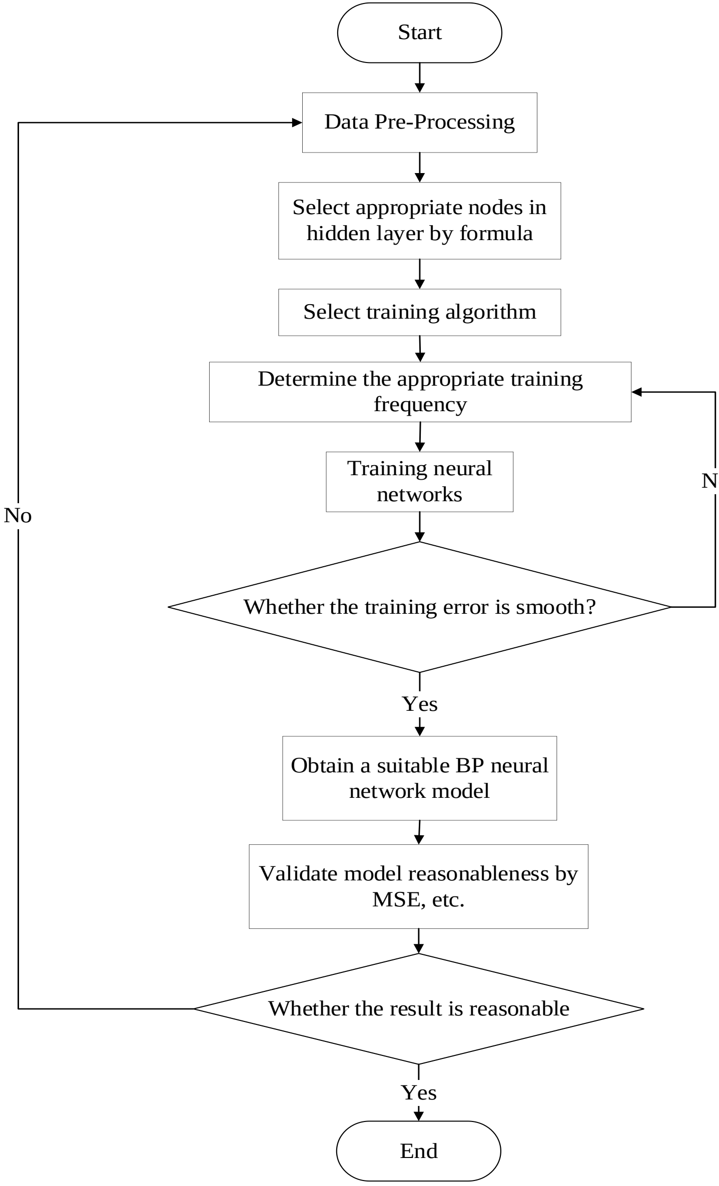

Start
Data Pre-Processing
Select appropriate nodes in hidden layer by formula
Select training algorithm
Determine the appropriate training frequency
Training neural networks
Whether the training error is smooth?
No
Yes
Obtain a suitable BP neural network model
Validate model reasonableness by MSE, etc.
Whether the result is reasonable
No
Yes
End

Figure 2: Process of BR-BP Neural network

4. Solving the model

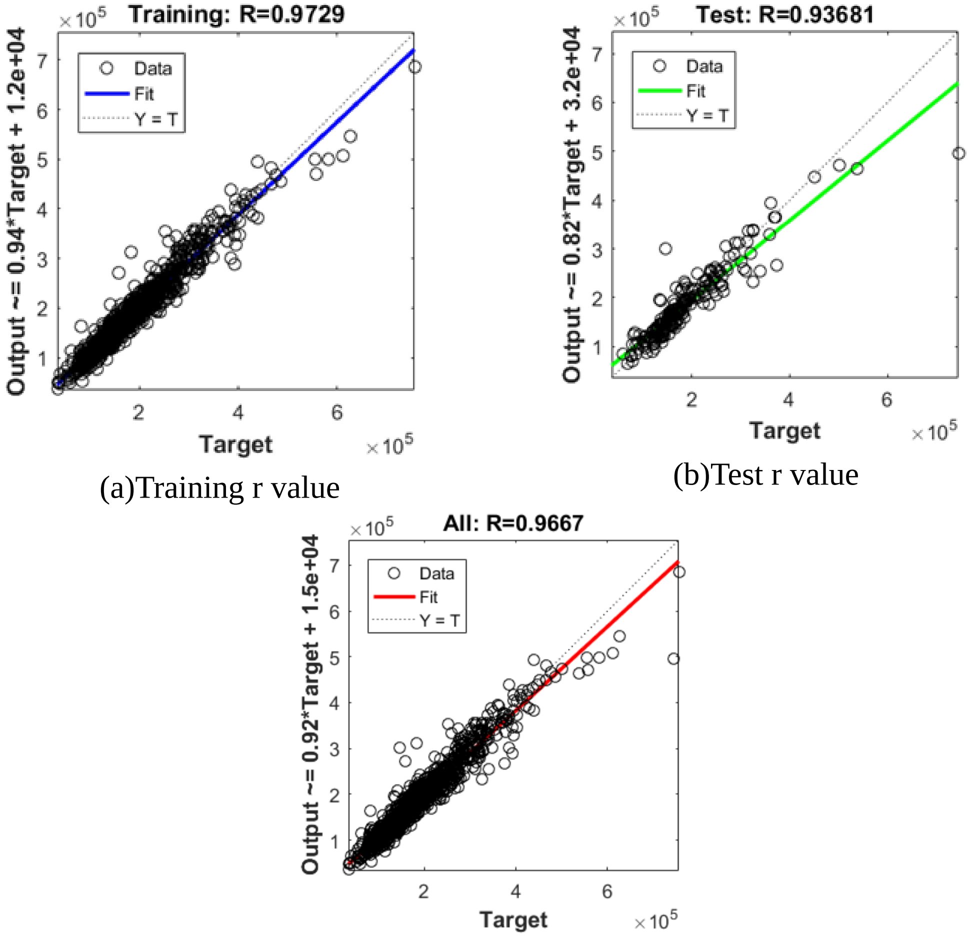
4.1 Brief description of the output

This study verifies the reliability of the prediction results by using R coefficient, mean square error, and error histogram. r coefficient is a statistical indicator used to measure the correlation between two variables, which takes values from -1 to 1, where -1 indicates a completely negative correlation, 1 indicates a completely positive correlation, and 0 indicates no correlation. The mean squared error is used to measure the difference between the predicted and actual values, where a smaller value means the difference between the predicted and actual values is smaller.

The error histogram measures the difference between the predicted and actual values. It divides the difference between the predicted and actual values into certain intervals, and then counts the number of differences within each interval, so that the distribution of the differences between the predicted and actual values can be clearly seen.

4.2 Analysis of the model solution

The results obtained R coefficients, mean squared errors, and error histograms are shown in Figure 3 in (a), (b), (c) and Figures 4 and 5. The samples were curve-fitted, and the R values in the training set and test set were 0.9729 and 0.93681 from the analysis in Figure 3, and the overall R value was 0.9667, which was a better overall result.



(a)Training r value

(b)Test r value

(c)Overall r value

Figure 3: R values

From the mean square error variation graph in Fig. 4, we can observe the degree and trend of error decline in the training and testing processes. In general, the error in both processes decreases as the number of iterations increases and starts to level off at the 35th Epoch, reaching the optimal value at the moment of termination of training. The error histogram shown in Fig. 5 conforms to the normal distribution, which indicates that the collected data and the proposed model method are both highly reasonable.

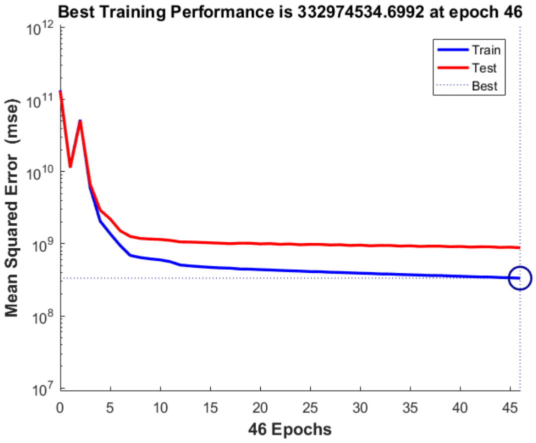


Figure 4: MSE chart

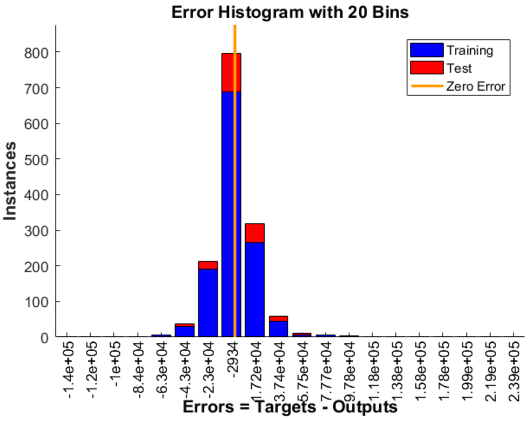


Figure 5: Error Histogram chart

5. Conclusion

In this paper, the BP neural network was optimized by Bayesian regularization and then used in the field of house price prediction, and the reliability of the results was verified by R-coefficient, MSE and error histogram. This paper analyzes the prediction results based on Ames city house price data, and then further evaluates the prediction results. It is shown that the improved neural network can reasonably predict the trend of house attributes on house price changes with higher accuracy.

In addition, this study can also be improved by further analyzing the trend of influencing factors through quarterly data collection to analyze the current property market changes from a temporal perspective. Similarly, the analysis of house price influencing factors can also be performed by combining integrated learning with a modified BP neural network.

References

[1] *Wu Xiuli, Zhang Feng. Application of time series analysis in house price prediction —Take the data of Guangzhou*

city as an example [J]. Science, Technology and Engineering, 2007,7 (21): 5631-5635.

[2] Wang Dongxue, Guo Xiujuan. House-price prediction model based on the XGBoost algorithm [J]. North Building, 2021,6 (3): 79-82.

[3] Zhang Baichuan, Zhao Baiting. Lightweight convolutional neural network classification algorithm combined with batch normalization [J]. Journal of Harbin University of Commerce (Natural Science Edition), 2021,37 (3): 300-306.

[4] Wu Mingshan, Wang Bing, Qi Yaning, Zheng Piao. Research on cigarette sales portfolio forecasting model [J]. (Chinese Journal of Tobacco, 2019, 25(03): 84-91. doi:10.16472/j.chinatobacco.2019.039.)

[5] Sun T T, Shen Y, Zhao L. A House Price Forecasting Model Based on BP Neural Network[J]. Computer Knowledge and Technology, 2019.