International Journal of u- and e- Service, Science and Technology
Vol.8, No. 7 (2015), pp.109-118
http://dx.doi.org/10.14257/ijunesst.2015.8.7.11

Study on the Prediction of Real estate Price Index based on HHGA-RBF Neural Network Algorithm

Huan Ma[1], Ming Chen[1] and Jianwei Zhang[1]

[1]*Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450002, China*
songge19840416@163.com

Abstract

The traditional error of the back-propagation algorithm multilayer feed-forward network (BP neural network), there are the flaws of a slow convergence of forecast, getting local minimum solutions easily, and forecast accuracy rate is not high. This paper proposes a new approach which is the combination of hierarchical genetic algorithm and least squares method to optimize the RBF neural network such that we can predict the real estate price of the Real estate Price Index. And which overcomes the shortcomings of traditional Fourier analysis, has good localized characteristics in the time domain and frequency domain, and has important value. In signal processing, image processing, voice analysis and other fields. The hierarchical genetic algorithm is usually used to optimize the topology of the RBF neural network, the radial basis function center and width. Alternatively, the least squares method could play an important role in deciding the weights of the output layer. The experimental result shows that the feasibility of RBF neural network which could be optimized by the hybrid hierarchical genetic algorithm to predict real estate closing price, and the superiority of this approach in the aspect of prediction accuracy verified in comparing with the other two methods.

Keywords: *Real estate Price Index prediction, RBF neural network, hierarchical genetic algorithm, least squares method*

1. Introduction

The real estate is a product of the market economy. The real estate price is determined by its value, but influenced by economic, political and social many factors. By observing the movement of Real estate price index, people can grasp of the macro economic situation of the real estate market and the changes in the market, or from the micro level, people also can take an analysis according its investment trend, and the realization of these roles based on the analysis and forecast of real estate price index. In recent years, with the real estate market continued to heat up, problems about the risk of the real estate industry have become increasingly prominent. Using scientific methods to reflect the changes in real estate prices, and presenting a correct information guide to the main market has become very urgent.

Generally speaking, there exist many ways to predict real estate prices. For traditional forecasting methods, we adopt the application of regression analysis, time series analysis and Markov model. However, traditional forecasting methods are mostly based on statistical analysis of long-term, large sample and it requires high regularity of the data distribution and integrity of data.

Existing research indicates that intelligent forecasting models outperform traditional models, especially in short-term forecasting [1]. In recent years, artificial intelligence techniques of genetic algorithms, artificial neural network and supporting vector machine methods are applied to short-term prediction of the real estate market by many scholars.

ISSN: 2005-4246 IJUNESST

For example, Erkam Guresen, *et al.*, evaluated the effectiveness of neural network models which were known to be dynamic and effective in real estate -market prediction [2].

A single model could only reflect the piece of information of real estate prices which is difficult to dig the hidden variation of price of real estate data. Furthermore, there are obvious limitations when using a single model, such as the slow convergence speed and poor generalization. However, the combined model could utilize the information provided by various models to a large extension, which would improve the prediction accuracy, especially in economic, management and statistical research, a number of prediction approaches have become an important way to improve forecast accuracy. For example, Armano, *et al.*, optimized ANN with GA to forecast real estate indices [3]. Md. Rafiul Hassan, *et al.*, proposed a fusion model by combining the Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and Genetic Algorithms (GA) to predict real estate price [4]. Lee proposed a prediction model based on a hybrid feature selection method and SVM to predict the trend of real estate market [5]. Wei Shen, *et al.*, used the artificial fish swarm algorithm (AFSA) to optimize RBF to forecast real estate indices [6]. Ling-Jing Kao, *et al.*, proposed a methodology which was the integration of nonlinear independent component analysis and support vector regression for real estate price forecasting [7]. Yakup Kara, *et al.*, developed two efficient models with different approaches: artificial neural network and support vector machine. Moreover, they compared the performance of two models in predicting the direction of movement in the daily Istanbul Real estate Exchange (ISE) National 100 Index [8].

Therefore, in order to improve the accuracy of prediction, this paper presents a kind of method combined the hybrid hierarchical algorithms and RBF neural network, and builds the fusion model to predict the closing price of the Real estate Price Index (SCI).

2. RBF Neural Network

Radial Basis Function (RBF) neural network is a feed-forward network with good performance, and it can decide the appropriate network topology based on different issues, with a high approximation precision, a small-scale of network training, fast learning speed and non-existence of local minima problems. The structure of RBF neural network consists of three layers: the input layer, hidden layer and output layer. The structure of topology is showed in Figure 1.

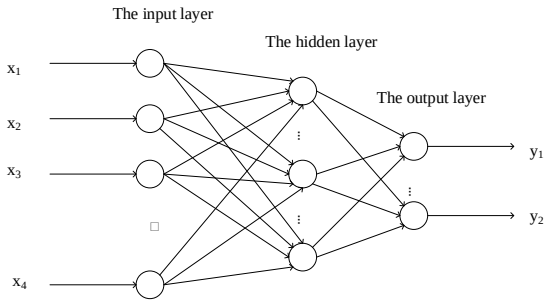

Figure 1. RBF Neural Network Structure

These nodes of input layer are only responsible for passing the input signals to the hidden layer, which including a group of non-linear radial basis function. Gaussian function is generally used as radial basis function, processing the input signal received with nonlinear transformation, and then transmitting the processed signal to the output layer. Gaussian function formula is given by

$$\varphi_j = \exp\left(-\frac{\left\|X - C_j\right\|}{2\sigma_j^2}\right),\ j = 1, 2, \ldots\ldots, \mathrm{L} \tag{1}$$

Where φ_j is the output of the hidden layer, $X \in R^n$ is the input of neural network, C_j is the center of Gaussian function, σ_j is the width of Gaussian function. L is the number of nodes in the hidden layer.

The output layer processes signals which have been processed by the hidden layer with linear weighted combination. Eventually, the predicted value that is obtained through the RBF neural network processing is one-dimensional vector y_j:

$$y_i = \sum_{j=1}^{L} w_{ij} \cdot \varphi_j,\ j = 12, \ldots\ldots, \mathrm{L} \tag{2}$$

Where w_{ij} is the connection weight between the j^{th} output of hidden layer and the i^{th} neuron of output layer.

3. The Prediction Model of RBF Neutral Network Optimized by HHGA

Hierarchical Genetic Algorithm (HGA) is a novel genetic algorithm introduced in recent years. For the case of determining the number of hidden nodes, traditional genetic algorithm optimizes the based function's center and width of hidden layer of RBF neutral network. The HGA is a learning algorithm proposed for RBF neural network features regarding the RBF network topology, center and variance of the hidden layer of basis function, and the connection weight from the hidden layer to the output layer are optimized simultaneously together. In the HGA, each chromosome consists of control genes using binary coding genes and parameter genes using real-coded genes, which is introduced according to the hierarchical structure of biological chromosome. Control genes use binary coding, binary "1" indicates that its corresponding parameter gene is activated, hence the node of hidden layer occurs. "0" indicates that its corresponding parameter gene is in dormant or inactive state which implies the node of hidden layer does not exist. Each control gene corresponds to a set of parameter genes in sequence. A set of parameter genes contain the center C_i and width σ_i of radial basis function of a hidden layer, and the weight w_i of the output layer. The hierarchical chromosome structure of RBF neutral network which is optimized by HGA is given below.

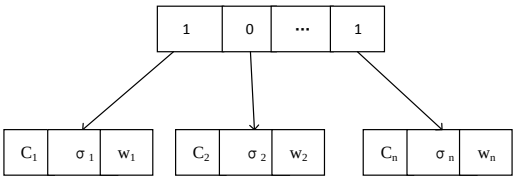

Figure 2. The Hierarchical Chromosome Structure of RBF Neutral Network Optimized by HGA

However, in the process of learning, the convergent speed of algorithm is quite slow. Since the output layer of RBF neural network is a linear neuron, hence if it is in the case that determining the center C_i and width σ_i, the weight w_i of output layer can be calculated by the least squares method to improve efficiency of the network's training. The hierarchical chromosome structure of RBF neutral network which is optimized by Hybrid Hierarchy Genetic Algorithm (HHGA) that is the combination of HGA and the least squares method is presented below.

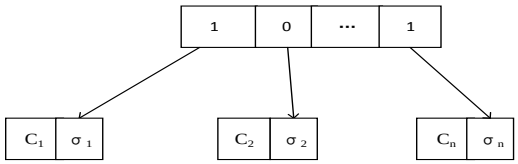

Figure 3. The Hierarchical Chromosome Structure of RBF Neutral Network Optimized by HHGA

Chromosome Coding

Each chromosome uses the hybrid encoding combining binary coding with real coding. The control gene using binary coding is in the upper layer, "1" indicates the number of hidden layer nodes. Using real-coded parameter gene is in the lower layer, including parameters of hidden layer: the center C_i and width σi of radial basis function.

Population Initialization

The size of population has an influence on the convergence of genetic algorithm. However, small population size is difficult to obtain the desired results. But if the population size is too large, it is easy to make the calculation more complicated. As much experience shown, when the size of population generally ranges from 20 to 160, the convergence of algorithm would be better. The population size in this paper is 40.

In HHGA, initialing population is to initialize control parameter genes. The command of genetic algorithm tool box is applied for initialization of control genes. The initialization of parameter genes has two parts, the first part is based on the training sample data, the center C_i of radial basis function corresponding to the parameter gene is obtained by fuzzy C-means clustering method, and the other part is that the width σ_i is determined by the minimum Euclidean distance of various data. Population P_1 is informed after its initialization.

Fitness Function

The purpose of training RBF neural network is to make the network structure simple with meeting the demand of certain precision, namely, which requires composite indicators which are approximation error accuracy and network complexity of RBF neural network are minimum respectively. The objective function of RBF neural network approximation error precision is represented by error sum of squares between network output and expected output.

$$F_1 = SSE = \sum_{i=1}^{N} \left(y_i - y_i^{'} \right)^2 \tag{3}$$

Where y_i is the expected output value, y_i ' is the real output value.

Network complexity is determined by the number of nodes n_c in the hidden layer.

$$F_2 = n_c \tag{4}$$

In order to make HHGA effectively train RBF neural network, this paper uses the fitness function of Akaike information criterion to reflect both the fitness of these two objectives.

$$f = -\left\{ N \left[\lg \frac{1}{N} \sum_{i=1}^{N} \left(y_i - y_i^{'} \right)^2 \right] + 4 n_c \right\} + d \tag{5}$$

Where N is the number of samples, n_c is the number of hidden layer nodes; y_i is the expected output value; y_i' is the output value of training RBF neural network. The value of d is 5000 to ensure that f> 0. According to the formula, when the squared error is smaller and the number of hidden layer nodes is smaller, then the value of f is lager.

4. Genetic Manipulation

Selection and Reproduction

The selection operation of HHGA is analogous to traditional genetic algorithms. If the fitness of an individual is larger, the greater probability of selected. This individual selection is used for the expectation value method. The expected value of the individual determines whether individuals of population could be divided into the next generation to optimize, the probability of individual j is copied directly proportional to f_j, the number of individuals being copied is the value of individual expectation v_j. The expected value of the individual used as the following formula.

$$v_j = \frac{f_j}{f_{avg}} = \frac{f_j}{f_{sum} / N} \tag{6}$$

Where f_j is the individual fitness, f_{avg} is the average fitness. f_{sum} is the total fitness of the population, N is the size of the population. After selection and reproduction operation, the initiated population P_1 becomes P_2.

Crossover and Mutation

The mutation operation is primarily used to maintain the diversity of population. Based on the population fitness, the population P_2 is conducted crossover and mutation operations. Crossover operation creates new gene combination to form group P_3. The crossover of control genes and parameter genes use single-point crossover. Due to the different encoding between control genes and parameter genes of each chromosome, the mutation of parameter genes selects real-value mutation. The mutation of control genes uses simulated binary mutation selecting randomly two individuals x_1 and x_2 from the parent population, then by linear combinations to produce new offspring.

$$\begin{cases} y_1 = a x_2 + (1 - a) x_1 \\ y_2 = a x_1 + (1 - a) x_2 \end{cases} \tag{7}$$

Where a is a random number, $a \in [0,1]$.

Calculate Weights of the Network Output

In the design of RBF neural network optimized by HHGA, the parameter information of output layer neurons in genetic algorithm encoding is redundant. If adding redundant information in the encoding, the efficiency of genetic algorithm will be reduced. Therefore, using HGA to determine the center C_i and width σ_i of radial basis function in the hidden layer, taking advantage of the least squares method to calculate the weights of output layer can study the final optimized RBF neural network.

5. Experiments and Analysis

In order to verify the RBF neural network optimized by HHGA which is feasible and effective in real estate prediction, comparing the performance of the method proposed in

this paper with the existing approaches, such as RBF neural network and BP neural network optimized by GA with conducting the same experiment.

This experimental data are the Real estate Price Index closing price of 125 trading days, from July 1, 2013 to December 31, 2013, collected on the Shanghai Real estate Exchange. The first 107 data is taken as training sample, and the remaining 18 data is presented as the testing sample from the 155th day to 185th day. In order to avoid a great range of data having a negative impact on the RBF neural network training, a real estate closing price x_i is normalized to interval [0, 1] by

$$x_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{8}$$

Where x_i is the input of neural network, max(x) and min(x) are the maximum and minimum value of the sample data, respectively.

Through the normalization of x_i, then imputing the sample data are regarded as the training sample of the neural network. After the network training and simulation of the sample, conducting anti-normalization process to revert the real closing price when outputting predicted results.

$$Y(x_i) = u(x_i) * (\max(x) - \min(x)) + \min(x) \tag{9}$$

Where $u(x_i)$ is the output of neural network.

The Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) are taken as the standard to assess the predictions accuracy of the model.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i' - y_i\right|}{y_i} \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i' - y_i\right|}{y_i}} \tag{11}$$

Where y_i' is the predicted value, y_i is the actual value, N is the number of samples.

When the value of MAPE and RMSE is smaller, the higher forecasting accuracy of the model is possible. To avoid the influence of random factors on the experimental results, three methods are conducted three times. The following is the result Tables.

Table 1. The MAPE Values of Three Models

Experiment	HHGA-RBFNN	GA-BPNN	RBF
1	0.0076	0.0093	0.0120
2	0.0077	0.0086	0.0120
3	0.0080	0.0097	0.0120

Table 2. The RMSE Values of Three Models

Experiment	HHGA-RBFNN	GA-BPNN	RBF
1	20.2072	26.1209	29.5758
2	20.9874	24.0441	29.5758
3	21.7370	25.5099	29.5758

6. Experimental Process

Step 1: obtain weight value and sequence of factors. There are lots of subjective weighting methods, such as Gulin method, AHP method, expert investment method, analytic hierarchy process (AHP) .in this paper, we use relative weight method.

r_{ij} is the score of i-th factor in the j-th questionnaire, then the original weight value of i-th factor in j-th investigation w_{ij} equals:

$$w_{ij} = \frac{r_{ij}}{\sum_{i=1}^{n} r_{ij}} \tag{12}$$

In formula (1), n is the number of affected factors in j-th investigation. So, for each investigation, $\sum_{i=1}^{n} w_{ij} = 1$.

Weight value, defined by different experts with variable acknowledgement towards each influence factors, may not be the same. in order to eliminate the deviation, we average w_{ij} from all questionnaire:

$$w_{ij}' = \frac{\sum_{j=1}^{m} w_{ij}}{m} \tag{13}$$

We get the final weight value of each influence factor by relative weight method. At the same time, the sequence about factors can be decided according to the weight value.

Step 2: program with MATLAB, build BP network, bring in MIV. Gets the sequence of influence factors using engineering data to train network

When using BP network, we should first provide a training set, where each sample is determined by the input mode and the desired output mode. Suppose there are q sample training set, then the sample can be formalized as a model of (X^k, Y^k), $k = 1, 2, ..., q$, X^k is the k-th input $(x_1^k, x_2^k, ..., x_n^k)$, Y^k is its output $(y_1^k, y_2^k, ..., y_n^k)$, and they correspond to n neurons of the input layer and the output layer, respectively. When the actual output of the network and the desired output are consistent, the learning process is over. Otherwise, the learning algorithm will make the actual output close to the desired by adjusting the connection weights of the network, according to the error between the two outputs.

Define w_{hi}^1 as a connection weight from the neurons h in the input layer to the neurons i in the hidden layer, and define w_{ij}^2 as a connection weight from the neurons i in the hidden layer to the neurons j in the output layer. Suppose for a (X^k, Y^k), the global error function in the output layer is:

$$E_k = \frac{1}{2}\sum_{j=1}^{m} (y_j - y_j^k)^2 \quad k = 1, 2, ...q \tag{14}$$

Among them, y_i is the actual output of output layer neurons j, j_j^k is the desired output. Then, error of the whole training set is:

$$E = \sum_{k=1}^{q} E_k \tag{15}$$

For the k-th sample, the weighted input of output layer neurons j is:

$$nety_j = \sum_{i=1}^{p} b_i * w_{ij}^2 \tag{16}$$

(p is the number of neurons in the hidden layer)

Then the actual output is:

$$y_j = f(nety_j) \tag{17}$$

In formula (17),

$$f(u) = \frac{1}{1+e^{-u}} \tag{18}$$

Among them, s means the number of iterations.

Thus, we can obtain the rank of the influence factors and the weight value.

The results in the Table 1 and Table 2 show that for the experiment of predicting the closing price of SCI, the values of MPAE and RMSE of RBF neural network optimized by HHGA are minimized, so the predicted accuracy is maximized. The following is comparison charts between actual and predicted values of three methods:

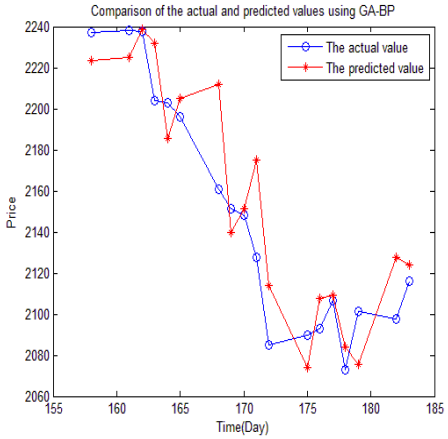

Figure 4. HHGA-RBF Approach

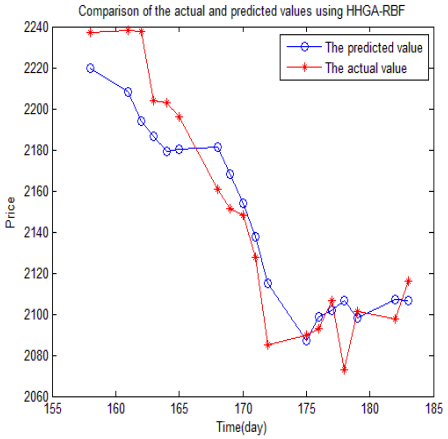

Figure 5. GA-BP Approach

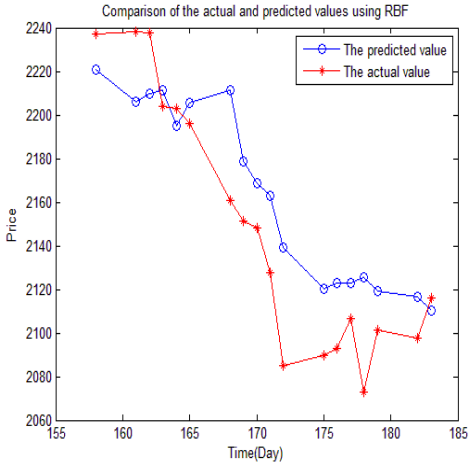

Figure 6. RBF Approach

7. Conclusion

A series of uncertain factors in real estate market have led to the difficulty of predicting real estate price, such as policy, economic environment and investor psychology. From numerous approaches studying neural network to predict the real estate price, this paper proposes and implements the combination of HHGA and RBF neural network to forecast the closing price of SCI. In addition, comparing with the prediction effect of using RBF neural network and BP neural network optimized by GA with the same experiment, the results confirms the prediction error is smaller using RBF neural network optimized by HHGA and the prediction accuracy is superior to other two methods.

References

[1] T. O. Hill, M. Connor and W. Remus, "Neural network models for time series forecasts", Management Science, vol. 42, no. 7, **(1996)**, pp. 1082-1092.
[2] E. Guresen, G. Kayakutlu and T. U. Daim, "Using artificial neural network models in real estate market index prediction", Expert Systems with Applications, vol. 38, **(2011)**, pp. 10389-10397.
[3] G. Armano, M. Marchesi and A. Murru, "A hybrid genetic-neural architecture for real estate indexes forecasting: Information Sciences, vol. 170, **(2005)**, pp. 3-33.
[4] R. Hassan, B. Nath and M. Kirley, "A fusion model of HMM, ANN and GA for real estate market forecasting", Expert Systems with Applications, vol. 33, **(2007)**, pp. 171-180.
[5] M.-C. Lee, "Using support vector machine with a hybrid feature selection method to the real estate trend prediction", Expert Systems with Applications, vol. 36, no. 8, **(2009)**, pp. 10896-10904.
[6] W. Shen, X. Guo and C. Wu, "Forecasting real estate indices using radial basis function neural networks optimized by artificial fish swarm algorithm", Knowledge-Based Systems, vol. 24, **(2011)**, pp. 378-385.
[7] L.-J. Kao, C.-C. Chiu and C.-J. Lu, "Integration of nonlinear independent component analysis and support vector regression for real estate price forecasting, Neurocomputing, vol. 99, **(2013)**, pp. 534-542.
[8] Y. Kara, M. A. Boyacioglu and Ö. K. Baykan, "Predicting direction of real estate price index movement using artificial neural networks and support vector machines", The sample of the Istanbul Real estate Exchange, Expert Systems with Applications, vol. 38, **(2011)**, pp. 5311–5319.

Authors

Huan Ma, born in June, 1981, Henan, P R china

Current position, grades: Lecturer at Zhengzhou University of Light Industry, China

University studies: Master degree in computer application technology from Huazhong University of Science and Technology in China

Scientific interest: Information processing, algorithm design and analysis

Ming Chen, born in April, 1983, Henan, P R china

Current position, grades: Lecturer at Zhengzhou University of Light Industry, China

University studies: PhD degree from Beijing University of Posts and Telecommunications in China

Scientific interest: big data

Jian-wei Zhang, born in April，1971，Henan，P R china

Current position, grades: Professor at Zhengzhou University of Light Industry, China

University studies: PhD degree from The PLA Information Engineering University in China

Scientific interest: broadband information network and network security