Scientific
Research

A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study

Itedal Sabri Hashim Bahia

Ministry of Higher Education and Scientific Research, Baghdad, Iraq
Email: itedal_bahia@yahoo.com

ABSTRACT

This paper aims to demonstrate the importance and possible value of housing predictive power which provides independent real estate market forecasts on home prices by using data mining tasks. A (FFBP) network model and (CFBP) network model are one of these tasks used in this research to compare results of them. We estimate the median value of owner occupied homes in Boston suburbs given 13 neighborhood attributes. An estimator can be found by fitting the inputs and targets. This data set has 506 samples. "ousing inputs" is a 13×506 matrix. The "housing targets" is a 1×506 matrix of median values of owner-occupied homes in $1000's. The result in this paper concludes that which one of the two networks appears to be a better indicator of the output data to target data network structure than maximizing predict. The CFBP network which is the best result from the Output_network for all samples are found from the equation output = 0.95 * Target + 1.2. The regression value is approximately 1, (R = 0.964). That means the Output_network is matching to the target data set (Median value of owner-occupied homes in $1000's), and the percent correctly predict in the simulation sample is 96%.

1. Introduction

Today real estate market has become very popular. Though the near future of real estate is still in question, investors have been hungry for a fast way to play the market or to hedge against their volatile portfolios. Futures contracts have been an extremely popular method of balancing a portfolio in other markets, and real estate is, with a little knowledge, now in the same boat [1].

Futures contracts that trade at a centralized exchange allow market participants more financial leverage and flexibility and are guaranteed by the exchange so there is no risk of counterparty default. They are also in and of themselves leveraged investments, which allow investors a way to benefit on movements in housing prices as well as provide them with the opportunity for a liquid short-term real estate investment. These futures also allow investors a way to speculate on housing prices with much lower capital requirements [1]. An accurate prediction on the house price is important to prospective homeowners, developers, investors, appraisers, tax assessors and other real estate market participants, such as, mortgage lenders and insurers [2]. Traditional house price prediction is

based on cost and sale price comparison lacking of an accepted standard and a certification process. Artificial Neural Network (ANN) is a neurobiological inspired paradigm that emulates the functioning of the brain based on the way that neurons work, because they are recognized as the cellular elements responsible for the brain information processing [3]. ANN models can detect patterns that relate input variables to their corresponding outputs in complex biological systems for prediction [4]. Methods for improving network performance include finding an optimum network architecture and appropriate number of training cycles, using different input combinations [5]. Therefore, the availability of a house price prediction model helps fill up an important information gap and improves the efficiency of the real estate market [6].

2. Related Works

The quest for patterns in data has been studied for a long time in many fields, including statistics, patterns recognition and exploratory data analysis [7,8]. Analyzing data can provide further knowledge about a business by going beyond the data explicitly stored to derive knowledge

about the business. This is where data mining has obvious benefits for any enterprise. Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. Data mining has been applied in a great number of fields, including retail sales, bioinformatics, and counterterrorism. In recent years, there has been increasing interest in the use of data mining to investigate scientific questions within educational research, an area of inquiry termed educational data mining. [9] The stages involved in data mining as in figure [10]. There are many research that related works with this paper.

Dongsong Zhang and Lina Zhou [11], they are describing data mining in the context of financial application from both technical and application perspectives. In addition, we compare different data mining techniques and discuss important data mining issues involved in specific financial applications. Finally, we highlight a number of challenges and trends for future research in this area.

G. Grudnitski, A. Quang Do and J. D. Shilling [7], they are applying a neural network analysis to supply evidence that answers this question. they find evidence that the characteristics of a borrower's net worth, marital status and education level and whether a co borrower is involved contribute in a significant way to the neural network's ability to determine mortgage choice.

Kaihla, Paul, Copeland, Michael V., Hawn, Carleen, Lappin, Todd, Lev-Ram, Michal, Sloan, Paul [12], they are presents information related to the current status of real estate in the U.S. The nationwide housing slump was most evident with the fact that the median sales price for existing U.S. homes slipped to $225,000 in August 2006. It is however opined that the impact of housing downturn on the American residential real estate assets will not be much extensive.

Kontrimas, Vilius; Verikas, Antanas [13], used the ordinary least squares (OLS) linear regression is the classical method used to build models in this approach. The method is compared with computational intelligence approaches-support vector machine (SVM) regression, multilayer perceptron (MLP), and a committee of predictors. The performance of the committee using the weights based on zones obtained from the SOM was also higher than of that exploiting the real estate value zones provided by the Register center.

Landers, Jay [14], produce offers information on the market forecasts for the nonresidential construction sector in the US in 2008. A report titled "Construction Outlook 2008," by the company McGraw-Hill Construction says that more rigid lending standards have begun to affect commercial real estate and have reduced the volume of property purchases. According to a forecast by FMI Corp., there will be a marginally better result for total construction in 2008.

Ruben D. Jaen [15], is presents the insights gained from applying data mining techniques, in particular neural networks for the purposes of developing an intelligent model used to predict real estate property values based on variety of factors. A dataset of over one thousand transactions in real estate properties was used. The dataset included 15 variables obtained from the multiple listing system (MLS) database and captured information on transactions taking place during a period of three years. The results from applying data mining techniques to predict real estate values are promising. Future plans and recommendations for further expanding the study are given.

Sean Zdenek [16], suggest, produces AI systems by rhetorical means; it does not merely describe AI systems or reflect a set of prevailing attitudes about technology. To support this argument, I analyses a set of research articles about an "embodied conversational agent" called the Real Estate Agent (REA).

3. Methodologies/Data Mining Process

3.1. Data Preparations

The data was created by a house price as a data set to test the data mining intelligent system, which will perform the predict system. A data is available from the UCI Machine Learning Repository in Irvine, CA: University of California, Department of Information and Computer Science, and the StatLib library which is maintained at Carnegie Mellon University. The main idea was to construct the neural network model by using two types of neural networks first feed forward neural network, second Cascade forward neural network, and compare to find the best perform predict house price. For better understanding of the problem let us consider definitions of house price. A house price is characterized by Median value of owner-occupied homes in $1000's as a target values. This dataset contains 13 predicting variables in 506 numbers of samples. **Table 1** presents the variables of data which are considered as 13 predicting variables.

3.2. Data Selection and Transformation

The study of feature selection finds its practical in machine learning in which a learning algorithm constructs a description of a function from a set of input/output instances through the interaction with the world [17].

In this step set of samples (records) and variables (fields) are selected which were required for data mining. All the predictor and response variables which were derived for the data mining activity are presented in **Figure 1**.
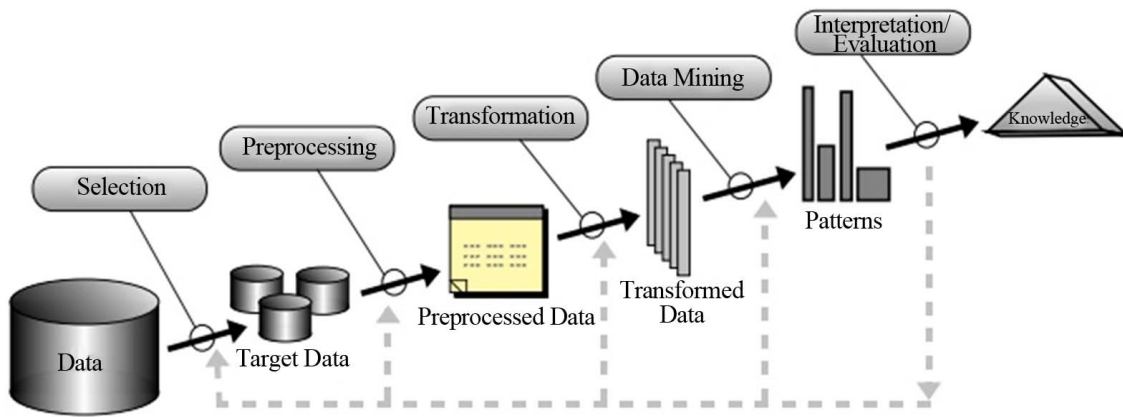
Figure 1. The steps of extracting knowledge from data [10].

Table 1. Predicting variables of data.

No	variables
1-	per capita crime rate by town
2-	Proportion of residential land zoned for lots over 25,000 sq.ft.
3-	proportion of non-retail business acres per town
4-	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
5-	nitric oxides concentration (parts per 10 million)
6-	average number of rooms per dwelling
7-	proportion of owner-occupied units built prior to 1940
8-	weighted distances to five Boston employment centers
9-	index of accessibility to radial highways
10-	full-value property-tax rate per $10,000
11-	pupil-teacher ratio by town
12-	$1000 (Bk - 0.63)^2$ where Bk is the proportion of blacks by town
13-	% lower status of the population

3.3. Artificial Neural Network

Neural network is an artificial intelligence model originally designed to replicate the human brain's learning process. The model consists of three main layers: input data layer (example the property attributes), hidden layer(s) (commonly referred as "black box"), and output layer [18]. Neural network is an interconnected network of artificial neurons with a rule to adjust the strength or weight of the connections between the units in response to externally supplied data **Figure 2** [18,19]. Each artificial neuron (or computational unit) has a set of input connections that receive signals from other computational units and a bias adjustment, a set of weights for input connection and bias adjustment, and transfer func-
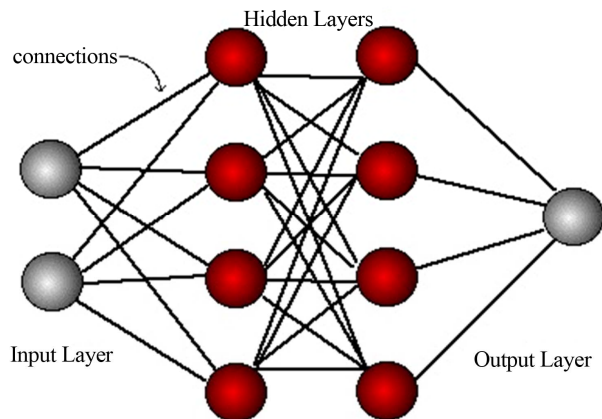

Figure 2. Typical neural networks [19].

tion that transforms the sum of the weighted inputs and bias to decide the value of the output from computational unit [18].

In supervised training, we present a pattern to the neural network, it makes a prediction, and we compare the predicted output to the desired output. Thus we have explicit information about the performance of the network. The major parameters used in supervised training have to do with how the error is computed and how big a step we take when adjusting the connection weights in the direction of the desired output [20].

Learning rate almost all neural network models have a learning rate parameter associated with them. In a typical supervised training case, a pattern is presented to the neural network; it makes an incorrect prediction, and the difference between the desired output and the actual output is used to adjust the weights [20].

There are many applications where prediction can help in setting priorities. For example, the emergency room at a hospital can be a hectic place. To know who needs the most time critical help can enable a more successful operation. Basically, all organizations must establish priorities which govern the allocation of their resources. This
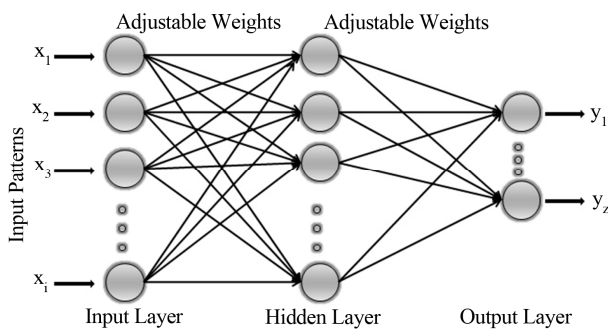

projection of the future is what drove the creation of networks of prediction [21].
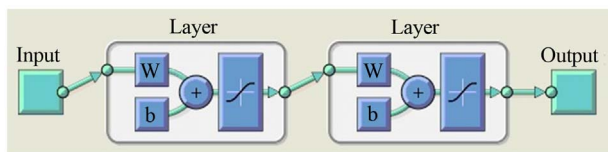
3.3.1. The ANN (FFBP) Network

Feed forward networks often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors. The linear output layer lets the network produce values outside the range –1 to +1. The first step in training a feed forward network is to create the network object. It requires three arguments and returns the network object. The first argument is a matrix of input vectors. The second argument is a matrix of target vectors. The sample inputs and outputs are used to set up network input and output dimensions and parameters. The third argument is an array containing the sizes of each hidden layer. (The output layer size is determined from the targets.) [22,23]. The typical and structure of the feed forward network is show in the **Figures 3** and **4**.

3.3.2. The ANN (CFBP) Network

The cascade forward network, these are similar to feed forward network, but include a weight connection from the input to each layer and from each layer to the successive layers. For example, a three-layer network has connections from layer 1 to layers 2, layer 2 to layer 3, and layer 1 to layer 3. The three-layer network also has connections from the input to all three layers. The additional connections might improve the speed at which the network learns the desired relationship [21,22]. CF artificial intelligence model is similar to feedforward backpropagation neural network in using the backpropagation algo-

rithm for weights updating, but the main symptom of this network is that each layer of neurons related to all previous layer of neurons [20]. Tan-sigmoid transfer function, log-sigmoid transfer function and pure linear threshold functions were used to reach the optimized status [22,23]. The performance of cascade forward backpropagation and feedforward backpropagation were evaluated using Root Mean Square Error (RMSE) Equation (1), Mean Square Error (MSE) Equation (2) and R^2 Equation (3) technique.

$$MSE = \left[\sum_1^N \left(\frac{X_{\exp} - X_{cal}}{X_{\exp}^2} \right)^2 \right] \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\left[\sum_1^N \left(\frac{X_{\exp} - X_{cal}}{n} \right)^2 \right]} \tag{2}$$

$$R^2 = 1 - \left[\sum_1^N \left(\frac{X_{\exp} - X_{cal}}{X_{\exp}^2} \right)^2 \right] \tag{3}$$

where, $X_{\exp}$ = Observed value; X_{cal} = Predicted value; $X_{\exp}$ = Mean predicted value; n = Number of observations in dataset. MSE, RMSE) and R^2 were used in order to compare the prediction performance of the developed models. The best score for R^2 measure is 1 and for other measures is zero. The typical and structure of the cascade forward network is show in the **Figures 5** and **6**.

Cascade Forward Back Propagation CFBP is similar to FFBP network in using the BP algorithm for weights up-

Figure 3. Typical FFBP.

Figure 4. The structure of the feed forward back propagation network.
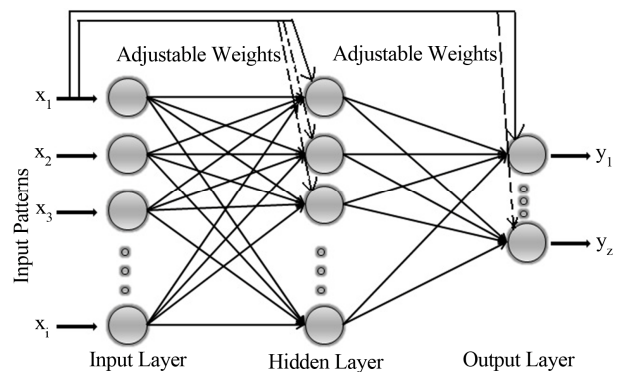
Figure 5. Cascade-forward ANN.
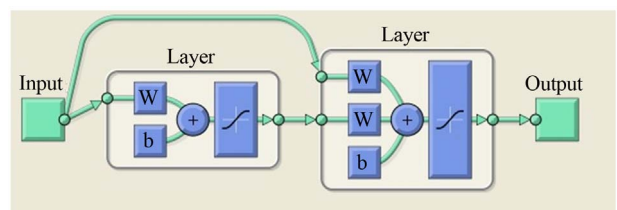
Figure 6. The structure of the cascade forward back propagation network.

dating, but the main symptom of this network is that each layer neurons relates to all previous layer neurons.

4. Results

4.1. Relative Importance of Inputs

We first discuss the results obtained from the two neural networks modeling to predict the housing price. The neural networks were created using the neural network toolbox from Matlab 7.9. Training network automatically stops when generalization stops improving, as indicated by an increase in the mean square error (MSE) of the validation samples. The results of apply by using the artificial neural networks methodology to predict be-

tween input data (actual data) and target data. The dataset were divided 80% for training, and 20% for testing. Weights and biases were randomly initialized. The network was trained with up to 100 epochs. Weight is information used by neural network to solve a problem. In neural network analysis the results, depending on value of house price is performed, we have three figures for each neural network model, **Figures 7-9** for FFBP neural network and **Figures 10-12** for CFBP neural network.

We will focus on which one of those neural network model is best to improving the neural network predict performance for house price. **Table 2** gives the training, validation, and test predicting performance of the FFBP neural network model. It shows measures of regression
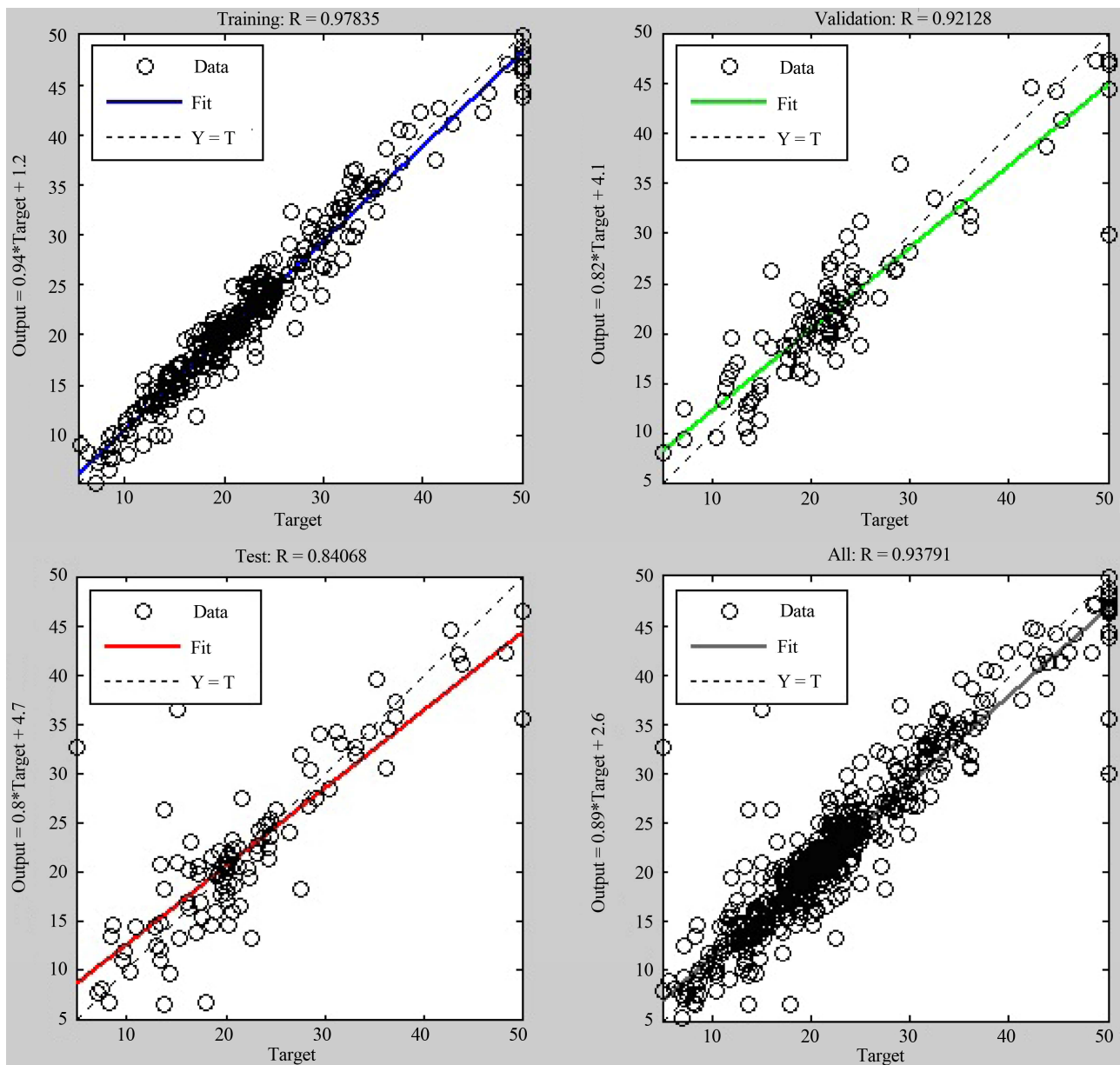

Figure 7. The regression values between the actual value and target values (FFBP).
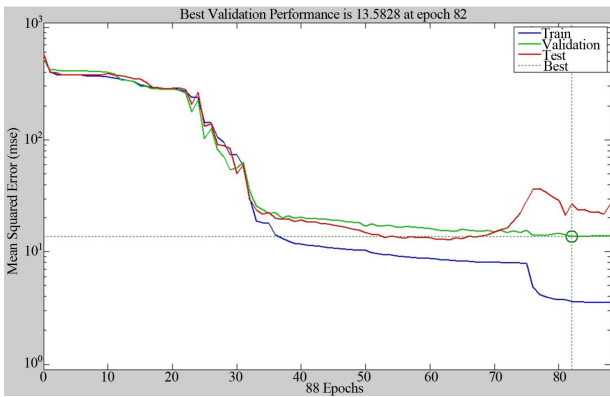
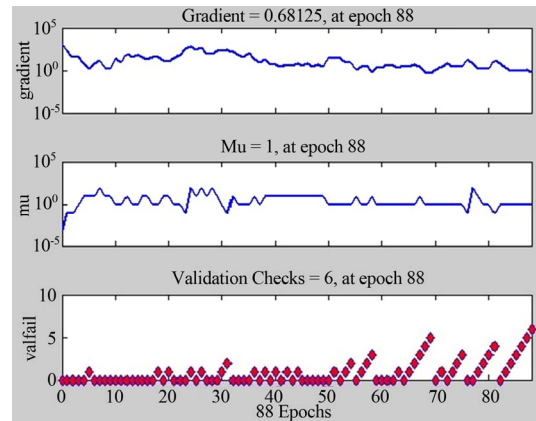**Figure 8. The best validation performance (MSE) at 82 epochs (FFBP).**

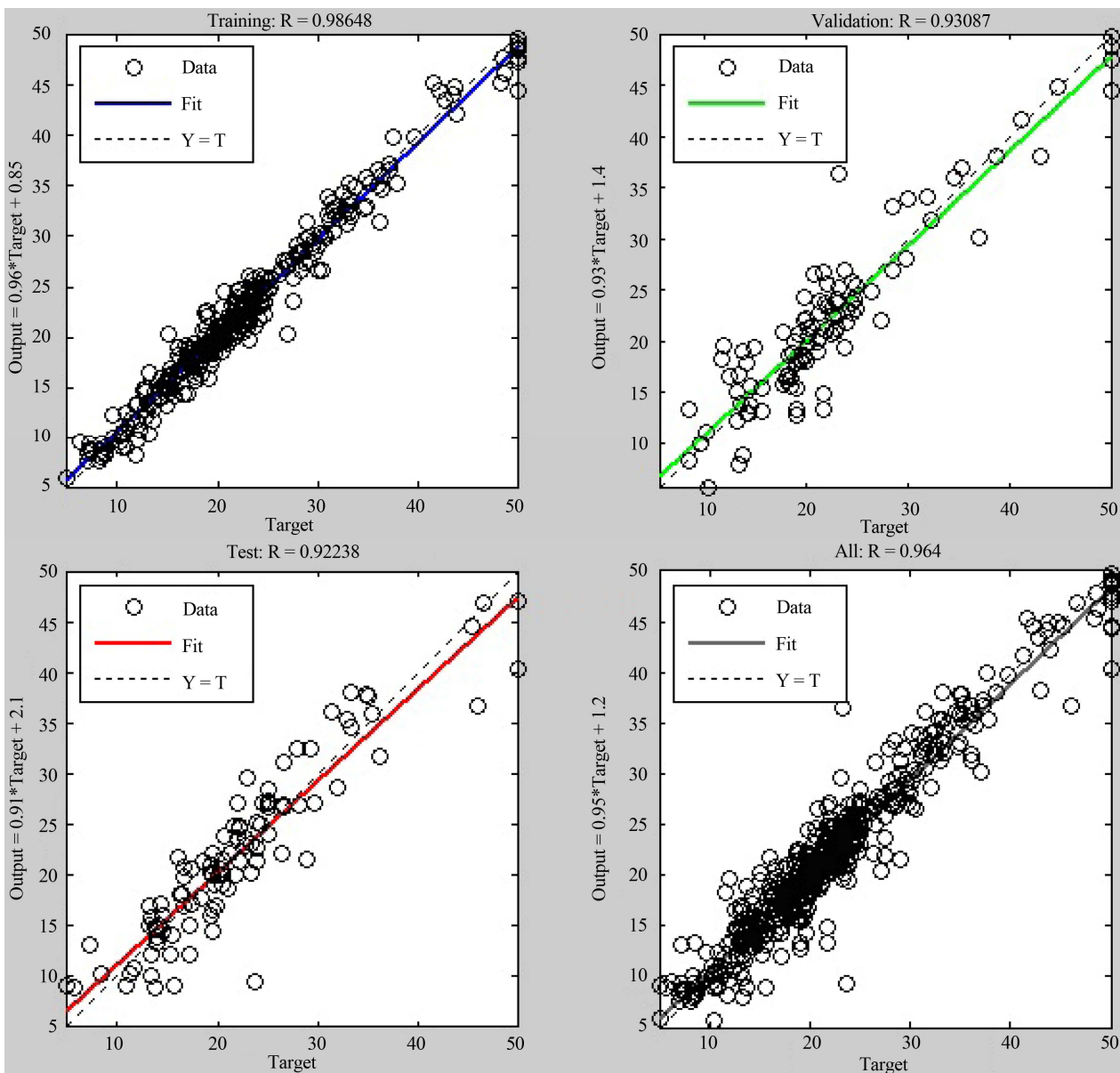**Figure 9. The training states for FFBP network.**

Figure 10. The regression values between the actual value and target values (CFBP).
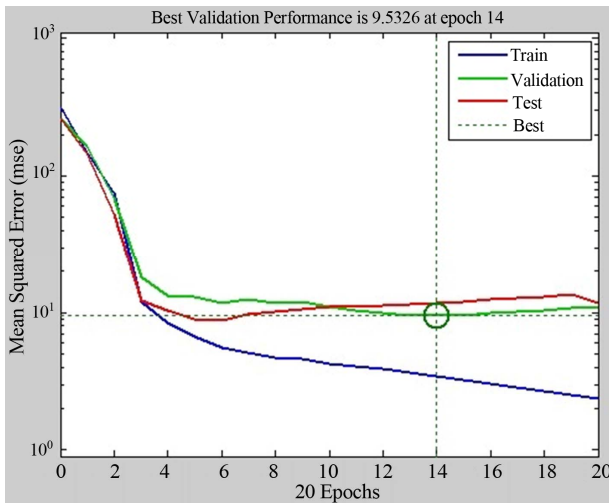
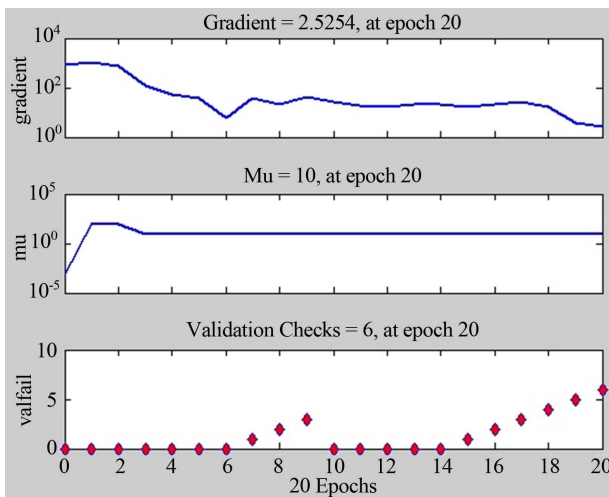**Figure 11. The best validation performance (MSE) at 14 epochs (CFBP).**

Figure 12. The training states for CFBP network.

Table 2. FFBP network statistics.

Type	Value	Equation
R.-training	0.98	$Output = 0.94Target + 1.2$
R-validation	0.92	$Output = 0.82Target + 4.1$
R-test	0.841	$Output = 0.8Target + 4.6$
Gradient	0.681	
Mu	1	

(R), training state (Gradient, Mu, and validation) and the best validation performance (MSE) is 13.583 at epoch 82.The statistics shows the measure of performance of the network with regard to the input variables presented to it. The regression of all samples of data set is 0.94 in epoch 88 and the predict values are obtained from the equation Output = 0.89Target + 2.6 in **Figure 5**. The

FFBP values obtained from the network in epoch 88 iterations is presented in **Table 2**.

Table 3 gives the training, validation, and test predicting performance of the CFBP neural network model. It shows measures of R, training state (Gradient, Mu, and validation) and the best validation performance (MSE) is 9.533 at epoch 14. The statistics shows the measure of performance of the network with regard to the input variables presented to it. The regression in the **Table 3** for all samples of data set is 0.96 in epoch 20 and the predict values are obtained from the equation Output = 0.96Target +1.2 in **Figure 8**. The CFBP values obtained from the network in epoch 20 iterations is presented in **Table 3**.

4.2. Discussion of Results

As can be observed from the results in the **Tables 2** and **3**, models with Cascade Forward Back Propagation neural network structure gives the best results because the validation performance value (MSE) mean square error is less than of the validation performance value (MSE) of Feed Forward Back Propagation neural network.

The regression graph is getting it from the scatter plot in **Figure 10**, the relation between the output and the target. Output_network for all samples are result from the equation Output = 0.95Target + 1.2. The regression value is approximately 1, R = 0.964. That mean the output_ network is matching to the target data set (Median value of owner-occupied homes in $1000's), and the percent correctly predict in the simulation sample is 96%.

5. Conclusion

An accurate prediction on the housing price is important to prospective homeowners and everything belongs to a real estate market. Data mining has come out of the research lab and into the real world to do just such tasks. The FFBP and CFBP neural networks are one of these tasks used to predict housing price and compare the result between them. When apply the CFBP neural networks methodology to predict housing price, the result is better in CFBP network in the predicting based upon selected parameters, showing abilities of the network to

Table 3. CFBP network statistics.

type	value	Equation
R.-training	0.99	$Output = 0.96Target + 0.85$
R-validation	0.931	$Output = 0.93Target + 1.4$
R - test	0.922	$Output = 0.91Target + 2.1$
Gradient	2.525	
Mu	10	

learn the patterns, and in the CFBP structure each layer neuron relates to all previous layer neurons that give for CFBP network more training to adjust the weight and gives accurate predict results depending on the output of network that matches the target. Artificial neural networks showing significant results on housing price prediction in the simulation samples is above 96%.

REFERENCES

[1] "An Introduction to Real Estate Futures," 2011. http://www.investopedia.com/articles/optioninvestor/11/an-introduction-to-real-estate-futures.html

[2] C. A. Calhoun, "Property Valuation Models and House Price Indexes for The Provinces of Thailand: 2000," *Housing Finance International*, Vol. 18, No. 3, 2003, pp. 31-41.

[3] A. Araque, E. D. Martin, G. Perea, J. I. Arellano and W. Buno, "Synaptically Released Acetylcholine Evokes Ca2+ Elevations in Astrocytes in Hippocampal Slices," *Journal of Neuroscience*, Vol. 22, No. 7, 2002, pp. 2443-2450.

[4] P. D. Wasserman, "Advanced Methods in Neural Computting," Van Nostrand Reinhold, New York, 1993, p. 255.

[5] R. S. Parmer, R. W. McClendon, G. Hoogenboom, P. D. Blankenship, R. J. Cole and J. W. Dorner, "Estimation of Aflatoxin Contamination in Preharvest Peanuts Using Neural Networks," *Transaction ASAE*, Vol. 40, No. 3, 1997, pp. 809-813.

[6] H. Demuth and M. Beale, "Neural Network Toolbox for Matlab-Users Guide Version 4.1," The Mathworks Inc., Natrick, 2003.

[7] G. Grudnitski, A. Quang Do and J. D. Shilling, "A Neural Network Analysis of Mortgage Choice," *International Journal of Intelligent Systems in Accounting*, *Finance and Management*, Vol. 4, No. 2, 1995, pp. 127-135.

[8] R. M. Heristev, "The ANN Book, GNU Public License," 1998. ftp://ftp.funet.fi/pub/sci/neural/books/

[9] M. Stanley, A. Alastair, M. Dylan and D. Patterson, "Neural Networks: The Prediction of Residential Values," *Journal of Property Valuation & Investment*, Vol. 16, No. 1, 1998, pp. 57-70. http://dx.doi.org/10.1108/14635789810205128

[10] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," American Association for Artificial Intelligence (© AAAI), *AI Magazine*, Vol. 17 No. 3, 1996, pp. 37-54.

[11] D. S. Zhang and L. N. Zhou, "Discovering Golden Nuggets: Data Mining in Financial Application," *IEEE Transactions on Systems*, *Man*, *and Cybernetics—Part C*: *Applications and Reviews*, Vol. 34, No. 4, 2004, pp. 513-522. http://dx.doi.org/10.1109/TSMCC.2004.829279

[12] P. Kaihla, M. V. Copeland, C. Hawn, T. Lappin, M. Lev-Ram and P. Sloan, "The New Rules of Real Estate," *Business* 2.0 *The New Rules of Real Estate Survival Tips for a Sluggish Market how to Buy how to Sell the* 10 *Best Places to Invest*, Vol. 7, No. 10, 2006, p. 80.

[13] V. Kontrimas and A. Verikas, "The Mass Appraisal of the Real Estate by Computational Intelligence," *Applied Soft Computing Journal*, Vol. 11, No. 1, 2011, pp. 443-448. http://dx.doi.org/10.1016/j.asoc.2009.12.003

[14] J. Landers, "Market Forecasts See Mixed Conditions for Nonresidential Construction," *Civil Engineering*, Vol. 78 No. 4, 2008, pp. 14-16.

[15] R. D. Jaen, "Data Mining: An Empirical Application in Real Estate Valuation, FLAIRS-02 Proceedings," American Association for Artificial Intelligence, 2002.

[16] S. Zdenek, "Artificial Intelligence as a Discursive Practice: The Case of Embodied Software Agent Systems," Springer-Verlag London Limited, AI & Society, 2003, pp. 340-363.

[17] L. Huan and M. Hiroshi, "Feature Selection for Knowledge Discovery and Data Mining," *The Springer International Series in Engineering and Computer Science*, 1998.

[18] J. Coakley and C. Brown, "Artificial Neural Networks in Accounting and Finance: Modeling Issues," *International Journal of Intelligent Systems in Accounting*, *Finance and Management*, Vol. 9, No. 2, 2000, pp. 119-144.

[19] G. Papadourakis, "Introduction to Neural Networks," Technological Educational Institute of Crete, Department of Applied Informatics and Multimedia, 2004.

[20] P. B. Joseph, "Data Mining with Neural Networks Solving Business Problems," McGraw-Hill Companies, Inc., 1996.

[21] A. Dave and M. George, "Artificial Neural Networks Technology," Kaman Sciences Corporation, Utica, 1992

[22] R. A. Chayjan and M. Esna-Ashari, "Comparison between Artificial Neural Networks and Mathematical Models for Equilibrium Moisture Characteristics Estimation in Raisin," *Agricultural Engineering International*: *The CIGR E-Journal*, Vol. 12, 2010.

[23] M. H. Beale, M. T. Hagan and H. B. Demuth, "Neural Network Toolbox™ 7 User's Guide," Math Works, Inc., Natick, 2010.