INVESTIGATING DATA MINING BY ARTIFICIAL NEURAL NETWORK: A CASE OF REAL ESTATE PROPERTY EVALUATION

[1]Rajat Pradhan, [2]Satish Kumar

[1,2] Dept. of Electronics & Communication Engineering, A.S.E.T.,
Amity University Uttar Pradesh, Lucknow, (India).

ABSTRACT

Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cut costs, forecasting, decision support etc. In this paper we have considered centralized database which is located and maintained locally unlike the distributed database. We have used MATLAB as a platform to carry out data mining and investigated how it can be effectively used to analyze data from a database making use of the artificial neural networks. Back propagation algorithm has been used to train the neural network with a set of data samples of real estate properties.

Keywords: Back Propagation, Data Mining, MATLAB, Neural Network.

I INTRODUCTION

Evaluation of a property requires consideration of several parameters. This study explores the use of neural network for decision support in property evaluation by presenting to it a database of 400 properties containing various parameters such as geographical area of the properties which includes sub parameters such as open space area, parking area, covered area etc. Other attributes in the database are the circle rates of the properties, the number of years since the properties were constructed and so on. All these parameters were chosen to be presented to the neural network as inputs to train it corresponding to certain target outputs for each specific property, so that it can evaluate the correct pricing for the properties, thus, giving a general idea to the end user about them.

II ARTIFICIAL NEURAL NETWORK

Artificial Neural Network (ANN) is the mathematical model of the biological neural network[1] of the human brain. It is an extremely simplified model of the human brain, essentially a function approximator composed of many neurons that cooperate to perform that desired function[2]. The Back propagation neural network is the most influential neural network in pattern classification[3]. Neural Networks are potentially useful for studying the complex relationships (possibly nonlinear) between input and output variables in a system[5]. Structurally, it is a layered type network with input layer, hidden layer and output layer of the 3 tier structure as shown in Fig. 1.

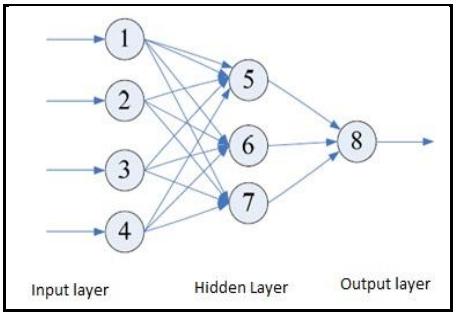

Figure 1: ANN Layered Structure

The input layer of the ANN is the one where the inputs are presented to the network in the form of input vectors in MATLAB. The hidden layer of the network is the one where the computational work of weight correction is done by back propagation algorithm by making use of the gradient descent method[3]. The output layer of the ANN is the one where we get the final output from the network. There is no restriction in choosing the number of neurons in all the three layers. An ANN consists of interconnected processing units, the general model of which consists of a summing part followed by an output part. The summing part receives input values from the input layer, weights each value and computes a weighted sum[6]. The number of neurons can be selected as per the requirement of the user according to the problem to be solved.

The Back Propagation learning process can be described as follows [1]:

1. The weight matrices initialized with random numbers which are small non zero values.
2. The bias vectors initialized with random numbers.
3. With the initialized values the output vector is computed corresponding to the input vector.
4. Weights adjusted between the hidden and output layer as well as between input and the hidden layer.
5. Bias vectors adjusted.
6. Above mentioned procedure repeated until the particular sum squared error reached.

The commonly used activation functions used in MATLAB are *logsig*, *tansig*.

$$\log sig(x) = \frac{1}{1+\exp(-x)}$$

$$\tan sig(x) = \frac{2}{1+\exp(-2x)}$$

These are called the *sigmoid functions*. The process of obtaining the weight matrix and bias vector is called as *Training.*

III DATA MINING

Data Mining is also called as data archaeology, data dredging, data harvesting, is the process of extracting hidden knowledge from large volumes of raw data and using it to make crucial business decisions[4]. The term

has been stretched beyond its limits to apply any form of data analysis. The use of computer technology in decision support is now widespread and pervasive across a wide range of business and industry. This has resulted data in immense volume and proportion.

The data are typically a collection of records where each individual record may correspond to a transaction or a customer and the fields in the record correspond to attributes. Very often, these fields are mixed types with some being numerical and some symbolic [4]. Data mining and knowledge discovery in database are concerned with extracting models and patterns of interest from large databases. It has been mostly used by statisticians, data analysts and the management information system (MIS) communities [7].

IV DATA SUMMARY

4.1 Data Sample

The data of the real estate properties used here in this study was arranged in the form of four hundred sets of thirteen element input vector and one element target vector. The attributes that were considered for the evaluation of the real estate properties were their geographical area, circle rate, years since the property was constructed, the market rate etc. The target values that were presented for the purpose of training the network contained an estimated cost of the property.

4.2 Training and Testing the Data Sets

For the purpose of analyzing the data, first a neural network was created using *newff* command in the neural network toolbox of MATLAB. This network consisted of the 18 hidden layer neurons. This network was given the input that was a set of 400 samples of real estate properties having 12 attributes in all. After creating the neural network, it was subsequently trained with the given sets of input and target vectors using the command line "*train(net, pInput, pTarget)"* in MATLAB. After the training process of the created neural network, the network was simulated using the command line *sim* in MATLAB, which generated the results in accordance to the input and target vectors.

As a property of the neural network toolbox presets in MATLAB, 60% of the input samples were used to train the neural network, 20% of the input samples were used for testing while the remaining 20% were used for the purpose of validation.

V RESULTS

5.1 After Training the Network

In all, the network took a total of 11 epochs to reach to the least mean squared error value, hence generating the results for the input data set with 6 validation checks. The post training analysis gave the value of slope 0.9109 which is equal to 1 in ideal case of perfect fit (output = target) and that of the correlation between the target and output to be 0.9455 which is also 1 in ideal case of perfect correlation, thus, depicting that the trained network reached close to that of the ideal case. The following figure 2 shows the post training analysis of the network

that most of the target fitted into the line of the targets that were presented to the network, showing a good regression. The dotted line shows the ideal case of mapping (outputs generated = target outputs) while the red line shows the one actual mapping of the targets to that of outputs generated by the network.

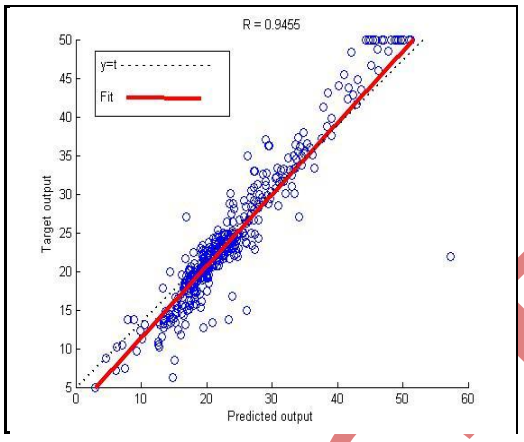

Figure 2: Post Training Analysis

5.2 Estimation of Property Value

After training the neural network with its performance depicted in figure 2, results were analyzed and verified based on various visualizations in MATLAB. The following figure 3 shows the clustering of the prices of the real estate properties that gives a good idea of the range of values in which most of the properties fall in consideration of the same kind of attributes for all the properties. Figure 3 shows that most of the properties lie in the range of 15 to 25 lakhs INR.

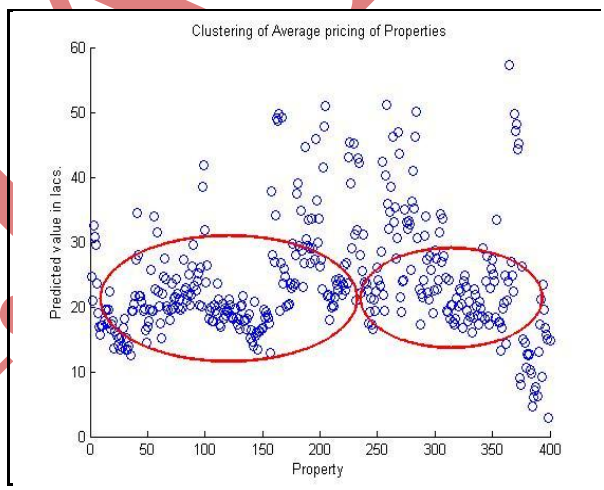

Figure 3: Common Range of Values of Properties

A more accurate estimate can be made from the results depicted in figure 4. It can be observed clearly that the properties that had the most common kind of attributes had the same prices and from figure it can be seen that majority of the properties lie in the range of 20 to 25 lakhs INR and an overall generalized result can be stated that most of the properties prices lie in the range of 15 to 25 lakhs INR.

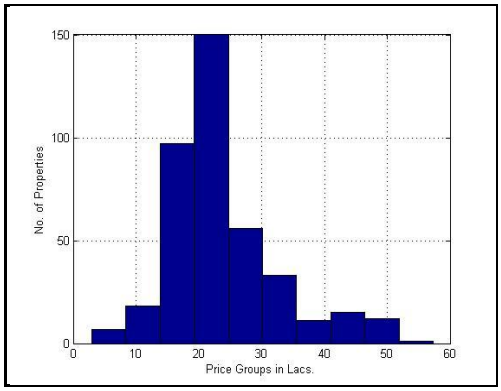

Figure 4: No. Of Properties Vs Prices

An average pattern of the property pricing can be seen from the following figure 5. This gives a good idea of the fact that what is average range of the majority of the properties.

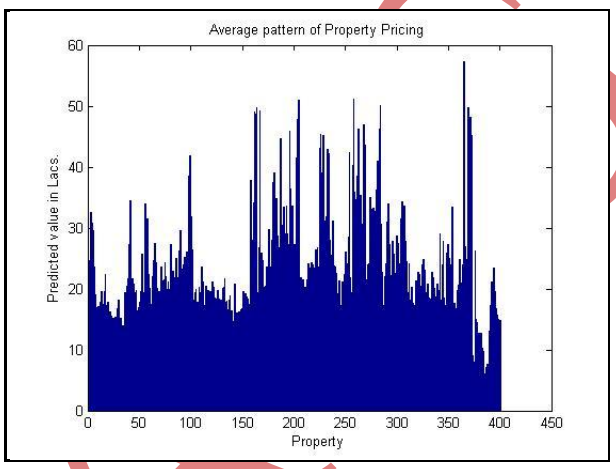

Figure 5: Average Pattern of Property Pricing

VI SUMMARY OF RESULTS

From the analysis mentioned in preceding sections shows that when a feed forward neural network was trained with a given set of input and target vectors, the network learned with a good precision, depicting results that were in compliance to the set of target outputs as shown in Fig. 2. Further results about the property pricing were analyzed through Fig. 3, Fig. 4 and Fig. 5. That majority of the real estate properties ranged between 15 to 25 lakhs INR.

VII CONCLUSION

The results from this study indicate that when data mining is done by a neural network model implemented through MATLAB, proves to be a good aid in carrying out reliable analysis of raw data. The ability to learn from given set of samples with a good precision and then generating the results in accordance to the target outputs being presented to the network shows that back propagation is an efficient method for studying the relationship between the input and output variables.

It is therefore concluded that the toolbox of neural network in MATLAB is an easy and robust aid for all the computational work in data mining, without any requirement of good coding skills. Also, MATLAB provides a great amount of flexibility for the data miners.

REFERENCES

[1] E.S. Gopi, "Back Propagation Neural Network", in *Algorithm Collections for Digital Signal Processing Using MATLAB*, Dordrecht, The Netherlands: Springer, 2007, pp. 24-29.

[2] V. Cheung, K. Cannons, "*An Introduction to Neural Networks*", Signal & Data Compression Laboratory, University of Manitoba, Manitoba, Canada, 2002.

[3] S. Mao, W. Wan, Y. Wang, Z. Wang, H. Yu, "*The application of an improved BP artificial neural network in data mining*", IET Int. Conf. On Smart & Sustainable City (ICSSC 2011), 2011, pp. 60-64.

[4] S. Prabhu, N. Venkatesh, "Data Mining and Warehousing Concepts", in *Data Mining and Warehousing*, New Age Int. Publishers, New Delhi, India, 2008, pp. 1-10.

[5] H.L. POH, T. JASIC, "Forecasting and Analysis of Marketing Data using Neural Network: A Case of Advertising and Promotion Impact", *Artificial Intelligence for Applications*, IEEE, L.A., USA, 20-23 Feb 1995, pp. 224-230.

[6] B. Yegnanarayanna, "Artificial Neural Networks: Terminology", in *Artificial Neural Networks*, PHI Learning Pvt. Ltd., New Delhi, India, 2009, pp. 24.

[7] W. Jian, S. Baohui, W. Yanwei, "Data Mining Technology and Its Possible Applications in City Planning", Proc. 7th IEEE Int. Conf. On Fuzzy Systems and Knowledge Discovery (FSKD 2010), Yantai, Shandong, 10-12 Aug 2010, pp. 2801-2804.