Using Neural Network with Statistical Analysis for Forecasting in E-Business

Thannob Aribarg, Siriporn Supratid

Department of Information Technology, Rangsit University, Pathumtani, 12000, Thailand
iamaribarg@yahoo.com, siri_sup@yahoo.com

ABSTRACT

A large number of relevant factors in forecasting problems is one of non-trivial difficulties in several e-business. In order to relieve this severe problem, the intelligent system: neural network based on main factors (NNMF) is proposed in this paper. The aim is to use only main factors, which are found by using statistical analysis for the forecast. Such a forecast relates to Boston housing data (BHD) problem. The comparison between NNMF and the model of neural network based on all factors is done. The result indicates the usage of significant reduction of the number of factors still maintains the accuracy as well as the use of all factors does. The experiments also denotes a very few bits of the decrease of accuracy when using the small number of hidden layer in NNMF.

Index Terms— Neural network, factor analysis, Rotated Component

1. INTRODUCTION

The forecasting problem is often involved with such several e-business as forecast pure water demand in a week for an E-business website [1], predicting the performance of dynamic e-Commerce systems on heterogeneous servers [2], predicting e-commerce consumer expenditure in European Commission(EC) countries [3], prediction of future facts enhances the decision making and the fulfillment of e-banking goals [4]. One of non-trivial difficulties in such many e-business is a large number of relevant factors involved with the prediction. In order to relieve this severe problem, the intelligent system: neural network based on main factors (NNMF) is proposed in this paper. Such a proposed method combines the statistical analysis with the neural network. First, various factors are evaluated by the statistical analysis; then the main factors are selected to construct a neural network model for the forecast. The forecasting problem applied here refers to Boston housing data (BHD) [5]. Such a problem relates to predict the median price of owner-occupied homes in Boston. There are many factors involved with forecasting such a median price. The factors such as crime rate, nitric oxides concentration, lower status, pupil-teacher ratio, property tax, number of rooms and so on are not independent and conflicted with each other. The comparison is made between NNMF and the model of neural network based on all factors. The experiments also concern the number of nodes in the hidden layer of neural network.

The rest of the paper is organized as follows: Section 2 introduces the principles and algorithm of neural network model. Section 3 conducts statistical analysis of the factors involved with a BHD median price of owner-occupied homes forecasting problem. Section 4 constructs the structure of the neural network model based on main factors. Section 5 shows the experiment result. Finally in section 6 some concluding remarks are drawn from this study.

2. ARCHITECTURE OF NEURAL NETWORK

Neural network have been widely used as a competitive tool in processing multivariable input-output hardware implementation because of their learning capacity, fault tolerance and 'model free' characteristics [6][7]. A neural network with one hidden layer can be shown in Fig.1

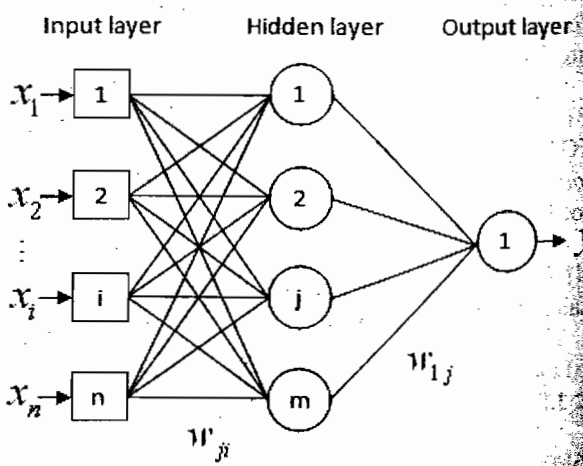

Fig. 1 Structure of neural network

The neural network is activated by applying inputs $x_1(t), x_2(t), \bullet\ , x_n(t)$, desired outputs $y_{d,1}(t)$ and t is time step.

Calculate the outputs of the neurons:

$$y_j(t) = sigmoid\left[\sum_{i=1}^{n} x_i(t) \times w_{ij}(t) - \theta_j\right] \quad (a)$$

$$y(t) = linear\left[\sum_{j=1}^{m} y_j(t) \times w_{1j}(t) - \theta\right] \quad (b)$$

Sigmoid, linear activation function is employed in the hidden and output layer as consecutively defined in eq. (a) and eq. (b)

where n is the number of inputs of neuron j in the hidden layer, m is number of inputs in the output layer, θ_j refers to neuron weights of hidden layer, θ refer to neuron weight of output layer.

Assume that all input values are given in advance the error value function can be expressed as:

$$E = \frac{1}{2}\sum_{i=1}^{r}(y_{d,i} - y_i)^2 \quad (c)$$

where r is the number of hidden neurons

Based on the BP algorithm, the parameter adjustment in neural network for backward propagation is shown in eq. (d-g):

Calculate the error gradient for the neurons in the output layer:

$$\delta_1(t) = y_1(t) \times [1 - y_1(t)] \times e(t) \quad (d)$$

Update the weights at the output neurons:

$$w_{1j}(t+1) = w_{1j}(t) + \alpha \times y_j(t) \times \delta_1(t) \quad (e)$$

Calculate the error gradient for the neurons in the hidden layer:

$$\delta_j(t) = y_j(t) \times [1 - y_j(t)] \times \delta_1(t) \times w_{1j}(t) \quad (f)$$

Update the weights at the hidden neurons:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha \times x_i(t) \times \delta_j(t) \quad (g)$$

where $0<\alpha<1$ is a leaning rate.

3. DATA SOURCE AND STATISTICAL ANALYSIS

3.1. Data Source

The data of samples used in this study refers to Boston Housing Data (BHD). The data relates to the housing values in suburbs of Boston which is maintained at Carnegie Mellon University. The factors, which affect BHD median price of owner-occupied home problem, such as per capita crime rate by town, proportion of residential land zoned for lots over 25,000 sq.ft., proportion of non-retail business acres per town, Charles River dummy variable, nitric oxides concentration, average number of rooms per dwelling, tax, pupil-teacher ratio by town and so on. 10 samples out of 506 samples is selected to show in this study.

All the data is summarized and shown in Table 1.

Table 1. 10 Samples out of 506 samples data selected

Index	1	2	3	4	5	6	7	8	9	10
CRIM	0.02731	0.02729	0.03237	0.06905	0.02985	0.08829	0.14455	0.21124	0.17004	0.22489
ZN	0	0	0	0	0	12.5	12.5	12.5	12.5	12.5
INDUS	7.07	7.07	2.18	2.18	2.18	7.87	7.87	7.87	7.87	7.87
CHAS	0	0	0	0	0	0	0	0	0	0
NOX	0.469	0.469	0.458	0.458	0.458	0.524	0.524	0.524	0.524	0.524
RM	6.421	7.185	6.998	7.147	6.43	6.012	6.172	5.631	6.004	6.377
AGE	78.9	61.1	45.8	54.2	58.7	66.6	96.1	100	85.9	94.3
DIS	4.9671	4.9671	6.0622	6.0622	6.0622	5.5605	5.9505	6.0821	6.5921	6.3467
RAD	2	2	3	3	3	5	5	5	5	5
TAX	242	242	222	222	222	311	311	311	311	311
PTRATIO	17.8	17.8	18.7	18.7	18.7	15.2	15.2	15.2	15.2	15.2
B	396.9	392.83	394.63	396.9	394.12	395.6	396.9	386.63	386.71	392.52
LSTAT	9.14	4.03	2.94	5.33	5.21	12.43	19.15	29.93	17.1	20.45
MEDV	21.6	34.7	33.4	36.2	28.7	22.9	27.1	16.5	18.9	15

3.2. Statistical Analysis

Note from Table 1 that there 13 attributes in the sample data, which are independent variables and denoted by X1, X2,..., X13 respectively. On the other hand, there is only one dependent variable that is the median price of owner-occupied homes of the boston housing data, denoted by y.

To select the main attributes from 13 attributes, a data pretreatment is done by means of statistical analysis in this study. The results obtained from SPSS indicate that there exist strong relations among 13 attributes, which can be shown by the Table2 and Table3 and component plot in rotated space in fig 2.

Table 2. Factors analysis

	Total	% of Variance	Cumulative %		Total	% of Variance	Cumulative %
1	6.127	47.130	47.130	8	.396	3.047	92.954
2	1.433	11.025	58.155	9	.277	2.130	95.084
3	1.243	9.559	67.713	10	.220	1.694	96.778
4	.858	6.597	74.310	11	.186	1.431	98.209
5	.835	6.422	80.732	12	.169	1.302	99.511
6	.657	5.057	85.789	13	.064	.489	100.000
7	.535	4.118	89.907				

Table 3. Rotated Component Matrix [a]

	Component		
	1	2	3
X1	.146	.744	.169
X2	-.773	-.002	-.252
X3	.711	.455	.205
X4	.321	-.087	-.539
X5	.764	.472	.013
X6	-.329	.001	-.763
X7	.813	.279	.056
X8	-.845	-.312	.021
X9	.296	.850	.157
X10	.368	.815	.210
X11	.180	.297	.633
X12	-.112	-.657	-.037
X13	.542	.351	.517

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

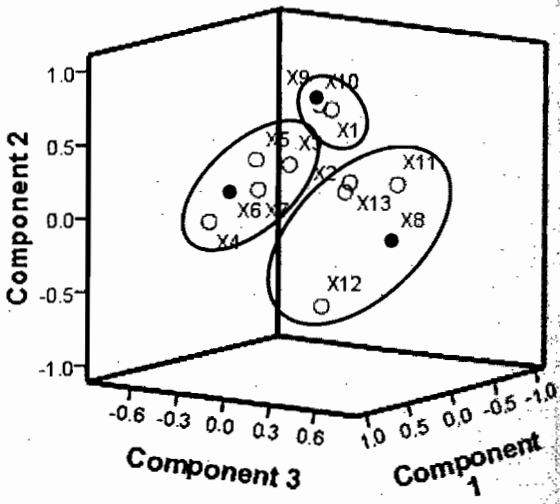

Fig 2. Component plot in rotated space

It is clearly seen from Fig. 2 that the 13 attributes are divided into 3 groups based on their relativities: group 1 includes X2, X8, X11, X12, X13; group 2 includes X1, X9, X10; and group 3 includes X3, X4, X5, X6, X7. Finally, the main indices are confirmed as: X6, X8 and X9 as a result of statistical analysis.

4. MODEL STRUCTURE OF NEURAL NETWORK BASED ON MAIN FACTOR (NNMF)

4.1. Data Pretreatment

After statistical analysis, the data information can be summarized in Table 4.

Table 4. Data information of BHD forecasting problem

Index	**1**	**2**	**3**	**4**	**5**
RM	6.421	7.185	6.998	7.147	6.43
DIS	4.9671	4.9671	6.0622	6.0622	6.0622
RAD	2	2	3	3	3
MEDV	21.6	34.7	33.4	36.2	28.7
Index	**6**	**7**	**8**	**9**	**10**
RM	6.012	6.172	5.631	6.004	6.377
DIS	5.5605	5.9505	6.0821	6.5921	6.3467
RAD	5	5	5	5	5
MEDV	22.9	27.1	16.5	18.9	15

Since there exist big differences among the data values, a normalized procedure is selected and used to improve the efficiency of NNMF model in this study, which yields the results shown in Table 5. .

Table 5. Normalized data information

Index	**X1**	**X2**	**X3**
1	0.5084	0	0
2	1	0	0
3	0.8797	0.6739	0.3333
4	0.9755	0.6739	0.3333
5	0.5142	0.6739	0.3333
6	0.2452	0.3652	1
7	0.3481	0.6052	1
8	0	0.6862	1
9	0.24	1	1
10	0.4801	0.849	1

4.2 Structure of NNMF

The neural network based on main factor determined by statistical classified into 4 models: NN13:13 with represents neural network 13 inputs and 13 hidden layer, NN13:3 13 inputs and 3 hidden layer, NNMF3:13 3 inputs and 13 hidden layer and NNMF3:3 refers to 3 inputs and 3 hidden layer as shown in figure 3 – 6 respectively.

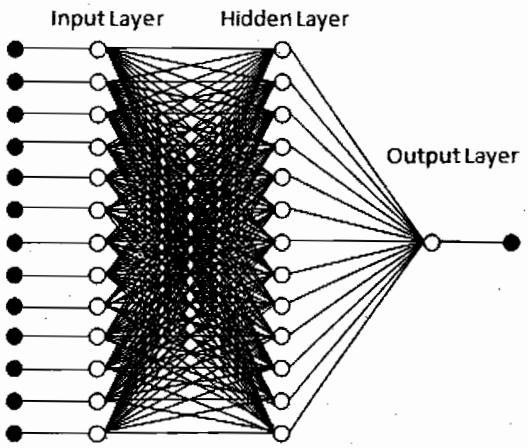

Fig 3. The structure of NN13:13

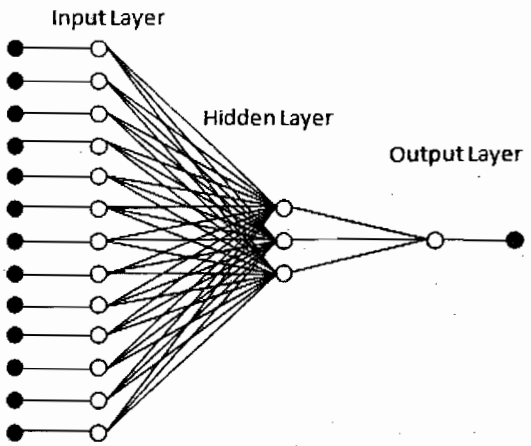

Fig 4. The structure of NN13:3

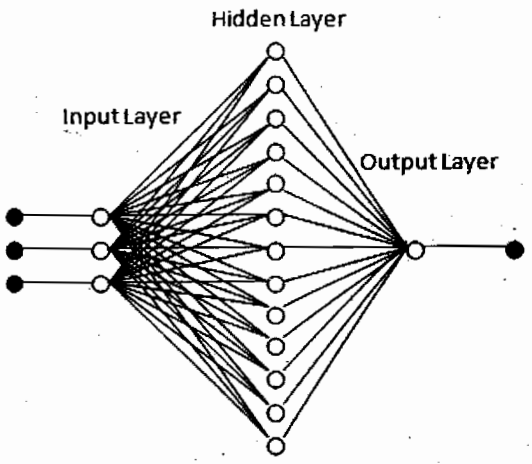

Fig 5. The structure of NNMF3:13

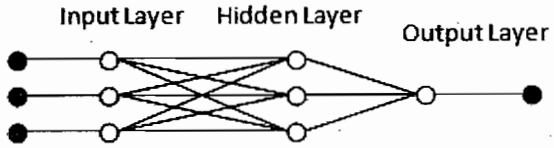

Fig 6. The structure of NNMF3:3

5. RESEULTS ANALYSIS

The forecasting performance of NN13:13 NN13:3 NNMF3:13 NNMF3:3 are illustrated as follows:

Table 6. Experiment results of 4 models

	Testing Error	Std. Error	Performance
NN13:13	0.058545	0.02817	99.941455
NN13:3	0.061786	0.020477	99.938214
NNMF3:13	0.057913	0.023519	99.942087
NNMF3:3	0.063803	0.024253	99.936197

The result indicates the usage of significant reduction of the number of factors still maintains the accuracy as well as the use of all factors does. The experiment denotes a very few bits of accuracy decrease: 0.003243% when reducing number of nodes in NN; and The accuracy decrease: 0.005893%when reducing number of nodes in NNMF. The performances of the 4 models are very close to each other.

6. CONCLUDING REMARKS

One of non-trivial difficulties in forecasting problem in a large number of e-business refers to a vast amount of involved factors. By using statistical method, the few number of main factors is extracted from the whole several numbers of factor. The results of this paper point that the usage of significant reduction of the number of factors with the small number of nodes in hidden layer could maintain the accuracy as well as the use of all factors with the large number of nodes in hidden layer. However, the results still depends much on how good the initial status of neural network is. In future works, such deficiency of initial state in neural network should be made less tedious than ever.

7. REFERENCES

[1] Qisong Chen, Yun Wu, Xiaowei Chen, "Research on Customers Demand Forecasting for E-business Web Site Based on LS-SVM," *isecs*, pp. 66-70, 2008 International Symposium on Electronic Commerce and Security, 2008

[2] David A. Bacigalupo, Stephen A. Jarvis, Ligang He, Graham R. Nudd, "An Investigation into the Application of Different Performance Prediction Techniques to e-Commerce Applications," *ipdps*, p. 248a, 18th International Parallel and Distributed Processing Symposium (IPDPS'04) - Workshop 14, 2004

[3] KovaÄ iÄ‡, Zlatko, 2004. "A predictive model for e-commerce consumer expenditure in EC countries", *MPRA Paper 5308*, University Library of Munich, Germany.

[4] *Vasilis Aggelis, Panagiotis Anagnostou,* "e-banking Prediction using Data Mining Methods",4th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED 2005), Salzburg, Austria.

[5] 、UCI Repository on machine learning database, www.ics.uci.edu/~mlearn/mlrepository.html

[6] K. Hornik, "Approximation capabilities of multi-layer feed forward networks", *Neural Networks*, Vol. 4, No. 2, 1991, pp. 251-257.

[7] B. A Jain., and B. N. Nag, "Performance evaluation of neural network decision models", *Manage Information Systems,* Vol. 14, No. 2, Fall 1997, pp. 201-216.