International Journal of the Physical Sciences Vol. 6(25), pp. 6082-6090, 23 October, 2011
Available online at http://www.academicjournals.org/IJPS
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

An evaluation of Iranian banking system credit risk: Neural network and logistic regression approach

Mahdi Salehi[1]* and Ali Mansoury[2]

[1]Accounting Department, Ferdowsi University of Mashhad, Mashhad, Iran.
[2]Management Department, Zanjan University, Zanjan, Iran.

Accepted 25 July, 2011

The current study seeks to provide a new solution for evaluation of banking system customers risk by integrating different scientific methodology. Evaluation of banking system customers risk in Iranian banks relies on experts judgment and fingertip rule. This type of evaluation resulted in high rate of postponed claims; therefore, designing new intelligent model for credit risk evaluation will be helpful, thus in this paper, we formulated an intelligent model by neural network and logistic regression that evaluated all individual customers credit risk without prejudice and discrimination. The result revealed that neural network and logistic regression have the same ability in predicting customer credit risk. Their ability in customer credit risk correct evaluation was nearly 79.50%. We suggested that both models could be used by all financial system as consultant model for customer credit risk prediction. The study also involved only one banking system credit customers, which concerns just Tehran city customers and its sample includes only individual customers, thus cannot be for institutional customers. Offering a case study, this paper presents a guide for banking system to predict any customer credit risk and regulate any customer loan in the light of customer risk that was extracted by neural network, and logistic regression employed different scientific methodologies in their service quality development efforts. Intending to offer scientific approaches to risk evaluation as a tool of customer credit risk assessment in banking system loan allocation procedures, this paper tries to bridge the current gap between academicians and practitioners; adds to the relatively limited theoretical literature.

Key words: Credit allocation, neural network, multilayer perceptron, logistic regression.

INTRODUCTION

Banking system is one of the most important effective factors that affect any country development (Salehi, 2008). However, in this way the correct performance of banking system plays an important role. Meantime banking system could allocate financial resources; suitably the economical growth can be expected. Then the correct banking performance depends on allocating financial resources to valid customer which share in GNP (Gross National Product) (Namazi and Salehi, 2010; Salehi et al., 2011). In another way, if banking system could allocate gathered financial resources suitably to the customer, allocated loans will be returned and banking system will succeed in short and long term. Customer credit risk is an important factor that guarantees both of these intentions.

Credit risk in banking system can be defined as probability of allocated credit return. Most commercial banks in Iran and other countries evaluate customer credit risk by subjective and judgment methods. Because of limitation of human ability in simultaneous processing of multi factor that affects credit risk, these methods does not have considerable efficiency comparison to statistical and artificial intelligence ones (Bernnan, 2005; Salehi and Yousefi, 2011).

In recent years, neural network and statistical analysis attain special position on financial market, evaluation of movable properties and even in health care (Lowe, 1994; Altman, 1994; Worzala et al., 1995; Ward System Group, 1995). However, neural network in comparison with other

*Corresponding author. E-mail: mahdi_salehi54@yahoo.com. Tel: +989121425323.

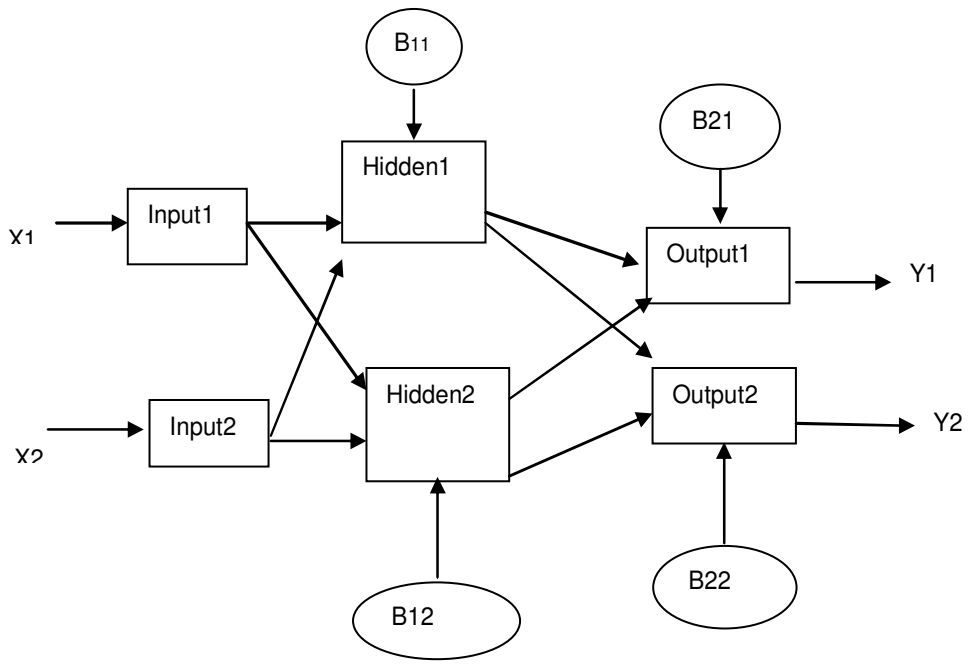

Figure 1. Multilayer perceptron neural network with two input and two output variables.

model such as discriminate analysis have particular position on estimation and forecasting (John, 2003). This type of model can be classified as generalized nonlinear regression model (Mcmenmin, 2004). Therefore, neural network is used in discrimination pattern of data and producing knowledge of data expansionary exhausted (Liao, 1999).

Extra flexibility of these models is another factor that has affected the application of this model in any branches of knowledge. For this reason, in addition to application to neural network in assessing credit risk of credit customer, result of these models was compared with logistic regression.

Neural network

One of the most important applications of neural network is predicting and forecasting of special variable by some inputs (John, 2003). This means that a series of input data is used to predict output variable by specific process. In some applications, neural network models get a series of data as input and output variable, and then such nonlinear regression estimate the relation between input and output variable (David, 2006). Figure 1 demonstrates this relation by multilayer perceptron neural network two input, two output and two hidden layers.

In this model, all data are divided into 2 groups including training and test data. The test data eliminates and training data is used to carry out model fitting process. After model fitting, the accuracy of model is assessed by test data. Most researchers consider 10 to 20% of all data as test data (Ward System Group, 1995), even though in some neural network models, the size of test data reaches 25%. It is important to notice that in time series model, the test data must be selected from latter data but in ordinary model, it must be selected randomly. The basic goal of test data is to forecast solutions for these data and then compare them with actual solutions. It is obvious that the differences between forecasted solutions and actual solutions for dependent variable demonstrate error quantity which is calculated by absolute error or mean squared error. Sometimes, test data set is defined as product set (Ward System Group, 1995). Figure1 demonstrates one of the important neural networks called multi layer perceptron.

As observed, this network has an input, one output layer and one hidden layer. In this network, x_1 and x_2 demonstrate input vector while y_1 and y_2 is output vector.

Then this network including 2 input and 2 output variable and $w^i jk$ specify the weight of j variable in k neuron in layer i.

Any neurons in the hidden layer include a transformation function thus, the output of one of the hidden layer as an example can be represented as follows:

$$Y^H = g(\sum_{i=1}^{2} w_{jk}^i x_j) + w_{jb}$$

Knowing hidden layer function quantity, output layer quantity can be calculated as follows:

$$Y^O = \sum_{h=1}^{2} w^2{}_{kj}(g(\sum w^i_{jk} x_j) + w^1{}_{jb}) + w^2_{jb}, k = 1,2$$

This is important in neural network that is related to calculation of suitable weight. It is obvious that after determining network weight, we will able to compute output vector easily. Weight detection in neural network applies an efficient algorithm that is defined as learning algorithm.

Learning algorithm

Network weight can be calculated in many ways, but we have described back propagation algorithm as one of the most important and applicable methods which often is used in multilayer perceptron neural network (Cudill, 1995). As the name of the algorithm implies, the error quantity returns to neural network and moderates the weight.

Thus using these concepts, we can define weight moderator equation. For this purpose assume that $x_j^{[s]}$ be j neuron's input at S layer and $f(I_i^{[s]})$ be rhythmic sum for I neuron's at S layer and $e_k^{[s]}$ be local error in k neuron's at S layer. Then weight vector moderator equations can be defined as follows (Hayking, 1994):

$$\Delta wij^{[s]} = \eta f(I_i^{[s]}) \sum_k (e_k^{[s+1]} wki) X_i^{[s-1]} + \theta \Delta wij \quad (1)$$

And without output layer, learning function can be defined as follows:

$$\Delta wij^{[s]} = \eta (Y_i - O_i).f[I_i] X_i^{[s-1]} + \theta \Delta wij \quad (2)$$

In these equations, η indicate learning equations and θ indicate momentum terms. Thus applying Equations 1 and 2 depend on network type.

For this purpose, research divides neural network into two groups; supervised and unsupervised network. When the model specifically have both input and output variable, the model can determine error quantity, and to the bases of this, error can improve weight vector. This group of model is defined as supervised network such as generalized regression neural network model, but sometimes input vector is completely specific but output vector is not specific. This type of model is defined as unsupervised network such as classifying and optimization of linear and nonlinear models.

In optimization model, the quantity of optimal solution is not specific but improvement amounts in iterations are completely specific. This means that model move to optimal solution. Thus in this type of model, learning loop will be continued till it reaches optimal solution. In learning process, neural network may stop in local minima, to avoid this difficulty we can use random training data and selecting suitable amount for learning rate. Selection of small rate leads to decrease in learning process and from another angle, section of big learning rate leads to divergence. However, selection of small coefficient would eliminate oscillations; still, sometime it is better to give large rate to learning rate, in this way, learning rate will be decreased gradually up to model conversancy. While we are permitted to iterate an algorithm, there are three ways in stopping an algorithm.

This includes (Mcmenmin, 2004):

1. Stop algorithm after specific iteration.
2. Stop algorithm after decreasing error up to determined limit.
3. Occurrence stability on neural network by means that after some alterations no improvement can be seen.

Size of neural network

There are some important questions in designing neural network models. First, how many neurons (processor) are needed in specific network? Secondly, assuming specific neurons, how can we determine hidden layer number?

Up till now, there is no obvious response to these questions. In 1986, a statistical method known as discriminating approach was suggested for determining size of neural network but finally, it did not represent an applicable pattern (Rumelhart et al., 2006). In 1993, a series of procedures were developed that optimized fitting measures (Refenes et al., 1993). But in 1996, these procedures were criticized and rejected by many researchers (Swinger, 2008). At that time to recent years, many techniques about topology of neural network have been presented but none of them have necessary comprehensiveness. Thus, up till now, often heuristic methods were developed for selection of neuron's number and hidden layer amount. However, while number of neurons increases the flexibility of network rise, surplus flexibility in network does not suggest anywhere because it may lead to over fitting. Over fitting means that the model in the learning process, amount of training error and generalization of model for new data decreases. Figure 2 demonstrates a sample model of over fitting (Liao, 1999; Back and Back, 1995).

About the number of hidden layers present in a neural network, most authors believe that a hidden layer is sufficient, but in this case the learning rate will be decreased, and adding a second hidden layer, learning speed will be raised. Sometimes, step by step methods similar to regression model are used to determine neuron's number. In any case, using network with greatly/numerous neuron is efficient than rigid network with constant neurons (Rumelhart et al., 2006). In assay,

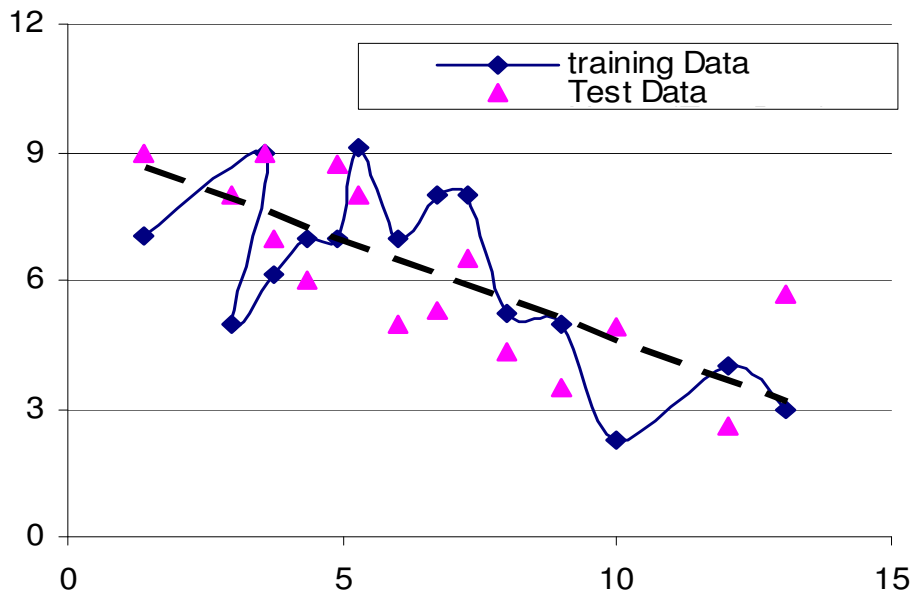

Figure 2. A sample model of over fitting.

we used step by step method to detect the number of neurons.

Logistic regression

One of the most important classic models that are often used in anticipating model refers to regression model. When dependent variable would have binary form, we can use Logistic regression. Its relation is defined as in Equation 3.

$$\pi(x) = \frac{e^{\sum \beta_i x_i + \beta 0}}{1 + e^{\sum \beta_i x_i + \beta 0}} \qquad (3)$$

When actual quantity of dependent variable is 1, we would expect that $\pi(x)$ get close to 1 and vice versa. In this assay, the credit risk of customer is defined as response variable with binary (0 or 1) value. Therefore we apply logistic regression for anticipating dependent variable (0) for waste or expired deadline claims/ requisition and 1 for live requisition. Then in relation of Equation 3, x_i denote dependent variable and β_i demonstrate coefficient of x_i that will be estimated by model and $\pi(x)$ is anticipated risk for customer credit risk.

RESEARCH METHODOLOGY

In this research, we selected 330 Mellat Bank customers' randomly. Then 51 customers were eliminated because of conflict and defective data, thus, the model was designed for 279 credit customer cases. Among these cases, there were 231 live claim and48 were expired deadline or waste claim.

In order to evaluate neural network models efficiency at beginning, the data were analyzed by classical statistic method. For this purpose, after recognizing logistic regression as favorite classic model using Pearson coefficient, we detected the important effective variables that had significant effects on credit risk (response variable) and eliminated those variables that had no significant effects. Then we used step by step logistic regression to finalize model and eliminate some variables because of co-linearity. Therefore at this step, we detected all important variables. In other word, at the first step we surveyed available 50 independent variables and then recognized 11 significant variables for each customer as estimator variable. These variables include:

1. Type of company activity as indicator variables
a. productive x_{1_1}
b. agricultural x_{1_2}
c. commercial x_{1_3}
d. service x_{1_4}
2. Record or activity history x_2
3. Recorded capital x_3
4. Work history of company management x_4
5. Sum of current properties x_5
6. Sum of fixed properties x_6
7. Sum of debt x_7
8. Sum of debit account circulation next to bank x_8
9. Sum of creditor account circulation next to bank x_9
10. Return of capital x_{10}

These variables were defined as independent variable in logistic regression model and as input variable in neural network model.

We defined Y as credit risk of customers. This variable was defined as response variable in logistic regression and as output variable in neural network.

Then in the next step, we conducted three distinct data base (3 different file) with the same data then in each group using systematic sampling selected and put aside 38 cases (almost 14%)

Table 1. Weight vector of input variable and bias in first layer.

H1_06	H1_05	H1_04	H1_03	H1_02	H1_01	
-1.485	-7.577	-3.092	0.4259	-8.8	-2.682	Bias
0.63	2.126	2.888	-1.091	0.9455	-0.4591	x1_1
-0.1985	-0.3579	-1.203	-1.915	-0.2781	0.3505	x1_2
-0.1944	-1.068	-0.1683	-2.326	0.5368	1.096	x1_3
1.24	-0.1947	-2.002	-1.469	-0.1761	2.058	x1_4
2.37	1.414	-0.8578	9.255	3.335	-3.81	x2
1.017	-1.445	-0.1611	-0.1683	-0.5752	0.2579	x3
-0.757	6.061	1.502	-1.453	-29.03	-0.8135	x4
1.639	-1.311	-1.563	0.5444	-3.256	-2.936	x5
0.1768	0.62	-0.8266	1.673	-0.9697	0.9887	x6
0.4466	-3.995	-3.069	0.8214	7.485	-1.949	x7
1.622	-3.534	-1.07	6.022	-1.287	0.6473	x8
1.251	-6.967	-1.537	7.823	-8.103	-0.3174	x9
-0.2302	-4.25	-0.00938	0.5329	-2.138	-0.1497	x10

Table 2. Weight vector of bias variables and hidden layer neurons.

h1_06	h1_05	h1_04	h1_03	h1_02	h1_01	BIAS
-4.417	17.11	2.468	-8.831	14.36	2.879	3.066

as test data (for comparing the efficiency of two models after modeling) and fit logistic and neural network model for reminder data distinctly in each group. In other word, 38 cases were not used in fitting model by two methods and model design to the reminder data (241 cases in each group). It is necessary to remember that three layer perceptron neural network was used in designing neural network model in each group.

DATA ANALYSIS AND DESIGNING MODEL

In this paper, in order to design logistic regression model, we used SPSS 15 and statistical neural network for designing neural network model. Therefore the models parameters-independent variable coefficient and constant quantity in logistic regression model and weight vector plus bias variable of each layer in neural network- for each group using Training Data was estimated and the model designed. Then three regressions and three neural network models were designed. We bring neural network and logistic regression model for group one.

Logistic regression

The logistic regression model that was designed for this group to estimate credit neural network model is as follows:

$$\ln\left(\frac{y}{1-y}\right) = 1.1333x_{1\_1} - 1.5792x_{1\_2} - 0.9657x_{1\_3} + 1.034x_{1\_4} - 0.273x_2 - 0.0003x_3 - 0.2446x_4 - 0.00005x_5 - 0.0001x_6 - 0.000006x_7 - 0.00002x_8 - 0.0009x_9 - 0.0017x_{10} + 4.9299$$

In this model, Y denotes customer credit risk. Thus we can show that:

$$p = \pi(x) = \frac{e^{\sum \beta ixi + \beta 0}}{1 + e^{\sum \beta ixi + \beta 0}}$$

In this model, $p \in (0,1)$ and customer credit risk is specified; as much as P gets close to 0, it shows that credit risk is high. In other word, the probability of returning loan by corresponding customer is low and as much as P get close to 1, it shows that credit risk is low, in other word, the probability of returning loan by corresponding customer is high.

Neural network model

The neural network used for this case was a three layer perceptron with a hidden layer. After designing the model by step wise method, it determined that the final neural network has six neurons in hidden layer, and the function used in each neuron was a radial base function. The following tables demonstrate weight vector plus bias variable in each layer. Table 2 demonstrates each neuron coefficient in hidden layer for output variable.

Comparing logistic regression model and neural network evaluation of neural network, and classical model will be possible when the estimated parameters in two models are applied to estimate dependent or output variable with test data. Then response variable will be

Table 3. Comparing result of neural network and logistic regression with observed value for group 1.

Number of test cases	Observed value	Neural net work	Logistic regression
1	1	Right	1
2	1	Right	1
3	1	Right	1
4	1	Right	1
5	1	Unknown	1
6	1	Right	1
7	1	Right	1
8	1	Right	1
9	1	Right	0
10	1	Right	1
11	1	Right	1
12	1	Right	1
13	1	Right	1
14	1	Right	0
15	1	Right	1
16	1	Right	1
17	1	Right	1
18	1	Right	0
19	1	Right	1
20	1	Right	1
21	1	Right	1
22	1	Unknown	1
23	1	Right	1
24	0	Unknown	0
25	0	Right	0
26	0	Right	1
27	0	Right	0
28	0	Wrong	0
29	0	Right	0
30	0	Right	1
31	0	Right	1
32	0	Right	0
33	0	Right	0
34	0	Unknown	0
35	0	Unknown	0
36	0	Right	0
37	0	Right	1
38	0	Unknown	0
Number of correct	-	31	32
Incorrect	-	2	6
unknown	-	5	-
Correct prediction percent		81.5	84.2

compared with actual value to determine error quantity in each of the three groups.

It is obvious that any model that has had low error will be efficient in front of another model. For this purpose using neural network and logistic regression model, we estimate credit risk (response variable) for test data in each group and then compare with actual value. The result is summarized in Tables 3 and 4.

As shown in Tables 3 and 4, the ability of multilayer perceptron neural network and logistic regression is nearly the same. The correct prediction percent in Table 3 for logistic regression is greater than correct prediction

Table 4. Comparing result of neural network and logistic regression with observed value for group 2.

Number of test cases	Observed value	Neural net work	Logistic regression
1	1	Right	0
2	1	Wrong	1
3	1	Wrong	1
4	1	Right	1
5	1	Right	1
6	1	Right	1
7	1	Right	1
8	1	Unknown	1
9	1	Unknown	1
10	1	Right	1
11	1	Right	1
12	1	Right	1
13	1	Unknown	1
14	1	Right	0
15	1	Unknown	0
16	1	Right	1
17	1	Right	1
18	1	Unknown	1
19	1	Right	1
20	1	Right	1
21	1	Unknown	1
22	1	Unknown	1
23	1	Right	1
24	0	Right	0
25	0	Right	0
26	0	Right	0
27	0	Right	0
28	0	Wrong	0
29	0	Right	1
30	0	Right	1
31	0	Unknown	0
32	0	Right	1
33	0	Right	1
34	0	Right	1
35	0	Right	1
36	0	Right	0
37	0	Right	0
38	0	Right	0
Number of correct	-	29	28
Incorrect	-	3	10
Unknown	-	6	_
Correct prediction percent		76.3	73.6

percent for neural network (84.3 in front of 81.4) while in Table 4, percent of correct prediction for neural network is greater than for logistic regression and Table 5 demonstrate same correct prediction for both model.

Conclusion

Regarding the importance of correct allocation of credit in banking system, we attempted to design a mathematical model for evaluating credit risk of customer. For this purpose, neural network and logistic regression were compared and analyzed to obtain an efficient model for credit risk evaluation. At the first the logistic regression was designed and applied to test data, then, neural network was modeled and applied to the test data. The result revealed that both models have same efficiency.

Table 5. Comparing result of neural network and logistic regression with observed value for group 2.

Number of test cases	Observed value	Neural net work	Logistic regression
1	1	Wrong	0
2	1	Right	1
3	1	Unknown	0
4	1	Right	1
5	1	Right	1
6	1	Wrong	0
7	1	Right	1
8	1	Right	1
9	1	Right	1
10	1	Unknown	0
11	1	Unknown	1
12	1	Right	1
13	1	Right	1
14	1	Right	1
15	1	Right	1
16	1	Unknown	1
17	1	Right	1
18	1	Right	1
19	1	Right	1
20	1	Right	1
21	1	Right	1
22	1	Right	1
23	1	Unknown	1
24	0	Right	0
25	0	Right	1
26	0	Right	0
27	0	Right	0
28	0	Right	0
29	0	Right	0
30	0	Right	0
31	0	Right	0
32	0	Right	0
33	0	Right	0
34	0	Right	1
35	0	Right	1
36	0	Unknown	1
37	0	Right	0
Number of correct	-	29	29
Incorrect	-	2	8
Unknown	-	6	—
Correct prediction percent		78.3	78.3

Thus, it can be said that the multilayer perceptron and logistic regression have the same ability in estimating binary dependent variable. However at the practice, the neural network use black box process in model designing since logistic regression parameter is meaningful and simply can be used as factor to determine direct and indirect effect of independent variable to dependent one. Thus regarding same ability in credit risk prediction and logistic regression perfects; it is very important to note that logistic regression is more efficient than Altman model. In logistic model, the response variable demonstrate the customer credit risk in [0,1] range that can interpret easier because it is very near to financial concepts but in Altman discriminate analysis, model z-

score value have unbounded range and result interpreting is very difficult. Other advantages of this model in front of another model are related to its independent variable that can be detected in an easy way in financial system. Paying attention to the models ability of both models can be used in all financial system as consultant software.

It is very important to remember that the ability of model depend on correction of input data that will be used in the model, which is very important. If input data be false then the result will be incorrect.

REFERENCES

Altman EI (1994). Corporate distress diagnosis: Comparisons using discriminate analysis and neural network (The Italian Experience). J. Bank Finance, 18: 505-529.

BacK B, Back RJ (1995). Financial statement planning in the presence of tax constraints, J. Operat. Res., 85(1): 66-81.

Bernnan PJ (2005). Promise of artificial intelligence remains elusive in banking today, Bank Manag., pp. 49-53.

Buchmen TG, Kubos K.L, Seidler AJ, Siog MJ (1994). A comparison of statistical and connectionist models for the prediction of Chronicity in a Surgical intensive Care unit, J. Crit. Care Med., 22(5): 750-762.

David W (2006). Neural network credits scoring models, Comput. Operat. Res., 27: 1181-52.

Hayking S (1994). Neural network: a comprehensive foundation, MacMillan, New York: pp.144-152.

John C, Cooper B (2003). Artificial neural network versus multivariate statistics: an application from Economics, J Appl. Statistics, 26 (8): 909-921.

Liao TF (1999). A modified GMDH approach for Social Science research: exploring Patterns of relationships in the data, Quality and Quantity, 26: 19-38.

Lowe D (1994). Novel explanation of neural network methods in financial markets, IEEE, Transactions on Neural Networks: pp. 3623-3628.

Mcmenmin JS (2004). A primer on neural network for forecasting, J. Bus. Forecast, 16 (3): 17-22.

Namazi M, Salehi M (2010). The Role of Inflation in Financial Repression: Evidence of Iran. World Appl. Sci. J., 11(6): 653-661.

Refenes AN, Zapranis AD, Francis G (1993). Neural network applications in financial asset management, J. Neural. Comput. Appl., 1 (1): 52-59.

Rumelhart DE, Hintom GE, Williams RJ (2006). Learning internal Representation by error Propagation, in Rulhart, D.E. (Ed) Parallel Distributed Proc., 1: 818-62.

Salehi M (2008). The Role of Financial Intermediaries in Capital Market, Zagreb Int. Rev. Econ. Bus., 6(1): 97-109.

Salehi M, Hematfar M, Khatiri M (2011). Entrepreneurship Gap in Islamic Banking Sector: Empirical Evidence of Iran, Afri. J. Bus. Manag., 5 (8): 3322-3329.

Salehi M, Yousefi Z (2011). Factors Affecting to Entrepreneurship in Banking Sector: Empirical Evidence of Iran. Int. J. Acad. Res., 3(2): 916-921.

Swinger K (2008). Financial prediction: some pointer and common errors, neural computing and applications. 4: 42-53.

Ward System Group (WSG) (1995). Neuroshell users guide, Frederick, MD: pp.101-108.

Worzala E, Lenk M, Silvia A (1995). An exploration of neural networks and its application to real state valuation, J. Real State Res., 10 (2): 185-201.