Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 241690, 10 pages
doi:10.1155/2012/241690

Research Article

A New Hybrid Method Logistic Regression and Feedforward Neural Network for Lung Cancer Data

Taner Tunç

Department of Statistics, Faculty of Science and Lecture, Ondokuz Mayis University, 55139 Samsun, Turkey

Correspondence should be addressed to Taner Tunç, ttunc@omu.edu.tr

Received 24 July 2012; Revised 3 October 2012; Accepted 6 November 2012

Academic Editor: Marek Lefik

Logistic regression (LR) is a conventional statistical technique used for data classification problem. Logistic regression is a model-based method, and it uses nonlinear model structure. Another technique used for classification is feedforward artificial neural networks. Feedforward artificial neural network is a data-based method which can model nonlinear models through its activation function. In this study, a hybrid approach of model-based logistic regression technique and data-based artificial neural network was proposed for classification purposes. The proposed approach was applied to lung cancer data, and obtained results were compared. It was seen that the proposed hybrid approach was superior to logistic regression and feedforward artificial neural networks with respect to many criteria.

1. Introduction

Data classification problems can be encountered in many fields such as medicine and economy. The basic statistical method used for data classification problem in the literature is logistic regression. Logistic regression is a white-box method which can test the efficacy of explanatory variable in the model. Another technique used for classification problem is artificial neural networks (ANNs). Although artificial neural network, a black-box method, has a strong modeling ability, it cannot interpret the coefficients in the model. However, the advantage of artificial neural network against logistic regression is that it is a data-based approach not a model-based. In their study, Dreiseitl and Ohno-Machado [1] compiled logistic regression and artificial neural networks literature about the analysis of data classification. In their study [1], Dreiseitl and Ohno-Machado compared logistic regression to artificial neural networks by analyzing the study results in the literature. In the study by Paliwal and Kumar [2], the literature about the use of artificial neural networks in medicine was outlined. Kurt et al. [3] compared LR to ANN for coronary artery disease. In the study by Oner et al. [4], LR and ANN techniques were compared in terms of lung cancer data. Chang and Hsu [5] compared logistic regression, ANN, and logistic regression based on genetic algorithm for pancreatic cancer data. In this study, we proposed a new method based on hybrid approach of logistic regression and artificial neural networks. We compared the proposed method with logistic regression and artificial neural networks in terms of lung cancer data. In the second chapter, brief information about logistic regression method was given. The third chapter dealt with feedforward artificial neural networks. In the fourth chapter, the new proposed hybrid method was introduced. In the fifth chapter, the method was applied to lung cancer data, and the proposed method was compared with the other methods in the literature. In the last chapter, obtained results were interpreted.

2. Logistic Regression

LR is a regression method for predicting a binary dependent variable. The dependent variable takes 0 or 1 values. The conditional probability for dependent variable is given as follows:

$$P\left(Y = \frac{1}{X}\right) = \pi(X) = \frac{e^{\beta' X}}{1 + e^{\beta' X}}, \tag{2.1}$$

where $\beta' X = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$, and k is number of independent variables. This formula is implied that $\pi(X)$ increases or decreases as an S-Shaped function of independent variables. The probability distribution of dependent variables is given as follows:

$$P(Y_i = y_i) = \begin{cases} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} & y_i = 0 \text{ or } 1 \\ 0, & \text{other wise.} \end{cases} \tag{2.2}$$

The likelihood function is the product of these probabilities, and the logarithm of likelihood function is given as follows:

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\beta' X_i) - \sum_{i=1}^{n} \log_e(1 + \exp(\beta' X_i)). \tag{2.3}$$

The parameters of logistics regressions are estimated via maximizing logarithmic likelihood function. The nonlinear optimization methods are used for maximizing logarithmic likelihood function. Another problem in logistic regression is selecting independent variables. The stepwise method, backward and forward selection methods are generally preferred in the literature.

3. Feedforward Neural Networks

Artificial neural network is a data processing mechanism generated by the simulation of human nerve cells and nervous system in a computer environment. The most important feature of artificial neural network is its ability to learn from the examples. Despite having a simpler structure in comparison with the human nervous system, artificial neural networks

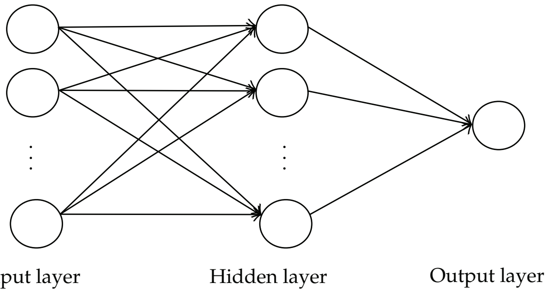

Figure 1: Multilayer feedforward artificial neural network with one output neuron.

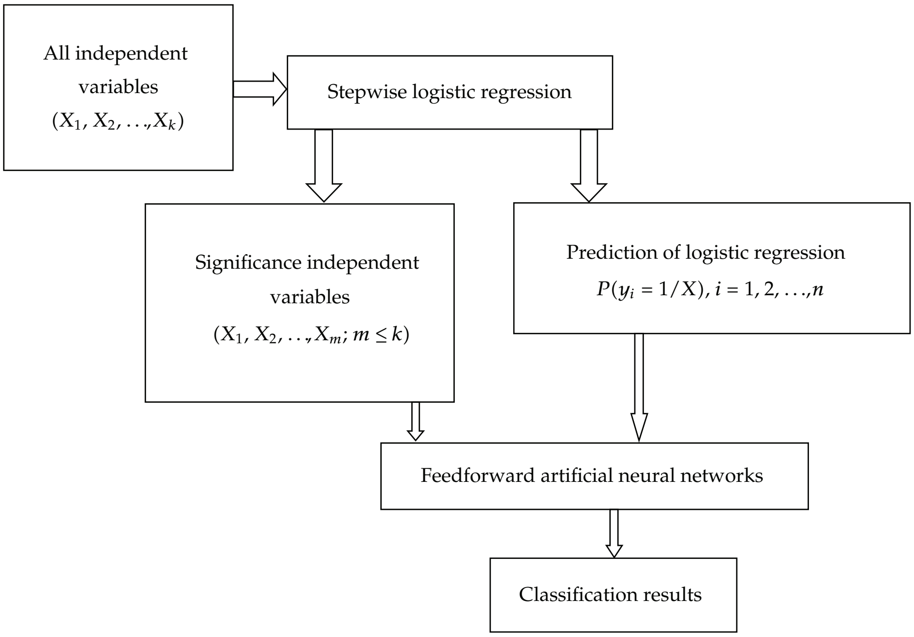
All independent variables $(X_1, X_2, \ldots, X_k)$
Stepwise logistic regression
Significance independent variables $(X_1, X_2, \ldots, X_m;\ m \le k)$
Prediction of logistic regression $P(y_i = 1/X),\ i = 1, 2, \ldots, n$
Feedforward artificial neural networks
Classification results

Figure 2: Flow diagram of proposed hybrid method.

provide successful results in solving problems such as forecasting, pattern recognition and classification.

Although there are many types of artificial neural networks in the literature, feedforward artificial neural networks are frequently used for many problems. Feedforward artificial neural networks consist of input layer, hidden layer(s), and output layer. An example of feedforward artificial neural network architecture is shown in Figure 1. Each layer consists of units called a neuron, and there is no connection between neurons, which belong to the same layer. Neurons from different layers are connected to each other with their weights. Each weight is shown with directional arrows in Figure 1. Bindings shown with directional arrows in feedforward artificial neural networks are forward and unidirectional. Single activation function is used for each neuron in hidden layer and output layer of

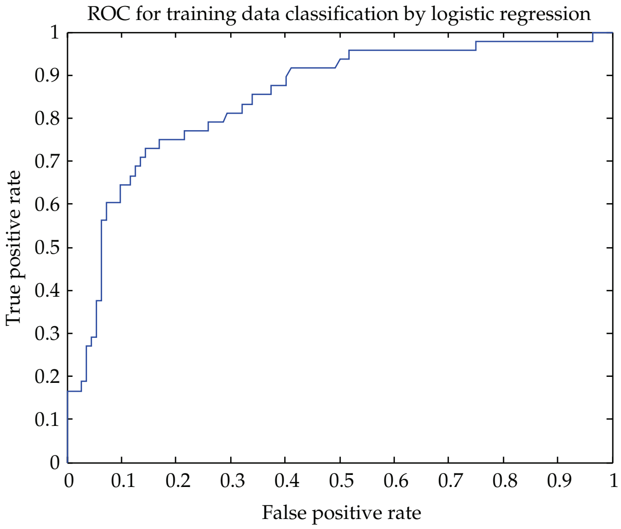

Figure 3: ROC Curve for training data in logistic regressions.

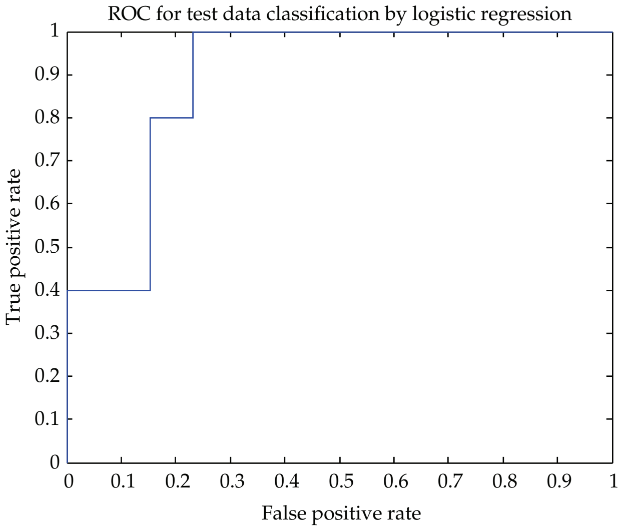

Figure 4: ROC Curve for testing data in logistic regressions.

feedforward artificial neuron network. Inputs incoming to neurons in hidden and output layer are made up multiplication and addition of neuron outputs in the previous layers with the related weights. Data from these neurons pass through the activation function and neuron output are formed. Activation function enables curvilinear match-up. Therefore, nonlinear activation functions are used for hidden layer units. In addition to a nonlinear activation function, linear (pure linear) activation function can be used in output layer neuron.

In feedforward artificial neural networks, learning is the determination of weights generating the closest outputs to the target values that correspond with the inputs of artificial neural network. Learning is achieved by optimizing the total errors with respect to weights. There are several types of training algorithms in the literature used for learning

Table 1: Estimation results of logistic regression.

Variables	Estimations	Standard errors	*t*-stat	*P*
Constant	−5,19	1,02	−5,07	0,00
Age	0,06	0,01	4,03	0,00
Time of HM	0,03	0,01	2,36	0,01
Number of HM	1,44	0,45	3,20	0,00
RAL	−1,26	0,49	−2,56	0,01

Table 2: Optimal weights of neural network.

Hidden layer neurons	Input neurons				Output neuron	Bias (input-hidden)	Bias (hidden-output)
	1	2	3	4			
1	0,32	0,32	−1,57	−6,13	6,371389	−13,34	0,50
2	−0,52	−0,19	−0,82	3,89	8,982728	−5,31	
3	−0,13	−1,64	−10,51	−2,97	−5,06077	−8,56	
4	1,04	1,36	5,78	6,87	−7,04415	−11,90	
5	0,16	0,62	−8,57	1,71	−4,95642	12,34	
6	0,31	−0,08	3,75	−3,69	4,679567	−15,32	
7	−1,14	3,94	12,95	−7,24	3,461384	7,62	
8	−1,50	−0,46	−2,81	2,81	−10,4971	−1,13	
9	−0,46	0,22	6,29	−2,44	−9,7104	−4,79	
10	−2,66	1,57	−4,24	−5,29	−1,1065	9,20	

of feedforward artificial neural networks. One of the widely used training algorithms is the Levenberg-Marquardt (LM) algorithm which was also used in this study. Matlab Package Program: Neural Network Toolbox is used for the ANN solutions.

4. A New Logistic Regression and Feedforward Neural Network Hybrid Method

Logistic Regression is a model-based white-box method in which coefficients can be interpreted. Therefore, in logistic regression method, the efficacy of explanatory variables in the model can be tested, and variable selection can be done easily. In logistic regression, forward, backward, and stepwise selection methods have been used for variable selection. Additionally, Bayesian variable selection methods were applied to logistic regression model in Chen and Dey [6] and Ghosh and Yuan [7] studies. Stacey and Kildea [8] selected variables using genetic algorithm, whereas Pacheco et al. [9] used tabu search algorithm in logistic regression. As artificial neural network is a black-box model, coefficients cannot be interpreted and thus statistical techniques such as stepwise technique cannot be used for variable selection. The use of genetic algorithm, tabu search algorithm, and artificial intelligence for variable selection is possible in artificial neural networks. In this study, a new hybrid approach which takes advantages of logistic regression and artificial neural networks was proposed. The proposed new approach is given below as an algorithm consisting of two steps. The flow diagram of proposed hybrid method is given in Figure 2.

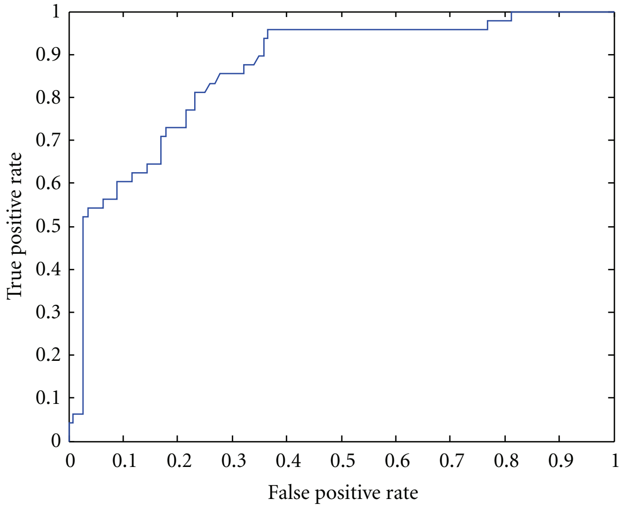

Figure 5: ROC Curve for training data in ANN.

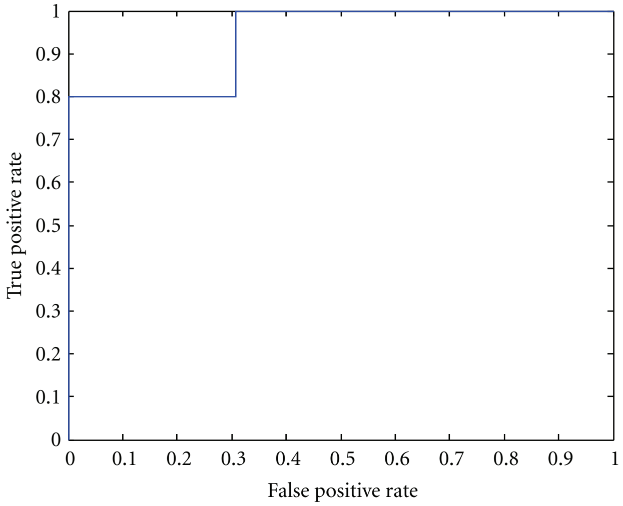

Figure 6: ROC Curve for test data in ANN.

Algorithm 4.1. Step 1. For all possible explanatory variables $(X_1, X_2, \ldots, X_k)$, significance explanatory variables $(X_1, X_2, \ldots, X_m; m \leq k)$ and logistic regression forecasts $P(y_i = 1/X)$, $i = 1, 2, \ldots, n$ are obtained by applying stepwise variable selection and logistic regression method.

Step 2. Classification is done by means of feedforward artificial neural networks using important explanatory variables which were obtained in the first step and forecasts of logistic regression as input of artificial neural networks.

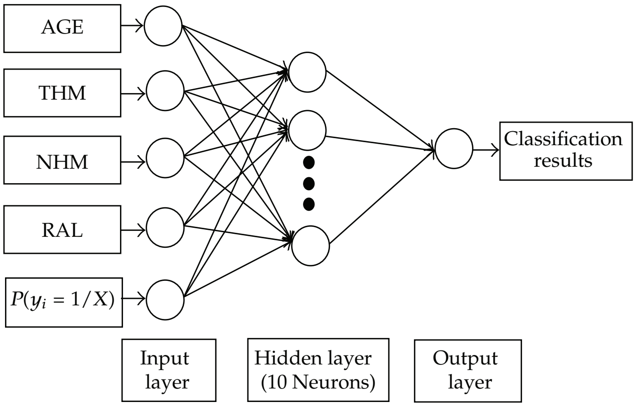

Figure 7: Artificial neural network architecture of proposed hybrid method.

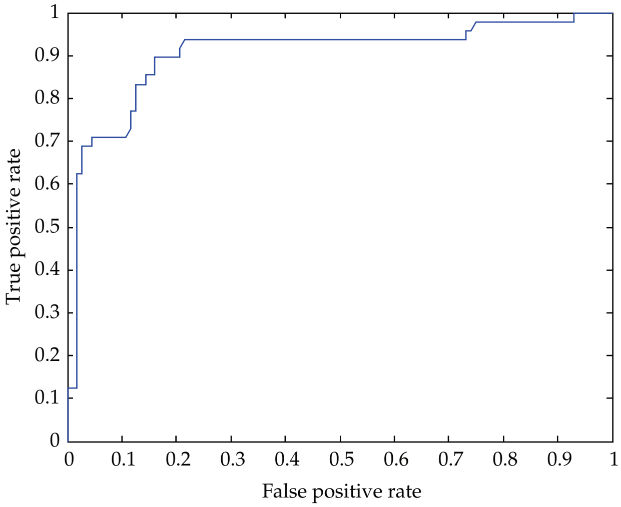

Figure 8: ROC curve for training data in proposed hybrid LR-ANN method.

5. Application to Lung Cancer Data

The data consist of 178 observations. Patient, presented with hemoptysis to Ondokuz Mayıs University Department of Chest Diseases, prospectively evaluated between November 2003 and September 2006. Posteroanterior chest radiography, complete blood count, and renal and hepatic function tests were performed for each patient. Another examination like thorax computer tomography, bronchoscopy, and different laboratory and pathological diagnostic modalities was done if needed. 160 observations used as training data and randomly selected 18 observations used as test data. The LR method, firstly, applied to data. Stepwise variable selection method is applied to data and four significance independent variables (age, time of the hemoptysis (THM), and number of hemoptysis (NHM) and RAL) are selected. The summarized information about these variables are given below.

Hepatic functions test. The hepatic function test, also known as liver function tests (LFTs), is used to evaluate the liver for injury, infection, or inflammation. This test measures

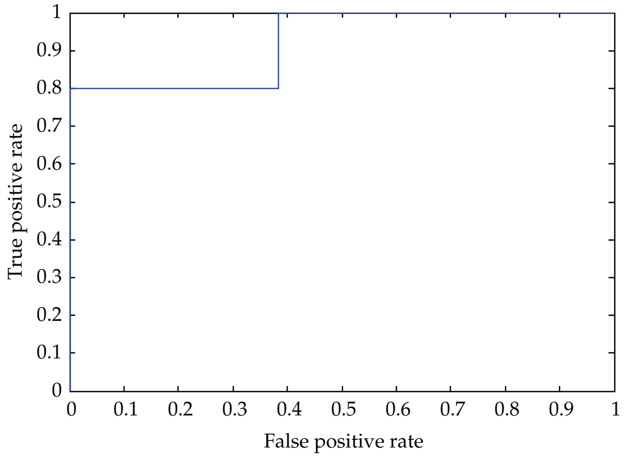

Figure 9: ROC Curve for test data in Proposed Hybrid LR-ANN Method.

Table 3: Optimal weights of neural network in hybrid method.

Hidden layer neurons	Input neurons					Output neuron	Bias (input-hidden)	Bias (hidden-output)
	1	2	3	4	5			
1	−6,61	2,65	−8,75	−1,33	5,44	−11,65	6,31	−72,62
2	−0,78	−2,55	5,75	−5,05	0,03	0,10	−7,15	
3	0,82	−1,39	11,42	15,77	1,74	95,32	−4,06	
4	−11,75	8,74	2,75	3,24	7,74	10,75	−8,36	
5	1,32	0,58	1,51	−5,39	5,72	−60,09	−8,33	
6	−0,29	−0,59	0,25	−3,59	−5,53	−1,63	−5,80	
7	0,68	8,74	2,30	4,72	−0,58	−66,45	−8,72	
8	0,70	0,13	−16,37	−39,04	140,65	94,49	−58,77	
9	−0,05	0,32	3,01	4,68	−6,32	100,77	−4,97	
10	10,60	−10,53	−5,43	0,85	−5,27	8,80	7,26	

the blood levels of total protein, albumin, bilirubin, and liver enzymes. Enzymes that are often measured in LFTs include gamma-glutamyl transferase (GGT); alanine aminotransferase (ALT or SGPT); aspartate aminotransferase (AST or SGOT); alkaline phosphatase (ALP). LFTs may also include prothrombin time (PT), a measure of how long it takes for the blood to clot. High or low levels may mean that liver damage or disease is present.

Time of hemoptysis (THM). Hemoptysis is the expectoration, or coughing up of blood, from the lower respiratory tract. There are three types: minor, moderate, and massive hemoptysis. Distinguishing between minor, moderate, and massive hemoptysis is important, as severity determines the need for emergency treatment. Minor hemoptysis is defined as small specks of blood or clots in the patient's sputum and is generally not life threatening. Moderate hemoptysis can include larger clots up to the loss of 200 mL of blood within a 24 hour period. Finally, massive hemoptysis is anything greater than a loss of 200 mL of blood within 24 hours and is always a medical emergency, as patient asphyxiation can occur rapidly.

Table 4: Results of LR, ANN, and LR-ANN hybrid method for lung cancer data.

	AUC		CCR		Sensitivity		Specificity	
	Training data	Test data	Training data	Test data	Training data	Test data	Training data	Test data
LR model	0,8468	0,8923	0,8125	0,7778	0,5208	0,6000	0,9375	0,8462
ANN model	0,8631	0,9185	0,8375	0,8889	0,5417	0,6000	0,9643	1,0000
Proposed hybrid LR-ANN	0.9050	0.9231	0.8812	0.9444	0.6875	0,8000	0.9645	1,0000

Number of hemoptysis (NHM). Hemoptysis can be classified as mild, moderate, or massive, depending on the amount of blood expectorated: < 100 mL in 24 h (mild); 100–600 mL in 24 h (moderate); > 600 mL in 24 h or > 30 mL/h (massive).

RAL: RAL is a breath sound, like a scrunch, from the lungs.

CBC: complete blood count (CBC) is often used as a broad screening test to determine an individual's general health status. It can be used to screen for a wide range of conditions and diseases.

Parameter estimations, standard errors of estimation, and significance values of these estimations are given in a Table 1. The ROC curves for logistic regression training and test data results are given in a Figures 3 and 4.

When Table 1 is examined, all parameters in the LR model are statistically significant. The (AUC) in values for the LR solutions are given in a Table 4.

The feedforward artificial neural network method secondly applied to data. The inputs of ANN are age, time of HM, and numbers of HM and RAL independent variables. Target of ANN is the diagnosis of lung cancer. The architecture of used ANN is given in a Figure 1. Hidden layer neuron number of ANN is varied 1–20. The best results of ANN are obtained from 10 hidden layer neurons. The optimal weights of ANN are given in a Table 2. The ROC curves for logistic regression training and test data results are given in Figures 5 and 6.

The proposed method was applied to the data. In the first step of hybrid approach, important explanatory variables and forecasts of logistic regression were obtained using stepwise logistic regression. Explanatory variables (age, time of the hemoptysis (THM), number of hemoptysis (NHM), and RAL) obtained from stepwise logistic regression and forecasts of $P(y_i = 1/X)$, $i = 1, 2, \ldots, n$ probability obtained from logistic regression were taken as inputs of feedforward artificial neural networks. In the proposed method, $k = 28$ and $m = 4$. The number of neurons in one hidden layer in feedforward artificial neural network was determined as 10 using trial and error methods. Artificial neural network model for the hybrid approach was given in Figure 7.

Artificial neural network which was given in Figure 7 was trained using Levenberg Marquardt algorithm, and training set and the weights obtained were presented in Table 3. The ROC curves for proposed hybrid method training and test data results were given in Figures 8 and 9.

The comparisons of proposed hybrid method, logistic regression, and artificial neural network are made in Table 4. In Table 4, under area ROC curve (AUC), correct classification rate (CCR), and sensitivity and specificity values are given for all methods.

6. Conclusions

Logistic regression and artificial neural networks which have both advantages and disadvantages are two methods used for classification. Whilst selection of meaningful variables is done via statistical techniques in logistic regression, as it is a black-box model variable selection cannot be done with statistical techniques in feedforward artificial neural network. In this study, a new method using both logistic regression and feedforward artificial neural network was proposed. In the proposed method, variable selection is done by stepwise logistic regression. Additionally, explanatory variables which were obtained from logistic regression and forecasts are taken as inputs of feedforward artificial neural networks. Thus, the proposed method has advantages of both LR and ANN. When the results given in Table 4 were analyzed, it was seen that the proposed method gave better results in regard to criteria for lung cancer data in comparison with LR and ANN.

References

[1] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics*, vol. 35, no. 5-6, pp. 352–359, 2002.

[2] M. Paliwal and U. A. Kumar, "Neural networks and statistical techniques: a review of applications," *Expert Systems with Applications*, vol. 36, no. 1, pp. 2–17, 2009.

[3] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," *Expert Systems with Applications*, vol. 34, no. 1, pp. 366–374, 2008.

[4] Y. Oner, T. Tunc, E. Egrioglu, and Y. Atasoy, "Comparisons of logistic regression and artificial neural networks in lung cancer data," submitted to *Scientific Research and Essays*.

[5] C. L. Chang and M. Y. Hsu, "The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10663–10672, 2009.

[6] M.-H. Chen and D. K. Dey, "Variable selection for multivariate logistic regression models," *Journal of Statistical Planning and Inference*, vol. 111, no. 1-2, pp. 37–55, 2003.

[7] D. Ghosh and Z. Yuan, "An improved model averaging scheme for logistic regression," *Journal of Multivariate Analysis*, vol. 100, no. 8, pp. 1670–1681, 2009.

[8] A. Stacey and D. Kildea, "Genetic Algorithm search for Large Logistic Regression models with significant variables," in *Proceedings of the 22nd International Conference Information Technology Interfaces (ITI '00)*, Pula, Croatia, June 2000.

[9] J. Pacheco, S. Casado, and L. Núñez, "A variable selection method based on Tabu search for logistic regression models," *European Journal of Operational Research*, vol. 199, no. 2, pp. 506–511, 2009.