American Journal of Intelligent Systems 2013, 3(2): 71-74
DOI: 10.5923/j.ajis.20130302.03

Comparisons of Logistic Regression and Artificial Neural Networks in Lung Cancer Data

Yuksel Oner[1], Taner Tunc[1], Erol Egrioglu[1,*], Yildiz Atasoy[2]

[1]Department of Statistics, University of OndokuzMayis, Samsun, 55139, Turkey
[2]Chest Diseases Hospital, Samsun, 55139, Turkey

Abstract In the recent years, there have been many studies rely on medical data classification with artificial neural networks and logistic regression. The logistic regression has been commonly used as a statistical method. The logistic regression is the nonlinear method and the nonlinear optimization methods are used parameter estimation in logistic regression. The logistic regression is the model based approximation, and it is the not data-based approximation. Another classification method is the artificial neural network which has been commonly used in the literature. There are a lot of kinds of artificial neural network; the feed forward neural networks are generally preferred in the literature. In this study, feed forward artificial neural network and logistic regression are compared by classifying lung cancer data. In the result of application, the satisfied accurate classification percentage is obtained from either method.

Keywords Logistic Regression, Feed Forward Artificial Neural Networks, Classification, Lung Cancer Data

1. Introduction

Data classification task is important in many science disciplines. The logistic regression (LR) and artificial neural networks (ANN) are commonly used for classification tasks. The LR is a statistical method. The ANN is an artificial intelligence method. In the literature, many studies have obtained different results for comparisons of these methods. Dreiseitl and Ohno-Machado[1] reviewed the literature for LR and ANN. In Dreiseitl and Ohno-Machado[1],72 papers were examined and it was found that ANN was better than LR in % 18 percentages of papers, LR was better than ANN in % 1 percentage of papers;there was no difference between them in % 42 percentages of papers according to statistical hypothesis tests. The ANN was better then LR in % 6 percentages of papers, LR was better than ANN in % 6 percentages of papers, there was no difference between them in % 0 percentage of papers without using statistical tests.

Although the logistic regression has got some assumptions, it is white-box models and coefficients of logistic regression are interpretable. The ANN is a black-box model, but ANN not needs any assumptions. Numerous articles involving applications of ANN and LR in medicine have been published over the years. The literature related to using of ANN in medicine was given by Paliwal and Kumar[2]. Kurt et al.[3] compared the performances of LR and ANN for the coronary artery disease. Chang and Hsu[4] used LR and ANN for pancreatic cancer. Upadhyaya et al.[5] and Morteza et al.[6] compared the performances of LR and ANN for classification type-2 diabetic patients.

In this study, LR and ANN are used classification of the lung cancer patients and the results of ANN and LR are compared. In the second section of paper, the logistic regression and ANN methods are briefly given. In the third section, results of LR and ANN for lung cancer data are given. The obtained results are discussed in fourth section.

2. Materials and Methods

2.1. 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 below.

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

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 below.

$$P(Y_i=y_i)=\begin{cases}\pi_i^{y_i}(1-\pi_i)^{1-y_i} & y_i=0 \text{ or } 1\\ 0 & ,o.w.\end{cases}$$

The likelihood function is the product of these probabilities and the logarithm of likelihood function is given below.

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

* Corresponding author:
erole1977@yahoo.com (Erol Egrioglu)
Published online at http://journal.sapub.org/ajis

The parameters of logistics regressions are estimated via maximizing logarithmic likelihood function. The non-linear 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. In this study, we use stepwise method in the SPSS package program.

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

Although there are many types of artificial neural networks in literature, feed forward artificial neural networks are frequently used for many problems. Feed forward artificial neural networks consist of input layer, hidden layer(s) and output layer. An example of feed forward 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 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 feed forward artificial neural networks are forward and unidirectional. Single activation function is used for each neuron in hidden layer and output layer of feed forward 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, non-linear activation functions are used for hidden layer units. In addition to a non-linear activation function, linear (pure linear) activation function can be used in the output layer neurons.

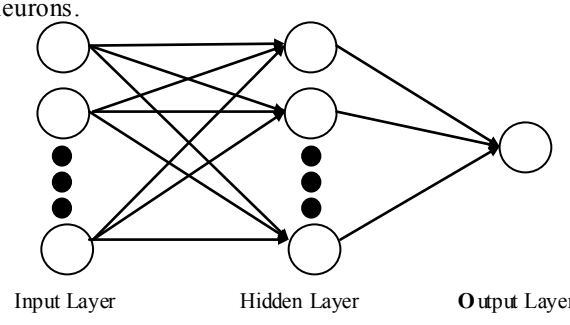

Figure 1. Multilayer feed forward artificial neural network with two output neuron

In feed forward 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 literature used for learning of feed forward 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.

3. Application to Lung Cancer Data

The lung cancer data consist of 178 observations. Patient, presented with hemoptysis to OndokuzMayıs University Department of Chest Diseases, prospectively evaluated between November 2003 and September 2006. Posteroanterior chest radiography, complete blood count, 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), number of hemoptysis (NHM) and RAL) are selected. Parameter estimations, standard errors of estimation and significance values of these estimations are given Table 1. The ROC curves for logistic regression training and test data results are given Figure 2 and Figure 3.

Table 1. Estimation Results of Logistic Regression

Variables	Parameter 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

When Table 1 is examined, all parameters in the LR model are statistically significant. The AUC (Area under curve) values for the LR solutions are given in Table 2.

The feed forward artificial neural network method secondly applied to lung cancer data. The inputs of ANN are selected as age, time of HM, numbers of HM and RAL independent variables. Target of ANN is the diagnosis of lung cancer. The architecture of used ANN is given Figure 1. Hidden layer neuron number of ANN is varied 1-20. The best results of ANN are obtained from the architecture which has 10 hidden layer neurons. The optimal weights of ANN are given Table 3.

Table 2. Comparisons of Logistic Regression and ANN

	AUC		CCR	
	Training Data	**Test Data**	**Training Data**	**Test Data**
LR Model	0,8468	0,8923	0,8125	0,7778
ANN Model	0,8631	0,9385	0,8375	0,8889
	Sensitivity		Specificity	
	Training Data	**Test Data**	**Training Data**	**Test Data**
LR Model	0,5208	0,6	0,9375	0,8462
ANN Model	0,5417	0,6	0,9643	1

The ROC curves for ANN training and test data results are given Figure 4 and 5. The comparisons of logistic regression and artificial neural network are made Table 2. In table 2, under area ROC curve (AUC), correct classification rate (CCR), sensitivity and specificity values are given for logistic regression and artificial neural network.

Table 3. Optimal weights of Neural Network

Hidden Layer Neurons	**Input Neurons** 1	2	3	4	**Output Neuron**	**Bias (Input-Hidden)**	**Bias (Hidden-Output)**
1	0,32	0,32	-1,57	-6,13	6,37	-13,34	0,50
2	-0,52	-0,19	-0,82	3,89	8,98	-5,31	
3	-0,13	-1,64	-10,51	-2,97	-5,06	-8,56	
4	1,04	1,36	5,78	6,87	-7,04	-11,90	
5	0,16	0,62	-8,57	1,71	-4,95	12,34	
6	0,31	-0,08	3,75	-3,69	4,67	-15,32	
7	-1,14	3,94	12,95	-7,24	3,46	7,62	
8	-1,50	-0,46	-2,81	2,81	-10,49	-1,13	
9	-0,46	0,22	6,29	-2,44	-9,71	-4,79	
10	-2,66	1,57	-4,24	-5,29	-1,10	9,20	

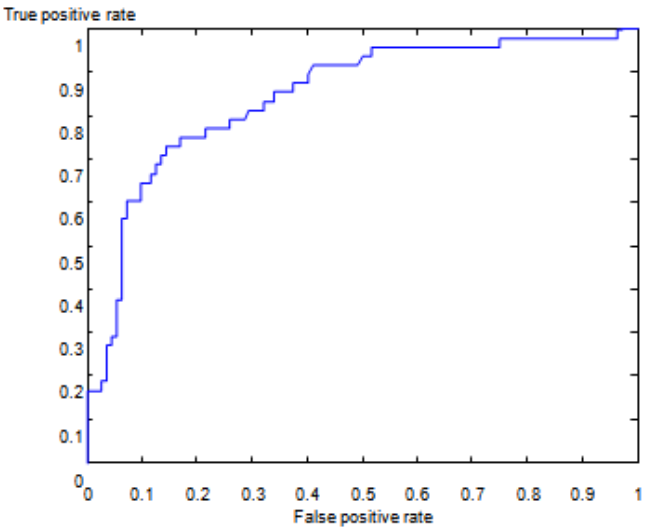

Figure 2. ROC Curve for training data in logistic regression

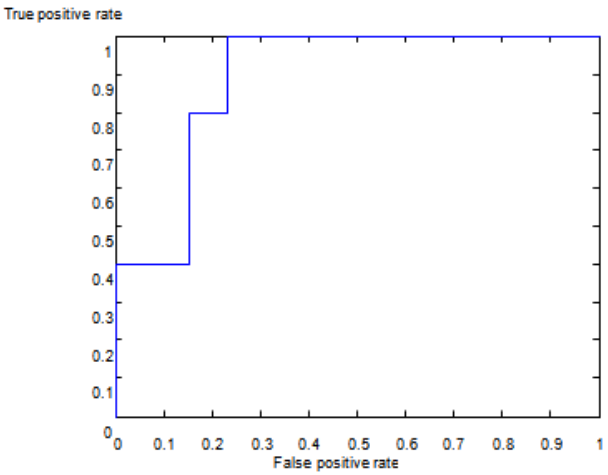

Figure 3. ROC Curve for testing data in logistic regression

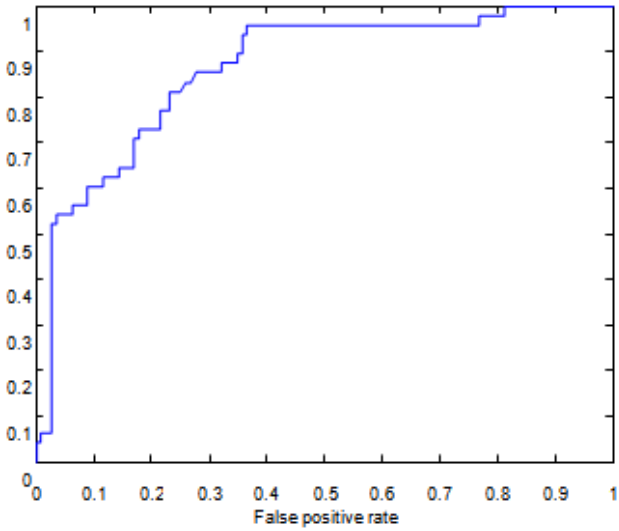

Figure 4. ROC Curve for training data in ANN

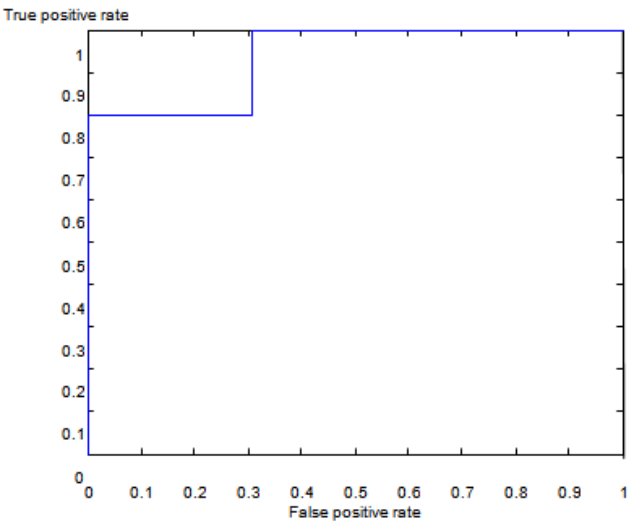

Figure 5. ROC Curve for testing data in ANN

4. Conclussions

The logistic regression (LR) and artificial neural networks (ANN) are commonly used for classification tasks. In this study, logistic regression and artificial neural networks are compared according to AUC, CCR, sensitivity and specificity criteria. ANN outperforms logistic regression for all criteria. The sensitivity values for either method in testing data are equal. According to obtained results, ANN method can be preferable for lung cancer data classifications.

REFERENCES

[1] Dreiseitl S., Ohno-Machado L., Logistic regression and artificial neural network classification models: a methodology review, Journal of Biomedical Informatics, 35, 352-359, 2002.

[2] Paliwal M., Kumar U.A., Neural networks and statistical techniques: A review of Applications, Expert Systems with Applications, 36, 2-17, 2009.

[3] Kurt I, Ture M., Kurum A.T., Comparing performances of logistic regression, classification and regression tree and neural networks for predicting coronary artery disease, Expert Systems with Applications, 34, 366-374, 2008.

[4] Chang C.L. , Hsu M.Y., The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer, Expert Systems with Applications, 36, 10663-10672, 2009.

[5] Upadhyaya S., Farahmand K., Baker-Demaray F., Comparison of NN and LR classifiers in the context of screening native American elders with diabetes Original Research Article, Expert Systems with Applications, 40(15), 1,5830-5838, 2013.

[6] Morteza A., Nakhjavani M., Asgarani F., CarvalhoFilipe L.F., Karimi R., Esteghamati A., Inconsistency in albuminuria predictors in type 2 diabetes: a comparison between neural network and conditional logistic regression, Translational Research, 161(5),397-405,2013.