ARTIFICIAL NEURAL NETWORK APPROACH FOR REGRESSION ANALYSIS

T. M. D. K. Bandara and R. D. Yapa

Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya

Introduction

Regression analysis is one of the predominant tools for analyzing and testing the relationships between independent and dependent variables. On the other hand Artificial Neural Network (ANN) is a powerful tool that can be used to approximate any functions to any level of accuracy. This article highlights the features and behavior of artificial neural networks as an attractive alternative to statistical regression analysis. Artificial Neural Network based solutions have already been proposed in the area of regression analysis. But most of them were focused on different specialized application areas (Hashem *et al.*, 2007) and (Eskandari *et al.*, 2004).

The objective of this study is to develop a general regression analysis tool based on ANN method. The important characteristic of the proposed method is that, it can be used to fit statistically significant regression models for any given dataset (with only one predictor variable) by a person even without a strong knowledge on regression analysis. The fitted model will be validated by the tool itself.

Methodology

The main three regression model types, simple linear regression, multiple linear regression and non linear regression have been considered for the proposed ANN tool. The three regression model types are handled using three different ANN models. If the model to be fitted is simple linear regression or multiple linear regression, the proposed system will automatically validate linearity conditions between the dependent variable (Y) and independent variable (X) and select suitable variables for the target model.

If the target model is a simple linear regression and if the linearity condition is satisfied, then the ANN model with two layers of neurons is used to find the intended model. Input layer includes two neurons in order to find the two coefficients of the regression model and there are only one output nodes. Weight values of the ANN model are initialized using smallest and the largest independent observations of the dataset. Then the ANN is trained using back propagation, gradient descent technique. Weight updating is done with the use of batch processing method. After the ANN is trained, the corresponding weight values represent coefficients of the target model.

If the target model is multiple linear regression having k independent variables and if independent variables having significant correlation with the dependent variable has been selected then the principle component analysis (PCA) is used as a variable reduction

method (URL: http://support.sas.com/publishing/pubcat/chaps/55129.pdf) PCA also used as a method to eliminate multi-co-linearity. Then the proposed system will select optimum number principle components (PCs) which adequately describe variability of observations. Selected PCs are used as inputs for the ANN model. Structure of the ANN for multiple linear regression is similar to that of the simple linear regression, it contains more input nodes which is equal to the number of PCs. Small random values are assigned as the initial weight values and then the training cycle is performed with the gradient descent batch mode.

In non linear regression, the function relating the response to the predictors is not necessarily linear. The associated artificial neural network is designed including an additional layer of neurons called hidden layer. According to our observations, the best way to choose the number of hidden nodes is selecting the hidden nodes equal to the sample size. To reduce the complexity when the sample size increases, the number of hidden nodes is limited to 20 if the number of observations exceeds that value. Small random values are assigned as the initial weight values. The non linear artificial neural network is trained using the online gradient descent mode.

Then, residual analysis is used to assess the quality of the regression model obtained by neural network training. To be an optimal regression model, the residuals should hold the three properties namely; Mean of the residuals is equal to zero, the residuals have a constant variance and the residuals are uncorrelated.

If the residual analysis proves that the model is inadequate, then the solution involves either transforming response or predictors or both. If the linearity condition is violated, the explanatory variable transformation increases the linear relationship between two variables which means it changes the correlation between them. The independent variable transformation is performed by looking at the shape of the scatter plot. The response variable is transformed when the residual analysis suggests that the variance of the residuals violates the constant variance or residuals are uncorrelated. The transformation method is chosen based on the trend represented by the residual plots.

Once the weight values of the ANN are initialized, the training process is improved by starting training process with a large learning rate (0.1) and keeps reducing until the best gain term is found.
All neural network architectures like multilayer perceptron prone to over fitting. Our research work proposes a technique to apply the concept of 'early stopping' criteria in order to solve the problem of over fitting. Different early stopping criterions have defined to suit for the error convergence paths occur in the real time situations. (Bhadeshia *et al.*, (n.d.) and URL: http://page.mi.fuberlin.de/prechelt/Biblio/stop_tricks_1997\.pdf).

Results and Discussion

The proposed ANN model was tested using several different datasets that suits for simple linear regression, multiple linear regression and non linear regression. (Table 1)

Conclusion

The proposed method obtains optimal solutions for simple linear regression since it uses the methods for both weight initialization and selecting gain term. However sometimes it converges into a local minimum in the problems of multiple linear regression since it does not use any weight initialization technique. One of the limitations of non linear neural network is it has no way to represent the final regression model in a simple mathematical format. But it will obtain best generalized results on a given data set.

The results of this study indicate that the system that has been developed could be used successfully for interpretation of relationships which exist between independent and dependent variables. Having identified these relationships, future trends of a given application could be determined.

References

Bhadeshia, H.K.D.H. and Sourmail, T. Lecture 1: Neural Networks; Master of Philosophy, Materials Modelling,

Eskandari, H., Rezaee, M.R. and Mohammadnia, M. (2004). Application of Multiple Linear Regression and artificial neural network techniques to predict share wave velocity from wire line log data for a carbonate reservoir, South West Iran"

Hashem, M. and Karkory, H. (2007). Artificial Neural Networks as alternative approach for predicting trihalomethane formation in chlorinated waters proceedings of the Eleventh International Water Technology Conference, IWTC11 2007 Sharm El-Sheikh, Egypt.

http://support.sas.com/publishing/pubcat/chaps/55129.pdf

http://page.mi.fu-berlin.de/prechelt/Biblio /stop_tricks 1997.pdf (1997),

Table 1. Results and discussion

Datasets	Proposed Technique
Dataset 1	Final regression model is, Y=9.434 + 1.491 x. Residual analysis proves that all three conditions are satisfied by the obtained regression model.
Dataset 2	Final regression model is, Y=-9.701 + 8.975 x. Once the residual Analysis proves that the obtained model is not adequate then transforming the dependant variable will obtain, Sqrt(y) = -0.661+ 1.057 x. Residual analysis prove that the transformed regression model provides a better relationship.
Dataset 3	Final regression model is, Y= 10.247 + 0.289 x1 + 0.91 x2 + 0.009 x3 Residual Analysis proves that all three conditions are satisfied by the obtained regression model.
Dataset 4	Once the system obtained the regression model, residual analysis proves that the obtained regression model is optimal.