American-Eurasian J. Agric. & Environ. Sci., 10 (1): 60-64, 2011
ISSN 1818-6769

Comparison of Multiple Linear Regressions (MLR) and Artificial Neural Network (ANN) in Predicting the Yield Using its Components in the Hulless Barley

[1]Mohammad Zaefizadeh, [2]Majid Khayatnezhad and [2]Roza gholamin

[1]Islamic Azad University, Ardabil Branch, Ardabil, Iran
[2]Young Researchers Club, Islamic Azad University, Ardabil Branch, Ardabil, Iran

Abstract: In this study 40 genotypes in a randomized complete block design with three replications for two years were planted in the region of Ardabil. The yield related data and its components over the years of the analysis of variance were combined.Results showed that there was a significant difference between genotypes and genotype interaction in the environment. MLR and ANN methods were used to predict yield in barley. The fitted model in a yield predicting linear regression method was as follows:

$$\hat{Y} \text{ Reg} = 1.75 + 0.883\ X1 + 0.05017X2 + 1.984X3$$

Also, yield prediction based on multi-layer neural network (ANN) using the Matlab Perceptron type software with one hidden layer including 15 neurons and using algorithm after error propagation learning method and hyperbolic tangent function was implemented, in both methods absolute values of relative error as a deviation index in order to estimate and using duad t test of mean deviation index of the two estimates was examined. Results showed that in the ANN technique the mean deviation index of estimation significantly was one-third (1 / 3) of its rate in the MLR, because there was a significant interaction between genotype and environment and its impact on estimation by MLR method.Therefore, when the genotype environment interaction is significant, in the yield prediction in instead of the regression is recommended of a neural network approach due to high yield and more velocity in the estimation to be used.

Key words: ANN • MLR • Hulless Barley

INTRODUCTION

There are important limitations, including the necessity of observing regression assumptions, significant non-linear relationships and multiple callinearity between independent variables and the presence of genotype and environment interaction, also replication interactions within the environment (E1) in the combined analysis of variance in national uniform tests, led to be predicting the MLR model non efficient [1]. However, the prediction in an artificial neural network method (ANN) always takes place according to any data situation (without limitation) based on initial training [2]. Each artificial neural network of an input layer, one or more hidden layers and an output layer has been established as the following general model is defined [3]:

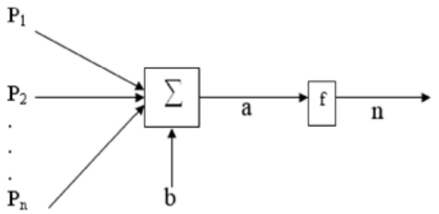

Fig. 1: Artificial neural net work models

However, networks can be different depending on type and number of layers and having control feed back and, etc. [2]. Hitherto almost seven different methods for artificial neural networks have been developed and each one can be used in a different analysis rather than classical statistical methods [4].When the non-linear relationship between dependent and independent

Corresponding Author: Majid Khayatnezhad, Young Researchers Club, Islamic Azad University, Ardabil Branch, Ardabil, Iran. E-mail: Khayatneghad@yahoo.com.

variables exists, the ANN method will be more efficient [2]. There is no need in ANN method to observe some of the regression assumptions. Other studies have compared two methods to show that in predicting the dependent variable, the ANN method results are more accurately than MLR [2, 5- 9]. Although the ANN is considered as a powerful technique for non-linear models [10]. But some researchers in this linear model have also applied and reported it better than the regression model [4, 9, 11, 12]. According to the present constraints on regression techniques to predict yield (Ŷi) based on its components and its importance in the reform election of superior genotypes as for high impact of environment on the multiple gene characteristics, this research on a real data from farm experiments of hulas barely to compare two methods of regression and artificial neural network approach was used.

MATERIALS AND METHODS

Farm Experiments and Numerical Resources (Database): An experiment in a randomized complete block design with forty genotypes of hulless barely was planted (advanced lines) in three replications of two agricultural years (1983-1985) in the research farm of Islamic Azad University of Ardabil. Each experimental plot included five 2.5 m culture lines with distance of 20 cm were formed, eliminating the margin effect, 5 samples were selected so that for each trait in total of plots and years totally 1200sample was selected. In addition to yield per unit plot, important yield components such as cluster length, total grain in the main cluster, grain weight and plant height were measured.

Statistical Methods: Each year, Simple variance analysis for traits derived from a simple farm experiment and also the combined analysis of variance test was performed for two years where the F test based on the mathematical expectancy mean squares (EMS) was conducted. MLR analysis for yield using independent variables to the stepwise variable election method was conducted and based on variables remaining in the equation model of the line was estimated then in 20 percent of samples (test), all of ${}_i$?s based on this model again was estimated and the average absolute value relative error: | Ŷi-Ŷi | was calculated. Artificial neural network (ANN) in Matlab software using three input variables, the numbers of fertile tillers and grains per main spike and grain weight was taught. The used Model was of a multi-layer Perceptron. Total data for predicting the yield of hulless barely were 1200 samples that the training data and testing and verification of total data randomly were selected and respectively made up of 60% and 20% and 20% of the total data. In the network, the changes of error recovery in terms of the epoch and error absolute value (AE) of learning data and authentication was also calculated and figured out. After training the network, also the Yi values using Xs available in test samples in the network re-calculated and the absolute value of deviation in this method | Ŷi-Ŷi | was obtained. Mean absolute value of deviation of Ŷi which was calculated in two ways by Ŷi using ANN and MLR methods to the way of paired t-test were compared.

RESULTS AND DISCUSSION

Multiple Linear Regression Models: Analysis of Combined variance results showed that there was significant differences between the interaction of genotype and genotypes in the year (G * E) in terms of yield, number of fertile tillers, spike and one thousand grain weight. Final model Regression analysis in the Stepwise method for grain yield after the multiple callinearity tests, Watson and Gold Camera Field Cont was fitted to the following:

$$\hat{Y}_{reg} = 1.75 + 0.883_{X1} + 0.05017_{X2} + 1.984_{X3}$$

Where fertile tillers = X1, the main spike and one thousand grain weight = X2 = X3 were.

According to the regression variance analysis table (Table 1), this model in a probability level of 0/0001 was significant.

Explanation coefficient in the above model was equal with R^2=0.678. Thus that nearly 68% of the yield changes by three variables of fertile tillers, main spike and one thousand grain weight can be justified.

Artificial Neural Networks: Based on comparing methods and changing the layers and number of neurons in each network, the best neural network was selected with the following characteristics due to the minimum of root mean square (RMS).Neural network structure used as (1-15-3) that has three neurons in the input layer (number of fertile tillers, number of grains in the main (cluster) spike and one thousand grains weight), 15 neurons in the hidden layer and one neuron in output layer (yield), (Figure 1). Of a stimulating function of the hyperbolic tangent in hidden layer and linear activator function in output layer was used.

Table 1: Regression variance analysis for the yield through the remaining variables in linear regression models

S.O.V	df	(MS)	F	Prob
Regression	3	365.25	126.2	0.0001
Remaining	236	2.49		

Table 2: results of the comparison of deviation absolute value average of actual Y which was brought on by Y in the two methods of MLR and ANN

Groups	Mean	Standard deviation	Mean difference	t	Prob
| Ŷi-Ŷi | in regression	1.334	1.074	-0.975	18.37	0.0001
| Ŷi-Ŷi | in ANN	0.358	0.597			

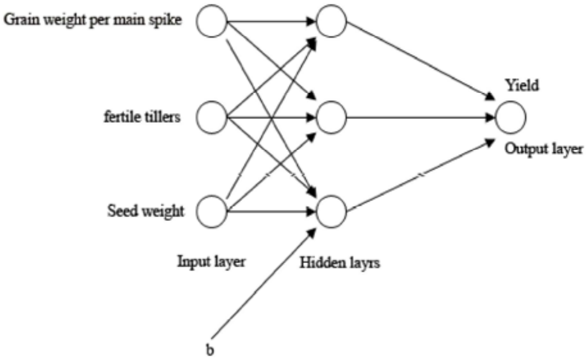

Fig. 2: The designed Network diagram.

In the network, the error recovery changes based on epoch is mapped and shown in Figure 2. The error absolute value (AE) and authentication and learning data are given in Figure 3. The error of these data in the replication 500 was converged means that epoch = 500 is the best learning mode of network that if most of this learning continues, over train in the network will occur.

Comparison Between the Two Methods for Estimating Yield: Using estimated Ŷi in both of regression and artificial neural network methods, the rate of the deviation absolute value from the regression line from the actual yield (Ŷi) were calculated and compared and its comparison results using paired t test was listed in Table 2.

The results of this comparison showed that with a lower probability level in 0.0001 of the mean absolute deviations from the regression line (as an indicator of the diversion in the estimation) in the regression method significantly was more than the artificial neural network method. This means that the amount of error in the estimation by the regression method was more than the error in artificial neural network method, thus can be stated that the neural network for estimating in the barely genotypes yield was more effective than regression approach. However, that the correlation between the two estimates for every x existing in the artificial neural network models and multiple linear regression was significant and positive.

The multiple linear regression models are the most important method of predicting the yield by its components. The estimated multiple linear regression models for yield based on the three remaining variables indicate that the selection of genotypes with high yield through its components can be used. Mohammadi [13] also reported the number of fertile tillers and the number of grains in the spike as the important trait in showing the yield in the barley genotypes, although these two traits and traits of grain filling time, days to appearing of the spike and plant (bush) height totally were showing the yield rate of 49% in the regression model. In the present model, the yield explanation rate by three traits was 68% that represents the best selection of independent

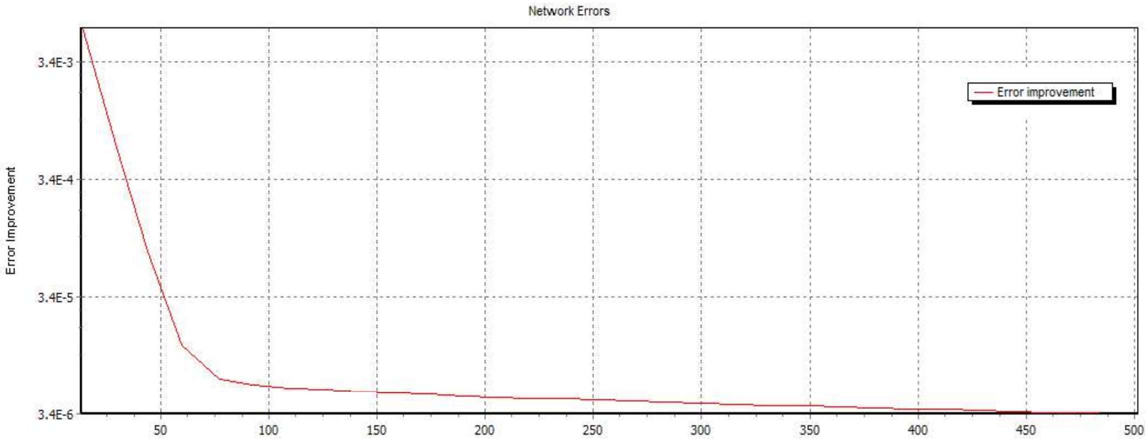

Fig. 2: Error recovery changes based on epoch in the network

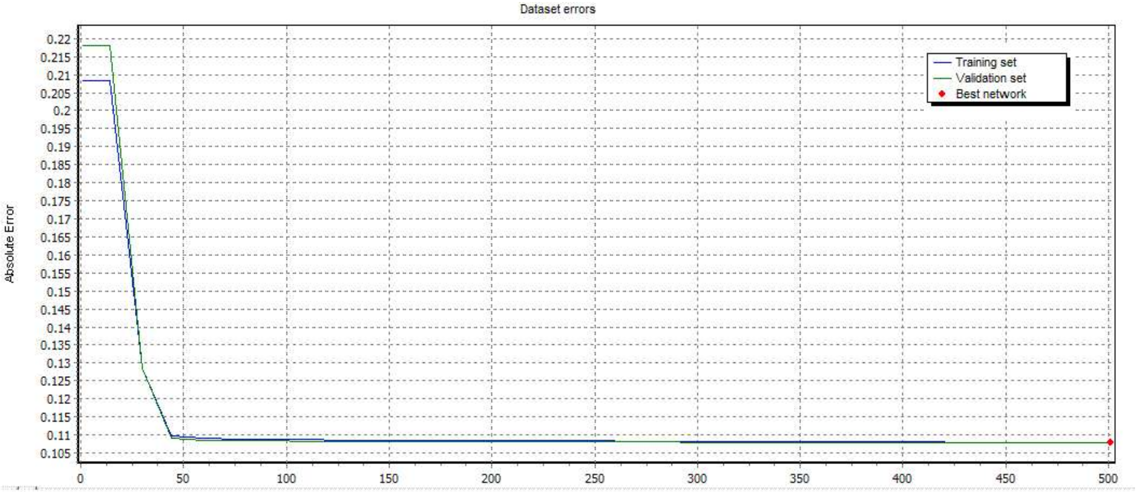

Fig. 3: Error absolute value (AE) learning data and authentication

variables for yield. The findings were consistent with the results of Alizadeh Gharib [14]. Existence and necessity of observing the multiple defaults (or assumptions) for design of regression models ex parte and the impact of other conditions, including the existence of a significant interaction between genotype and environment (G * E) in the regression models from other hand causes in the predictions of Each unit with available Xs in the model deviation rate vary. However, the cause of this deviation in addition to genotype environment interaction effect could be related to violating regression assumptions are including multiple callinearity and dependency of errors, etc. which is generally observed less by users.Estimates and forecasts based on artificial neural network as an efficient method of learning based on samples towards regression method in the linear and non linear relationships have been recommended [6, 15- 18] because the smart artificial neural networks considering all the circumstances can do the estimates with lesser skew based on the single samples.

The comparison of skewed values |Yi-Ŷ| showed in both experiments that the skewed values in the ANN method are significantly less than the MLR method. That can be related to the lack of neural network limitations than other statistical methods, especially parametric method. So that more yield and accuracy of neural network forecasting method is achieved than the correlation and multiple regressions [7, 13, 16, 18]. General health indicators than the sequential method of least squares logistics regressions have been reported but in comparing both multiple regression and neural network for formulation optimization of lypoprolid acetate in liposomes, two methods yield reported the same [8].

So regarding the lack of restrictions on artificial neural network application (ANN) and the extent of its application in biological sciences and also a fast and easy prediction it can be rather than regression techniques as an alternative method in plant breeding and Forecast quantitative traits in barley polygenic in hulless barely used. Because most of the relationships between variables in the hulless barely are not necessarily linear constraints and it is using constraints of linear regression with this method is disappeared.

REFERENCES

1. Molazem, D., M. Valizadeh and M. Zaefizadeh, 2002. North West of genetic diversity of wheat. J. Agricultural Sci., 20: 353-431.
2. Adielsson, S., 2005. Statistical and neural networks analysis of pesticide losses to surface water in small agricultural catchments in Sweden. M.Sc. Thesis, Sweden University, Sweden.
3. Gevery, M., I. Dimopoulos and S. Lek, 2003. Review and comparison of methods to study the contribution of variable in artificial neural network models. Eco. Modeling, 160: 249-264.
4. Miao, Y., D. Mulla and P. Robert, 2006. Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. Springer, 7: 117-135.
5. Pastor, O., 2005. Unbased sensitivity analysis and pruning techniques in ANN for surface ozone modeling. Ecological modeling, 182: 149-158.

6. Gail, B., C. Viswanthan, T.R. Nelakantan, L. Srinivasa, R.Girones, D. Lees, A. Allard and A. Vantarakis, 2005. Artificial neural networks prediction of viruses in shellfish. Appl. and Environ. Microbiol., 31: 5244-5253.
7. Diane, M.L. and P.A. David, 2007. For predicting facial cal form concentrations. Hydrological sci. J., 52: 713-731.
8. Salt, D.W., N. Yildiz and D.J. Livingstone, 1999. The use of artificial neural networks in QSAR. Pesticide Sci., 36: 161-170.
9. Starett, S.K., Y. Najjar, S.G. Adams and J. Hill, 1998. Modeling pesticide leaching from golf courses using artificial neural networks. Communications in Soil Science and Plant Analysis, 29: 3093-3106.
10. Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga and A. Aulagnier, 1996. Application of neural networks to modeling nonlinear relationships in Ecology. Ecological Modeling, 90: 39-52.
11. Manel, S., S.M. Dias and S.J. Ormerod, 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river Bird. Ecological Modeling, 120: 337-347.
12. Ozesmi, S.L. and U. Ozesmi, 1999. An artificial neural network approach to spatial habitat modeling with interspecific interaction. Ecological Modeling, 116: 15-31.
13. mohammadi, M., 2002. Physiological traits associated with the performance of two barley genotypes in normal conditions and drought stress. Seed and Plant Research Magazine, 17: 61-72.
14. Alizadeh Gharib, F., 2006. Effect of dose and time-consuming and yield components in the DMS naked barley. M.Sc. Thesis, Azad University, Ardabil, Iran.
15. Chen, Y.K., D.Z. Chen, C.Y. He and S.X. Hu, 1999. Quantative structure-activity relationships study of herbicides using neural networks and different statistical methods. Chemo metrics and Intelligent Laboratory System, 45: 267-276.
16. Arulsundar, N., N. Subramanian and R.S.R. Murth, 2005.Comparison of artificial neural networks and multiple liner regression in optimization of formulation parameters of leuprlide acetate loaded leptosomes. J. Pharm. Sci., 8: 243-258.
17. Dahnal, V., K. Kuca and D. Jun, 2005. What are artificial neural networks and what they can do?. Fac. Biomed. Papmed. Palacty olomuac Univ., Czech. Republic, 149: 221-224.
18. Nikolopoulos, K., P. Goodwin, A. Patelis and V. Assimatopoulos, 2007. Forecasting with cue information: A comparison of multiple regressions with alternative forecasting approaches. European J. Operational Res.,180: 354-368.