Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 131–137

Statistical and Artificial Neural Network Approaches for the Classification of Rice Genotypes based on Morphological Characters

Shavi Gupta[1], Manish Sharma[1], S. E. H. Rizvi[1], R. K. Salgotra[2] and S. K. Gupta[3]
[1]*Division of Statistics and Computer Science, SKUAST- Jammu*
[2]*School of Bio Technology, SKUAST- Jammu*
[3]*Division of Silviculture and Agroforestry, SKUAST- Jammu*

Received: 12 July 2023; Revised: 28 July 2023; Accepted: 30 July 2023

Abstract

Classification is that arena of science which deals with grouping of objects on the basis of information available about those objects. It plays a significant role for planning purposes in agriculture system. The aim of this study was to classify the rice genotypes using statistical methods like discriminant analysis and artificial neural network (ANN), such as multilayer perceptron neural network for different classes of yield. These methods are fitted to primary data recorded for 100 genotypes of rice for five morphological variables and the data has been collected from the trial laid in SKUAST, Jammu. The class variable grain yield was categorized into 3 classes and was considered as dependent variable and all morphological characters as independent variables. The ability measures of classification such as Accuracy Rate and Kappa Statistics were used for testing samples. Number of days for full maturity was found to be important attributing character followed by number of effective tillers per plant for classification. Artificial Neural Network model (85 %) performed better than Discriminant Analysis (75 %) for classification of genotypes for different classes of yield of rice genotypes.

Key words: Classification; Rice; Discriminant analysis; Multilayer perceptron neural network; Accuracy rate; Kappa statistics.

1. Introduction

Agriculture sector plays a very crucial role in the economy of developing countries and it is the main source of income, employment and food for their population. With the aim of producing more and better crops, the agricultural sector has gone through many new technologies. According to UN Report 2017, the world population is expected to have an increase of 9.8 billion in 2050 and 11.2 billion in 2100. So, there should be need to increase world food production by 50% to feed the estimated world production.

Rice is an important staple food in Jammu and Kashmir as well as all over the world and its production plays an important role in the life of all farmers. Agriculture and Food security policymakers all over the world should give their attention in promoting the research work and projects for studying the processing, food manufacturing, improvement in nutritive

Corresponding Author: Manish Sharma
Email: manshstat@gmail.com

values and potential health benefits of rice by considering its different varieties to promote their utilization as food in respective places.

Classification is playing a very important role in the field of research in agriculture sector. It is a data mining technique used for prediction of class of objects and is an example of supervised learning as suggested by Kumar *et. al*. 2012. Classification predicts categorical label either discrete or ordered. Classification problems can be done using either statistical methods or machine learning methods or both. For classification through statistical methods, discriminant analysis and through machine learning methods, artificial neural network can be used. The classification of genotypes for different classes of yield, can helps to create genetic variability among the genotypes with respect to a particular character.

The primary goal of the research work is to provide a best approach to classify the rice genotypes for different classes of yield on the basis of different characters.

2. Material and method

The primary data collected on yield and attributing characters of rice genotypes such as average plant height (X_1), number of effective tillers per plant (X_2), number of days for 50 percent maturity (X_3), number of days for full maturity (X_4) and 1000 grains weight (X_5) of 100 rice genotypes from the trail laid in SKUAST Jammu. The yield of rice genotypes is considered as dependent variable which has been classified into three categories as given below

Low : Yield less than 150 grams
Medium : Yield between 150 & 300 grams
High : Yield greater than 300 grams

and all other physical characters are considered as independent variables. The data set is divided randomly into training data consists of 80 percent of data and test data consists remaining 20%. Discriminant Analysis is a multivariate technique introduced by Fisher (1936) to differentiate between groups. The maximum number of discriminant functions that can be computed is equal to minimum of *K*-1 and *t*, where *K* is the number of groups and *t* is the number of variables. Suppose the first discriminant function is

$$D_1 = A_{11}X_1 + A_{12}X_2 + \cdots + A_{1t}X_t$$

where the A_{1j} is the weight of the *j*th variable for the first discriminant function. The weights of the discriminant function are such that the ratio is

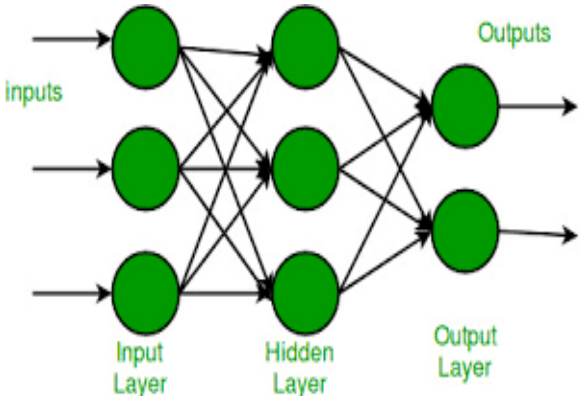
$$\lambda_1 = \frac{Between\ groups\ SS\ of\ D_1}{Within\ groups\ SS\ of\ D_1}.$$

Suppose the second discriminant function is given by $D_2 = A_{21}X_1 + A_{22}X_2 + \cdots + A_{2t}X_t$.

The weights of above discriminant function are estimated such that the ratio is

$$\lambda_2 = \frac{Between\ groups\ SS\ of\ D_2}{Within\ groups\ SS\ of\ D_2}.$$

The above λ_i's are maximized subject to the condition that D_i and $D_{i-1}$are uncorrelated. The procedure is repeated until all possible discriminant functions are identified. Once the identification of discriminant functions is done, the next step is to determine a rule for classifying the future observations.



The other technique used is multilayer perceptron (MLP), the most known and most frequently used type of neural network for classification problems. In this neural network there are multiple layers of neurons that are present between input and output. MLPs are also known as feedforward neural networks which means that data flow in one direction from the input to the output layer. Layers that are present in between the input and output layers are referred as hidden layers. The hidden layer performs useful intermediary computations before directing the input to the output layer. The input layer neurons are linked to the hidden layer neurons through some weights known as input-hidden layer weights. Similarly, the hidden layer neurons are linked to the output layer neurons by hidden-output layer weights.

In order to check the classification ability of statistical models and artificial neural network model we use measures like Accuracy rate and Kappa statistics given as

$$\text{Accuracy Rate} = \frac{\text{correctly classified data}}{\text{total data}} \times 100$$

$$\text{Kappa Statistics} = \frac{N\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{ir}x_{ic}}{N^2 - \sum_{i=1}^{k} x_{ir}x_{ic}}$$

where x_{ii} is the count of diagonal elements of the confusion matrix; x_{ir} and x_{ic} are the total of rows and columns of confusion matrix respectively and N is the total number of observations.

3. Results and discussions

The Discriminant analysis and Multilayer Perceptron Neural Network (MLPNN) used for classification of research data and the results of these methods have been discussed.

Table 1: Classification table of discriminant analysis for yield of rice

Sample	Observed (Number of genotypes)	Predicted (Number of genotypes)			
		Low	Medium	High	% Correct
Training (80 %)	Low (9)	8	0	1	88.9
	Medium (50)	10	24	16	48.0
	High (21)	1	8	12	57.1
	Overall %				55.0
Testing	Low (5)	5	0	0	100.0

(20%)	Medium (13)	3	8	2	61.5
	High (2)	0	0	2	100.0
	Overall %				75.0

The classification of rice genotypes using discriminant analysis is given by Table 1 which represents that in training dataset of discriminant analysis, 8 out of the 9 Low yield genotypes with 88.9 percent of accuracy, 24 out of the 50 Medium yield genotypes with 48.0 percent of accuracy, 12 out of 21 High yield genotypes with 57.1 percent of accuracy are correctly classified and overall, 55.0 percent of the training cases are classified correctly. In testing dataset of discriminant analysis, 5 out of 5 low yield genotypes are classified correctly with 100 percent accuracy, 8 out of 13 medium yield genotypes are correctly with 61.5 percent of accuracy, 2 out of 2 high yield genotypes are correctly with 100 percent of accuracy and overall, 75 percent of the testing cases are classified correctly.

Table 2: Tests of equality of group means

Variable	Wilks' Lambda	*F*	DF1	DF2	*p*-value
X_1	0.865	1.326	2	17	0.292^{ns}
X_2	0.893	1.014	2	17	0.384^{ns}
X_3	0.576	6.250	2	17	0.009**
X_4	0.557	6.755	2	17	0.007**
X_5	0.971	0.256	2	17	0.777^{ns}

ns: non-significant **: significant at 1% level of significance

The Table 2 represents that the Wilks' Lambda statistics for variables average plant height, number of effective tillers per plant, number of days for 50 percent maturity, number of days for full maturity and 1000 grains weight which was 0.865, 0.893, 0.576, 0.557 and 0.971 respectively. As per values of Wilks' Lambda, the smaller the value of Wilks' Lambda, the more important the independent variable. Therefore, it indicates that the important independent variable is number of days for full maturity followed by number of days for 50 percent maturity, average plant height, number of effective tillers and 1000 grains weight for yield classes of rice genotypes. Also, it is concluded that the variables such as number of days for 50 percent maturity and number of days for full maturity are highly significant and these regressors are the main contributors for differences in means of three classes for yield of rice.

The architecture of Multilayer Perceptron Neural Network (MLPNN) in Figure 1 depicts that there are 5 input nodes and 4 hidden nodes for yield of rice, the lines with light colour represents weights greater than zero and the dark colour lines display weights less than zero.

Table 3 depicts that in training dataset of MLPNN, 7 out of the 9 Low yield genotypes are correctly classified with 77.8 percent of accuracy, 39 out of 50 Medium yield genotypes are classified correctly with 78.0 percent of accuracy, 3 out of 21 High yield genotypes are correctly classified with 14.3 percent of accuracy and overall, 61.3 percent of the training cases are classified correctly. In testing dataset of MLPNN, 4 out of the 5 Low yield genotypes are correctly classified with 80 percent of accuracy, 13 out of 13 Medium yield genotypes are classified correctly with 100 percent of accuracy, 0 out of 2 High yield genotypes are correctly

classified with 0 percent of accuracy and overall, 85 percent of the testing samples are classified correctly.

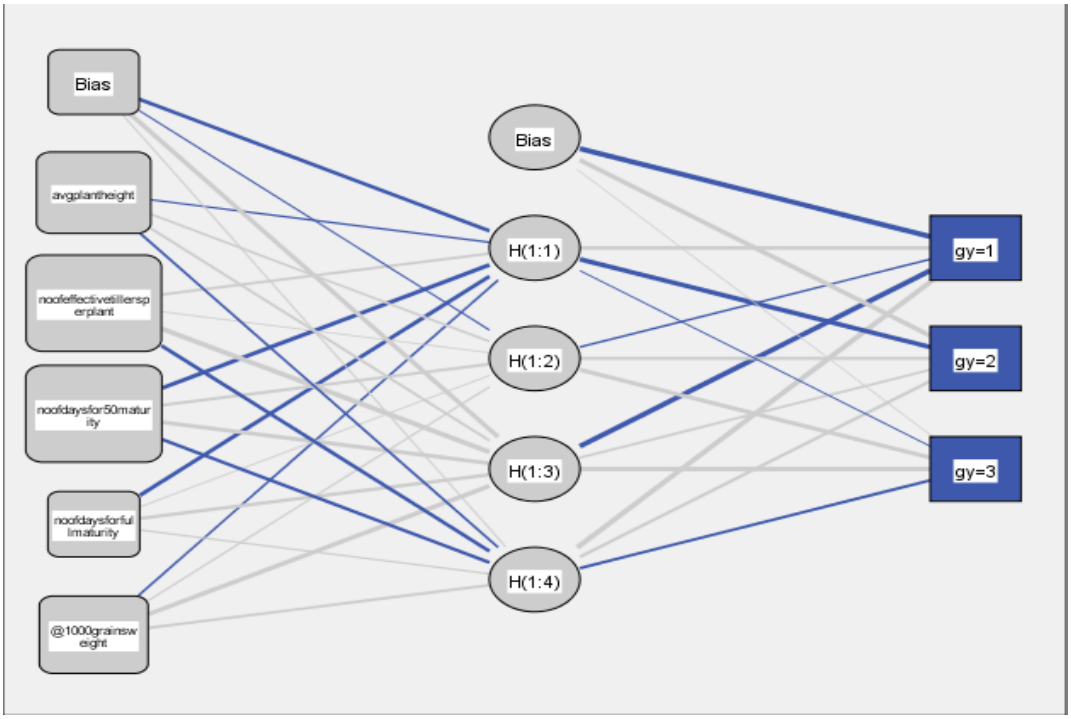


Figure 1: Architecture of MLPNN for yield of rice genotypes

Table 3: Classification table of MLPNN for yield of rice genotypes

Sample	Observed (Number of genotypes)	Predicted (Number of genotypes)			
		Low	Medium	High	% Correct
Training (80 %)	Low (9)	7	2	0	77.8
	Medium (50)	7	39	4	78.0
	High (21)	1	17	3	14.3
	Overall %				61.3
Testing (20 %)	Low (5)	4	1	0	80.0
	Medium (13)	0	13	0	100.0
	High (2)	0	2	0	0.0
	Overall %				85.0

Figure 2 represents the importance of independent variable through MLP neural network for classification of rice genotypes for different classes of yield and depicts that number of days for 50 percent maturity is the most important independent variable for classification (100 percent) followed by number of effective tillers per plant (99.1 percent), average plant height (76 percent), 1000 grains weight (69.8 percent) and number of days for full maturity (51.2 percent).

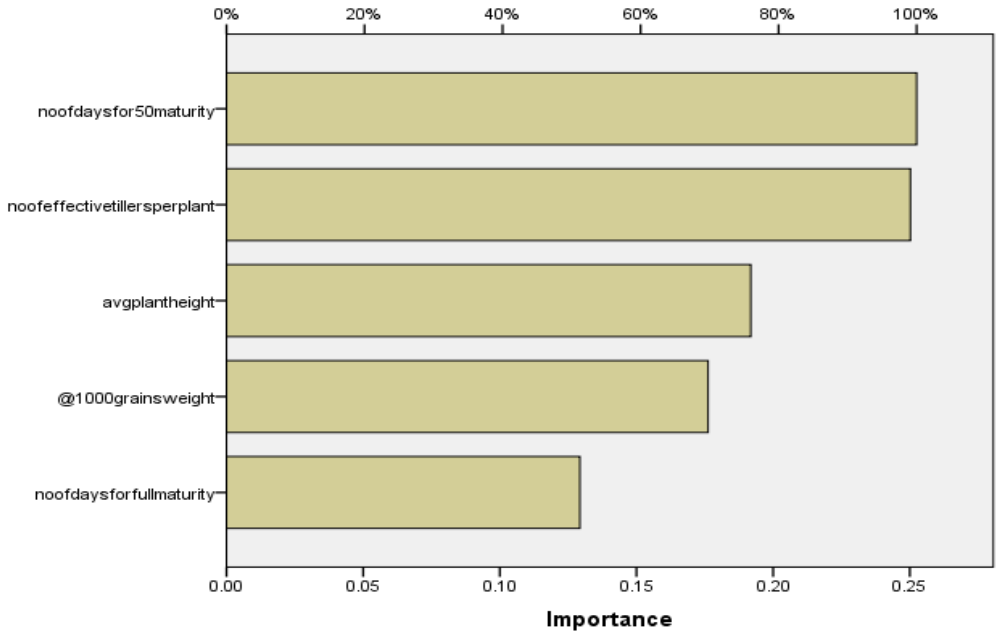


Figure 2: Normalized independent variable importance

Table 4: Classification ability measure

Criteria	Measures	Discriminant Analysis	Multilayer Perceptron
Classification Ability	Accuracy Rate	75	85
	Kappa Statistics	0.59	0.65

Table 4 represents the value of different classification ability measures and these measures will help to select the best model for classification. The accuracy rate for Discriminant Analysis is 75 percent whereas for MLPNN it is 85 percent. Also, value of kappa statistics for discriminant analysis is 0.59 but 0.65 for MLPNN.

4. Conclusion

The MLPNN method performed better as compare to Discriminant Analysis method for classification of rice genotypes for different classes of yield of rice genotypes as it has larger values of classification ability measures. The important attributing character for classification of rice genotypes through MLPNN is number of days for 50 percent maturity followed by number of effective tillers per plant, average plant height, 1000 grains weight and number of days for full maturity whereas for Discriminant Analysis important attributing variable is number of days for full maturity followed by number of days for 50 percent maturity, average plant height, number of effective tillers and 1000 grains weight for classification of rice genotypes for yield.

References

Fisher, R. A. (1936). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Inc.publication.

Galdon, B. R., Mendez, E. M., Havel, J., and Diaz, C. (2010). Cluster analysis and artificial neural networks multivariate classification of onion Varieties. *Journal of Agricultural and Food Chemistr*y, **58**, 11435–11440.

Halagundegowda, G. R., Singh, Abhishek, and Meenakshi, H. K. (2017). Discriminant analysis for prediction and classification of farmers based on adoption of drought coping mechanisms. *Agriculture Update*, **12**, 635-640.

Khan, M., and Hooda, B. K. (2021). Potential of artificial neural networks as compared to discriminant analysis in the classification of mustard accessions using grain yield. *International Journal of Statistics and Applied Mathematics*, **6**, 20-23.

Kumar, R., and Verma, R. (2012). Classification algorithms for data mining: A survey. *International Journal of Innovative Engineering Technology (IJIET)*, **1**, 7-14.

Nagraja M. S., and Singh, Abhishek (2018). Statistical models for classification of genotypes for yield of little Millet. *International Journal of Agriculture Sciences*, **10**, 5593-5597.

Nagraja, M. S., and Singh, A. (2018). Use of ordinal logistic regression and multiclass discriminant model for classification of genotypes for maturity of little millet. *International Journal of Pure Applied Biosciences*, **6**, 248-258.

Pazoki, A. R., Farokhi, F., and Pazoki, Z. (2014). Classification of rice grain varieties using two artificial neural networks (mlp and neuro-fuzzy). *The Journal of Animal & Plant Sciences*, **24**, 336-343.

Praveen, S., and Gayatri, V. (2005). Discriminant analysis for rice-wheat system having the same attributing characters towards grain yield. *Indian Journal Agricultural Research*, **39**, 203-207.

Savakar, D. (2012). Identification and classification of bulk fruits images using artificial neural networks. *International Journal of Engineering and Innovative Technology*, **1**, 36-40.