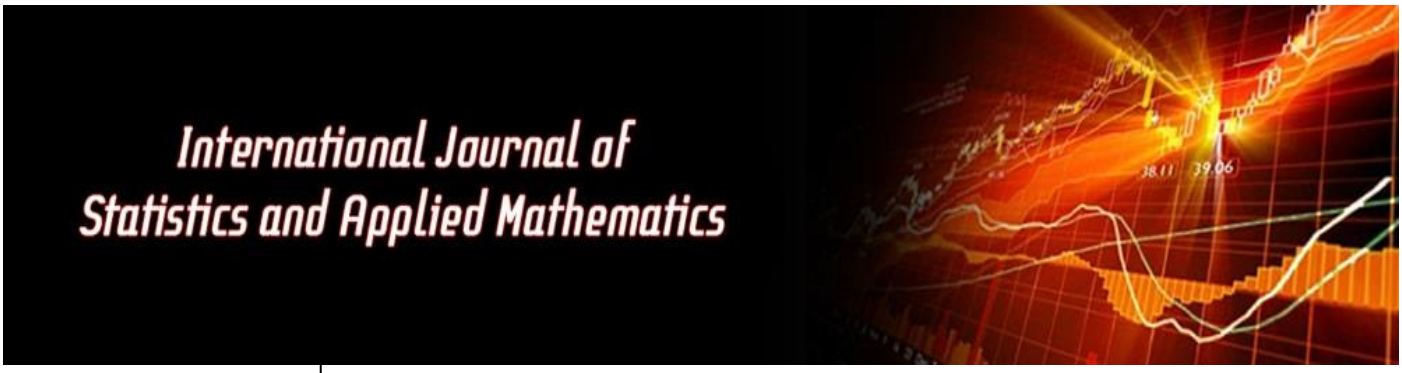
International Journal of Statistics and Applied Mathematics 2021; 6(3): 20-23



ISSN: 2456-1452
Maths 2021; 6(3): 20-23

www.mathsjournal.com
Received: 13-03-2021
Accepted: 15-04-2021

Mujahid Khan
a) Department of Mathematics and Statistics, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India
b) Agricultural Research Station (S.K.N. Agriculture University, Jobner), Fatehpur-Shekhawati, Sikar, Rajasthan, India

BK Hooda
Department of Mathematics and Statistics, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India

Corresponding Author:
Mujahid Khan
Agricultural Research Station (S.K.N. Agriculture University, Jobner), Fatehpur-Shekhawati, Sikar, Rajasthan, India

Potential of artificial neural networks as compared to discriminant analysis in the classification of mustard accessions using grain yield

Mujahid Khan and BK Hooda

Abstract
The possibility of using four different models (LDA, QDA, RDA and MLP neural network trained by the back-propagation algorithm) for the classification of mustard accessions was investigated and performances of the optimal models were compared. The secondary data of 870 mustard accessions for 13 morphological attributes was collected from the Department of Genetics and Plant Breeding, Chaudhary Charan Singh Haryana Agricultural University, Hisar. The attribute grain yield (g/plant) was used as class attribute in this study. When considering predictive accuracy over an independent testing dataset, the MLP neural network trained by the back-propagation algorithm, being able to correctly predict about 91% of mustard accessions. The corresponding predictive accuracies for LDA, QDA and RDA were 87.0%, 88.9% and 88.9%, respectively.

Keywords: Accuracy, artificial neural network, classification, discriminant analysis, mustard

1. Introduction
Agriculture is one of the most extensive activities in the world. It is the most traditional of all productive activities and has gone through many technological evolutions and transformations over time with the aim of producing more and better. However, this sector is now facing major challenges. On the one hand, according to a United Nations report being launched in June 2017, an increase of the world population is expected to attain 9.8 billion in 2050 and 11.2 billion in 2100 (United Nations, 2017) [13]. In the next 20 years, world food production is required to increase by 50% to feed the projected world population. So, agricultural intensification is required to feed the growing and increasingly demanding human population. On the other hand, agricultural intensification can have some profound impact on the environment like soil degradation due to wind and water erosion, air and water pollution due to excessive nutrients and agrochemicals, loss of biological and ecological diversity.
To reduce the negative effects of productive but intensive agriculture, it is urgent to transform agricultural production processes in a more sustainable way by properly allocating resources, using smart agriculture practices and making all productive systems resilient to climate change (Issad *et al.*, 2019) [5]. Resources optimization for sustainable production with controlled costs is the basic principle of smart agriculture practices. Smart and precision agriculture systems have arisen as new scientific fields that use data intense approaches. Machine learning has emerged to create new opportunities to resolve, quantify and understand these data intensive processes in agricultural sectors (Liakos *et al.*, 2018) [7]. The use of machine learning techniques has led to several tasks in the agricultural field such as classification and prediction of crop diseases, fertilizer suggestion, input management (planning of pesticides and irrigation), pest identification, predicting soil moisture in real time, yield prediction etc.
A variety of techniques for classification are available in literature including the discriminant analysis, artificial neural networks (Huang *et al.*, 2005) [4], Bayesian networks (Marcot & Penman, 2019) [9], k-nearest neighbour classifier (Zhang *et al.*, 2019) [15], decision trees (Trabelsi *et al.*, 2019) [12] and support vector machines (Vapnik, 1995) [14]. As simplicity and fast speed of the classification process are essential factors for a real-time system, four relatively simple and fast classification approaches were selected:

The linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularized discriminant analysis (RDA) and artificial neural network (ANN). These models were implemented for the classification of mustard accessions based on grain yield attribute, and their accuracies were compared.

2. Materials and Methods

2.1 Dataset

The mustard data consists of 1167 mustard (*Brassica spp*.) accessions including 5 checks repeated in 11 blocks. The data were not available for 247 entries, so they were eliminated from further analyses. The 5 checks were averaged individually which makes it a complete dataset of 870 accessions. The secondary data were taken from the experiment "Mustard germplasm for detailed evaluation under CRP" conducted by the Department of Genetics and Plant Breeding, Chaudhary Charan Singh Haryana Agricultural University, Hisar during *Rabi* season of 2015-16. The data were obtained for the following 13 morphological attributes viz., Days to 50% flowering (DF), Days to 80% maturity (DM), Leaf width in centimetre (LW), Leaf length in centimetre (LL), Plant height in centimetre (PH), Number of primary branches (NPB), Number of silique on main stem (NSMS), Number of silique on basal primary branches (NSBPB), Number of silique/plant (NSP), Number of seeds/siliqua (NSS), Silique length in centimetre (SL), Test weight in gram (TW) and Grain yield/plant in gram (GY).

2.2 Pre-processing of data

The attribute grain yield (g/plant) was used as class attribute in this study. To change the continuous attribute into class attribute, the Jenks natural breaks optimization method (Jenks, 1967) [6] was used. It is a data clustering method designed to determine the best arrangement of values into different classes. This is done by seeking to minimize average deviation of each class from the class mean, while maximizing deviation of each class from means of the other classes. In other words, the method seeks to reduce the variance within classes and maximize the variance between classes. After grouping the class attributes in their respective number of classes, the descriptive statistics were calculated class-wise as well as overall for all the morphological attributes of mustard accessions.

2.3 Discriminant analysis

The fundamental problem inherent to the discriminant analysis lies in assigning an unknown subject to one of two or more classes on the basis of a multivariate observation. The discriminant analysis procedure permitted the development of a predictive model of class membership based on characteristics observed in each case. The procedure originated a discriminant function corresponded to the number of classes minus one, based on linear combinations of the independent variables. A simple and popular discrimination method is LDA which is typically carried out using Fisher's (1936) [2] method. This method relies on the sample averages and covariance matrices computed from the different classes. In LDA, we assume that the different classes are distributed as multivariate normal and have equal covariance matrices. In QDA, we relax the assumption of equality of the covariance matrices, which means the covariances are not necessarily equal. Because of the quadratic decision boundary which discriminates the different classes, this method is named QDA. When covariance matrices are equal, the decision boundary will be linear and QDA reduces to LDA. The RDA builds a classification rule by regularizing the group covariance matrices (Friedman, 1989) [3] allowing a more robust model against multi-collinearity in the data. This might be very useful for a large multivariate data set containing highly correlated predictors. RDA is a kind of a trade-off between LDA and QDA. RDA shrinks the separate covariances of QDA toward a common covariance as in LDA.

2.4 Artificial neural network

Classifying mustard accessions was done using ANN based on morphological attributes using R version 4.0.2. A multi-layer perceptron (MLP) network which is commonly used to classify observations was designed. MLPs often have one or more hidden layers of non-linear or linear neurons followed by an output layer. Several layers of neurons with non-linear activation functions allow the network to learn non-linear and linear relationships between input and output vectors. Based on concepts developed by various researchers to find the optimum number of neurons in hidden layer, networks with one hidden layer with different number of neurons for hidden layers were developed. Number of neurons in input layer was equal to number of input attributes (i.e. 12) and number of neurons in output layer was two. In the developed MLP model, the sigmoid activation function was used for hidden as well as output layers. One of the most popular methods for MLP training is the Back-propagation (BP) algorithm, used in this study. 80% of data were used as training dataset for model training and 20% of data were used as test dataset for the performance assessment of the classification model. Before training the classification model, attributes were normalized to zero mean and unit variances. The widely applied performance measures like confusion matrix, accuracy, Kappa value, sensitivity, specificity, predictive values, balanced accuracy and F-measure were used in this study.

3. Results and Discussions

The Jenks method categorized the attribute grain yield into three classes: low yield class (3.3 – 16.0 g/plant) having 360 mustard accessions, medium yield class (16.1 – 26.0 g/plant) having 330 mustard accessions and high yield class (26.1 – 38.7 g/plant) with 180 accessions. Only low and high classes were kept for further analyses by deleting the medium classes due to disturbances created by medium class in classification process. Therefore, only 540 accessions were utilized in case of grain yield based classification (by removing the effect of 330 medium yield accessions). The descriptive statistics of the morphological attributes with respect to individual classes as well as overall are presented in Table 1 for grain yield attribute.

The standardized discriminant coefficients obtained from the linear discriminant analysis on the basis of grain yield are presented in Table 2. The attributes days to 50% flowering (-0.185), plant height (-0.147) and number of silique on main stem (-0.047) showed negative standardized loadings, while, rest of the attributes (days to 80% maturity, leaf length, leaf width, number of primary branches, number of silique per plant, number of silique on basal primary branches, silique length, number of seeds per siliqua and test weight) showed positive standardized loadings for grain yield based classification.

Table 1: Mean, standard deviation and coefficient of variation for various mustard attributes on the basis of grain yield

Attributes	Low Yield (360 Accessions)			Medium Yield (330 Accessions)			High Yield (180 Accessions)			Overall (870 Accessions)		
	Mean	SD	CV	Mean	SD	CV	Mean	SD	CV	Mean	SD	CV
DF	44.6	7.7	17.3	45.6	6.5	14.3	48.4	7.5	15.5	45.8	7.3	15.9
DM	139.2	5.5	4.0	139.6	4.9	3.5	142.5	5.5	3.9	140.0	5.4	3.9
PH	175.0	25.9	14.8	183.2	21.1	11.5	188.5	17.3	9.2	180.9	23.2	12.8
LL	5.8	1.4	24.1	6.3	1.1	17.5	6.7	0.7	10.4	6.2	1.2	19.4
LW	6.6	0.9	13.6	6.6	0.9	13.6	6.7	0.8	11.9	6.6	0.9	13.6
NPB	6.8	1.6	23.5	7.3	1.6	21.9	7.8	1.6	20.5	7.2	1.6	22.2
NSP	322.1	157.9	49.0	449.0	181.0	40.3	611.3	241.0	39.4	430.0	215.5	50.1
NSBPB	29.7	6.8	22.9	33.3	6.2	18.6	35.1	6.7	19.1	32.2	6.9	21.4
NSMS	44.1	8.8	20.0	47.6	8.7	18.3	49.4	7.7	15.6	46.5	8.8	18.9
SL	3.4	0.4	11.8	3.4	0.4	11.8	3.5	0.4	11.4	3.4	0.4	11.8
NSS	12.3	1.6	13.0	12.9	1.5	11.6	13.1	1.5	11.5	12.7	1.6	12.6
TW	3.4	1.1	32.4	3.7	1.0	27.0	3.8	1.0	26.3	3.6	1.1	30.6

Table 2: Standardized coefficients obtained from linear discriminant analysis in mustard accessions

Attributes	Standardized coefficients
Days to 50% flowering	-0.185
Days to 80% maturity	0.362
Plant height (cm)	-0.147
Leaf length (cm)	0.352
Leaf width (cm)	0.074
Number of primary branches	0.151
Number of silique per plant	1.042
Number of silique on basal primary branches	0.220
Number of silique on main stem	-0.047
Silique length (cm)	0.045
Number of seeds per siliqua	0.244
Test weight (g)	0.696

The comparative classification performances of LDA, QDA and RDA in the training as well as testing datasets for the mustard accessions are given in Table 3. Results given in table indicated that the performance of RDA was superior to LDA and QDA. The training accuracies (and Kappa values) of 85.7% (66.5%), 84.9% (65.2%) and 86.9% (69.5%) were obtained for LDA, QDA and RDA, respectively for classification on the basis of grain yield. The 88.9% (74.3%) prediction accuracies (Kappa values) were attained for QDA and RDA both, but prediction accuracy (Kappa value) of 87.0% (69.1%) was achieved for LDA in this dataset. Feldesman (2002) [1] also obtained the 86.9% classification accuracy for LDA using typical morphometric data. For the Parma ham dataset, Silva *et al.* (2016) [11] obtained similar classification accuracy of about 86% by using LDA.

Table 3: Comparative performance of LDA, QDA and RDA for classification of mustard accessions

Discriminant Analysis	Training		Testing	
	Accuracy	Kappa	Accuracy	Kappa
Linear	0.857	0.665	0.870	0.691
Quadratic	0.849	0.652	0.889	0.743
Regularized	0.869	0.695	0.889	0.743

Table 4 shows the obtained confusion matrix and performance measures results of LDA, QDA and RDA for classification based on grain yield. In LDA, 69 of low yield accessions were correctly predicted as low, 3 of low yield accessions were misclassified as high yield, 11 of high yield accessions were wrongly predicted as low yield and 25 high yield accessions were accurately classified. While for QDA and RDA, the low yield class obtained the highest accuracy of 94.4% and high yield class accurately classified the 77.8% accessions.

Table 4: Confusion matrix and performance measures for LDA, QDA and RDA based classification of mustard accessions

Performance statistics		Actual					
		LDA		QDA		RDA	
		L	H	L	H	LL	HH
Prediction	L	69	11	68	8	68	8
	H	3	25	4	28	4	28
Sensitivity		0.958		0.944		0.944	
Specificity		0.694		0.778		0.778	
Positive Predictive Value		0.863		0.895		0.895	
Negative Predictive Value		0.893		0.875		0.875	
Balanced Accuracy		0.826		0.861		0.861	
F–measure		0.908		0.919		0.919	

The results of MLP neural networks with back-propagation algorithm to classify mustard accessions on the basis of grain yield are given in Table 5. The results showed that the accuracy of the classification on the basis of grain yield was best when 33 neurons (87.4% accuracy) were used in hidden layer for training dataset while 105 neurons (90.7% accuracy) were used in hidden layer for testing dataset. The training and testing accuracies were in the range of 75.3-87.4% and 66.7-90.7%, respectively. The testing accuracy of BP trained MLP neural networks marked a nearly 2% improvement in comparison with classification results yielded by regularized discriminant analysis. Paliwal *et al.* (2001) [10] also obtained 88% accuracies for barley and rye with various neural network models for cereal grain classification. Lu *et al.* (2016) [8] also demonstrated that the optimized single hidden layer feed-forward neural network achieved 89.5% accuracy. Zhang *et al.* (2016) [16] also proposed a fruit classification system by optimized feed-forward neural networks and the results showed that the method yielded an accuracy of 89.11%.

Table 5: Performance of Back-propagation MLP neural networks for classification of mustard accessions

Method	No. of Neurons	Training		Testing		Error
		Accuracy	Kappa	Accuracy	Kappa	
Marchandani & Cao (1989)	105	0.868	0.700	**0.907**	**0.789**	0.763
Li *et al*. (1995) / Sheela & Deepa (2013)	4	0.862	0.691	0.852	0.662	0.596
Tamura & Tateishi (1997)	11	0.872	0.707	0.870	0.700	0.624
Lawrence & Fredrickson (1998)	7	0.860	0.683	0.870	0.704	0.626
Zhang *et al*. (2003)	315	0.753	0.420	0.667	0.000	5.654
Jinchuan & Xinzhe (2008)	33	0.874	0.712	0.843	0.638	0.571
Shibata & Ikeda (2009)	5	0.868	0.702	0.852	0.652	0.581
Hunter *et al*. (2012)	13	0.871	0.707	0.880	0.727	0.575
Ranganayaki & Deepa (2016)	9	0.865	0.692	0.880	0.731	0.615

4. Conclusion

In this paper, four different classification models (LDA, QDA, RDA and MLP neural network trained by the back-propagation algorithm) have been used to classify 870 mustard accessions on the basis of attribute grain yield, considering as inputs 12 morphological attributes. The MLP neural networks resulted in a predicted accuracy of 90.7% which was higher than all the discriminant analyses models (87.0%, 88.9% and 88.9% for LDA, QDA and RDA, respectively). When comparing the classification models, it was shown that the best MLP neural network trained by the back-propagation algorithm outperforms discriminant analysis approaches, providing a predictive ability which was about 2-4% higher. Therefore, artificial neural networks, and particularly back-propagation architecture, allow to reliably predict the mustard accessions.

5. References

1. Feldesman MR. Classification trees as an alternative to linear discriminant analysis. American Journal of Physical Anthropology 2002;119: 257-275.
2. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics 1936;7(2):179-188.
3. Friedman J. Regularized discriminant analysis. Journal of American Statistical Association 1989;84:165-175.
4. Huang DS, Ip HHS, Law KCK, Chi Z. Zeroing polynomials using modified constrained neural network approach. IEEE Transactions on Neural Networks 2005;16:721-732.
5. Issad HI, Aoudjit R, Rodrigues JJPC. A comprehensive review of data mining techniques in smart agriculture. Engineering in Agriculture, Environment and Food 2019;12(4):511-525.
6. Jenks GF. The data model concept in statistical mapping. International Yearbook of Cartography 1967;7:186-190.
7. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: A review. Sensors 2018;18:1-29.
8. Lu S, Lu Z, Phillips P, Wang S, Wu J, Zhang Y. Fruit classification by HPA-SLFN. 8th International Conference on Wireless Communications and Signal Processing (WCSP) 2016, 1-5.
9. Marcot BG, Penman TD. Advances in Bayesian network modelling: Integration of modelling technologies. Environmental Modelling and Software 2019;111:386-393.
10. Paliwal J, Visen NS, Jayas DSAE: Automation and emerging technologies. Journal of Agricultural Engineering Research 2001;79(4):361-370.
11. Silva AC, Soares SFC, Insausti M, Galvao RKH, Band BSF, de Araujo MCU. Two-dimensional linear discriminant analysis for classification of three-way chemical data. Analytica Chimica Acta 2016;938:53-62.
12. Trabelsi A, Elouedi Z, Lefevre E. Decision tree classifiers for evidential attribute values and class labels. Fuzzy Sets and Systems 2019;366:46-62.
13. United Nations. World population projected to reach 9.8 billion in 2050, and 11.12 billion in 2100 2017. https://www.un.org/development/desa/en/news/population/ world-population-prospects-2017.html.
14. Vapnik VN. The nature of statistical learning theory. Springer Verlag, New York 1995, 188.
15. Zhang Y, Cao G, Wang B, Li X. A novel ensemble method for k-nearest neighbor. Pattern Recognition. 2019;85:13-25.
16. Zhang Y, Phillips P, Wang S, Ji G, Yang J, Wu J. Fruit classification by biogeography-based optimization and feedforward neural network. Expert Systems 2016;33(3):239-253.