Available online at www.ijpab.com

DOI: http://dx.doi.org/10.18782/2320-7051.7792

ISSN: 2582 – 2845

Ind. J. Pure App. Biosci. (2019) *7*(5), 342-352

Research Article

Indian Journal of Pure & Applied Biosciences

Variable Selection for Discrimination between Low and High Yielding Populations of Indian Mustard

Poonam Godara[1*], B. K. Hooda[1] and Ram Avtar[2]

[1]Department of Mathematics & Statistics,
[2]Department of Genetics and Plant Breeding,
CCS, Haryana Agricultural University Hisar-125004 (Haryana), India
*Corresponding Author E-mail: poonamsinghsinghmar@gmail.com

Received: 4.08.2019 | Revised: 15.09.2019 | Accepted: 26.09.2019

ABSTRACT

Variable Selection is an important problem in classification and discriminant analysis. The selection of important variables for the purpose of discrimination between populations is important from the point of view of time and resources required for the experimentation. Keeping this in view, the present study has been designed to find important characters of Indian mustard which can discriminate between high and low yielding genotypes. Secondary data set on 310 genotypes of Indian mustard recorded for 12 characters was used for discrimination between populations of low and high yielding genotypes of Indian mustard. Three variable selection methods (Univariate t-test, Rao´s F test for additional Information and Random Forests Algorithm) for classification and discrimination were used and compared. Performance of the methods was assessed in terms of leave one out cross-validation error and out of bag error rate for classification. The Four most important variables for discrimination among genotypes based on seed yield per plants were secondary branches, primary branches, days to maturity and siliqua number on main shoot.

Keywords: *Classification, Discriminant analysis, Error rates, Gini index, Random Forests.*

INTRODUCTION

Variable selection is the process of selecting the most relevant original variables from the set of given variables that have predictive efficiency in any context. Selection of relevant variables is important because it reduces the complexity of a model and makes it easier to interpret. It reduces the training time and effort and also reduces over-fitting.

For developing better genotypes/hybrids, the choice of suitable parents is a matter of great concern to the plant breeders. For this purpose, breeders conduct experiments and record data on large number of variables. These variables do not have equal importance. Many of these variables are irrelevant and redundant to the investigator.

Cite this article: Godara, P., Hooda, B. K., & Avtar, R. (2019). Variable selection for discrimination between low and high yielding populations of Indian mustard, *Ind. J. Pure App. Biosci. 7*(5), 342-352. doi: http://dx.doi.org/10.18782/2320-7051.7792

The analysis and interpretation of such data sets is often difficult and causes several problems. Different variables in the data carry different amounts of information. Some will be more informative in some sense than others. So, the researchers may wish to reduce the number of variables for the final decision making while maintaining high performance by discarding those least useful. Identification of redundant variables and selection of important variables is, thus an important area of research in multivariate analysis involving large number of variables. Variable selection is, also employed in order to find most important or useful variables for various data mining tasks such as classification and discriminant.

Several criteria and algorithms have been purposed for feature selection in classification and discrimination problems. McCabe (1975) adopted Furnival's algorithm to obtain all possible subsets of variables of any size, using Wilk´s Λ criteria. Farver & Dunn (1979) considered a forward stepwise discrimination procedure and stepwise procedure preceded by a preliminary screening of variables on the basis of individual *t* statistics. Munita et al. (2006) provided stopping rules to identify the redundant variables using stepwise discriminant analysis. Random forests algorithm which uses Decision Trees as base classifier for classification was originally conceived as a method of building a forest of uncorrelated trees by combining several classification and regression type randomized decision trees using bagging (Breiman, 1996). Genuer et al. (2010) proposed a variable selection method based on random forests (Breiman, 2001) and described the associated R package called **VSURF** to illustrate its use on real datasets. Painsky & Rosset (2016) proposed a framework for splitting using leave-one-out (LOO) cross-validation (CV) for selecting the splitting variable, then performing a regular split for the selected variable.

The present study has been designed to find important characters of Indian mustard which can discriminate between high and low yielding genotypes. For this purpose, three variable selection methods (Univariate t-test, Rao´s F test for Additional Information and Random Forests Algorithm) for classification and discrimination were used and compared. Performance of the methods was assessed in terms of leave one out cross validation error and out of bag (OOB) error rate for classification.

MATERIALS AND METHODS

The study was conducted on Indian mustard (*Brassica juncea*). Secondary data on 310 Indian mustard genotypes were obtained from an experiment conducted by Oilseeds Section of the Department of Genetics and Plant Breeding, CCS HAU, Hisar during rabi season of 2015-16. The observations were recorded on three plants per row per character per plot. The genotypes were recorded for the 12 characters; *viz.* Days to flowering (DF), Number of primary branches (PB), Number of secondary branches (SB), Main shoot length (MSL), Plant height (PH), Siliqua length (SL) in centimetres, Siliqua number on main shoot (SNOMS), Seeds per siliqua (SPERS), Days to maturity (DM), Thousand seed weight (TSW) in grams, Seed yield (SY) in gram/plant, Oil content (OC) in per cent. The genotypes were divided into two Groups (G_1 and G_2) for low and high yielding genotypes on the basis of the following criterion: G_1: Seed yield < mean-standard deviation/2 and G_2 : Seed yield ≥ mean+ standard deviation/2. Accordingly 118 genotypes contained in G_1 were found low yielding. Seed yield of 118 genotypes in this group was less than 12.71 g/plant. These genotypes were considered as individuals from low yielding populations of Indian mustard. In group 2, there were 80 genotypes whose mean seed yield was found high as compared to set benchmark mean (19.69). These genotypes were considered as individuals from high yielding populations of Indian mustard.

2.1 Test for Homogeneity of Covariance Matrices (Box-M Test)

Box (1949) proposed the statistic for testing the hypothesis of equal covariance matrices. Let S_i is the unbiased estimate of the variance covariance $\sum$.

Null Hypothesis $H_0: \sum^{(1)} = \sum^{(2)} = \cdot\cdot\cdot = \sum^{(k)}$

Alternative Hypothesis H_1: At least one of the equality does not hold good

$$M = (N\text{-}k)\ln|S| - \sum_{i=1}^{k} (N_i\text{-}1) \ln|S_i|$$

$$C^{-1} = 1 - \frac{(2p^2+3p\text{-}1)}{6(p+1)(k\text{-}1)} \left[\sum_{i=1}^{k} \frac{1}{N_i\text{-}1} - \frac{1}{N\text{-}k}\right]$$

Where $S = \frac{1}{N\text{-}k} \sum_{i=1}^{k} (N_i\text{-}1)S_i$ and $N = \sum N_i$ for i=1, 2,…, k

S is the pooled sample covariance matrix.

Test statistic is given by

$$MC^{-1} = (N\text{-}k)\ln|S| - \sum_{i=1}^{k} (N_i\text{-}1) \ln|S_i| \left[1 - \frac{(2p^2+3p\text{-}1)}{6(p+1)(k\text{-}1)} \left[\sum_{i=1}^{k} \frac{1}{N_i\text{-}1} - \frac{1}{N\text{-}k}\right]\right]$$

Box MC^{-1} has a Chi-square distribution with $\frac{1}{2}p(p+1)(k-1)$ degrees of freedom. H_0 is rejected if the MC^{-1} is greater than tabulated Chi-square.

2.2 Variable Selection Methods for Classification and Discrimination

In this section we describe various variable selection methods which were applied on the secondary data available for 12 variables of mustard crop:

2.2.1 Rao´s F test for Additional Information

Let Π_1: $N_p(\mu^{(1)}, \sum)$ and Π_2: $N_p(\mu^{(2)}, \sum)$ be two populations assumed to be p variate normal with same covariance matrix. Let $X^{(1)}: (N_1 \times p)$ and $X^{(2)}: (N_2 \times p)$ be the samples of size N_1 and N_2 from Π_1 and Π_2 respectively. The null and alternative hypotheses that discarded (p-q) variables do not discriminate between these populations or do not provide additional discrimination are:

$$H_0: \Delta^2_{(p-q)} = 0$$

and $$H_1: \Delta^2_{(p-q)} \neq 0$$

Rao´s F statistic (1973) to test H_0 is:

$$F = \frac{N_1 + N_2 - p - 1}{p - q} \cdot \frac{N_1 N_2 (D_p^2 - D_q^2)}{(N_1 + N_2)(N_1 + N_2 - 2) + N_1 N_2 \Delta_q^2}$$

where

$$D_p^2 = (\bar{x}_p^{(1)} - \bar{x}_p^{(2)})' S^{-1} (\bar{x}_p^{(1)} - \bar{x}_p^{(2)}) \text{ and } D_q^2 = (\bar{x}_q^{(1)} - \bar{x}_q^{(2)})' S^{-1} (\bar{x}_q^{(1)} - \bar{x}_q^{(2)})$$

are estimates of Mahalanobis distance between two populations based on the original set of p and q variables respectively. Rao’s F statistic follows F-distribution with (p-q, N_1+N_2-p-1) degree of freedom. Reject H_0 if $F_{cal} > F_{p\text{-}q,\, N_1+N_2\text{-}p\text{-}1}(\alpha)$.

2.2.2 Random Forests

The Random forests, a classification technique introduced by Breiman (2001) is a substantial improvement of bagging that builds large collection of de-correlated trees, and then averages them. The method combines bagging and the random selection of features. In Random forests different subsets of equal sizes, are selected with replacement (bootstrapping) from the training data, to train each tree and the remaining testing data is used to estimate the error and importance of variable. About two-thirds of the total dataset is included in each random subset. The other one-third of the data is not used to build the trees, and this part is called the out-of-the-bag data. This part is later used to evaluate the model. This technique uses a user-defined number of variables selected at random from all of the variables to determine node splitting. A randomly selected subset of variables is used to split each node. Splits are chosen according to a purity measure called Gini index. Nodes with the greatest decrease in impurity start the trees, while nodes with the least decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node, a subset of the most important features is created. Random forests develop many classification trees, and to add a new classification tree to the forest, add it down to the each of the trees in the forest. Each tree provides its classification and we consider it as its vote for that class. The forest considers the classification receiving the most votes from all the trees in the forest.

Outlines of Random forests algorithm

i) From the Training of n samples draw ntree bootstrap samples.

ii) For each of the bootstrap samples, grow classification tree. At each node, randomly sample mtry of the predictors. The tree is grown to the maximum size and not pruned back. Bagging can be thought of as the special case of random forests obtained when mtry = p, the number of predictors.

iii) The out-of-bag prediction is obtained through a majority vote across trees whose observation was not included in the bootstrap sample.

Variable Importance using Gini Importance

Random Forest implementations provide variable importance measures. One such measure is based on mean decrease in impurity (or gini importance). Random Forest uses Gini Index based impurity measures for building a decision tree. Gini index is the inaccuracy measure of the decision trees. The Gini inaccuracy criterion for the parent node is always higher than the two descendent nodes that are split from the parent node. It assigns a score and ranks the features for feature importance. Improvement in the Gini decrease of each individual attribute for every tree in the forest provides surplus variable importance that is often quite consistent with the permutation importance measure. At each node'tʾ, decreases in Gini impurity are recorded for all variables used to form the split. Gini impurity Δgini(t) is defined as follows:

$$\Delta gini(t) = p(t)\ gini(t) - gini_{split}(t)$$

where $gini_{split}(t) = p_L gini\ (t_L) + p_R gini\ (t_R)$

and $gini(t)=1-\sum_K p(K \mid t)^2$

$p(K \mid t)$ is the rate at which class K is discriminated correctly at node t. gini(t_L) is a Gini index on the left side of the node, gini(t_R) is a Gini index on the right side of the node, p(t) is the number of observations before the split, p_L is the number of observations on the left side after the split, and p_R is the number of observations on the right side after the split. The Gini criterion is used to select the split with the highest impurity at each node. The average of all decreases in Gini impurity yields the Gini Importance or Mean Decrease in Impurity (MDI).

Error Rate Estimation

Out-of-bag error estimation was proposed by Tibshirani (1996) as an important ingredient for the calculation of generalization error. It is not required to cross-validate or a separate test to calculate an unbiased error estimate of the validation set in the random forest, since it performs eternally during the execution. Each tree is built using a random sampling with replacement from the original data. About one-third of the cases are left out as OOB data that are not used in the built of the p^{th} tree. Put each case from n OOB data in the build of the p^{th} tree down to the p^{th} tree to get a classification. With this process, a test classification is achieved for each case in about one-third of the trees. Eventually, consider q to be the class variable with maximum votes every time from m cases of OOB. The OOB error is estimated with the factor that q is not equal to the true class of m averaged over all cases.

The performance of a discriminant criterion in the classification of new observations in the validation data could be evaluated by estimating the probabilities of misclassification or error rates. To reduce the bias in the apparent error rate, the methods used is cross validation (Lachenbruch & Mickey, 1968). In cross validation, n-1, out of n training observations in the calibration sample are treated as a training set. It determines the discriminant functions based on these n-1 observations and then applies them to classify the one observation left out. We repeat this procedure for each observation, so that, in a sample of size $N = \sum_i N_i$ each observation is classified by a function based on the other N − 1 observations. Let n_{11} = number of correctly classified observations in group G_1, n_{22} = number of correctly classified observations in group G_2, n_{12} = number of observations misclassified in group G_2, n_{21} = number of observations misclassified in group G_1. Let N_1 observations are from group G_1 and N_2 observation are from group G_2.$N_1 = n_{11}+ n_{12}$, $N_2= n_{21}+ n_{22}$ and $N=N_1+ N_2$.

Error-rate estimates of the conditional misclassification probabilities can be calculated by the proportion misclassified in the validation sample given as:

$$\hat{P}(2|1)=\frac{n_{12}}{N_1}$$

$$\hat{P}(1|2)=\frac{n_{21}}{N_2}$$

and the total proportion misclassified is the unbiased estimate of the expected actual error rate, E(AER)

$$\hat{E}(AER) = \frac{n_{12} + n_{21}}{N_1 + N_2}$$

Sensitivity:

Sensitivity of a binary classification test with respect to some class is a measure of how well this test identifies a condition and expresses the probability of a case being classified in that class. It is the proportion of true positives of all positive cases in the group.

$$\text{Sensitivity} = \frac{n_{11}}{n_{11}+n_{21}}$$

Specificity:

Specificity, on the other hand, expresses the proportion of the true negative classified cases of a binary classification test of all the negative cases in the group.

$$\text{Specificity} = \frac{n_{22}}{n_{12}+n_{22}}$$

3. RESULTS AND DISCUSSION

Box-M test was applied to test the homogeneity of group covariance matrices. According to the Box-M test, Chi-Sq (χ^2) (approx.) = 50.30 at degree of freedom (df) = 55 and p-value = 0.65. The decision failed to reject the null hypothesis of equal covariance matrices and we concluded that the low and high seed yielding groups have covariance homogeneity. The test for equality of mean vectors (Hotelling T^2) and corresponding F-value were found to be 137.93 and 13.16 respectively. The mean vectors for low and high yielding populations were found to be significantly different at 5% level of significance, which meant that data are suitable for discrimination.

Pooled variance-covariance matrix for two groups was found to be:

$$S = \begin{bmatrix} 16.25 & 1.24 & 3.30 & -11.88 & 49.86 & -0.27 & -0.63 & 0.47 & 4.00 & -1.47 \\ 1.24 & 1.56 & 3.08 & -2.89 & 2.56 & -0.16 & 0.54 & 0.03 & 0.20 & -0.55 \\ 3.30 & 3.08 & 25.18 & -12.02 & 14.18 & -0.98 & 4.87 & 0.01 & 0.65 & -3.22 \\ -11.88 & -2.89 & -12.02 & 163.49 & 43.77 & 0.69 & 83.25 & -4.26 & -3.22 & 2.98 \\ 49.86 & 2.56 & 14.18 & 43.77 & 464.33 & -2.33 & 73.73 & -6.56 & 13.94 & -4.47 \\ -0.27 & -0.16 & -0.98 & 0.69 & -2.33 & 0.30 & -1.37 & 0.29 & 0.08 & 0.36 \\ -0.63 & 0.54 & 4.87 & 83.25 & 73.73 & -1.37 & 90.54 & -4.19 & -1.67 & -1.63 \\ 0.47 & 0.03 & 0.01 & -4.26 & -6.56 & 0.29 & -4.19 & 3.27 & -0.05 & -0.26 \\ 4.00 & 0.20 & 0.65 & -3.22 & 13.94 & 0.08 & -1.67 & -0.05 & 8.44 & 0.17 \\ -1.47 & -0.55 & -3.22 & 2.98 & -4.47 & 0.36 & -1.63 & -0.26 & 0.17 & 1.24 \end{bmatrix}$$

Univariate measures such as mean, standard deviation and coefficient of variation of ten characters for groups are given in Table 3 along with the t-values for testing the significance of the difference in individual variable means. Based on univariate t–test, the variables primary branches, secondary, siliqua number on main shoot and days to maturity are found to have significant differences in the two group means whereas the variables main shoot length, siliqua length and thousand seed weight are found least discriminatory variables. According to t–values, criterion, secondary branches is the most contributing variable for discrimination followed by primary branches, days to maturity and siliqua number on main shoot. The least discriminatory variable is main shoot length.

Table 1: Discriminatory variable selection using univariate independent sample t-test for equality of means

Variables	Std. Deviation		CV		Mean		t-value
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	
DF	3.97	4.12	8.21	8.43	48.41	48.86	-0.78
PB	1.16	1.37	23.7	23.52	4.88	5.83	-5.28*
SB	4.97	5.08	33.66	24.79	14.78	20.5	-7.87*
MSL	13.17	12.19	16.39	15.08	80.37	80.84	-0.25
PH	23.61	18.07	11.09	8.31	212.98	217.39	-1.41
SL	0.57	0.52	16.48	15.2	3.47	3.43	0.59
SNOMS	10.1	8.59	19.4	15.63	52.04	54.94	-2.10*
SPERS	1.79	1.84	14.33	14.24	12.46	12.93	-1.77
DM	2.74	3.14	1.87	2.12	146.44	148.13	-4.00*
TSW	1.06	1.19	28.72	31.77	3.68	3.76	-0.5

*t- values are significant at 0.05 levels (2-tailed)

For variable selection in discriminant analysis the backward elimination technique begins with the computation of the F-statistic for examining whether or not a particular variable supplies additional information independently of the remaining p-1 variables. The elimination technique deletes one variable at a time. Table 2 contains the Mahalanobis distance and Rao´s F-statistic values resulting from the backward elimination procedure for variable selection in mustard data. The largest value of Δ_q^2 = 2.89 and corresponding smallest value of F = 0.00001 corresponds to the variable days to flowering. The computed F-value was compared with critical value F (9, 187; 0.05) = 1.93. Hence days to flowering is eliminated and implies that selected nine variables provide equal distance between two groups as is given by all ten variables. The smallest F-value, in second step, 0.0001 corresponds to the variable main shoot length. The comparison of this with F (8, 187; 0.05) = 1.99 leads to the elimination of variable main shoot length.

Table 2: Variable selection using Rao´s F test for no additional Information

Step	Variable Selected	Variable discarded (p-q)	subset size(q)	Mahalanobis distance for selected variables (Δ_q^2)	Rao's F	F critical
1	PB,SB,MSL,PH,SL,SNOMS,SPERS,DM,TSW	DF	9	2.89	0.00001	1.93
2	PB,SB,PH,SL,SNOMS,SPERS,DM,TSW	MSL	8	2.89	0.00016	1.99
3	PB,SB,SL,SNOMS,SPERS,DM,TSW	PH	7	2.89	0.002	2.06
4	PB,SB,SNOMS,SPERS,DM,TSW	SL	6	2.79	1.01	2.15
5	SB,SNOMS,SPERS,DM,TSW	PB	5	2.66	1.91	2.26
6	SB,SPERS,DM,TSW	SNOMS	4	2.46	3.00	2.42
7	SB,DM,TSW	SPERS	3	2.30	3.59	2.65
8	SB , TSW	DM	2	2.10	4.26	3.04
9	SB	TSW	1	1.30	7.78	3.89

Continuing in the same way, the variables plant height, siliqua length and primary branches were eliminated at third, fourth and fifth steps respectively. The procedure terminated at sixth step when Rao´s F statistic value 3.00, exceeded the F (4, 187; 0.05) = 2.42. The null hypothesis that discarded variables siliqua number on main shoot do not discriminate between populations is rejected. Hence siliqua numbers on main shoot provides extra information, and should not be rejected.

The Random forests method was also applied to the low and high seed yield data, using the gini index to evaluate the importance of the predictors. For this purpose, total sample was divided randomly into two samples viz. the training sample and the test sample. Split ratio of 75:25 per cent was chosen for dividing the sample. Test samples (out-of-bag) samples were used to get an error rate for each bootstrap tree. Twenty five per cent of randomly selected samples were left out in modeling each bootstrap tree. The fitted tree was then used to get a predicted value for the OOB samples. The final classification of each OOB sample was determined by counting the number of times it was classified as a certain class every time it was an OOB sample. The analysis was carried out using 'randomForest' package in R. To select the important variables, the analysis starts by taking all ten variables. The data set contained 198 observations, so the training samples comprised of 148 which were used to construct the random forests while the left out 50 observations were used for assessing the performance of random forests (test set). The number of trees was set to 500. The OOB error rate for all ten variables came out to be 30 per cent. The variable having lowest gini index was discarded. In Table 3, at 1st step, the variable siliqua length has the lowest gini index (3.85) and hence siliqua length is the first variable to be discarded. At step 2, out of remaining nine variables, days to flowering with lowest gini index of 4.16 is deleted. Similarly, at third step main shoot length is removed followed by plant height, seeds per siliqua, thousand seed weight and so on. Variables which are removed from the analysis in the later stages considered as important. So, according to the gini index, secondary branches is the most important variable followed by primary branches, siliqua number on main shoot and days to maturity.

Table 3: Variable Selection using Random Forest Algorithm

Step	Variable Selected	Variable discarded	subset size(q)	Mean decrease gini	OOB error rate
1	DF,PB,SB,MSL,SL,PH,SNOMS,SPERS,DM,TSW	SL	10	3.85	30%
2	DF,PB,SB,MSL,PH,SNOMS,SPERS,DM,TSW	DF	9	4.16	28%
3	PB,SB,MSL,PH,SNOMS,SPERS,DM,TSW	MSL	8	6.08	28%
4	PB,SB,PH,SNOMS,SPERS,DM,TSW	PH	7	7.33	30%
5	PB,SB,SNOMS,SPERS,DM,TSW	SPERS	6	8.31	24%
6	PB,SB,SNOMS,DM,TSW	TSW	5	11.58	26%
7	PB,SB,SNOMS,DM	DM	4	11.98	26%
8	PB,SB,SNOMS,	SNOMS	3	19.25	26%
9	PB,SB	PB	2	20.62	36%
10	SB		1		32%

The first four variables led to reasonably stable and small error for classification. So, four most important discriminatory variables are used to compare the methods for variable selection. Table 4 shows misclassification error rates for these variable subsets selected by t-test, Rao's F-test, and Random Forest algorithm for the Indian mustard data. Among all the methods, t-test is performing better. The four variables selected based on t values criterion are: No. of primary branches, No. of secondary branches, siliqua number on main shoot and days to maturity with error rate of 21.72 %. In case of Rao’s additional criterion, the selected variables are No. of secondary branches, seeds per siliqua, days to maturity, and thousand seed weight with error rate of 23.23%. Random forests provide maximum error rate 26% and selected No. of primary branches, No. of secondary branches, siliqua no main shoot and days to maturity as the best discriminators.

For the evaluation of the variable selection methods sensitivity and specificity were also measured. The sensitivity criterion indicates the percentage of the observations that are correctly detected as actually belonging to a particular class. The specificity criterion among the samples which are not related to a class determines the percentages that have been recognized correctly as false. So the results of Table 4 revealed that in terms of the sensitivity and specificity, the best classifier performance was given by t-value with 83% and 70%, respectively. The degree of effectiveness of classifier performance is categorized by the accuracy statistical criterion. The highest accuracy of the classification was obtained by t-test with value of 78.28 per cent.

Table 4: Overall comparison of variable selection methods

Methods	**Variables Selected**	**Error Rate (%)**	**Sensitivity**	**Specificity**
t-value	PB, SB, SNOMS, SPERS	21.72	0.83	0.70
Rao's F	SB, SNOMS, SPERS, TSW	23.23	0.84	0.65
Random Forests	PB, SB, SNOMS, SPERS	26.00	0.75	0.70

4. Relative importance of the variables of the Indian Mustard

To study the relative importance, Magnitude of Linear Discriminant function coefficient, Correlation between variables and discriminant scores and Variable importance using mean decrease in Gini index methods were used.

Criteria of magnitude of discriminant function coefficients was applied for computing relative importance of individual characters of Indian mustard groups formed for classification and discrimination. The discriminant function is a linear combination of independent variables that will discriminate between the categories of the dependent variable. It enables to examine whether significant differences exist among the groups, in terms of the predictor variables. It is also used to evaluate the accuracy of the classification. Magnitudes of the coefficients were indicators of the relative importance of variables, as variables with large coefficients contribute more to the overall discriminant function. Variables ranks orderings according to Discriminant function coefficient using all variable are presented in Table 5.

Based on the coefficient of the linear discriminant function, variables thousand seeds weight, siliqua length, seed per siliqua and primary branches are observed as important variables. These variables contribute more to the discriminant score for discriminating between the groups partitioned on the basis of seed yield. Variables plant height, days to flowering and main shoot length are the least important variables. Criteria of correlation between each variable and discriminant scores was also applied for computing relative importance of individual characters of Indian mustard groups formed for classification and discrimination. Table 5 presents the correlation of various characters of Indian mustard with the discriminant function score. Secondary branches, primary branches, days to maturity, siliqua number on main shoot have higher correlation with the discriminant score and hence are considered as more important variables.

All the characters were considered in Random forests algorithm for computing relative importance of individual characters of Indian mustard groups for classification and discrimination. The mean decrease in Gini coefficient measures, how each variable contributes to the homogeneity (or purity) of the nodes and leaves in the resulting random forest. Higher the value of mean decrease in Gini index, better is the variable for prediction and hence greater impurity is removed from the model. The variables wise mean decrease in Gini index (Table 5). Application of Random forests indicated that if variable secondary branches is removed from the model then the mean decrease in Gini (or decrease in impurity) will be around 18.76. Therefore, number of secondary branches is the most important variables for discrimination purpose. The second most important variable after secondary branches is primary branches having Gini index values as 9.30 followed by siliqua number on main shoot and days to maturity. The variable which is least important for discrimination on the basis of the seed yield of mustard is days to flowering.

Table 5: Relative Importance of variables determined by different methods

Linear discriminant function coefficient			Correlation with discriminant score			Mean Decrease Gini		
Variables	Linear discriminant function coefficient	Ranks	Variables	Correlation with discriminant score	Ranks	Variables	Mean Decrease Gini	Ranks
DF	0.00	9	DF	0.09	7	DF	3.87	9
PB	0.20	4	PB	0.55	2	PB	9.30	2
SB	0.19	5	SB	0.76	1	SB	18.76	1
MSL	0.00	10	MSL	0.03	10	MSL	4.94	8
PH	0.00	8	PH	0.16	6	PH	5.45	7
SL	-0.48	2	SL	-0.07	8	SL	3.85	10
SNOMS	0.03	7	SNOMS	0.23	4	SNOMS	6.63	3
SPERS	0.23	3	SPERS	0.19	5	SPERS	5.60	6
DM	0.09	6	DM	0.43	3	DM	6.56	4
TSW	0.83	1	TSW	0.06	9	TSW	5.95	5

To measure the strength of relationship in ranking behaviour of the three methods (Linear discriminant function coefficient, Correlation with discriminant score and Mean Decrease Gini) Spearman rank-order correlation was used. Spearman Correlation Coefficients in Table 6 displayed the correlation and the t-value under the null hypothesis of zero correlation. It showed that correlation between Mean Decrease Gini and Correlation with discriminant score is 0.81 which is found significant with t-value as 3.85 at 5% level of significance. This indicates a strong positive relationship between the ranks of these two methods.

According to the ranks obtained under two methods (Correlation with discriminant score and Mean Decrease Gini) first four variables (Secondary branches, primary branches, days to maturity, siliqua number on main shoot) are same. From the ranks of both methods it is clearly indicated that secondary branches is most important variable. Second most important variable is primary branches followed by days to maturity and siliqua number on main shoot as third and fourth important variable.

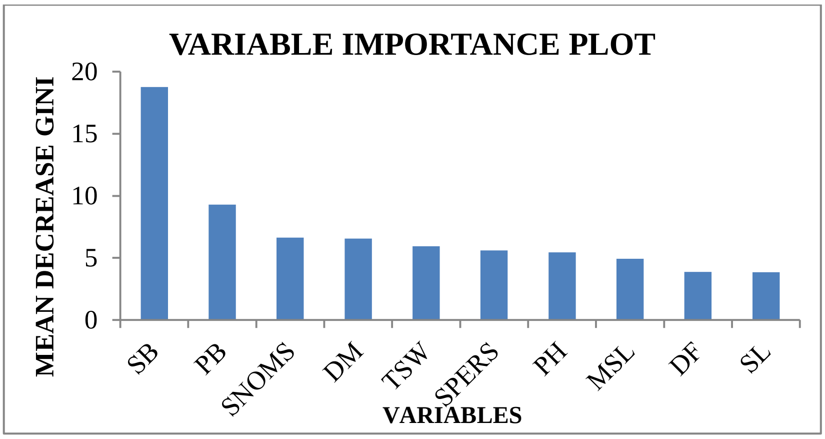
Table 6: Spearman Correlation Coefficients for the ranks of three methods

	Linear discriminant function coefficient	Correlation with discriminant score	Mean Decrease Gini
Linear discriminant function coefficient	1.00	0.09 (0.26)	0.19 (0.54)
Correlation with discriminant score	0.09 (0.26)	1.00	0.81* (3.85)
Mean Decrease Gini	0.19 (0.54)	0.81 (3.85)	1.00

* indicate significant t-value at 5% level of significance

Relative importance of variables using Gini index plot

Fig 1 depicts the variable importance by measuring the decrease in mean Gini. Variables are ranked and displayed in the Variable Importance Plot created for the Random Forest by this measure. It was observed that the secondary branch is a key classifier. Three variables in descending order of importance are primary branches, siliqua no on main shoot and days to maturity. However, siliqua length and days to flowering are the two least important variables.



(c)

Fig. 1: variable importance plot for the variable using mean decrease gini

CONCLUSION

Three methods; *viz.* univariate t-test, Rao´s additional information and Random forest were used for selection of variables for the purpose of classification and discrimination between low and high seed yield population of Indian mustard. The purpose of these methods was compared in term of classification error rates. Univariate t-test method was observed to be best with least error rate 21.72%. The optimum size of four using this method included the characters: No. of primary branches, No. of secondary branches, siliqua number on main shoot and days to maturity. Relative importance and ranking of variable was studied using magnitude of discriminant function coefficient, Mean Decrease Gini index and correlation with discriminant score. The Mean Decrease Gini index and correlation with discriminant score method provided similar pattern of ranking. The most important variable selected by these are number of primary branches, number of secondary branches, siliqua number on main shoot and days to maturity.

REFRENCES

Box, G. E. (1949). A general distribution theory for a class of likelihood criteria. Biometrika, *36*(3/4), 317-346.

Breiman, L. (1996). Bagging predictors, Machine Learning, *24*(2), 123–140.

Breiman, L. (2001). Random forests. Machine learning, *45*(1), 5-32.

Farver, T. B., & Dunn, O. J. (1979). Stepwise variable selection in classification problems. Biometrica, *21*, 145-153.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. Pattern Recognition Letters, *31*(14), 2225-2236.

Hastie, T., & Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification and regression. In Advances in Neural Information Processing Systems, 409-415.

Lachenbruch, P. A., & Mickey, M. A. (1968). Estimation of error rates in discriminant analysis, Technometric, *10,* 1-10.

McCabe, G. P. (1975). Computations for variable selection in discriminant analysis. Technometrics, *17*, 103-109.

Munita, C. S., Barroso, L. P., & Oliveira, P. M. S. (2006). Stopping rule for variable selection using stepwise discriminant analysis. *Journal of Radioanalytical and Nuclear Chemistry*, *269*(2), 335–338.

Painsky, A., & Rosset, S. (2016). Cross-validated variable selection in tree-based methods improves predictive performance. IEEE transactions on pattern analysis and machine intelligence, *39*(11), 2142-2153.

Rao, C. R. (1973). Linear statistical inference and its applications. Willey, New York.