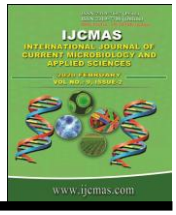
Int.J.Curr.Microbiol.App.Sci (2020) *9*(2): 798-807

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 9 Number 2 (2020)**
Journal homepage: http://www.ijcmas.com



Original Research Article https://doi.org/10.20546/ijcmas.2020.902.097

Feature Selection for Discrimination between Low and High Oil Content Genotypes of Indian Mustard

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

[1]Department of Mathematics & Statistics, CCS, Haryana Agricultural University Hisar-125004 (Haryana), India
[2]Department of Genetics and Plant breeding, CCS, Haryana Agricultural University Hisar-125004 (Haryana), India

**Corresponding author*

Keywords

Discriminant analysis, Error rates, Gini index, Random Forests, Wilk's Lambda

Article Info

Accepted:
08 January 2020
Available Online:
10 February 2020

A B S T R A C T

Variable selection in discriminant analysis may be used to identify those variables which are most relevant for use in allocating future observation. It is also expected to reduce the cost of experimentation and conditional error rate by increasing the ratio of the training sample size to the dimension. Thus, feature Selection has become important task in classification and discriminant analysis. Three variable selection methods (Univariate t-test, Wilk's lambda Criterion and Random Forests Algorithm) were used and compared in the present study for classification and discrimination to find important characters of Indian mustard. Secondary data set on 310 genotypes of Indian mustard recorded for 12 characters was used for discrimination between populations of low and high oil content genotypes of Indian mustard. Performance of the methods was assessed in terms of leave one out cross-validation error and out of bag error rate for classification. The important variables for discrimination which significantly affected the oil content were siliqua length, Secondary branches, primary branches and days to maturity with least error rate of 33.90 per cent.

Introduction

The application of variable selection methods in big datasets has been greatly increased in last few years. This is because most datasets have large number of samples of high dimensional variables. This makes it impractical, computationally expensive and causes reduction in classification accuracy when an entire input set is used. It is useful to identify and ignore the variables which simply complicate the analysis and do not provide any extra information. For developing better genotypes, 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. It becomes difficult to

analyze and interpret such data set. The identification of least important or poor variables that are rather unnecessary, irrelevant or even distracting is the afore-mentioned goal for the Plant breeders. These can be removed from the data set which can be achieved through variable selection. Variable selection helps in understanding data, reducing the curse of dimensionality and improving the prediction performance without incurring much loss of information. The selection of important variables for the purpose of discrimination between populations is also important from the point of view of cost of recording and processing a large set of variables.Variable selection methods are, thus employed in order to find most important or useful variables for various data mining tasks such as classification and discriminant.

Several methods to select variables that are subsequently used in discriminant analysis are proposed and analysed. McCabe (1975) proposed an algorithm for computing statistics for all possible subsets of variables for a discriminant function analysis. McLachlan (1976) developed a method for selecting variables for the linear discriminant function in case of two multivariate normal populations. McKay and Campbell (1982a) reviewed the variable-selection methods in discriminant function analysis. McKay and Campbell (1982b) addressed the problem where the aim was to allocate future entities of unknown origin to the groups. Rencher (1993) examined the effect of each variable on the Hotelling's T^2, Wilk´s Lambda and R^2 statistic and stated the contribution of each variable. McLachlan (2004) gave a detailed account of discriminant analysis and statistical pattern recognition. Breiman (2001) proposed random forests method, by adding an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. Munita *et al.,* (2006) proposed stopping rules for the identification of redundant variables, maintaining low probabilities of misclassification using sequence of standard F test. Genuer *et al.,* (2008) examined statistical method using random forests, for classification and regression problems. Han *et al.,* (2016) proposed a new method based on Random Forest to select variables using Mean Decrease Accuracy and Mean Decrease Gini. Chavent *et al.,* (2019) evaluated a novel methodology for dimension reduction and variable selection, which combines clustering of variables and feature selection using random forests. The present study has been designed to find important characters of Indian mustard which can discriminate between high and low oil content genotypes. For this purpose, three variable selection methods (Univariate t-test, Wilk's lambda Criterion 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 oil content genotypes on the basis of the following criterion: G_1 : Oil content < mean - standard deviation and G_2 : Oil content ≥ mean + standard deviation. Accordingly 44 genotypes were found to have low oil content and 74 genotypes with high oil content. Genotypes in G_1 were considered as samples from the low oil content populations of Indian mustard. The genotypes in G_2 were considered as samples from high oil content populations of Indian mustard.

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

$$H_0 : \sum^{(1)} = \sum^{(2)} = \cdots = \sum^{(k)}$$

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

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

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

Where

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

S is the pooled sample covariance matrix.
Test statistic is given by

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

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.

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:

Wilk´s Lambda criterion

This criterion was initially derived by Wilk´s (1932) applying the principle of likelihood ratio in some special cases and later extended by Bartlett (1947) for general use in multivariate analysis. The Wilk´s statistic Λ is the ratio of the within generalized dispersion to the total generalized dispersion. The within generalized dispersion is the determinant of the within-group sum of squares and cross-products matrix W and the total generalized dispersion is the determinant of the total sum of squares and cross-products matrix T (Johnson and Wichern, 2007).

Consider a data set with N observations measured on x_1, x_2, …x_p variables. These observations are shared by qualitative variables into K groups of sizes N_1, N_2, …,N_k. Let $X^{(i)}(N_i \times p)$ be the i^{th} data matrix [i=1,2, ….K] from $N(\mu^{(i)}, \sum^{(i)})$.The objective of the analysis is to test the significance of null hypothesis of equality of K groups means.

$$H_0 : \mu^{(1)} = \mu^{(2)} \ldots = \mu^{(K)}$$

Let us assume

$$\sum^{(1)} = \sum^{(2)} \ldots = \sum^{(K)}$$

For the set of X variables test criterion under homoscedastic condition is constituted by the ratio of within sum of cross-product matrix determinant and Total sum of cross product matrix determinant.

Wilk´s Lambda for Testing the Equality of Group Means

$$\Lambda = \frac{|W|}{|T|}$$

The corresponding Partial F is given by

$$F = \frac{1-\Lambda}{\Lambda} \cdot \frac{N-K-p+1}{K-1}$$

The partial F is distributed as F(K-1, N-K-p+1). Reject H_0 if $F_{cal} > F(K-1, N-K-p+1)$ at α level of significance.

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.

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 't', decreases in Gini impurity are recorded for all variables used to form the split. Gini impurity $\Delta gini(t)$ 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 Kis 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 Hastie and 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 and 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)

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

Results and Discussion

Box-M test was applied to test the homogeneity of group covariance matrices. According to the Box-M test, Chi-Sq (approx.) = 69.24, at df = 55, with p-value = 0.094. The decision failed to reject null hypothesis and we conclude that the low and high oil content groups have covariance homogeneity. Results of Table 1 gives the univariate measures such as mean, standard deviation and coefficient of variation (CV %) along with the t values for testing the significance of the difference in individual variable means, of ten characters for groups formed.

Results indicate that the variables siliqua length, primary branches, secondary branches, seeds per siliqua have non-significant but relatively large differences in the two group means and main shoot length, plant height, days to maturity are found to be the least discriminatory variables. Siliqua length is the most contributing variable for discrimination between low and high oil content groups followed by secondary branches, seeds per siliqua and primary branches. The Least discriminatory variable is observed as main shoot length.

Using Backward elimination procedure least discriminatory variables were sequentially eliminated (Table 2). The smallest value of Λ, 0.01 and largest value of F, 0.01 corresponds to the variable seeds per siliqua. The smallest value of Λ, 0.01 and largest value of F, corresponds to the variable seeds per siliqua. Hence seed per siliquais eliminated. At second step the smallest value of Λ, 0.04 and largest f value 0.089 corresponds to days to maturity. Hence days to maturity is eliminated. Continuing in the same way all the variables are discarded up to last step.

The Random forests method was applied to the low and high oil content sample, using the gini index to evaluate the importance of the predictors. Split ratio of 80:20 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 per cent of randomly selected samples were left out in modelling each bootstrap tree. The data set contained 118 observations, so the training and the test samples comprised 94 and 24 individuals respectively. To select the important variables the analysis starts by taking all ten variables. 94 observations were used to construct the random forests (training set) while the left 24 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 is 37.50 per cent. Table 3 represents variable selection using Random Forests.

At step 1, the variable days to flowering have lowest gini index (2.50). So, days to flowering are the first variable to be discarded. At step 2, out of remaining nine variables, days to maturity with lowest gini index 3.14 was deleted. Similarly, at third step primary branches was removed followed by secondary branches, main shoot length, siliqua number on main shoot and so on.

According to gini index siliqua length was observed to be the most important variable followed by plant height, seed per siliqua and siliqua number on main shoot for discrimination low and high oil contents population of Indian mustard.

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, Wilk´s lambda criteria (Backward), and Random Forest algorithm for the Indian mustard data. Initially univariate t-test was used to discriminate the two groups having low oil content and high oil content. Among all the methods, the highest accuracy of the classification was obtained by Wilk´s lambda (Backward) with value of 66.10 percent. Wilk´s lambda (backward) obtains 33.90 per cent error rate by selecting 4 variables: secondary branches, siliqua length, plant height and thousand seed weight. This clearly shows that Wilk´s lambda (backward) consistently find a significantly better subset of variables from the candidate variable set than from the t-test and Random Forest algorithm candidate feature sets.

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	3.18	4.33	6.68	9.03	47.57	48.01	-0.59
PB	1.3	1.28	25.46	24.04	5.1	5.31	-0.85
SB	4.54	4.4	27.7	25.1	16.4	17.54	-1.34
MSL	11.95	11.38	14.89	14.13	80.25	80.54	-0.13
PH	22.62	18.35	10.59	8.64	213.48	212.38	0.29
SL	0.45	0.41	12.51	12.21	3.58	3.37	2.67
SNOMS	11.02	8.66	20.97	16.03	52.55	54.01	-0.80
SPERS	1.75	1.8	13.55	14.35	12.9	12.58	0.96
DM	3.03	3.29	2.07	2.24	146.34	146.62	-0.46
TSW	1.21	0.97	31.77	26.6	3.8	3.66	0.68

Table.2 Variable selection using Wilk´s lambda criterion

Step	Variable Selected	Variable Discarded (p-q)	subset size(q)	Wilks Lambda (Λ)	F Value	F critical
1	DF,PB,SB,MSL,PH,SL,SNOMS,DM,TSW	SPERS	9	0.89	0.01	1.97
2	DF,PB,SB,MSL,PH,SL,SNOMS,TSW	DM	8	0.89	0.04	2.02
3	DF,PB,SB,MSL,PH,SL,TSW	SNOMS	7	0.89	0.11	2.09
4	DF,SB,MSL,PH,SL,TSW	PB	6	0.90	0.66	2.18
5	DF,SB,PH,SL,TSW	MSL	5	0.91	1.11	2.30
6	SB,PH,SL,TSW	DF	4	0.92	1.04	2.45
7	SB,SL,TSW	PH	3	0.92	0.71	2.68
8	SL,TSW	SB	2	0.93	1.32	3.08
9	SL	TSW	1	0.94	1.02	3.92

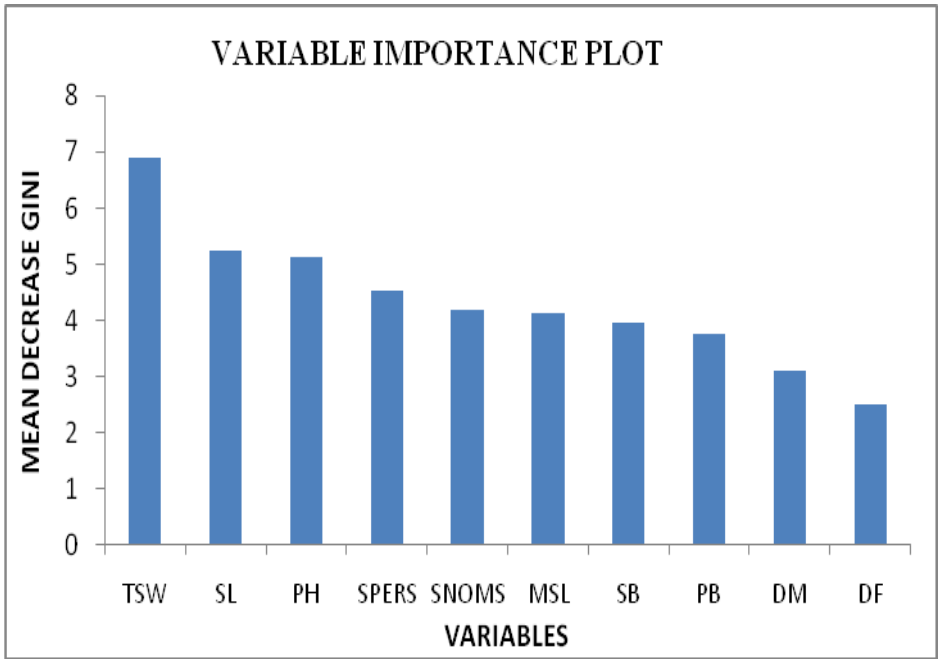
Table.3 Variable selection using random forest algorithm

Step	Variable Selected	Variable discarded	subset size	Mean decrease Gini of discarded variable	oob error rate
1	PB,SB,MSL,PH,SL,SNOMS,SPERS,DM,TSW	DF	10	2.50	37.50%
2	PB,SB,MSL,PH,SL,SNOMS,SPERS,DM,TSW	DM	9	3.14	41.67%
3	PB,SB,MSL,PH,SL,SNOMS,SPERS,TSW	PB	8	4.69	54.17%
4	SB,MSL,PH,SL,SNOMS,SPERS,TSW	SB	7	5.20	58.33%
5	MSL,PH,SL,SNOMS,SPERS,TSW	MSL	6	6.06	58.33%
6	PH,SL,SNOMS,SPERS,TSW	SNOMS	5	7.10	54.17%
7	PH,SL,SPERS,TSW	SPERS	4	9.48	54.17%
8	PH,SL,TSW	PH	3	14.08	54.17%
9	SL,TSW	SL	2	12.16	54.17%
10	TSW		1		54.17%

Table.4 Overall comparison of variable selection methods

Methods	Variables Selected	Error Rate (%)
t-value	SL, PB, SB, SPERS	35.59
Wilk´s lambda (Backward)	SL, TSW, SB, PH	33.90
Random Forests	PH, SL, SPERS, TSW	54.17

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



Relative importance of variables using Gini index plot

Figure 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. The most important variable is one with the largest Mean Decrease in Gini value because a higher mean decrease in Gini will imply a higher importance. It was observed that the thousand seed weight is the most important variable and can be considered as a key classifier. Three variables in descending order of importance are siliqua length, plant height and seeds per siliqua. However, days to maturity and days to flowering are the two least important variables for discrimination of low and high oil content populations.

It is concluded that, three methods; viz. univariate t-test, Wilk's lambda and Random forest were used for selection of variables for the purpose of classification and discrimination between low and high oil content population of Indian mustard. The purpose of these methods was compared in term of classification error rates.t-value based method obtains error rate of 35.59 percent. The variables selected from t-test are: primary branches, secondary branches, siliqua length, seeds per siliqua. Random forests provided maximum error rate of 54.17 per cent for selecting siliqua length, thousand seeds weight, plant height, seeds per siliqua. Wilk's lambda criterion method was observed to be best with least error rate 33.90%. The optimum size of four using this method included the characters: Siliqua length, thousand seed weight, secondary branches and plant height.

References

Bartlett, M. S. (1947). The use of transformations. Biometrics, 3(1), 39-52.

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

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

Chavent, M., Genuer, R., and Saracco, J. (2019).Combining clustering of variables and feature selection using random forests. Communications in Statistics-Simulation and Computation, 1-20.

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

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

Han, H., Guo, X., and Yu, H. (2016).Variable selection using mean decrease accuracy and mean decrease gini based on random forest. IEEE International Conference on Software Engineering and Service Science (ICSESS), 219-224.

Johnson, R. A., and Wichern, D. W. (2007). Applied multivariate statistical analysis. PrenticeHall International.INC., New Jersey.

Lachenbruch, P. A., and 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.

McLachlan, G. J. (1976). A criterion for selecting variables for the linear discriminant function. Biometrics, 32, 529–534.

McLachlan, G. (2004). Discriminant analysis and statistical pattern recognition. John Wiley and Sons.

McKay, R.J., and Campbell, N.A. (1982a).

Variable selection techniques in discriminant analysis, I. Description. British Journal of Mathematical and Statistical Psychology, 35(1), 1-29.

McKay, R.J., and Campbell, N.A. (1982b). Variable selection techniques in discriminant analysis, II. Allocation. British Journal of Mathematical and Statistical Psychology, 35(2), 30-41.

Rencher, A. C. (1993). The contribution of individual variables to Hotelling's T^2, Wilks' Λ, and R^2. Biometrics, 479-489.

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

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. Biometrika, 471-494.

How to cite this article:

Poonam Godara, B. K. Hooda and Ram Avtar. 2020. Feature Selection for Discrimination between Low and High Oil Content Genotypes of Indian Mustard. *Int.J.Curr.Microbiol.App.Sci.* 9(02): 798-807. doi: https://doi.org/10.20546/ijcmas.2020.902.097