C. 2020. Bioethanol production from oil palm empty fruit bunch with SSF and SHF processes using Kluyveromyces marxianus yeast. Cellulose, 27, 301-314.

Sun, F. and Chen, H. 2007. Evaluation of enzymatic hydrolysis of wheat straw pretreated by atmospheric glycerol autocatalysis, Journal of chemical technology and biotechnology, 82: 1039-1044.

Vaez, S., Karimi, K., Mirmohamadsadeghi, S. and Jeihanipour, A. 2021. An optimal biorefinery development for pectin and biofuels production from orange wastes without enzyme consumption. Process Safety and Environmental Protection, 152, 513-526.

Wang, Z., Lv, J., Gu, F., Yang, J. and Guo, J. 2020. Environmental and economic performance of an integrated municipal solid waste treatment: A Chinese case study. Science of the Total Environment, 709, 136096.

Yusuf, A. A. and Inambao, F. L. 2021. Effect of low bioethanol fraction on emissions, performance, and combustion behavior in a modernized electronic fuel injection engine. Biomass Conversion and Biorefinery, 11(3), 885-893.

Zhou, J., Wang, Y. H., Chu, J., Zhuang, Y. P., Zhang, S. L. and Yin, P. 2008. Identification of the main components of cellulases from a mutant strain of Trichoderma viride T 100-14. Bioresour Technol., 99: 6826-33.

Zhu, J. Y., and Pan, X. J. 2010. Woody biomass pretreatment for cellulosic ethanol production: technology and energy consumption evaluation. Bioresource Technology, 101, 4992-5002.

J. Agric. Res. Technol., Special Issue (1) : 100-105 (2022) DOI: https://doi.org/10.56228/JART.2022.SP116

Importance of Variable using Gini Index and Discriminant Score in Indian Mustard Genotypes

Poonam Godara[1*], Suman Verma[2], Sachin Kumari[3], Shrawan Kumar[4]
Department of Mathematics, GJUS&T, Hisar, Haryana, India
Email : poonamsinghsinghmar@gmail.com

Abstract

In any crop improvement programme, for developing better genotypes the choice of suitable parents is a matter of great concern to the plant breeders. Therefore, it is imperative to know the role of important characters in the selection experiments aimed to increase seed yield. The present study has been conducted using Linear Discriminant function coefficient, Correlation between variables and discriminant scores and mean decrease in Gini index were used for computing relative importance of individual characters of Indian mustard discrimination between low and high oil content population of Indian mustard. Using the correlation between variables and discriminant score, the most important variables affecting the seed yield were secondary branches, primary branches and days to maturity. The three most important variables discriminating between oil content were siliqua length, secondary branches and seeds per siliqua.

Key words :

India is the largest producer of oilseed in the world covering approximately 12 per cent of total cropped area of the country. Over the past few decades, breeding programs, in India, have followed pure line breeding methods for development of new varieties primarily through the exploitation of genetic variability that existed among the adapted pool of elite germplasms. It

1. *Dept. of Mathematics, 2. Faculty of Agricultural Sciences, SGT University, Gurugram, Haryana, India, 3. Dept. of Genetics and Plant Breeding, CCSHAU, Hisar and 4. Dept. of Statistics, Kirori Mal College, University of Delhi

has resulted in only marginal improvement in productivity. To realize further gains in productivity, it is important to utilize new sources of variation which would lead to broadening the genetic base of the existing varieties. Productivity can also be substantially increased by heterosis breeding. To boost up further productivity of Indian mustard hybridization and exploitation of heterosis may play significant role in coming years. 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. 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. It is useful to identify and ignore the variables which simply complicate the analysis and do not provide any extra information. Variable selection is, 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. 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. Keeping in view the above points in mind, the present study was planned with the objectives to study relative importance of Indian mustard characters for classification and discrimination of genotypes. The 310 genotypes were divided into two Groups according to the low and high yielding genotypes on the basis of seed yield and oil content.

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.

Following methods were used to study the relative importance of the variables of the Indian Mustard.

(i) Magnitude of Linear Discriminant function coefficient

(ii) Correlation between variables and discriminant scores

(iii) Variable importance using mean decrease in Gini index

Measures to identify the relative importance of variables that discriminate between two independent groups are based on discriminant analysis. It quantifies the relative importance of a variable based on its contribution to grouping effects and discriminant function scores (Huberty and Wisenbaker, 1992; Thomas, 1997). Discriminant Analysis (DA) measures variable importance and identify one or more linear combinations of the variables that maximize group separation. It is based on functions of the discriminant function coefficients and includes standardized discriminant function coefficients (Huberty and Wisenbaker, 1992; Thomas, 1992). The standardized discriminant function coefficient is one commonly adopted variable importance measure. Another method to find the importance of the discriminators is to find the correlation between discriminant function and discriminatory variables (Huberty, 1971).

Discriminant analysis (DA) is a multivariate statistical technique concerned with separating distinct sets of objects or observations and with allocating new objects to previously defined groups. Discrimination terminology was introduced by Fisher in the first discriminatory problems (Fisher 1936). Linear discriminant analysis is most widely used statistical methods for classification problems. The discriminant function is the linear combination of these p variables that maximizes the distance between two group mean vectors. A linear combination

$$D = a'X = a_1X_1 + a_2X_2 + \ldots + a_pX_p,$$

where $X' = (X_1, X_2, \ldots, X_p)$ transforms each observation vector to a scalar. The discriminant function coefficient vector is estimated by

$$\hat{a} = S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)})$$

where $\bar{x}^{(1)}$, $\bar{x}^{(2)}$ and S^{-1} denote the sample mean vectors and pooled sample covariance matrix computed from the samples from two multivariate normal populations $Np (\mu^{(1)}, \Sigma)$ and $Np (\mu^{(2)}, \Sigma)$. The variable importance measures based on discriminant function coefficients can be used to rank variables according to their contributions to group separation (Huberty and Wisenbaker, 1992).

Another approach to identify the relative importance of the variable is to examining correlation between each outcome variable and each linear discriminant function. For each object the value of discriminant function is computed, and then Pearson correlation coefficient between it and the original variables are found. These correlations are often called structure or loading. Only within group correlations are considered. The variables that correlate highly with linear discriminant function are considered important. For two groups there is single linear discriminant function. To access the relative contribution of the variable, the rank ordering maybe reasonable assessment approach. In this approach we find the correlation discriminant score and discriminatory variables (Huberty & Wisenbaker, 1992). The correlation coefficient of all the values of D and X_j (j=1, 2, …, p) is used to measure the contribution of j^{th} variable in discriminating the groups. The most contributing variable is one for which the above-mentioned correlation coefficient is maximum.

Results and Discussion

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 objective of discriminant analysis was to develop discriminant functions which are 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 also evaluates 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 under scheme: low (seed yield and oil content) and high (seed yield and oil content) are presented in Table 1.

Based on the coefficients of the discriminant function, variables thousand seeds weight, siliqua length, seed per siliqua and primary branches are observed as the most important variables. These variables contribute more to the discriminant score for discriminating between the groups partitioned on the basis of low seed yield with low oil content and high seed yield with high oil content. Variables plant height, days to flowering and main shoot length are the least contributing variables in the discrimination of groups as shown in Table 1.

The variables secondary branches, primary branches, siliqua number on main shoot and seeds per siliqua are considered as important variables. Clearly these variables contribute more towards the discrimination of the groups formed. Variables 1000 seed weight, days to flowering and main shoot length are least important variables for discrimination.

The results of Table 1 describe that secondary branches is found to be the most important variable. It means if this variable is removed from the model then the mean decrease in gini (or decrease in impurity) will be around 8.06. Higher the value of mean decrease gini index better is the variable for prediction. The second most important variable after secondary branches is siliqua number on main shoot having gini index values as 5.80 followed by primary branches and seeds per siliqua are. The least important variable is siliqua length.

In Table 2 the Spearman correlation and the t-value under the null hypothesis of zero correlation were presented. It showed that a correlation of 0.71 between Mean Decrease

Table 1. 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**	**Vari-ables**	**Mean decr-ease gini**	**Ranks**
DF	-0.03	9	DF	-0.06	9	DF	2.61	9
PB	0.29	4	PB	0.47	2	PB	4.67	3
SB	0.13	5	SB	0.61	1	SB	8.06	1
MSL	-0.03	8	MSL	0.00	10	MSL	3.14	7
PH	0.00	10	PH	0.11	7	PH	4.00	5
SL	-0.79	2	SL	-0.22	5	SL	2.35	10
SNOMS	0.08	6	SNOMS	0.30	3	SNOMS	5.80	2
SPERS	0.36	3	SPERS	0.22	4	SPERS	4.58	4
DM	0.05	7	DM	0.19	6	DM	2.67	8
TSW	0.96	1	TSW	0.1	8	TSW	3.47	6

Table 2. 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	0.38 (1.17)	0.09 (0.26)
Correlation with discriminant score	0.38 (1.17)	1.00	0.71* (2.84)
Mean Decrease Gini	0.09 (0.26)	0.71 (2.84)	1.00

Gini and Correlation with discriminant score, which is significant with t-value 2.84 at 5% level of significance. This indicates the ranks of these two methods are in agreement with each other.

According to the ranks obtained under two methods (Correlation with discriminant score and Mean Decrease Gini) first four variables (Secondary branches, primary branches, seeds per siliqua, siliqua number on main shoot) are same. Ranks showed that variable secondary branches is most important variable, primary branches and siliqua number on main shoot as second and third important variable whereas seeds per siliqua was on fourth rank.

Conclusion

Three methods *viz*., Linear Discriminant function coefficient, Correlation between variables and discriminant scores and mean decrease in Gini index were used for computing relative importance of individual characters of Indian mustard discrimination between low and high oil content population of Indian mustard. Using the correlation between variables and discriminant score, the most important variables affecting the seed yield were secondary branches, primary branches and days to maturity. The three most important variables discriminating between oil content were siliqua length, secondary branches and seeds per siliqua. Most important variables discriminating between low seed yield with low oil content and high seed yield with high oil content groups were secondary branches, primary branches and siliqua number of main shoot. The variable, number of secondary branches have been found to be the most important for classification and discrimination of Indian mustard genotypes for seed yield and oil content.

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.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. 1984. Classification and Regression Trees. Wadsworth Int., 37(15), 237-251.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179-188.

Fujikoshi, Y. 1983. A criterion for variable selection in multiple discriminant analysis. Hiroshima Mathematical Journal, 13, 203-214.

Fujikoshi, Y. 1985. Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. Journal of Multivariate Analysis, 17, 27-37. McCabe, G.P. (1975). Computations for variable selection in discriminant analysis. Technometrics, 17, 103-109.

McKay, R. J. 1976. Simultaneous procedures in discriminant analysis involving two groups. Technometrics, 18, 47-53.

McKay, R. J. 1978. A graphical aid to selection of variables in two-group discriminant analysis. Applied Statistics, 27, 259-263.

Rao, C. R. 1948. The utilization of multiple measurements