International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Obtaining Minimum Error of Classification with Modified Fisher's Cutoff Point

Ogbonna, Eric Nnamdi[1], Abam Ayeni Omini[2], Nsien, Edwin Frank[3]

[1]Department of Statistics, School of Science and Industrial Technology, Abia State Polytechnic, Aba Abia State Nigeria

[2]Department of Mathematics, Faculty of Science, Federal University Lafia Nigeria

[3]Department of Statistics, Faculty of Science University of Uyo, Akwa Ibom State Nigeria

Abstract: ***Fisher's method of adjusting the cutoff point to obtain a balance on the total probability of misclassification are disadvantaged in the sense that they inflate the classification error in the presence of outliers. The proposed method of Modified Fisher's Cutoff Point (MFCP) is aimed at addressing this problem. To accomplish this purpose, linear discriminant analysis was performed using Fisher's technique and the modified approach. Misclassification errors were computed for both methods and the modified procedure was observed to have minimum error than the Fisher's procedure. The results obtained in applying the cutoff point showed that the modified approach performs better than the Fisher's linear discriminant cut off point. The application of this new method reduces the effect of outliers to the barest minimum.***

Keywords: Cutoff point, outliers, misclassification, apper, and discriminant

1. Introduction

Fisher's linear discriminant analysis is a conventional multivariate technique for dimension reduction and classification (Pohar etal, 2004 & Sugiyama, M. 2007;). Discriminant analysis deals with the problems of discrimination and classification (Johnson and Wichern, 2007) Fisher's discriminant analysis is concerned with the problem of classifying an object of unknown origin into one or more distinct groups or population on the basis of observations made on it. Hawkins (1982). These observations form the training sample which is used to construct a discriminant rule for the allocation of new individual objects into one of the groups. The basic objective of discriminant analysis is to classify and predict problems when the dependent variables are in numerical form, Alvin (2002).

An automobile engineer may classify an auto mobile engine into grade I, grade II, or grade III on the basis of measurements of its output, shape, size and shape. Nutritionist may classify food items into carbohydrate, protein, minerals, fat and oil based on measurements observed about the food composition. These examples illustrate range of problems that can be solved through the use of discriminant analysis.

Fisher (1936) proposed the transformation of multivariate x to univariate y such that the y observations from π_1 and π_2 has maximum separation and that the mean difference determines this separation. His work was based on the assumption of equal covariance matrices. Fisher's technique is determined by the optimal transformation of minimizing the within and between class separation.

The proposed method of Modifying Fisher's cutoff point has the advantage of reducing the classification error or the probability of misclassification. Theorems that will be developed will be used to solve the proposed method.

We are therefore proposing to develop a robust Fisher's linear discriminant function by modifying the Fisher's cutoff point using the moving average method. In carring out this study we shall consider only two groups discriminant analysis. The emphasis is on Fisher's two group discrimination.

2. Linear Discriminant Function

Discriminant analysis is performed by making the weight of each variable to maximize the between and within group variance. The linear discriminant function is given by

$$D = B + C_1X_1 + C_2X_2 + ... + C_jX_j \qquad 1$$

D is the discriminant score
B is the discriminant constant
C_j, $j = 1,2,...,k$ is the discriminant weight or coefficient,
$X_i, i = 1,2,...,k$ is the independent variable or predictor

Chen and Muirhead (1994) proposed two methods- the first is a projection pursuit index on the Fisher's discriminant ratio of between class variation and within class variation. The second is the total probability of misclassification. Robust linear discriminant functions were constructed by applying projection pursuit optimization algorithm and the rank cut off point for robust location estimates. Randles, Brofitt, Ramberg and Hogg (1978) modified the work by improving the balance between two classification rates for linear discriminant and quadratic functions. Chen (1989) applied the cutoff point which minimizes the error rate in classifying the training samples. See Anderson (1984) and Gnadadesikan (1988) for other procedures of cutoff points.

3. Methodology

The major aim of discriminant analysis is to distinguish between two known populations, group G_1 and G_2. The purpose is to create a classification rule which can be used

for classifying individual units that belong to any of the two groups based on specified characteristics.

We assume that the we have n x p observations of data matrix from G groups and the data matrix X contains the observed values x_{ij} of the jth variable of the ith individual, $i = 1,2,.....,n$, $j = 1,2,....,p$. It also has a variable $g_i \in (1,2,.....,G)$, whose value indicates the group the observation belongs to.

$$X = \begin{pmatrix} g_1 & x_{11} & x_{12} & . & x_{1p} \\ g_2 & x_{21} & x_{22} & . & x_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ g_n & x_{n1} & x_{n2} & . & x_{np} \end{pmatrix} \quad 2$$

Here we have n individual items and on each item we have measurements on p variables. Each individual item belongs to only one group and many individuals will be allocated to a specified group.

Let n_g be the number of individual items that belong to group g.
g =1,2...,G.
Let x_{gi} be the ith individual item in group g, i=1, 2,… n_g.
Let G be the number of group and each group has m-dimensional samples.

4. Determination of Fisher's cutoff point

Let $(\bar{x}_1 - \bar{x}_2)^2$ and $(\bar{y}_1 - \bar{y}_2)^2$ be the separation of the two populations which represents the distance that measures the variation between means of the discriminant scores. We find a vector a that maximizes the standardized difference $\frac{(\bar{y}_1 - \bar{y}_2)}{s_y}$

$\frac{\bar{y}_1 - \bar{y}_2}{s_y}$ measures the difference between the transformed means $\bar{y}_1 - \bar{y}_2$ with respect to the sample standard deviation s_y
The separation of the projected y is measured by the ratio of the squared difference.

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(a'\bar{x}_1 - a'\bar{x}_2)^2}{a'Sa} = \frac{[a'(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)'a]}{a'Sa} \quad 3$$

and

$$D^2 = (\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2) \quad 4$$

Is the maximum of this ratio and is known as the Mahalanobi's squared distance. The vector of coefficient a that maximizes the standardized difference and the squared distance is the ratio $\frac{[a'(\bar{x}_1 - \bar{x}_2)]^2}{a'Sa}$ (refer to eq.70) and the separation is maximized for $a' = (\bar{x}_1 - \bar{x}_2)'S^{-1}$ The linear combination

$$y = a'x = (\bar{x}_1 - \bar{x}_2)S^{-1}x \quad 5$$

maximizes the ratio.

We proceed and apply the allocation rule as follows;
Allocate

$$x_o \text{ to } \pi_1 \text{ if } y_o = (\bar{x}_1 - \bar{x}_2)S_p^{-1}x_o \geq \hat{m} \quad 6$$

Allocate x_o to π_2 if $y_o = (\bar{x}_1 - \bar{x}_2)S_p^{-1}x_o < \hat{m}$ 7

Where $m = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)'S_p^{-1}(\bar{x}_1 + \bar{x}_2)$ 8

This rule is known as Fisher's linear discriminant function and m is the CUTOFF POINT. That is

$$m = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \quad 9$$

$$\bar{y}_1 = (\bar{x}_1 - \bar{x}_2)'S_p^{-1}\bar{x}_1 = \hat{a}'\bar{x}_1 \quad 10$$

$$\bar{y}_2 = (\bar{x}_1 - \bar{x}_2)'S_p^{-1}\bar{x}_2 = \hat{a}'\bar{x}_2 \quad 11$$

$$. Y_c = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \quad 12$$

Y_c

The cutoff point is the projection of the midpoint between the two population sample means into the same subspace. The classification rule based on the cutoff point is
Assign an observation x_o to π_1 if $Y \geq Y_c$
Assign an observation x_o to π_2 if $Y < Y_c$
This is called the linear discriminant analysis. (Fisher, 1936,).

Graphical display of separation by discriminant analysis.

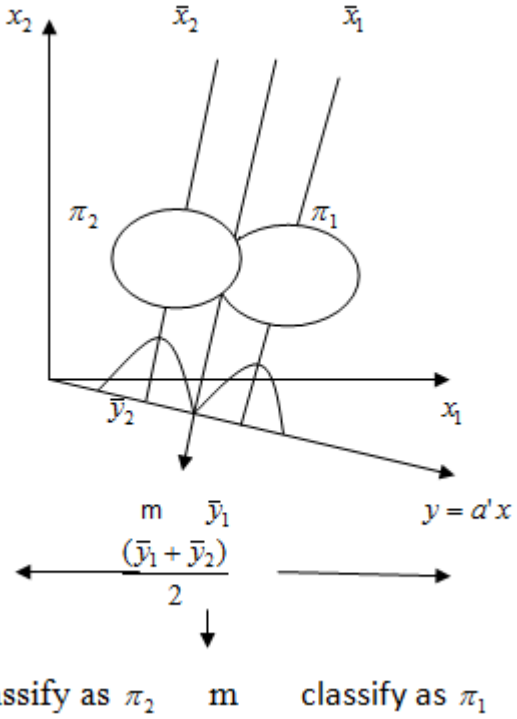

Figure 1: Two group discriminant analysis

The samples are projected very close to each other but the means are far apart from each other.

The Apparent error rate (APER) will be evaluated using the confusion matrix as presented below. It indicates the number of correct and incorrect classified individuals in the data set.

Table 1: Confusion matrix table

		Predicted membership π_1	Predicted membership π_2	
Actual Membership	π_1	n_1c	n_2m	n_1
	π_2	n_2m	n_2c	n_2

Where

n_1c is the number of individuals from π_1 correctly classified as π_1

n_2c is the number of individuals from π_2 correctly classified as π_2

n_1m is the number of individuals from π_1 misclassified as π_2

n_2m is the number of individuals from π_2 misclassified as π_1

The apparent error rate (APER) is given as

$$\text{APER} = \frac{n_1m + n_2m}{n_1 + n_2} \quad 14$$

The hit ratio (HR) is given as

$$HR = \frac{n_1c + n_2c}{n_1 + n_2} \quad 15$$

$$= \hat{p}(2|1) = \frac{n_{1m}}{n_1} \quad \hat{p}(1|2) = \frac{n_{2m}}{n_2} \quad \hat{p}_1 = \frac{n_1}{n_1 + n_2}$$

$$\hat{p}_2 = \frac{n_2}{n_1 + n_2} \quad 16$$

The confusion matrix and APER is to justify how good or bad the rule is. The APER is an estimate of the probability that a classification procedure based on a given data will misclassify a future observation.

5. Modifying the Fisher's cutoff point

Theorem 1.

If $x_1, x_2, x_3, ..., x_n$ is a set of observations and

$y_{(1)} \le y_{(2)} \le ... \le y_{(n)}$ is the ordered n observations

Where

$$y_{(1)} = \min[x_1, x_2, ..., x_n] \text{ and } y_{(n)} = \max[x_1, x_2, ..., x_n] \quad 17$$

If $y_{(1)}, y_{(2)}, ..., y_{(n)}$ belong to two regions R_1 and R_2 where $R_1, R_2 \in R_\Omega$ and R_Ω is the sample space, then the cutoff point is the midpoint of the moving average of R_Ω.

Proof.

Let $x_1, x_2, ..., x_n \in R_1$ and R_2

And $R_\Omega = (y_{(1)}, y_{(2)}, ..., y_{(n)}) : (y_{(1)} \le y_{(2)} ... \le y_{(n)})$ 18

Then $\bar{R}_{\Omega MA} = \frac{y_{(1)} + y_{(2)}}{2}, \frac{y_{(2)} + y_{(3)}}{2}, ..., \frac{y_{(n)} + y_{(n+1)}}{2}$ 19

Hence, the nth term is

$$T_n = \frac{y_{(n)} + y_{(n+1)}}{2} \quad 20$$

Therefore, the midpoint is

$$\frac{T_n}{2} = \frac{y_{(n)} + y_{(n+1)}}{2} \quad /2 \quad 21$$

$$\Rightarrow \bar{Z}_C = \frac{T_n}{2} = \frac{y_{(n)/2} + y_{(n+1)/2}}{2} \quad 22$$

The cutoff point is

$$\bar{Z}_C = \frac{y_{(n)/2} + y_{(n+1)/2}}{2} \quad 23$$

6. Problem on FLDF

This problem on Fisher's linear discriminant function was chosen from one of the text books. Compute the Fisher's linear discriminant function for two data set.

Samples for n_1:

$X_1(x_1, x_2) = (6,7), (7,5), (9,10), (8,8), (8,9), (10,9)$

Samples for $n_2 : X_2(x_1, x_2) = (13,11), (15,16), (22,20), (11,16), (12,11), (13,14)$

The matrix form of the data sets is

$$X_1 = \begin{pmatrix} 6 & 7 \\ 7 & 5 \\ 9 & 10 \\ 8 & 8 \\ 8 & 9 \\ 10 & 9 \end{pmatrix} \quad X_2 = \begin{pmatrix} 11 & 13 \\ 15 & 16 \\ 22 & 20 \\ 11 & 16 \\ 12 & 11 \\ 13 & 14 \end{pmatrix}$$

Application of Fisher's cutoff point

The mean class for each group

$$\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X = \begin{pmatrix} 8 \\ 8 \end{pmatrix}$$

$$\mu_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X == \begin{pmatrix} 15 \\ 15 \end{pmatrix}$$

The covariance matrix for n_1 sample

$$S_1 = \sum_{i=1}^{6}(x-\mu_1)(x-\mu_1)' = \begin{pmatrix} 10 & 9 \\ 9 & 16 \end{pmatrix}$$

The covariance matrix for n_2 samples

$$S_2 = \sum_{i=1}^{6}(x_i-\mu_2)(x_i-\mu_2)' = \begin{pmatrix} 82 & 59 \\ 59 & 48 \end{pmatrix}$$

The within class scatter matrix

$$S_p = \frac{S_1+S_2}{n_1+n_2-2} = \begin{pmatrix} 9.2 & 6.8 \\ 6.8 & 6.4 \end{pmatrix}$$

$$S_p^{-1} = \begin{pmatrix} 0.5063 & -0.5379 \\ -0.5379 & 0.7278 \end{pmatrix}$$

The optimal direction 'a'

$$a = S_p^{-1}(\mu_1-\mu_2) = \begin{pmatrix} 0.2212 \\ -1.3293 \end{pmatrix}$$

Hence, the discriminant function is

$$Y = a'x = 0.2212x_1 - 1.3293x_2$$

Computation of Fisher's CUTOFF point

$$\overline{Y}_1 = a'\overline{x}_1 = -8.8648$$

$$\overline{Y}_2 = a'\overline{x}_2 = -16.6215$$

The Fisher's CUTOFF point is

$$Y_c = \frac{\overline{Y}_1+\overline{Y}_2}{2} = -12.74315$$

The classification rule is

Assign the variable x to π_1 if $Y > Y_c$

Assign the variable x to π_2 if $Y \le Y_c$

The discriminant scores for π_1

$Y_1 = -7.9779, -5.0981, -11.3022, -8.8648, -0.1941, -9.7517.$

$Y_2 = -14.8477, -17.9508, -21.7196, -18.8356, -11.9679, -15.7356.$

The discriminant scores for π_2

Results of FCP

The confusion matrix table constructed resulted to the following classification errors

$p(2/1) = 0,\ p(1/2) = 0.2$

$Aper = 0.083$

$HR = 91.7\%$

Problem on MFCP

Given the same Fisher's dataset above,Perform the linear discriminant analysis using the MFLDF.

7. Manual application MFCP analysis for equal sample size

The discriminant scores for π_1 and π_2 above are combined and arranged in ascending order.

-21.7196,-18.8356,-17.9508,-15.7356,-14.8477,-11.9679,-11.3022,-10.1941, -9.7517, -8.8648, -7.9779, -5.0981

The moving average is obtained as follows

-20.2776 -18.3932 -17.8432 -15.2917, - 13.0750, - 11.6351 -10.7482 -9.9729 -9.3083 -8.4214 -6.5380

The cutoff point is given as

$$z_c = \frac{x_{n/2} + x_{n+1/2}}{2}$$

Given that n=12,

$$Z_c = \frac{x_{12/2} + x_{12+1/2}}{2}$$

$$Z_c = \frac{x_6 + x_7}{2} = \frac{-11.9679 + (-11.3022)}{2} = -11.6351$$

The modified cut off point, $Z_c = -11.6351$

The allocation procedure becomes

Assign the variable x to π_1 if $Y > Z_c$

Assign the variable x to π_2 if $Y \le Z_c$

8. Results of MFCP

The confusion matrix table constructed resulted to the following classification errors.

$P(2/1) = 0, P(1/2) = 0$

$Aper = 0$

$HR = 100\%$

9. Summary

A theorem was developed which was used to solve problems on Fisher's Linear discriminant analysis and the modified procedure.

It was discovered that the modified solution yielded minimum error of classification than that of Fisher's technique as observed in table 2. The apparent error rate for Fisher is 0.083 while the modified is 0. The MFCP showed a 100% correct classification while that of FCP showed 91.7% correct classification. The results of the experiment

conducted suggested a comparable classification procedure to the Fisher's linear discriminant function.

10. Conclusion

A modified cutoff point was developed to solve problems in discriminant analysis. This proposed procedure reduces the effect of outliers and yields minimum error of classification when compared with the Fisher's discriminant function.

References

[1] Anderson, T.W. (1984). An introduction to Multivariate Statistical Analysis, 2nd edition Wiley, New York.
[2] Chen, Z. & Muirhead, R. (1994). A comparison of robust discriminant procedures using projection pursuit methods.
[3] Chen, Z.Y. (1989)robust discriminant procedures using projection pursuit Methods. Ph.D Desertation, Department of Statistics. University of Michigan, Ann Arbor.
[4] Fisher, R., (1936). The use of multiple measurements in taxonomic problems, Ann. Eugenics 7, 179 – 188.
[5] Gnanadesikan, R. (1988). Discriminant analysis and clustering. Board of Mathematical sciences, National Academy Press.
[6] Hawkins, D. M.(1982). Topics in Applied Multivariate Analysis. Cambridge University Press, first edition new York.
[7] Johnson R. A. & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis 6th edition, Pearson Hall, upper Saddle River, New jersey 07458.
[8] Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: A simulation study. Metodoloski zvezki, 1, 143 – 161.
[9] Randles, R.H.,Brofitt, J.D., Ranberg, J.S. & Hogg, R.V. (1978a). Discriminant Analysis based on ranks. Journal of American Statistic assoc., 73, 379 – 384
[10] Rencher, A.C.(2002). Methods of Multvariate Analysis. 2nd ed., John Wiley and sons, inc. New York.