Journal of Mathematics Research; Vol. 7, No. 4; 2015
ISSN 1916-9795 E-ISSN 1916-9809
Published by Canadian Center of Science and Education

Performance of Robust Linear Classifier with Multivariate Binary Variables

I. Egbo, M. Egbo & S. I. Onyeagu

[1] Department of Mathematics, Alvan Ikoku University of Education, Owerri

[2] Department of Statistics, Federal University of Technology, Owerri.

[3] Department of Statistics, Nnamdi Azikwe University, Awka

Correspondence: I. Egbo, Department of Mathematics, Alvan Ikoku University of Education, Owerri. E-mail: egboike@gmail.com

Received: July 30, 2015 Accepted: August 17, 2015 Online Published: November 3, 2015

doi:10.5539/ijsp.v4n4p104 URL: http://dx.doi.org/10.5539/ijsp.v4n4p104

Abstract

This paper focuses on the robust classification procedures in two group discriminant analysis with multivariate binary variables. A normal distribution based data set is generated using the R-software statistical analysis system 2.15.3 using Barlett's approximation to chi-square, the data set was found to be homogenous and was subjected to five linear classifiers namely: maximum likelihood discriminant function, fisher's linear discriminant function, likelihood ratio function, full multinomial function and nearest neighbour function rule. To judge the performance of these procedures, the apparent error rates for each procedure are obtained for different sample sizes. The results obtained ranked the procedures as follows: fisher's linear discriminant function, maximum likelihood, full multinomial, likelihood function and nearest neigbour function.

Keywords: Apparent error rates, fisher's linear discriminant, full-multinomial, likelihood function, maximum likelihood and nearest neigbour function

1. Introduction

Over the years, a considerable body of research has accumulated on classification analysis, with its usefulness demonstrated in various fields, including engineering, medical and social sciences, economics, marketing, finance and management (Anderson 1972, McLachlan 1992, Joachimsthaler and Stam 1988, 1990, Ragsdale and Stam 1992, Huberty 1994, Onyeagu, 2003, Okonkwo 2011, Ekezie 2012, Egbo, Onyeagu and Ekezie 2014). Most of the research in classification analysis is based on statistical methods (Dillon and Goldstein 1978, Hand 1981, McLachlan 1992, Onyeagu 2003). However, the classification performance of existing parametric and non parametric statistical methods has not been fully satisfactory. For instance, it is well documented that parametric statistical methods such as Fisher's linear discriminant function (LDF) (1936) and Smith's quadratic discriminant function (QDF), Smith (1947) may yield poor classification results if the assumption of multivariate normally distributed attributes is violated to a significant extent (McLachlan 1992, Huberty 1994).

A number of the statistical classification methods are based on distance measures, some involve probability density functions and variance covariance and have a Bayes decision theoretic probabilistic interpretation, while others have a geometric interpretation only. An example of a distance-based measure is the Euclidean distance measure, which obviously has a geometric interpretation. If the attribute variables are independent, the Euclidean distance measure is equivalent to the Mahalanobis distance, with the usual probabilistic interpretation. However, if the variables are correlated the Euclidean measure does not have a probabilistic justification, as it does not involve any function of the probability density functions.

In this paper, we focus on two-group classification problems with binary attribute variables. There are numerous real-life binary variable classification problems; e.g. in the field of medical disease diagnosis, where the medical conditions of patients is evaluated on the basis of the presence or absence of relevant symptoms. It is obvious that the multivariate distribution of the binary attributes is non-normal, and it appears promising to analyze such problems using some statistical discriminant approaches. The statistical classification methods either minimizes some function of the undesirable distances of the training sample observations from the separating surface or minimizes the number of misclassified observations directly.

Estimation of error rates has received considerable attention in the literature. The task of classification is to classify unknown objects into predefined classes based on their observed attributes using a classification model learned from a

set of training data. Many applications such as characters recognition, decision-making and disease diagnosis, can be viewed as extensions of the classification problem (Hen and Kamber 2001). A classification instrument can be modeled using different structures such as decision graphs, decision trees, neural networks and rules. Reducing the processing time and increasing the classification rate are the two main issues in the classification problem. We consider a classical problem of discriminant analysis: an individual is to be allocated to one k distinct classes $w_1,\ldots w_c$, whose members are described by an r-component vector of binary variables $X= (x_1,x_2\ldots x_r)$. These binary variables can be viewed equivalently as a single multinomial variable having $S = 2^r$ states. The problem of classification is that of assigning item(s) into one of k, $k \geq 2$ known populations assuming that the items actually belong to one of the populations. Suppose only two populations are admitted with infinite number of individual objects. Let there be r characteristics of interest with corresponding measurement variables $X_1, X_2\ldots X_r$, $r \geq 1$. Let the response vector of individual objects in π_1 be $X_1 = (X_{11}, X_{12}\ldots X_{1r})^1$ and in π_2 be $X_2 = (X_{21}, X_{22}\ldots X_{2r})^1$. Suppose we find an object 0 with measurement vector $X_0 = (X_{01}, X_{02}\ldots X_{0r})$ outside π_1 and π_2. The problem is how to classify 0 into π_1 and π_2 in an optimum fashion. The measurement vector X can be discrete or continuous. It can also be a mixture of discrete and continuous variables. In this study, our interest is about X whose arguments are discrete. The problem is to classify 0 with measurement vector X_0 into π_1 and π_2. In this inferential setting, the researcher can commit one of the following errors. An object from π_1 may be misclassified into π_2. Also an object from π_2 may be misclassified into π_1. If misclassification occurs, a loss is incurred. Let c(i/j) be the cost of misclassifying an object from π_j into π_i. The objective of the study is to find the 'Best' classification rule. "Best" here means the rule that minimizes the expected cost of misclassification (ECM). Such a rule is referred to as the optimal classification rule (OCR) in this study we want to find the OCR where X is discrete and to be more precise, Bernoulli. Whereas classification rules with optimal properties for discriminant problems with multivariate normally distributed attribute variables are well known (Wald 1944, 1949; Smith, 1947; Adebanji, Adeyemi and Iyaniwura, 2008; Oludare, 2011), alternative rules be more appropriate if some of the attributes are skewed. Most of the studies that compared non-normal classification methods with normality-based methods for various different data conditions have assumed equal misclassification costs across groups. Hence, it is not clear to what extent the conclusions in these studies can be generalized to typical problems with distributions that are skewed with unequal misclassification costs across groups. The purpose of the current study is to establish guidelines for choosing an appropriate classification method if the problem at hand is characterized by Bernoulli multivariate data. To achieve this objective, several Monte Carlo simulation experiments are conducted to compare the performance of some traditional classification methods designed specifically to handle problems with Bernoulli multivariate data. This study is limited to the two-group classification problem.

2. Classification Procedures

2.1 Maximum Likelihood Rule (ML, Rule)

The maximum likelihood discriminant rule for allocating an observation x to one of the population $\pi_1 \ldots \pi_n$ is to allocate x to the population which gives the largest likelihood to x. that is the maximum likelihood rule says one should allocate x to π_j when

$$L_i = \max L_i(x) \qquad \text{Anderson (1984)}$$

Theorem: if π_i is the $N_p(\mu_i, \sum)$ population, $i = 1 \ldots g$ and $\sum > 0$, then the maximum likelihood discriminant rule allocate x to π_j where $j \in \{1, \ldots n\}$ is that value of i which minimized the Mahalanobis distance $(x-\mu)^1 \sum^{-1} (x-\mu)$ where $g = 2$ the rule allocate x to π_1 if $\alpha^1(x-\mu) > 0$ and $\alpha^1\{x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\} > 0$, where $\alpha = \sum^{-1}(\mu_1 - \mu_2)$ and $\mu = (\mu_1 + \mu_2)$ and to π_2 or otherwise.

2.2 The Fisher's Linear Discriminant Function (FLDF rule)

The linear discriminant function for discrete variables is given by

$$\hat{L}(x) = \sum_j \sum_k (\hat{p}_{2j} - \hat{p}_{1j}) s^{kj} x_k - \tfrac{1}{2} \sum\sum (\hat{p}_{2j} - \hat{p}_{1j}) s^{kj} (\hat{p}_{2k} + \hat{p}_{1k}) \qquad (2.2.1)$$

where S^{kj} are the elements of the inverse of the pooled sample covariance matrix, $\hat{p}_{1j}$ and $\hat{p}_{2j}$ are the elements of the sample means in π_1 and π_2 respectively. The classification rule obtained using this estimation is: classify an item with response pattern X into v if

$$\sum_j \sum_k (\hat{p}_{2j} - \hat{p}_{1j}) s^{kj} X_k - \tfrac{1}{2} \sum_j \sum_k (\hat{p}_{2j} - \hat{p}_{1j}) s^{kj} (\hat{p}_{2k} + \hat{p}_{1k}) > 0 \tag{2.2.2}$$

and to π_2 or otherwise.

2.3 The Likelihood Function Rule (LF rule)

Consider the generalized ratio test for the hypothesis H0: X, $X_{11}...X_{1n} \sim f_1(x)$ and $X_{21}... X_{2n} \sim f_2(x)$ against H1: $X_{11}...X_{1n1} \sim f_1(x)$ and $X_{21}...X_{2n_2} \sim f_2(x)$. As was proposed by Anderson (1982), Pires & Bronco (2004) and Onyeagu et al (2013) found that the likelihood ratio criterion also handles the problem of zero frequency. For multinomial model, they proposed a test statistic that is a function of X and is given by:

$$L(x) = \frac{\left[1 + \frac{1}{n_1(x)}\right]^{n_1(x)} \cdot \left(n_1(x) + 1\right)}{\left[1 + \frac{1}{n_2(x)}\right]^{n_2(x)} \cdot \left(n_2(x) + 1\right)} x \frac{\left(1 + \frac{1}{n_2}\right)^{n_2} \left(n_2 + 1\right)}{\left(1 + \frac{1}{n_1}\right)^{n_1} \left(n_1 + 1\right)} \tag{2.3.1}$$

This rule fails to take account of several factors that may be important in practice. These factors are the differential prior-probabilities of observing individuals from the two populations and differential cost incurred by misclassification and a-prior probabilities and if $n_1(x) = 0$ and $n_2(x) = 0$, the classification rule becomes: Classify item with response pattern into π_1 if $L(x) > 1$ and into π_2 $L(x) < 1$. For $n_1 = n_2$, this rule falls back to the Full Multinomial Rule. The LF Rule also solves the zero frequency problem. A new observation X with $n_1(x) = 0$ will be classified in π_1 if and only if

$$\left(1 + \frac{1}{n_2(x)}\right)^{n_2(x)} \times n_2(x) + 1 < C \tag{2.3.2}$$

2.4 Full Multinomial Function Rule (FMF rule)

Suppose we have a d-dimensional random vector $x^1 = (x_1,...x_d)$ where each $x_j, j = 1,...d$ assumes one of the two distinct values: 0 or 1. The sample space then has a multinomial distribution consisting of the 2^d possible states. Given two disjoint populations, π_1 and π_2 with priori probabilities p_1 and p_2, the density is

$$f(x) = p_1 f_1(x) + p_2 f_2(x) \tag{2.4.1}$$

The two group problem attempts to find an optimal classification rule that assigns a new observation x to π_1 if

$$f_1(x) / f_2(x) > p_2 / p_1 \tag{2.4.2}$$

When x has only two states, it will be a binomial random variable with $n_i(x)$ observation from π_i and expected value $np_i f_i(x), i = 1,2.$ Estimates for prior probabilities can be obtained by $\hat{p}_i = \frac{n_i}{n}$, where $n = n_1 + n_2$ represents the total number of sample observations. The full multinomial model estimates the class-conditional densities by

$$f_i(x) = \frac{n_i(x)}{n}, \quad i = 1,2. \tag{2.4.3}$$

where $n_i(x)$ is the number of individuals in a sample of size n_i from the population having response pattern X . The classification rule is: classify an item with response pattern X into π_i if

$$q_1 \frac{n_1(x)}{n_1} > q_2 \frac{n_2(x)}{n_2} \tag{2.4.4}$$

and to π_2 if

$$q_1 \frac{n_1(x)}{n_1} < q_2 \frac{n_2(x)}{n_2} \tag{2.4.5}$$

and with probability

$$\frac{1}{2} \text{ if } q_1 \frac{n_1(x)}{n_1} = q_2 \frac{n_2(x)}{n_2} \tag{2.4.6}$$

The full multinomial rule is simple to apply and the computation of apparent error does not require rigorous computational formula. However, Pires and Bronco (2004) noted as pointed out by Dillon and Goldstein (1978) that one of the undesirable properties of the full multinomial Rule is the way it treats zero frequencies. If $n_1(x) = 0$ and $n_2(x) \neq 0$, a new observation with vector X will be allocated to π_2, irrespective of the sample sizes n_1 and n_2.

2.5 Nearest Neighbour Function Rule (NNF rule)

Hills (1967) introduced perhaps the simplest nearest neighbour estimator for binary data, which classifies a particular response vector x based on the number of cells in response vectors y that differ from x. Specifically, let k be the number of cells in which x and y differ. Then define $R_j = \{y_j / (x - y_j)^1 (x - y_j) \leq k\}$ to be a rule which classifies x if each of its cells differs by no more than k components. That is, classify x into π_1 if:

$$\sum_{R_j} \frac{n_1(y_j)}{n_1} > \sum_{R_j} \frac{n_2(y_j)}{n_2} \tag{2.5.1}$$

and into π_2 otherwise.

For example, with d = 3 and x = (111), the neighbours of order k = 1 are R_{111} = 110, 101, 011. Note that k = 0 reduces to the full multinomial model. In practice, one simply needs to construct the table of frequencies for all possible pattern of x and use a counting procedure over the set Rj to form the sample-based likelihood ratio for classification purpose. If the cell count for the jth cell is nij, then the nearest neighbour procedure assigns the observation to π_1 if

$$\frac{\left[n_{ij} + \sum_A n_{ij}\right] / n_1}{\left[n2j + \sum_A n_{2j}\right] / n_2} > \frac{p_2}{p_1} \tag{2.5.2}$$

where A is the set of neighbour of state j. Hills comments that the estimate of the likelihood ratio has less sampling variability than the simple method using cell frequencies.

2.6 Testing Adequacy of Discriminant Coefficient

Consider the discriminant problems between two multinomial populations with mean μ_1, μ_2 and common matrix $\sum$. The coefficient of the MLD discriminant function $a^1 x$ are given by $\alpha = \sum^{-1} \delta$ where $\delta = \mu_1 - \mu_2$. In practice of course the parameters are estimated by $\bar{x}_1, \bar{x}_2$ and $S = m^{-1}\{(n_1 - 1)s_1 + (n_2 - 1)s_2\}$, where $m = n_1 + n_2 - 2$.

Letting $d = \bar{x}_1 - \bar{x}_2$, the coefficients of sample MLDF given by $a = MW^{-1}d$

A test of hypothesis H_0: $\alpha_1 = 0$ using the sample Mahalanobis distances $D_p^2 = Md^1 W^{-1} d$ and $D_1^2 = Md_1^1 W_{11} d_1$ has been proposed by Rao (1965) this test statistics uses the statistic:

$$\left\{\frac{m - p + 1}{p - k}\right\} C^2 \left\{(D_p^2 - D_k^2) / (m + c^2 D_p^2\right\} \tag{2.6.1}$$

where $c^2 = \frac{n_1 n_2}{n}$. Under the null hypothesis (2.6.1) has $F_{p-k}, m - p + 1$ distribution and we reject H_0 for large value of this statistics.

2.7 Evaluation of Classification Functions

One important way of judging the performance of any classification procedure is to calculate the error rates or misclassification probabilities (Richard and Dean, 1988). When the forms of parent populations are known completely, misclassification probabilities can be calculated with relative ease. Because parent populations are rarely known, we shall concentrate on the error rates associated with the sample classification functions. Once this classification function is constructed a measure of its performance in future sample is of interest. The total probability of misclassification (TPM) is given as:

$$TPM = P_1 \int_{R_1} f_1 dx + P_2 \int_{R_2} f_2 dx$$

The smallest value of this quantity by a judicious choice of R_1 and R_2 is called the optimum error rate (OER).

OER= Minimum TPM

2.8 Probability of Misclassification

In constructing a procedure of classification, it is desired to minimize on the average the bad effects of misclassification (Onyeagu 2003, Richard and Dean, 1988, Oludare 2011). Suppose we have an item with response pattern x from either π_1 or π_2. We think of an item as a point in a r-dimensional space. We partition the space R into two regions R_1 and R_2 which are mutually exclusive. If the item falls in R_1, we classify it as coming from π_1 and if it falls in R_2 we classify it as coming from π_2. In following a given classification procedure, the researcher can make two kinds of errors in classification. If the item is actually from π_1, the researcher can classify it as coming from π_2. Also the researcher can classify an item from π_2 as coming from π_1. We need to know the relative undesirability of these two kinds of errors in classification. Let the priori probability that an observation comes from π_j be q_1, and from π_2 be q_2. Let the probability mass function of π_1 be $f_1(x)$ and that of π_2 be $f_2(x)$. Let the regions of classifying into π_1 be R_1 and into π_2 be R_2. Then the probability of correctly classifying an observation that is actually from π_1 into π_1 is

$$p(1/1) = \sum_{R_1} f_1(x)$$

and the probability of misclassifying such an observation into π_2 is

$$p(2/1) = \sum_{R_2} f_1(x) \tag{2.8.1}$$

Similarly, the probability of correctly classifying an observation from π_2 into π_2 is $p(2/2) = \sum_{R_2} f_2(x)$ and the probability of misclassifying an item from π_1 into π_2 is

$$p(1/2) = \sum_{R_1} f_2(x) \tag{2.8.2}$$

The total probability of misclassification using the rule is

$$TPMC(R) = q_1 \sum_{R_2} f_1(x) + q_2 \sum_{R_1} f_2(x) \tag{2.8.3}$$

In order to determine the performance of a classification rule R in the classification of future items, we compute the total probability of misclassification known as the error rate. Lachenbruch (1975) defined the following types of error rates.

(i). Error rate for the optimum classification rule, R_{opt}. When the parameters of the distributions are known, the error rate is $TPMC(R) = q_1 \sum_{R_2} f_1(x) + q_2 \sum_{R_1} f_2(x)$ which is optimum for this distribution.

(ii) Actual error rate: The error rate for the classification rule as it will perform in future samples.

(iii) Expected actual error rate: The expected error rates for classification rules based on samples of size n_1 from π_1 and n_2 from π_2.

(iv) The plug-in estimate of error rate obtained by using the estimated parameters for π_1 and π_2.

(v) The apparent error rate: This is defined as the fraction of items in the initial sample which is misclassified by the classification rule.

	π_1	π_2	
π_1	n_{11}	n_{12}	n_1
π_2	n_{21}	n_{22}	n_{11} n

The table above is called the confusion matrix and the apparent error rate is given by

$$\hat{P}(mc) = \frac{n_{12} + n_{21}}{n} \tag{2.8.4}$$

Hills (1967) called the second error rate the actual error rate and the third the expected actual error rate. Hills showed that the actual error rate is greater than the optimum error rate and it in turn, is greater than the expectation of the plug-in estimate of the error rate. Martin and Bradley (1972) proved a similar inequality. An algebraic expression for the exact bias of the apparent error rate of the sample multinomial discriminant rule was obtained by Goldstein and Wolf (1977), who tabulated it under various combinations of the sample sizes n_1 and n_2, the number of multinomial cells and the cell probabilities. Their results demonstrated that the bound described above is generally loose.

3. Simulation Experiments and Results

The five classification procedures are evaluated at each of the 118 configurations of n, r and d. The 118 configurations of n, r and d are all possible combinations of n = 40, 60, 80, 100, 200, 300, 400, 600, 700, 800, 900, 1000, r = 3, 4, 5 and d = 0.1, 0.2, 0.3, and 0.4. A simulation experiment which generates the data and evaluates the procedures is now described.

(i) A training data set of size n is generated via R-program where $n_1 = n/2$ observations are sampled from π_1 which has multivariate Bernoulli distribution with input parameter p_1 and $n_2 = n/2$ observations sampled from π_1, which is multivariate Bernoulli with input parameter $p_2, j = 1...r$. These samples are used to construct the rule for each procedure and estimate the probability of misclassification for each procedure is obtained by the plug-in rule or the confusion matrix in the sense of the full multinomial.

(ii) The likelihood ratios are used to define classification rules. The plug-in estimates of error rates are determined for each of the classification rules.

(iii) Step (i) and (ii) are repeated 1000 times and the mean plug-in error and variances for the 1000 trials are recorded. The method of estimation used here is called the resubstitution method.

The following table contains a display of one of the results obtained

Table 3.1(a). Effect of input parameters P_1 and P_2 on classification rules at various values of sample size and Replications (mean apparent error rates)

P_1 = (.3, .3, .3, .3) P_2 = (.4, .4, .4, .4)

Sample sizes	Full M.	LIK	NN	LD	ML
40	0.319000	0.321075	0.489112	0.381037	0.374825
60	0.343100	0.343550	0.478416	0.391000	0.387550
100	0.366815	0.367215	0.461320	0.398890	0.397110
140	0.378457	0.378846	0.454800	0.404271	0.402992
200	0.387760	0.388057	0.446400	0.407427	0.406502
300	0.397721	0.397241	0.438453	0.409610	0.409196
400	0.400930	0.401657	0.435726	0.411002	0.410721
600	0.406289	0.405407	0.432000	0.411500	0.411304
700	0.407172	0.407675	0.429813	0.411761	0.411586
800	0.408018	0.408408	0.429216	0.411552	0.411533
900	0.408085	0.408826	0.428669	0.411161	0.411172
1000	0.409524	0.409221	0.428876	0.411682	0.411710

p(mc) = 0.4117

Table 3.1(b). Effect of input parameters P_1 and P_2 on classification rules at various values of sample size and Replications (actual error rates)

P_1 = (.3, .3, .3, .3) P_2 = (.4, .4, .4, .4) $\left| p(mc) - \hat{p}(mc) \right|$

Sample size	Full M.	LIK	NN	LD	ML
40	0.050941	0.050731	0.080872	0.044290	0.045492
60	0.045663	0.045781	0.067813	0.038558	0.038499
100	0.034565	0.035548	0.055462	0.030836	0.030861
140	0.032885	0.032247	0.047209	0.027729	0.027813
200	0.027185	0.027286	0.038667	0.022319	0.021727
300	0.024630	0.023152	0.029343	0.018785	0.018688
400	0.0203367	0.020843	0.023498	0.0170960	0.017054
600	0.017737	0.017672	0.018107	0.013808	0.013752
700	0.016275	0.017144	0.016404	0.012829	0.012844
800	0.015970	0.015479	0.015434	0.012267	0.012249
900	0.014064	0.014624	0.014416	0.011348	0.011378
1000	0.013622	0.014593	0.014414	0.010406	0.010447

Tables 3.1(a) and (b) mean apparent error rates increases with the sample size in all the classification rules except in the Nearest neighbour rule where the mean apparent error rates decreases with the increase in sample sizes. The actual error rates decreases with the increase in the sample sizes. In terms of performance, Fisher's linear discriminant function ranked first followed by maximum likelihood function rule, full multinomial function, Nearest Neighbour Rule and likelihood ratio function came last.

Classification Rule	**Performance/Rank**
Fisher linear discriminant function rule	1
Maximum likelihood	2
Full Multinomial function rule	3
Nearest Neighbour function rule	4
Likelihood function rule	5

4. Conclusion/Recommendation

We considered eight population pairs for the case of four variables. On the average, fisher's linear discriminant function rule was the best in terms of estimating the probability of misclassification because it gives values closer to the actual probability of misclassification. The next is the maximum likelihood function rule which was better than the full multinomial function rule, the fourth is the Nearest Neighbour rule while the likelihood ratio occupied the last position and is the worst. This study, in addition to its mean structures characterized by marginal probabilities P_1 and P_2, we considered structures determined by the difference $d = (p_2 - p_1) \leq 0.4$. It was observed that as d increases from 0.1 to 0.4 the accuracy of the procedures also increased. This shows that accuracy increases with increasing d. It is important to note that Fisher's linear discriminant function (FLDF), maximum likelihood function Rule and Full multinomial function Rule performed also very well in situation where $d = 0.2$ in the three variables. It was also observed that the more the information or the number of variables, the lower the probability of misclassification. This implies that accuracy increases with increasing number of variables. Fisher's linear discriminant function outperformed other classification rules. From the analysis so far carried out, the procedures can be ranked as follows: Fisher's linear discriminant function rule, maximum likelihood function rule, Full multinomial function rule, Nearest Neighbour rule and likelihood function rule. Secondly, we conclude that it is better to increase the number of variables because accuracy increases with increasing number of variables. We recommended that the work be extended to the area of multiple group discrimination and classification.

References

Adebanji, A.O., Adeyemi, S. & Iyaniwura, J.O., (2008). Effect of unequal sample size Ratio on the performance of the Linear Dischminant Function: *International Journal of Modern Mathematics, 2*(1), 97 - 108.

Anderson, J.A. (1972). Separate Sample Logistic Discrimination. *Biometrika*, *59*, 19 - 35. http://dx.doi.org/10.1093/biomet/59.1.19

Anderson, T.W. (1981). An Introduction to Multivariate Statistical Analysis, 2nd Edition, Wiley, New York.

Anderson, T.W. (1984). An Introduction to Multivariate statistical methods 2nd edition. John Willey, New York.

Dillon, W.R., & Goldstain, M. (1978). On the performance of some multinomial classification rules, *Journal of American Statistical Association, 73*(362), 305 -320. http://dx.doi.org/10.1080/01621459.1978.10481574

Egbo, I., Onyeagu, S.I. & Ekezie, D.D. (2014). A Comparison of Multinomial Classification rules for Binary Variables. *International Journal of Mathematical Sciences and Engineering Applications (IJMSEA),* 8(V), 141 - 157.

Ekezie, D. D. (2012). Comparison of seven Asymptotic Error Rate Expansion for the sample Linear Discriminant Function. Unpublished Ph.D Thesis submitted to Department of Statistics, Imo State University, Owerri.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179 - 188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

Goldstein, M., & Wolf. (1977). On the problem of Bias in multinomial classification. *Biometrics, 33*, 325 - 331. http://dx.doi.org/10.2307/2529782

Hand, D. J. (1981). Discrimination and Classification. Wiley, New York, NY.

Hen, J. E., & Kamber, M. (2001). Data Mining: Concepts and Techniques. Academic press, San Diego, CA.

Hills, M. (1967). Discrimination and allocation with discrete data. *Applied Statistics, 16,* 237 - 250. http://dx.doi.org/10.2307/2985920

Huberty, C.J. (1994). *Applied Discriminant Analysis.* New York: John Wiley.

Joachimsthaler, E. A., & Stam, A. (1988). Four Approaches to the Classification Problems in Discriminant Analysis: An Experimental Study. *Decision Sciences, 19*, 322 - 333. http://dx.doi.org/10.1111/j.1540-5915.1988.tb00270.x

Joachimsthaler, E. A., & Stam, A. (1990). Mathematical Programming Approaches for the Classification Problems in Two_Group Discriminant Analysis, *Multivariate Behavioural Research, 25*, 427 - 454. http://dx.doi.org/10.1207/s15327906mbr2504_2

Lachenbruch, P. A. (1975). Discriminant Analysis, Hafrier press New York.

McLachlan, R. (1992). Discriminant Analysis and Statistical Pattern Recognition. John Wiley and sons Inc., New York. http://dx.doi.org/10.1002/0471725293

Martin, D. C., & Bradley, R. A. (1972). Probability Models, Estimation and Classification for Multivariate Dichotomous Populations, *Biometrics, 23*, 203 - 221. http://dx.doi.org/10.2307/2528968

Okonkwo, E. (2011). A comparative Performances of Several Robust discriminant analysis procedure. *JNSA, 23*, 55-64.

Oludare, S. (2011). Robust Linear classifier for equal Cost Ratios of misclassification, *CBN Journal of Applied Statistics, 2*(1).

Onyeagu, S. 1. (2003). Derivation of an optimal classification rule for discrete variables. *Journal of Nigerian Statistical Association.*

Onyeagu, S. 1., Osuji, G. A., Ekezie, D. D., & Ogbonna C. J. (2013). A Review of Classification Models Using Discrete Variables. *Research Journal of Mathematical and Statistical Sciences, 1*(8), 28 - 38.

Pires, A.M. & Bronco, J.A. (2004). Comparison of vol 4, 79-80 multinomial classification rules. *Journal of the American Statistics Association,* 73, 305 - 313.

Ragsdale, C. T., & Stam, A. (1992). Introducing Discriminant Analysis to the Business Statistics Curriculum. *Decision Sciences, 23*, 724 - 745. http://dx.doi.org/10.1111/j.1540-5915.1992.tb00414.x

Richard, A. J., & Dean, W. W. (1998). Applied Multivariate Statistical Analysis. 4th edition, Prentice Hall, Inc. New Jessey.

Smith, C. A. (1947). The robust estimation of classification error rates; some examples of discrimination. *Annals of Eugenics, 18*, 272 - 282.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Annals of Mathematical Statistics, 15*, 145 - 162. http://dx.doi.org/10.1214/aoms/1177731280

Wald, A. (1949). Statistical decision functions. *Annals of Mathematical Statistics, 20*, 165-205. http://dx.doi.org/10.1214/aoms/1177730030