Journal of the Nigerian Association of Mathematical Physics
Volume 35, (*May, 2016*), pp 163 – 174

Classification Into Two Groups With Different Cost of Misclassification Ratios

G. M. Oyeyemi and L. A. Oyebanji

Department of Statistics, University of Ilorin, Ilorin, Nigeria.

Abstract

Fisher's Linear Discriminant and Bayesian Classification procedures were compared when the assumption of equal cost of misclassification is violated. The comparison was carried out at various samples sizes and different misclassification cost ratios. Data were simulated to consist two groups (populations) of four variables each from two multivariate normal populations. The homogeneity of the variance-covariance matrices of the two groups was tested using Box's M-Test. The Apparent Error Rate as the estimate of the Actual Error Rate was used to judge the performances of both procedures at different misclassification cost ratios (1:1, 1:2, , , 4:5) and sample sizes (10, 20, 30, 40, , , 100). The results show that at equal cost ratio (1:1), both approaches produced almost the same error rate at different sample sizes. With difference in misclassification cost ratio, the Bayesian approach generally has higher proportion of misclassifications than the Fisher at various ratios and sample sizes. The Fisher performed better in small sample cases (n < 50) under all the cost ratios considered except 1:2 and 1:5. For large sample cases (n > 50), the performance was better at cost ratios 2:3, 2:4 and 2:5.

Keywords: Fisher's Linear Discriminant Function, Baye's Classification Rule, Apparent Error Rates, Cost of Misclassification, Cost Adjusted Prior Probabilities, Cost Matrix

1.0 Introduction

There are vast literatures on classification and discrimination. Anderson [1] viewed the problem of classification as the problem of decision theory. Anderson and Bahadur [2] viewed it as a problem of assigning an unknown observation to a group with low error rate.There are various approaches to classification which include Bayesian approach, Fisher's Linear Discriminant Function (Fisher's LDF), Maximum Likelihood Discriminant Function, Distance Base Discriminant Function and so on. The consideration of cost-sensitive studies in discriminant function has received growing attention in the past few years. Brefeld et al [3] discussed dependence of cost on the single sample and not on the class of the samples.Oyeyemi and Oyebanji [4] compared the performance of both the Fisherian and Bayesian approaches to classification and concluded that the Bayes' approach performed better. Ariyo and Adebanji [5] compared the performance of both linear and quadratic classifier under unequal cost of misclassification. Zandrozny and Elkan [6] looked into issues of cost-sensitivity when costs and prior probabilities are both unknown.

Bayesian approach to classification assigns an observed unit to a group with the greatest posterior probability. Bayesian approach to classification in the case of normally distributed observations with unknown parameters was discussed in[7]. Classification based on Bayes' formula was also discussed in[8].

Fisher's LDF is the most widely used method of classification because of its simplicity and optimality (it minimizes the expected cost of misclassification) properties. In classical discriminant problem [9], the number of groups or subpopulations was two. Discriminant function was obtained by choosing linear combination of the variables to maximize the ratio of the 'between-group' to that of the 'within-group' variance. Linear discriminant analysis is known to be optimal for two multivariate normal groups with equal covariance matrices. The robustness of linear discriminant analysis and effect of failure of assumption to hold have been studied in [10, 11]. The effect of unequal covariance matrices on linear discriminant analysis was studied by Gilbert [10] for large sample cases.

In this paper, a comparison of both Bayesian and Fisher's linear discriminant procedures was carried out at various samples sizes and different misclassification cost ratios.

Corresponding author: G. M. Oyeyemi, E-mail: gmoyeyemi@gmail.com, Tel.: +2348052278655

2.0 Evaluating Classification Rules

To judge the performance of a sample classification procedure, we want to calculate its misclassification probability or error rate.A measure of performance that can be calculated for any classification procedure is the Apparent Error Rate (APER) which is the defined as the fraction of observations in the sample that are misclassified by the classification procedure. Denote n_{1m} and n_{2m} the number of objects misclassified as $\pi_1$and $\pi_2$respectively, then

APER = $\frac{n_{1m}+n_{2m}}{n_1+n_2}$ (1)

The APER is intuitively appealing and easy to calculate. Unfortunately, it tends to underestimate the Actual Error Rate (AER) when classifying new objects. This underestimation occurs because we used the sample to build the classification rule (therefore we can call this the training sample) as well as to evaluate it. To obtain a reliable estimate of the AER we ideally consider an independent test sample of new objects from which we know the true class label. This means that we split the original sample into a training and test samples. The AER is then estimated by the proportion of misclassified objects in the test sample while the training was used to construct the classification rule. However, there are two drawbacks with this approach

(i) It requires large samples.

(ii) The classification rule is less precise because we do not use the information from the test sample to build the classifier.

An alternative is the (leave-one-out) cross-validation or jackknife procedure which works as follows.

1. Leave one object out of the sample and construct a classification rule based on the remaining n-1 objects in the sample.
2. Classify the left-out observation using the classification rule constructed in step 1.
3. Repeat the two previous steps for each of the objects in the sample.

Denote n_{1m}^{cv} and n_{2m}^{cv} the number of left-out observations misclassified in class 1 and 2 respectively.

Then a good estimate of the Actual Error Rate is given by

AER = $\frac{n_{1m}^{cv}+n_{2m}^{cv}}{n_1+n_2}$ (2)

3.0 Methodology

3.1 Fisher's Linear Discriminant Function

The objective of Fisher's LDF is finding projection to a plane such that samples from different groups (classes) are well separated, i.e. finding linear combination that maximizes the ratio of the between-group sum of squares and the within-group sum of squares such that a good separation is achieved. It is also known for maximizing the sample mahalanobis distance between the two sets of data. And maximizing difference between groups may lead to reducing probability of misclassification.

Fisher suggested that using a linear combination of observations and choosing coefficients so that the ratio of the difference of means of linear combination in the two groups to its variance is maximized. Fisher's linear discriminant function is known to be optimal (in the sense of expected number of misclassifications) for two multivariate normal populations with equal covariance matrices.

3.2 Incorporating cost of misclassification intoFisher's LDF

If the discriminant function

$Y = V'X_i = S_w^{-1}(\mu_1 - \mu_2)X_i,\ i = 1,2,\dots,p\ \ and\ \ X = (x_1 x_2 \dots\ x_p)$ (3)

Expected value of Y with respect to class j is given by

$\overline{Y_j} = S_w^{-1}(\mu_1 - \mu_2)\mu_j = \theta_j,\quad j = 1,2.$ (4)

Since $Y_{ij} = V'X_i = V_1X_{1i} + V_2X_{2i} + \cdots + V_pX_{pi},\quad i = 1,2,\dots,n_j,\qquad j = 1,2.$ (5)

Where X_{ij} represents the score of individual from j[th] class with measurement X_i.

Now, $\pi_1$and π_2 are the two classes that X must be assigned.

Also if $C = [C_1\ C_2]$ be the cost of misclassification matrix, where

$C_1$and C_2 are the vectors of cost associated with classes (groups) 1 and respectively

Furthermore, if we define $C_{kj}, kj = 1,2$ as the cost associated with classifying an individual into π_k when infact the correct decision is to classify it intoπ_j.

Hence possible costs in class 1 are C_{11} and C_{21}, and possible costs in class 2 are C_{12} and $C_{12.}$

Where,

C_{11} is the cost of classifying an individual from $\pi_1$as π_1.

C_{21} is the cost of classifying an individual from $\pi_1$asπ_2.

C_{12} is the cost of classifying an individual from $\pi_2$as π_1.

C_{22} is the cost of classifying an individual from $\pi_2$as π_2.
With C_1 and C_2 being cost vectors, C is the misclassification cost matrix given as

$$C = [C_1 \quad C_2]$$

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

If we define $Z_i = Y_{ij}C_{kj}, \ \ i = 1,2,\dots,n \ \ and \ k,j = 1,2.$ (6)
Where C_{kj} is as described above.
Then $Z = YC^{'}$ (7)
Where $Y = [Y_1 \quad Y_2]$ and $Y_{ij} = V^{'}X_{ij}$, $\ \ i = 1,2,\dots,n. \ and \ j = 1,2.$
So $Z_i^{'s}$ are the new observations that are to be classified into one of the two classes.

$$Y = [Y_1 \quad Y_2]$$

$$C^{'} = \begin{bmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \end{bmatrix}$$

Then $$Z = [Y_1 \quad Y_2]\begin{bmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \end{bmatrix}$$

$$= [Y_2C_{12} \quad Y_1C_{21}]$$

$$= [Z_1 \quad Z_2]$$

Note that $C_{11} = C_{22} = 0$, no cost of misclassification for correctly classified units.
i.e $C_{kj} = 0, \forall \ k = j$
Hence, $Z_i = Y_{ij}C_{kj}, \ \ i = 1,2,\dots,n_j \ and \ kj = 1,2$ (8)

$$\overline{Z_j} = \overline{Y_j}C_{kj} \ , k \neq j$$

$$= \frac{1}{n_j}\sum_{i=1}^{n} Y_{ij}C_{kj}$$

$$= C_{kj}\left\{\frac{1}{n_j}\sum_{i=1}^{n} Y_{ij}\right\}$$

So $\overline{Z_1} = \overline{Y_1}C_{21}$ and $\overline{Z_2} = \overline{Y_2}C_{12}$

$$\bar{Z} = [\bar{Z_1} \quad \bar{Z_2}]$$

Assignment Rule
$For \ i = 1,2, \ \dots, n_j \ \ and \ j = 1,2.$ Assign i[th] individual to class j if
$C^{'}\left\{Y_i - \frac{1}{2}(\bar{Y_1} - \bar{Y_2})\right\} > 0,$ and to class j' if otherwise. Where j≠ j'.
Simplifying the rule above gives;
Assign an individual to class 1 if $Z_1 > \frac{\overline{Z_1}+\overline{Z_2}}{2}$, $if \ C_{21} > C_{12} \ and \ to \ class \ 2 \ if \ otherwise.$
OR
Assign an individual to class 1 if $Z_2 < \frac{\overline{Z_1}+\overline{Z_2}}{2}$, $if \ C_{21} < C_{12} \ and \ to \ class \ 2 \ if \ otherwise.$

3.3 Baye's Classification Rule (BCR)

The Bayesian approach towards classification when all parameters are known (parameters are not always known but estimated from the data) and misclassification cost are equal would begin with evaluation of the posterior probability that $X \in \pi_j$ given X, for each $j = 1, 2, \dots, J$. Then the posterior odds or ratio are computed for each pair of population. Alternatively for $j > 2$, the population with the greatest posterior probability can be selected.
When the costs of misclassification are unequal, the Bayesian would select the population that produces a minimum cost when average with respect to the posterior probability. This result also holds for all $J \geq 2$ when all parameters are known.
Bayesian approach to classification is more generally applicable. Even the covariance matrices need not be equal for the approach to be applicable and it requires no complicated distribution theory, though it is much more difficult to apply.

4.0 Analysis and Comparison

4.1 Test for Normality

Test for normality was done using Mardia's test. The Mardia [12] test function is used to calculate the Mardia's multivariate skewness and kurtosis coefficients as well as their corresponding statistical significance. This function can also calculate the corrected version of the skewness coefficient for small sample (n<20).

Table 1: Mardia's Multivariate Normality Test

	Group 1	Group 2
G1p	1.8757	1.4792
Chi. Skew	31.2620	24.6533
p-value. Skew	0.0518	0.2150
G2p	26.2888	26.8068
z. Kurtosis	1.7240	2.0257
p-value. Kurtosis	0.0847	0.0628
Chi.Small. Skew	32.5871	25.6982
p-value. Small	0.0374	0.1760

4.2 Calculation of Apparent Error Rate.

We calculate the misclassification probability using

$$\hat{\boldsymbol{P}} = \frac{\boldsymbol{n_{1m} + n_{2m}}}{\boldsymbol{n_1 + n_2}}$$

and the Apparent Error Rate (APER) as the estimate of the Actual Error Rate (AER) using

$$APER = A\hat{E}R = \frac{n_{1m} + n_{2m}}{n_1 + n_2} \times 100$$

Where n_{1m} and n_{2m} are the number of observations misclassified as π_1 and π_2 respectively and n_1 *and* n_2 are number of samples in group1 and 2 respectively.

Table 2: Table of Apparent Error Rates at equal cost ratio but different sample sizes.

Apparent Error Rate at equal cost ratio but different sample sizes (%)		
Sample Sizes	Fisher's LDF	BCR
10	10	10
20	15	15
30	15	15
40	12.5	13.8
50	15	15
60	18.3	18.3
70	15.7	14.3
80	15.6	16.3
90	15	15.6
100	13.5	15.5

Table 3: Table of Apparent Error Rate at different cost ratios and different sample sizes.

	Apparent Error Rate at different cost ratios and different sample sizes (%)									
	n=10		n=20		n=30		n=40		n=50	
Ratio	FLDF	BCR	FLDF	BCR	FLDF	BCR	FLDF	BCR	FLDF	BCR
1:2	5.0	20.0	2.5	17.5	1.7	13.3	1.3	15.0	4.0	17.0
1:3	0.0	20.0	0.0	17.5	0.0	16.7	0.0	18.8	2.0	19.0
1:4	0.0	20.0	0.0	20.0	0.0	21.7	0.0	23.8	0.0	19.0
1:5	5.0	20.0	2.5	22.5	3.3	23.3	2.5	22.5	4.0	22.0
2:3	0.0	10.0	0.0	17.5	0.0	15.0	0.0	13.8	0.0	12.0
2:4	0.0	20.0	0.0	17.5	0.0	13.3	0.0	17.5	0.0	18.0
2:5	0.0	20.0	0.0	15.0	0.0	16.7	0.0	18.8	0.0	19.0
3:4	0.0	10.0	0.0	15.0	0.0	15.0	1.3	16.3	2.0	14.0
3:5	0.0	15.0	0.0	17.5	0.0	13.3	0.0	13.8	0.0	15.0
4:5	0.0	10.0	0.0	15.0	0.0	13.3	2.5	16.3	3.0	14.0

Table 3 continued.

	n=60		n=70		n=80		n=90		n=100	
Ratio	FLDF	BCR	FLDF	BCR	FLDF	BCR	FLDF	BCR	FLDF	BCR
1:2	5.0	18.3	3.6	17.1	6.3	18.1	5.6	16.7	7.0	16.0
1:3	0.8	20.8	0.7	19.3	2.5	20.6	1.1	16.7	2.0	23.0
1:4	0.8	21.7	0.7	20.0	1.9	23.1	0.6	23.3	1.0	24.5
1:5	0.0	22.5	0.0	20.7	0.0	23.8	0.0	24.4	7.0	25.5
2:3	0.0	16.7	0.0	15.7	0.0	17.5	0.0	12.2	0.0	15.5
2:4	0.0	18.3	0.0	17.8	0.0	18.1	0.0	16.1	0.0	11.5
2:5	0.0	20.8	0.0	18.6	0.0	21.3	0.0	18.9	0.0	19.5
3:4	2.5	18.3	1.4	15.7	1.3	16.9	1.7	16.7	0.5	15.0
3:5	0.0	17.5	0.0	16.4	0.0	16.9	0.0	17.8	0.0	16.5
4:5	3.3	18.3	2.9	15.7	2.5	16.3	0.6	16.1	1.5	16.0

Graphical presentation of results

Below are the graphical presentations of the results. Two kinds of graphs were presented here, firstly we plot the sample sizes against the error rate at constant cost ratio to see the effect of sample sizes, and secondly cost ratios was plotted against the error rates at constant sample sizes to see the effect of cost ratios.

Graphs of sample sizes plotted against error rates at differentcost ratio.

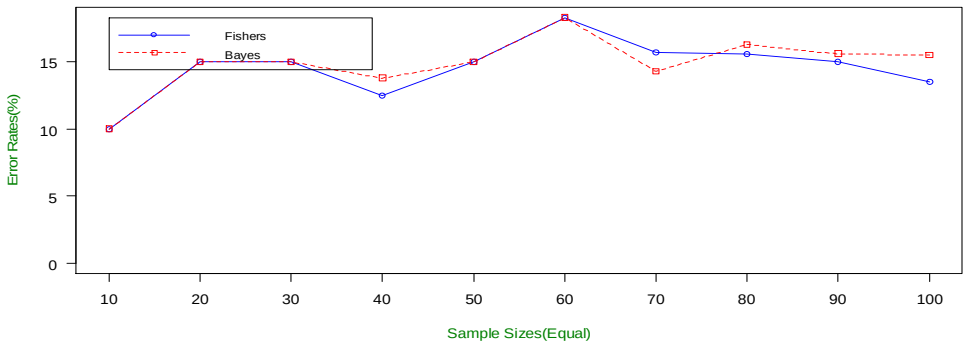

Figure 1: Graph of sample sizes against error rate at equal cost ratio.

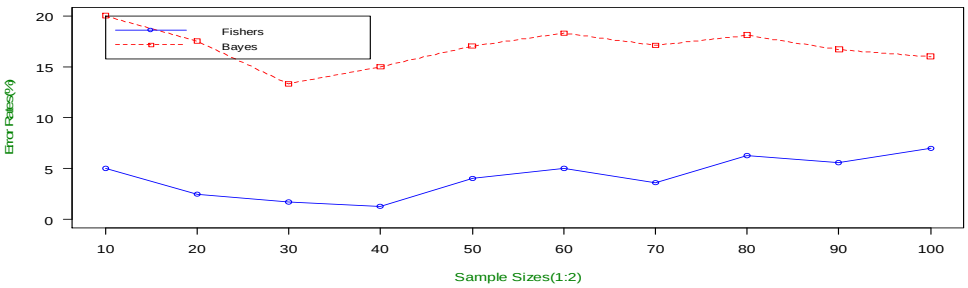

Figure 2: Graph of sample sizes against error rate at cost ratio of 1:2.

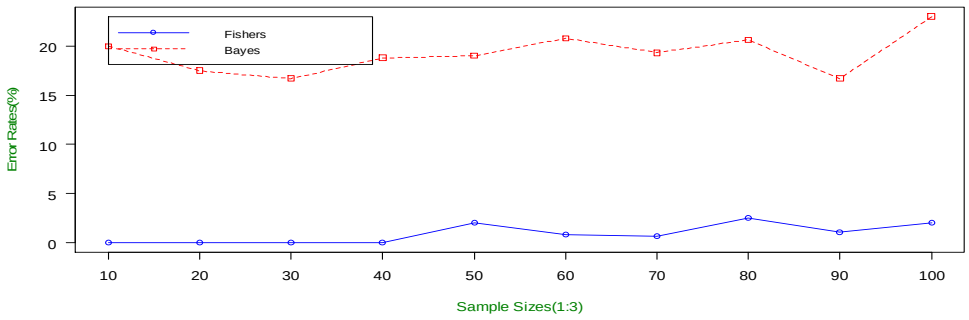

Figure 3: Graph of sample sizes against error rate at cost ratio of 1:3.

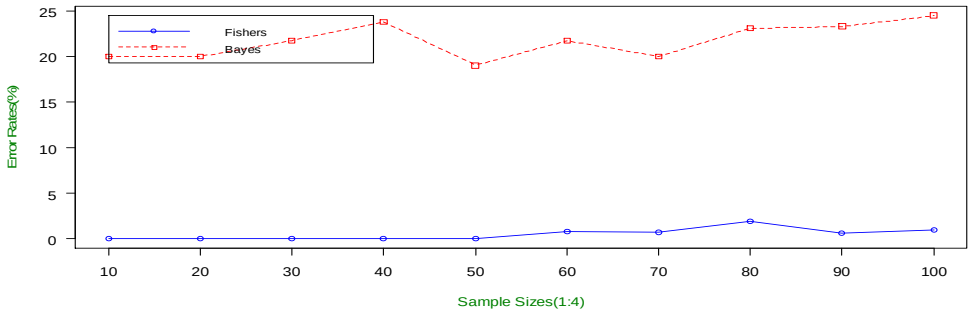

Figure 4: Graph of sample sizes against error rate at cost ratio of 1:4.

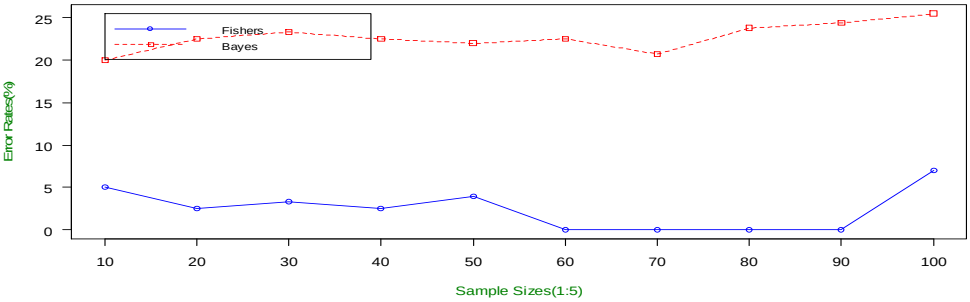

Figure 5: Graph of sample sizes against error rate at cost ratio of 1:5.

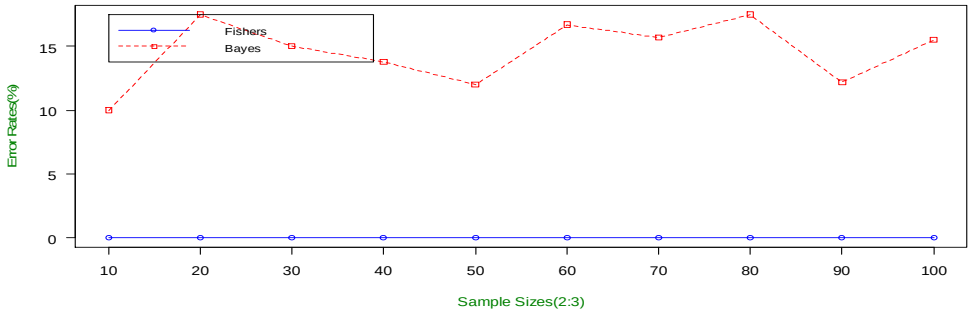

Figure 6: Graph of sample sizes against error rate at cost ratio of 2:3.

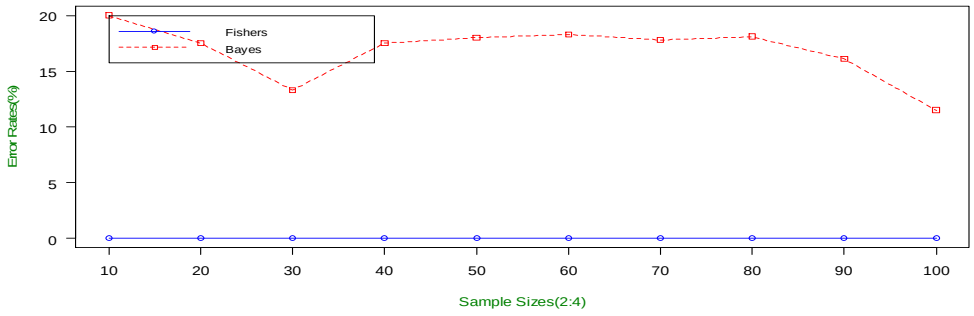

Figure 7: Graph of sample sizes against error rate at cost ratio of 2:4.

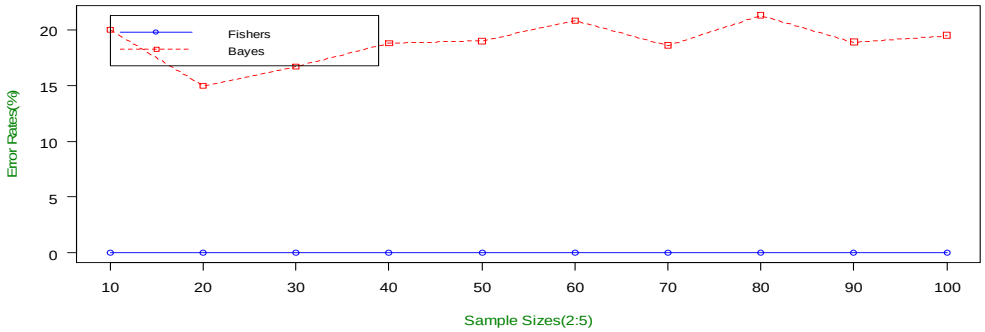

Figure 8: Graph of sample sizes against error rate at cost ratio of 2:5.

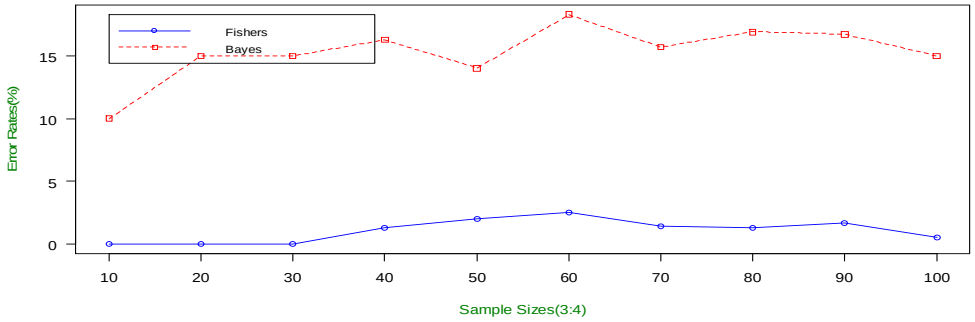

Figure 9: Graph of sample sizes against error rate at cost ratio of 3:4.

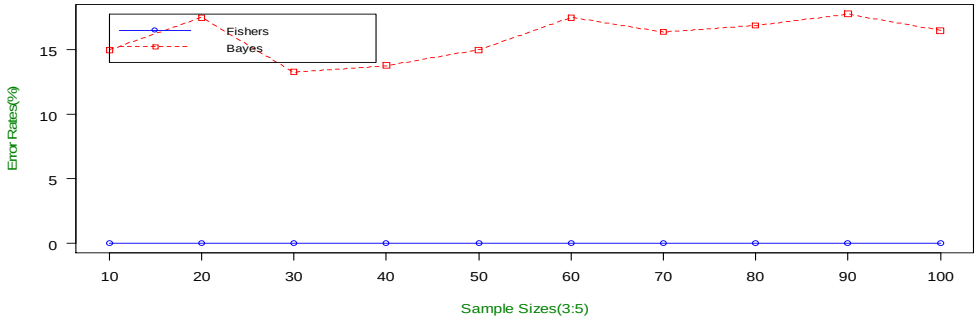

Figure 10: Graph of sample sizes against error rate at cost ratio of 3:5.

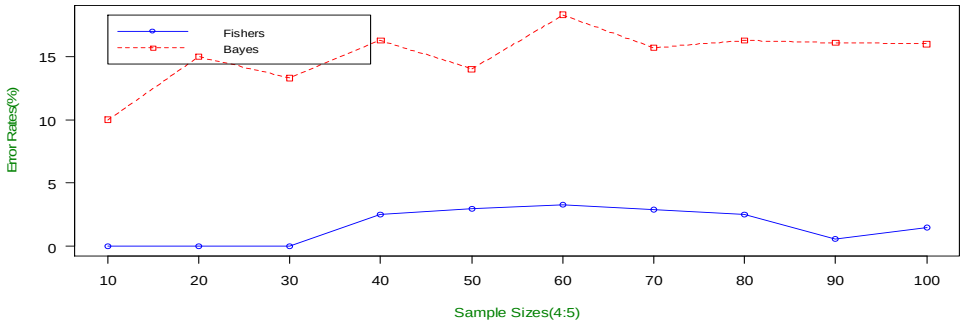

Figure 11: Graph of sample sizes against error rate at cost ratio of 4:5.

Graphs of cost ratios against error rate for different sample sizes.

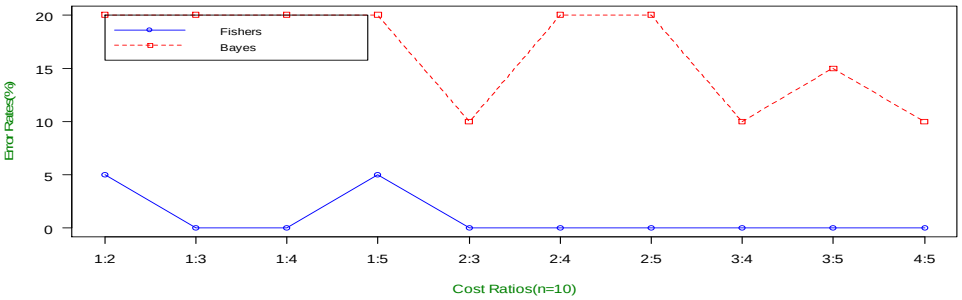

Figure 12: Graph of cost ratios against error rates at sample sizes 10.

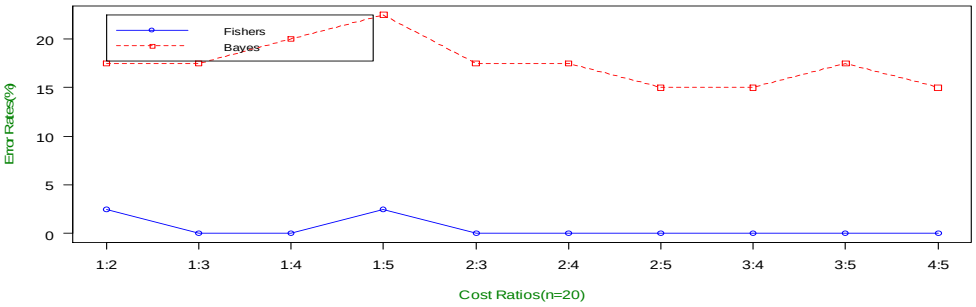

Figure 13: Graph of cost ratios against Error rates at sample sizes 20.

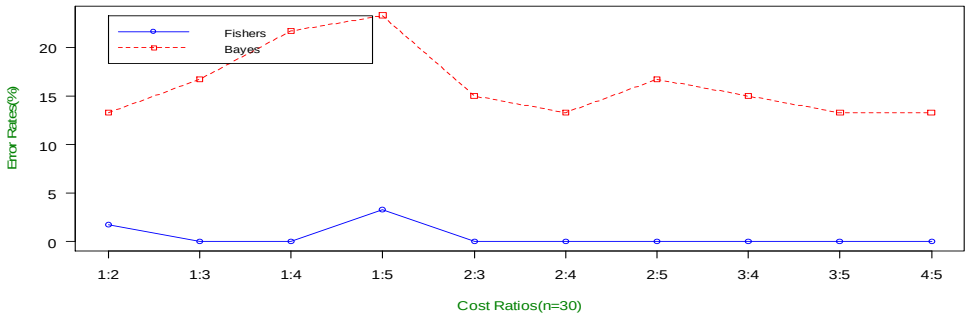

Figure 14: Graph of cost ratios against error rates at sample sizes 30.

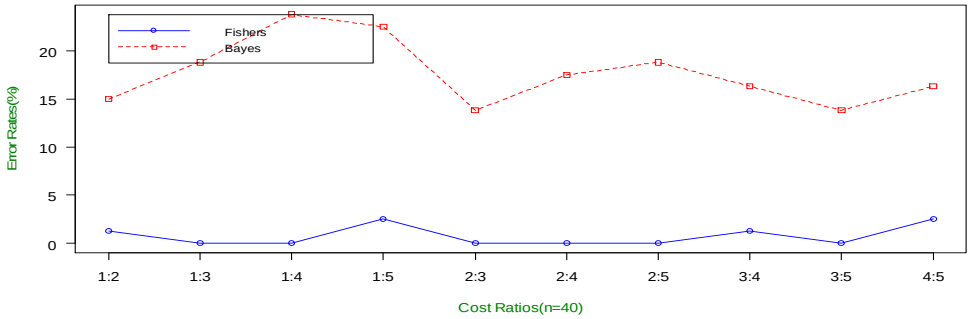

Figure 15: Graph of cost ratios against error rates at sample sizes 40.

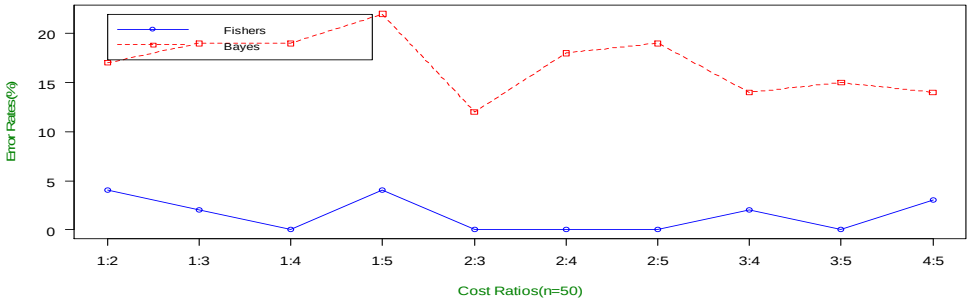

Figure 16: Graph of cost ratios against error rates at sample sizes 50.

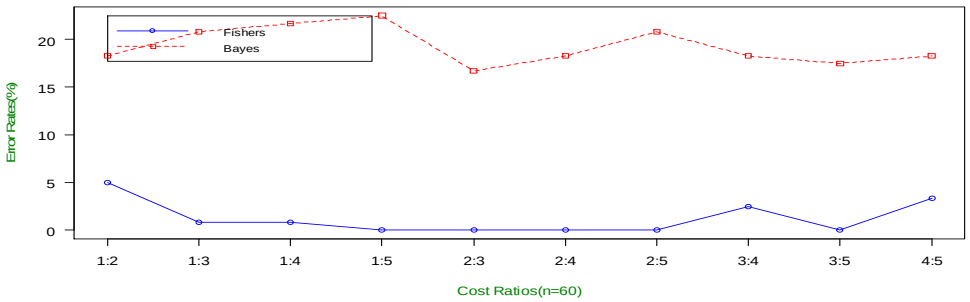

Figure 17: Graph of cost ratios against error rates at sample sizes 60.

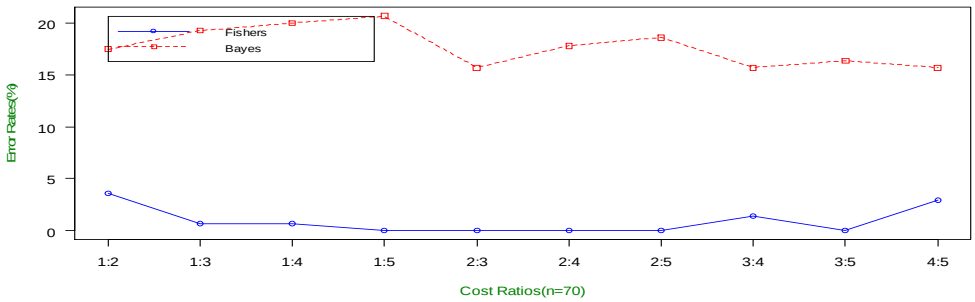

Figure 18: Graph of cost ratios against error rates at sample sizes 70.

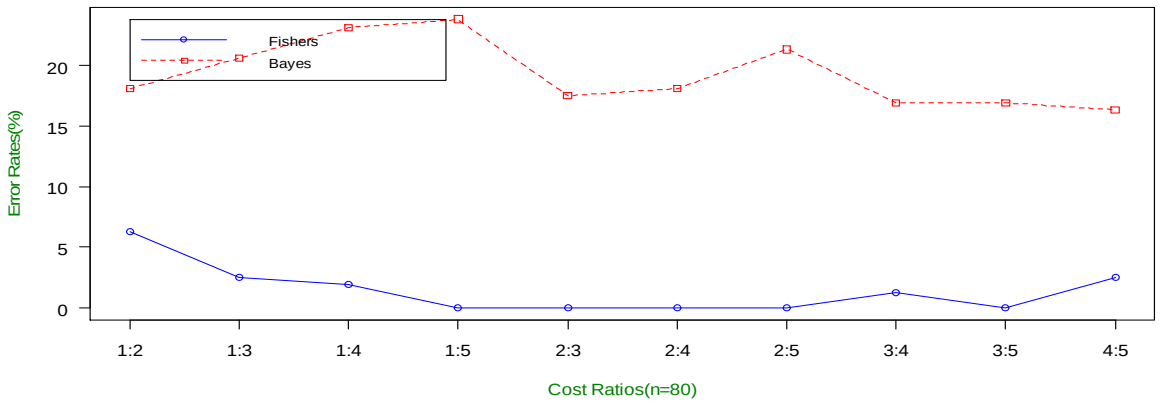

Figure 19: Graph of cost ratios against error rates at sample sizes 80.

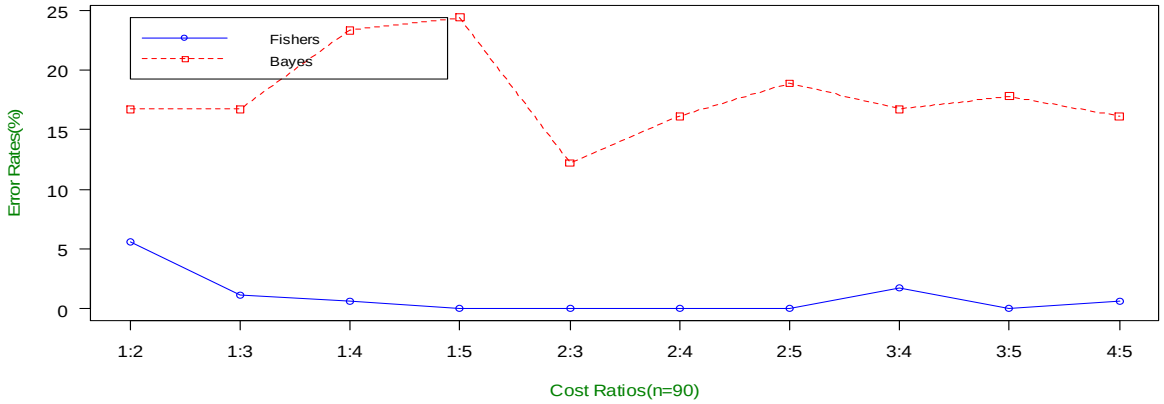

Figure 20: Graph of cost ratios against error rates at sample sizes 90.

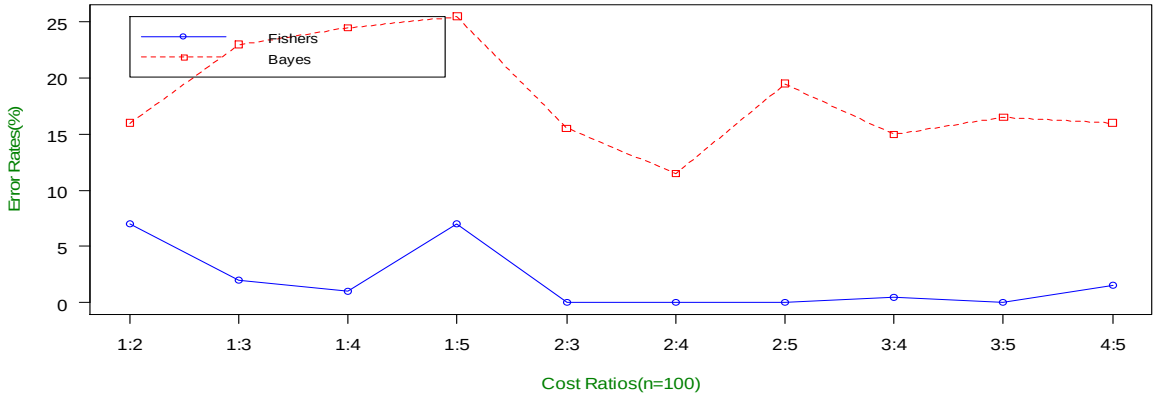

Figure 21: Graph of cost ratios against error rates at sample sizes 100.

5.0 Discussion of Results and Conclusions

We have considered incorporation of cost of misclassification into Fisher's LDF and the Bayes' Classification Rule to assign observations with measurement X into one of the 2 groups when the classical cost assumption was violated. It is quite clear from the results as shown in Table 2 and Figure 1, at equal cost ratio that both approaches produced almost the same error rate at different sample sizes (with little difference noticed when n=70 and 100). Introduction of different cost ratios caused imbalances in the proportion of misclassification error rate in each of the approaches as shown in Table 3. Figures 2 to 11 show the graphs of sample sizes against misclassification error rate for different cost ratios while figures 12 to 21 are the graphs of cost ratios against misclassification error rates for different sample sizes. The high error rate reduction noticed in the Fisherian approach which has its highest error rate of 7% when the sample size is 100 (at 1:2 and 1:5; Figure 21). On the other hand the Bayesian approach has its lowest error rate to be 10% at sample sizes 10 with cost ratios of 2:3, 3:4 and 4:5 (Figures 6, 9, 11 and 12).

The Fisherian approach provided only a decision that attempt to cope with the problem of risk associated with the classification decision but falls short in that it is then required that covariance matrices be equal, assumptions that are not always true. The Bayesian approach provided only a predictive distribution for placing a vector of observations into the second population (in addition to the predictive odds).

Since we know that a classification rule with lower probability of misclassification or error rate is better than the one with high probability of misclassification or error rate, we conclude that the Fisherian approach was greatly improved by this method even when the classical cost assumption was violated and performing twice better or more than the Bayesian approach.

6.0 References

[1] Anderson T.W. (1958). An Introduction of Multivariate Analysis.John Wiley & Sons, New York.

[2] Anderson, T. W and Bahadur, R. R. (1962).Classification into two Multivariate Distributions with Different Covariance Matrices. The Annals of Mathematical Statistics, 33(2), 420-431.

[3] Brefeld, U., Geibel, P. and Wysotzki, F. (2003). Support Vector Machine with example Dependent Costs. In proceedings of the 4th European Conference on Machine Learning, 23 – 24.

[4] Oyeyemi, G.M, Oyebanji L.A, Salawu, I. S andFolurunsho, A.I. (2014). Comparison between Fisherian and Bayesian Approach to Classification using Two Groups.Scientia Africana, 13 (1), 24-35.

[5] Ariyo, S. andAdebanji, A.O. (2010). Robust linear Classifiers for Unequal Cost Ratios of Misclassification. CBN Journal of Applied Statistics,2(1), 23-31.

[6] Zandrozny, B. and Elkan, C. (2001). Learning and Making decisions when costs and probabilities are both unknown. In proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 204 – 213.

[7] Dunsmore, I.R.(1966). A Bayesian Approach to Classification. Royal Statistical Society Journal, Series B, 28, 568-577.

[8] Birnbaum A. and Maxwell, A.E.(1960). Classification Procedures Based on Baye's formula.Applied Statistics, 9, 152–169.

[9] Fisher, R.A (1936). Use of multiple measurements in Taxononic problems. Ann Eugenics, 7, 179-188.

[10] Gilbert, E. S. (1985). The Effect of Unequal Variance- Covariance Matrices on Fishers Linear Discriminant Functions. Biometrics, 25, 323-357.

[11] Hardle, W. and Simar L. (2003). Applied Multivariate Statistical Analysis. Springer, New York.

[12] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3): 519–530.