Robust Linear Classifier for Unequal Cost Ratios of Misclassification

Oludare S, Ariyo[1] and A.O. Adebanji[2]

This paper focuses on the robust classification procedures when the assumption of equal cost of misclassification is violated. A normal distribution based data set is generated using the Statistical Analysis System (SAS) version 9.1. Using Barlett's approximation to chi-square, the data set was found to be homogenous and was subjected to three linear classifiers namely: Maximum Likelihood Discriminant Function (MLDF), Fisher's linear Discriminant Function and Distance Based Discriminant Function. To Judge the performances of these procedures, the Apparent Error Rates for each procedure is obtained for different cost ratios 1:1, 1:2, 1:3, 1:4 and 1:5 and sample sizes 5:5, 10:10, 20:20, 30:30, and 50:50. The results shows that the three procedures are insensitive to cost ratio exceeding ratio 1:2 and that MLDF was observed as robust discriminant function among classification functions considered.

Key Words: Apparent Error Rates, Maximum Likelihood Discriminant Function, Distance Based Discriminant Function, Fisher's linear

1.0 Introduction

Fisher (1936) was the first to suggest a linear function of variables representing different characters, hereafter called the linear discriminant function (discriminator) for classifying an individual into one of two populations. Fisher's linear discriminant function (LDF) method is well established for equal covariance multivariate normal predictors (Aderson, 1958).It optimally deteriorates, however, as the assumption of normality gets unrealistic (Krzanowski, 1988). Qian Du and Chein-I Chang (2001) used distance-based discriminant function (DBDF) that uses a criterion for optimality derived from Fisher's ratio criterion. It not only maximizes the ratio of inter-distance between classes to intra-distance within classes but also imposes a constraint that all class centers must be aligned along predetermined directions. A method of discrimination, based on maximum likelihood estimation, is described. On a variety of mathematical models, including and extending the models most commonly assumed in discriminant theory, the discriminant reduces to multivariate logistic analysis. Even when no simple model can be assumed, other considerations show that this method should work well in practice, and should be very robust with respect to departures from the theoretical assumptions. The method is compared with others in its application to a diagnostic problem. The consideration of Cost-sensitive Studies in linear discriminant function has received growing attention in the past years. (Elkan, 2001; Margineantu and Dietterich, 2000). One way to incorporate such costs is the use of a cost matrix, which specifies the misclassification costs in the

[1] National Horticultural Research Institute, P.M.B 5432 Jericho Research Area, Ibadan, Nigeria. (ariyodare @gmail.com, +234-08035206932)

[2] Deparment of Mathematics, Kwame Nkruma University of Science and Technology, Kumasi , PMB KNUST, Ghana. tinuadebanji@yahoo.com. +233241860372.

class dependent manner. (Elkan, 2001). Brefeld *et al*, (2003) discuses the ideal to let the cost depend on the single example and not only on the class of the example. Authors also presented a natural cost-sensitivities extension of the Support Vector Machine (SVM) and discused its relation to Bayes rule. Ariyo and Adebanji (2010) compared the performance of both Linear and Quadratic classifier under unequal cost of misclassification and concluded that both classifiers are insensitive to the cost ratio exceeding ratio 1:2. Adebanji et al (2008) investigated the performance of the homoscedastic discriminant function (HDF) under the non-optional condition of unequal group representation (prior probabilities) in the population and the asymptotic performance of the classification function under this condition. The results obtained showed that the misclassification of observation from the smallest group escalate when the sample size ratio 1:2 is exceeded and this increases in error rate is not corrected by increasing the sample size. They observed that the performance of the function is more susceptible to higher variability in the reported error rates. Several Authors had looked into issue of cost-sensitivity when costs and prior probabilities are both unknown (Zandrozny and Elkan, 2001) and its application in different areas especially in neural Network (Berardi and Zhang., 1999; Xingye, and Yufeng, 2008 and Zheng, *et al*, 2007). The issue of different misclassification costs for balanced data has not been given much attention. Hence, the study is motivated to evaluate the performance and robustness of selected linear classifiers when the assumption of equal cost of Misclassification is violated.

2.0 Methodology

A Simulated data from SAS 9.1 was used for this study. The data consists of two groups with four variables (x_1 , x_2, x_3, x_4). The Simulation process creates a data set by simulated random variables from two normal populations**.**

The above procedure was repeated for n = 5, 10, 20, 30, 50. For each value of n the, procedure returned 10, 20, 40, 60 and 100 sample sizes. To test the equality of mean by multivariate methods, Hotelling T^2 and Wilks's lambda was used. The Barlett's Likelihood ratio test was also used to test the homogeneity or other wise of the data sets and the data set was found to be homogenous and was subjected to three (3) selected linear classifiers namely: Maximum Likelihood Discriminant Function(MLDF), Fisher's linear Discriminant Function (FLDF) and Distance Based Discriminant Function (DBDF). To Judge the performances of these procedures, the Apparent Error Rates (APER) for MLDF, FLDF and DBDF under different cost ratio 1:1, 1:2, 1:3, 1:4 and 1:5 were obtained.

2. 1 Discriminant rules

A discriminant rule d corresponds to a division of R^p into disjoint region $R_1, \ldots, R_n$

$$(UR_i{=}R^n)$$

The rule d defined by allocate x to π_j if x ε R_i for j=1 ,…,n. Discriminant will be more accurate if π_j has most of its probability concentrated in R for each j.

2. 2 Maximum Likelihood rule (ML rule)

The maximum likelihood discriminant rule for allocating an observation x to one of the population $\pi_1, \dots, \pi_n$ 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

$$L_i = \max L_i(x) . \qquad \text{(Anderson, 1984)} \qquad (1)$$

Theorem 1 If π_i is the $N_p(\mu_i, \sum)$ population, i =1,…, g and $\sum > 0$, then the maximum likelihood discriminant rule allocate x to π_j where $j\epsilon\{1, \dots, n)$ is that value of i which minimized the mahalanobis distance $(x-\mu)^{'}\sum^{-1}(x-\mu_1)$ where g=2 the rule allocate x to π_1. If $\alpha^{'}(x-\mu) > 0$ and $a^{'}\left\{x-\frac{1}{2}(\bar{x}_1+\bar{x}_2)\right\} > 0$, where $\boldsymbol{\alpha}$=$\sum^{-1}(\mu_1-\mu_2)$ and $\mu = (\mu_1+\mu_2)$ and to π_2 otherwise.

2.3 Fisher's Linear Discriminant rule (FDL rule)

Once the linear discriminant function has been calculated, an observation x can be allocated to one of the n population on the basis of its "discriminant scores"$a^{'}x$. The samples $\bar{x}_i$ have scores $a^{'}\bar{x}_i$=$\bar{y}_i$. The x is allocated to that population where mean scores is closest to $a^{'}x$ that is allocate x to π_j if $|a^{'}x\text{-}a^{'}\bar{x}_j| < |a^{'}x\text{-}a^{'}\bar{x}_i|$ for $i \neq j$ (Giri,2004)

Fisher's discriminant function is most important in the special case of g=2 groups. Then B has rank one and can be written as $B = \left(\frac{n_1 n_2}{n}\right) dd^{'}$ where $d = \bar{x}_1 - \bar{x}_2$. Thus, $W^{-1}B$ has only one zero eigenvalue. This eigenvalue equals to $tr\ W^{-1}B = \left(\frac{n_1 n_2}{n}\right) d^{'} W^{-1} d$. The corresponding eigenvalue is $a = W^{-1}d$.Then the discriminant rule becomes; allocate x to π_1 if $d^{'} W^{-1}\left\{x-\frac{1}{2}(\bar{x}_1+\bar{x}_2)\right\} > 0$ and to π_2 otherwise.

2.4 Distance –based discriminant Function

This approach requires a definition of distance between the single observation x and each training sample. One possibility is to define a squared distance by the Mahalanobis qualities:

$$D_i^2 = (x-\mu_i)^{'} S^{-1}(x-\mu_i). \qquad (2)$$

Where μ_i is the mean of ith training set (i=1,2),and S is the covariance matrix pooled within the training set.

2.5 Testing Adequacy of discriminant coefficient

Consider the discriminant problems between two multinormal populations with mean μ_1, μ_2 and common matrix $\sum$. The coefficient of the MLD discriminat function $a^{|}x$ are given by $\boldsymbol{\alpha}$=$\sum^{-1}\delta$ where $\delta = \mu_1 - \mu_2$. In practice of course the parameters are estimated by $\bar{x}_1, \bar{x}_2$ and $S = m^{-1}\left((n_1 - 1)S_1 + (n-1)S_2\right)$, where $m = n_1 + n_1 - 2$. Letting $d = \bar{x}_1 - \bar{x}_2$,the coefficients of the sample MLDF given by $a = m\ W^{-1}d$.

A test of hypothesis $H_0; \alpha_i = 0$ using the sample Mahalanobis distances $D_p^2 = md^{|}W^{-1}d$ and $D_1^2 = md_1^{|}W_{11}^{-1}d_1$ has been proposed by Rao (1965) this test statistics uses the statistic:

$$\left\{\frac{m-p+1}{p-k}\right\} c^2(D_p^2 - D_k^2)/(m + c^2 D_p^2)\} \qquad (3)$$

Where $c^2 = \frac{n_1 n_2}{n}$. Under the null hypothesis (3) has $F_{p-k,m-p+1}$ distribution and we reject H_0 for large value of this statistics.

2.6 Evaluating of Classification Function

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:

$$TPM = P_1 \int_{R_1} f_1\ dx + P_2 \int_{R_2} f_2\, dx \qquad (4)$$

The smallest value of this quantity obtained by a judicious choice of R_1 andR_2 is called the optimum error rate (OER).

OER = Minimum TPM .

Probability of Misclassification

The probability of allocating an individual to population π_i , when in fact he comes from π_j is given by:

$$P_{ij} = \int \phi_i(x) L_j(x) dx \qquad (5)$$

If the parameters of the underlying distribution are estimated from the data, then we get estimated probability $\check{P}_{ij}$. Consider the case of two normal population $N_p(\mu_1,\sum)$ and $N_p(\mu_2,\sum)$. If $\mu = \frac{1}{2}(\mu_1 + \mu_2)$, then when x comes from π_1, $a'(x-\mu) \sim N_p(\frac{1}{2}a(\mu_1 - \mu_2), a'\sum\boldsymbol{\alpha})$.Since the disccriminant function is given by $l_2(x) = a'(x-\mu)$ with $a = \sum^{-1}(\mu_1 - \mu_2)$,we see that if x comes from π_1, $h_2(x) \sim N\left(\frac{1}{2}\Delta^2, \Delta^2\right)$,where:

$$\Delta^2 = (\mu_1 - \mu_2)'\sum^{-1}(\mu_1 - \mu_2) \qquad (6)$$

Equation (6) is the square Mahalanobis distance between the positions, similarly, if x comes from π_1, $h_2(x) \sim N\left(\frac{1}{2}\Delta^2, \Delta^2\right)$.Thus, the misclassification probabilities are given by:

$$\begin{aligned} P_{12} &= p(h(x) > 0/\pi_2 \\ &= \phi(-E(h)/\pi_2) \\ &= \phi(\tfrac{-1}{2}\Delta) \quad \text{(Giri,2004)} \end{aligned} \qquad (7)$$

where ϕ is the standard normal distribution function.

2.7 Error Rates

Optimal error rates (OER) are error rate associated with the best possible allocation rule that could be used, if all assumption made are appropriate. This error rate can be calculated when the population density functions are known it given by:

$$\text{OER} = \text{ minimum TPM} = \tfrac{1}{2}\phi(\tfrac{-1}{2}\Delta) + \tfrac{1}{2}\phi\left(\tfrac{-1}{2}\Delta\right) = \phi\left(\tfrac{-1}{2}\Delta\right) \qquad (8)$$

The performance of sample classification function can be evaluated by calculating the actual error rate (AER).

$$\text{AER} = P_1 \int_{R_2} f_1(x)\, dx + \int_{R_1} f_1(x) dx \qquad (9)$$

Where R_1 and R_2 represent the classification regions determined by sample size n_1, n_2 respectively.

The AER indicates how the sample classification function will perform in future samples. Like the OER, it cannot, in general, be calculated because it depends on the unknown density functions $f_1(x)$ and $f_2(x)$.There is a measure of performance that does not depends on the form of the parent populations and that can be calculated for any classification function procedure. This measure is called the apparent error rate (APER) is defined as the fraction of observation in the training sample that are misclassified by the sample classification function. It can be easily calculated from the confusion matrix which shows actual versus predicted group membership. For n_1 observation from π_1 and n_2 observations from π_2, the confusion matrix has the form.

Actual Membership	Predicted Membership	
	n_{1c}	n_{1M}
	n_{2c}	n_{2M}

Where

n_{1c}= Number of π_1 items correctly as π_1 items.

n_{2c}= Number of π_2 items correctly as π_2 items.

n_{1M} = Number of π_1 items misclassified as π_2 items.

n_{2M} = Number of π_2 items misclassified as π_1 items.

This is called the Apparent Error Rate (APER) and is defined as:

$$\text{APER} = \frac{n_{1M}+n_{2M}}{n_{1c}+n_{2c}} \quad \text{(Richard and Dean, 1998)} \qquad (10)$$

3.0 Results

Table1 shows that Likelihood test (Barlett's approximation to Chi-square) statistics is not signification at 5% level of significant, we accept the null hypothesis that the simulated data has equal variance covariance matrix.

Table 1: The Test results for testing Equality of Variance.

Data	Chi – square	Df	P-value (5%)	Null Hypothesis
A	6.411	10	0.7796	Equal Variance

Table 2 shows the error for the three classification procedures under different cost ratio, the APER for the three discriminant rules becomes unchanged as the cost assign C_{21} is tripled. For cost ratio 1:1 and 1:2 MLDF rule gave the least "error rate" compared to the classification rules considered.

Table 2 : Apparent Error rate for the classification rules under different cost ratio.

Cost ratio	ML rule	FL rule	DB rule
1:1	47.83	48.00	49.67
1:2	50.17	51.67	50.00
1:3	50.00	50.33	50.00
1:4	50.00	50.00	50.00
1:5	50.00	50.00	50.00

In Table 3, the value of Apparent Error rate (APER) for diffract Sample size under different cost ratios was presented. The result shows that on only the Distance based Discriminnat Function (DBDF) is sensitive to Small Sample sizes. At Sample size 5:5, DBDF had that least APER but this value increases as the sample size increases. It is also clear that sample sizes considered has little effect on the performance of classification functions considered under different cost ratio.

Table 3: The Apparent Error rate for the Classification rules for different Sample sizes and cost Ratio

Sample ratio															
	n; 5:5			**n; 10:10**			**n; 20:20**			**n; 30:30**			**n; 50:50**		
	ML	FL	DB	ML	FL	DB	ML	FL	DB	ML	FL	DB	ML	FL	DB
Cost ratio															
1:1	39.4	49.0	30.3	41.9	48.0	30.3	38.0	50.1	30.2	39.0	50.0	39.2	47.8	48.0	49.7
1:2	40.1	49.0	48.0	43.0	47.0	49.0	51.0	50.2	49.1	49.9	51.0	50.0	50.2	51.7	50.0
1:3	49.7	50.0	50.3	51.0	53.0	50.0	50.1	50.0	49.0	50.0	50.0	50.0	50.0	50.3	50.0
1:4	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
1:5	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0

4.0 Conclusion/Recommendation

Three Linear dicsriminant rules: MLDF, FLDF, and DBDF were studies when classical cost assumption is violated. In each allocation rule, introduction of different cost ratios causes imbalances in the proportion of misclassification also the error rates. At cost ratio 1:1, 1:2 all classification rules except MLDF gave equal misclassification proportion. The APER for the three classification rules under different cost ratio were also examined in this study, for cost ratio 1:1 and 1:2 MLDF gave the least error rate. At cost ratio exceeding ratio 1:3, the APER remain unchanged for all classification rules. We conclude that APER for all classifications considered is insensitive to cost ratio exceeding ratio 1:3.

Reference

Adebanji, A.O., Adeyemi, S. and Iyaniwura, J .O. (2008). Effect of sample size Ratio on the Performance of the Linear Discriminate function: International Journal of Modern Mathematics 3 (1), 97-108

Aderson,T.W (1958) . An Introduction to Multivariate Statistics Analysis. New York:Wiley.

Anderson, T.W. (1984). An Introduction to Multivariate Statistical Methods 2nd ed. John Willey, New York

Ariyo. O S and Adebanji. A. O (2010) Effect of Misclassification Costs on the Performance Functions. Paper Presented at the 45th Annual Conference of the Science Association of Nigeria (SAN) held Niger Delta University, Wilberforce, Island , Bayelsa State

Berardi, V.L. and Zhang, G.P.(1999). The Effect of Misclassification Costs on Neural Network Classifiers. Decision Sciences 30(3), 659 – 682.

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.

Elkan, C. (2001). The Foundations of Cost – Sensitive Learning In proceedings of the 7th International Joint Conference on Artificial Intelligence, 973 – 978.

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

Giri, M.C. (2004). Multivariate Statistical Analysis 2nd Edition, Revised and Expanded. New York Basel.

Krazonowski, W.J. (1988). Principles of Multivariate Analysis: users' perspective. John Willey and sons Inc New York.

Margineantu, D. and Dietterich, T.G. (2000). Bootstrap methods for the cost-sensitive evaluation of classifiers. In Proc.17th International Conf on Machine Learning, .583 – 590.

Rao, C.R (`1965). Linear Statistical Inference and Its Applications: John Willey New York .

Richard, A. J. and Dean, W.W. (1998). Applied Multivariate Statistical Analysis. 4th ed. Prentice Hall, Inc, New Jersey.

Qian Du and Chein-I Chang (2001). A linear constrained distance-based discriminant analysis For hyperspectral image classification Pattern Recognition 34 (2), , 361-373

Xingye, Q and Yufeng, L. (2008). Adaptive Weighted Learning for unbalanced Multicategory classification. Biometric 70(3), 629-642.

Zandrozny, B. and Elkan, C. (2001). Learning and Marking 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.

Zheng,Y. Zhao, J. Jie,S and Ge,S.S (2007).High Performance Quadratic Classifier and the application on Pendigits Recognition. Journals of Decision and Control 46th IEEE Conference, 3072-3077.