Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.5, No.9, 2015

A Review of the Limitations of Some Discriminant Analysis Procedures in Multi-Group Classification.

Anyanwu Paul E.
Department of Mathematics & Statistics, Federal Polytechnic, P. M. B. 55, Bida, Niger State, Nigeria.

Ekezie Dan Dan
Department of Statistics, Imo State University, P. M. B. 2000, Owerri, Imo State, Nigeria.

Onyeagu I. Sidney
Department of Statistics, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria.

Abstract
A review is given on existing work and result of the performance of some discriminant analysis procedures under varying conditions. Few of the developed methods (Fisher's Linear Discriminant Function, Logistic Regression and Quadratic discriminant function) were reviewed. Some new results are presented for the case involving allocation with more than two groups. Shortfalls in the reviewed procedures necessitated the need for an improved procedure that can classify observations into multiple groups with high efficiency (minimal error rate).

Keywords: Multivariate, Discrimnant function, Classification, Multi-groups, Optimal

1. Introduction

Multivariate analysis has been a major arm of statistics which has significantly solved problems in classifications of multivariable data. Discrimination analysis and Logistic Regression are tools that are used for classification and prediction. Press, and Wilson,(1978) defines classification into one of several populations is discriminant analysis, or classification. Relating quantitative variables to other variables through a logistic cdf functional form is logistic regression. Estimators generated for one of these problems are used in the other. According to Markowski and Markowski (1987) Fisher's approach to discriminant problem is parametric and relies on assumptions such as multivariate normality for optimality and, therefore, may be less effective on more realistic classes of problems.

Several methods for discriminant analysis have been proposed. Differences between methods arise because of the variety of distributional assumptions made about the variables describing each object or individual to be classified. The methods based on the assumption of normality are the ones most widely used in practice. If we are willing to assume that our groups are described by multivariate normal densities with different means but the same covariance matrix, then a rule exist (for a two groups case) that allocates an individual with vector scores $\mathbf{x}$ to group D_1 if

$$\mathbf{a}^{\mathrm{T}}\left[\mathbf{x}-\tfrac{1}{2}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)\right]>0 \qquad (1)$$

Where $\mathbf{a}=\mathbf{S}^{-1}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)$, $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ are the vector of means for group one and two respectively and $\mathbf{S}$ is the variance-covariance matrix assumed to be the same for the two groups. The above was suggested by Fisher (1936) whose idea was to find a linear combination of the p variables which separates the two training samples as much as possible, and he showed that for any such combination, $\mathbf{a}^{\mathrm{T}}\mathbf{x}$ **,** is maximized by taking $\mathbf{a}$ as defined above. In many practical situations, the population parameters, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are unknown, Wald (1944) and Anderson (1951) suggested replacing the unknown parameters by the estimates of their sample, then the allocation rule will therefore be to allocate any future observation to group D_1 if

$$\delta=(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^{\mathrm{T}}\mathbf{S}^{-1}\left[\mathbf{x}-\tfrac{1}{2}(\bar{\mathbf{X}}_1+\bar{\mathbf{X}}_2)\right]>\mathrm{K} \qquad (2)$$

Otherwise, allocate to group D_2.

The constant, K is the cutoff point depending on the relative costs of misallocation from each population and also on the a-priori probabilities of $\mathbf{x}$ coming from each population. When none of this information is available, this constant is taken as zero. Fisher's LDF have attracted a large amount of methodological research due to the popularity of its technique in the field of multivariate analysis. Areas of research have been majorly on the distribution of the classification statistic, estimation of probabilities of misclassification, variables selection, estimation of the parameters of the allocation rule, performance of Fisher's LDF under varying conditions

including when classification involves more than two groups. Fisher's LDF have attracted great patronage due to its simplicity and ease of computation, however, knowledge of the performance of Fisher's LDF under certain conditions would be significantly valuable even though users may feel that little damage done in using Fisher's LDF in such situation are negligible. Hills (1967) in Krzanowski (1977) pointed out that Fisher's LDF will provide a useful tool for discrimination under wide distributional conditions but may be quite unsuitable for allocating a particular observation to one of two populations which are not multivariate normal.

This brings us to the light of what this paper intends to achieve, that is, to review known and existing results of the performance of the Fisher's LDF vis-a-viz when some basic underlying assumptions are violated, when available data are of several forms and when discrimination involves more than two groups. Also, present some new result of the limitation of Fisher's LDF for the case of discrimination and allocation with more than two groups. Conditions for the success and/or failure of linear discriminant function will be investigated and presented as well.

2. Review of Some Violated Assumptions and the Consequences

2.1 Unequal Variance-Covariance Matrices

When the assumption that two populations of interest have equal covariance matrices fails, the allocation rule described in (2) becomes, assign a future individual with vector of scores $\mathbf{x}$ to group D_1 if

$$\mathbf{x}^{\mathrm{T}}(\mathbf{S}_2^{-1}-\mathbf{S}_1^{-1})\mathbf{x}-2\mathbf{x}^{\mathrm{T}}(\mathbf{S}_2^{-1}\bar{\mathbf{X}}_2-\mathbf{S}_1^{-1}\bar{\mathbf{X}}_1)+(\bar{\mathbf{X}}_2^{\mathrm{T}}\mathbf{S}_2^{-1}\bar{\mathbf{X}}_2-\bar{\mathbf{X}}_1^{\mathrm{T}}\mathbf{S}_1^{-1}\bar{\mathbf{X}}_1)\geq \mathrm{In}\left[\left(\frac{|S_2|}{|S_1|}\right)+2\mathrm{In}\left(\frac{\pi_2}{\pi_1}\right)\right] \qquad (3)$$

Otherwise to group D_2. π_1 and π_2 are probabilities of group membership for group one and two respectively. Since the left-hand side contains square and cross-product terms, it is termed the Quadratic disriminant function. Even if the assumption of multivariate normality is justified, use of Fisher's LDF may not be optimal for allocation due to heterogeneity of dispersion matrices (Krzanowski, 1977). The robustness of Hotelling's T^2 under heterogeneity of dispersions could provide an alternative to this problem. Gilbert (1969) investigated the behavior of Fisher's LDF by comparing it with that of the optimal quadratic form when the parameters of the two populations are assumed to be known. Gilbert restricted attention to the case where one variance-covariance matrix is a multiple, d, of the other. The result revealed that the Fisher's LDF may be satisfactory for classification but not for estimating risks of individuals belonging to a particular population. It becomes worse as number of variables increased.

2.2 Non-Normality.

One of the basic assumptions in discriminant analysis is that observations are distributed multivariate normal. In practical cases, this assumption is even more important in assessing the performance of Fisher's LDF in data which do not follow the multivariate normal distribution. Fisher's LDF has shown to be relatively robust to departure from normality. The non-normality of data could be as a result of the nature of the data. While some may be continuous but with joint distribution that is not normal, others may be discrete and each can assume only a finite number of values. Some could also be a mixture of both discrete and continuous. Although, various methods of data treatment such as logarithmic transformation, square root transformation, inverse transformation e.t.c. suggests normality yet optimality is not commonly met.

2.3 When Data are mixtures of Continuous and Categorical variables (Normal and otherwise).

Fisher's Linear Discriminant Function (FLDF) approach to dealing with mixtures of Continuous and Categorical variables is such that assigns codes or score to each state of the categorical variable and analysis performed with methods originally intended for continuous data. This implies that certain amount of distortion is likely to be introduced by treating the categorical variables as if they were continuous. Olkin et al (1961) and Krzanowski (1975, 1977, 1980, 1982) developed a method (henceforth referred to as Location Model) that handles this situation without making any such transformation to the categorical data. This procedure assumes that, $\mathbf{y}$, vector of continuous variables has a multivariate normal distribution with mean $\boldsymbol{\mu}_{i(m)}$ in cell m and population D_i (m=1,2,…,k; i=1,2), with common dispersion matrix $\boldsymbol{\Sigma}$ in all cells of both populations. The advantage of their proposed method over the conventional FLDF was evaluated using their respective average error rates under various scenarios. Scenarios such as varying sample sizes, varying number of discriminating variables (categorical and continuous), varying prior probabilities of group membership among others. The average error rate using FLDF was higher than that of using the Location model in all cases considered and for many parameter combinations. Thus, Krzanowski (1982) asserted that when there is evidence of interaction between categorical variables and populations, FLDF tends to give poorer result than the rule derived from the Location

model. An extreme case would arise when the continuous variable means differ between the populations for each cell but the marginal means are the same in the two populations. Oyeyemi et al (2013) investigated the performance of the FLDF (direct and stepwise), Location Model and Logistic Regression under various number of binary variables mixed with continuous variables. Oyeyemi et al revealed that the performance FLDF in such situation is very poor especially with few binary variables and large sample sizes.

2.4 When Data are Continuous but Non-normal

When continuous data are non-normal, an appropriate way of handling it is by making transformation to the data. The choice of the distributional transformation is relatively dependent on the situation at hand. Such approach could also be implored in discriminant analysis when continuous data are non-normal. Lachembruch et al (1973) considered three distributions generated from the normal distribution by using some non-linear transformation. Results indicated that Fisher's LDF was greatly affected by non-normality of the population. Error rates for one population were generally larger than the optimum values while the reverse was true in the other population and the sum of the two error rates increased for some distributions. Lachembruch et al concluded that, the use of Fisher's LDF in non-normal situations could be bad and misleading, and recommended that the data be transformed to approximate normality before the use of the LDF. Zhezhel (1968) considered the case of arbitrary distribution with equal covariance matrix when the continuous data are non-normal and calculated the maximum error rate from each population over this class of distributions. Results obtained shows that there are cases when the Fisher's LDF gives poorer result than the random classification. The performance of the Logistic regression has not been investigated in this situation possibly because the distribution makes no assumption about the observations.

2.5 When Data are Discrete

When data are discrete, Aitchson and Aitken (1976) and Titterington (1977) suggested the use of kernel density estimation, and conventionally, researchers assign arbitrary numerical score and proceed with procedures intended for continuous variables. However, the work by Gilbert (1968) and Moore (1973) using data generated from a first order interaction model indicated that care was needed in the selection of a discrimination procedure with binary variables. Log likelihood ration for the population was said to undergo traversal which Fisher's LDF didn't follow, hence, was found to have performed quite well.

2.6 A Case of Three Groups

With Fisher's LDF, when more than two groups are involved, the allocation rule described in (2) can no longer work. This has been extended following the already existing procedure of Fisher's LDF to provide an appropriate allocation rule. With, three groups, considering all possible combinations, without repetition, the allocation rule will be based on three functions:

$$h_{12}(x) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^{T}\mathbf{S}^{-1}\left[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2)\right] \quad (4)$$

$$h_{13}(x) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_3)^{T}\mathbf{S}^{-1}\left[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_3)\right] \quad (5)$$

$$h_{23}(x) = (\bar{\mathbf{X}}_2 - \bar{\mathbf{X}}_3)^{T}\mathbf{S}^{-1}\left[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{X}}_2 + \bar{\mathbf{X}}_3)\right] \quad (6)$$

The above functions assumes equal covariance matrix for the three groups but vector of means are different in each of the here groups. This is to avoid the violation of any of the underlying assumptions. Now, assuming no information is available about the cost of misclassification and a-priori probabilities, the classification rule derived from (3), (4) and (5) is to allocate an individual with vector of scores **x** to

$$D_1 \text{ if } h_{12}(\mathbf{x}) > 0 \text{ and } h_{13}(\mathbf{x}) > 0 \quad (7)$$

$$D_2 \text{ if } h_{12}(\mathbf{x}) < 0 \text{ and } h_{23}(\mathbf{x}) > 0 \quad (8)$$

$$D_3 \text{ if } h_{13}(\mathbf{x}) < 0 \text{ and } h_{23}(\mathbf{x}) < 0 \quad (9)$$

3. New Result (for case of more than two groups) And Its Discussion.

A natural extension of the m=3 groups to m=4 groups give rise to the following functions.

$$h_{12}(x) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^{\mathrm{T}}\mathbf{S}^{-1}\left[\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2)\right] \quad (10)$$

$$h_{13}(x) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_3)^{\mathrm{T}}\mathbf{S}^{-1}\left[\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_3)\right] \quad (11)$$

$$h_{14}(x) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_4)^{\mathrm{T}}\mathbf{S}^{-1}\left[\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_4)\right] \quad (12)$$

$$h_{23}(x) = (\bar{\boldsymbol{X}}_2 - \bar{\boldsymbol{X}}_3)^{T}\boldsymbol{S}^{-1}\left[\boldsymbol{x} - \tfrac{1}{2}(\bar{\boldsymbol{X}}_2 + \bar{\boldsymbol{X}}_3)\right] \quad (13)$$

$$h_{24}(x) = (\bar{\boldsymbol{X}}_2 - \bar{\boldsymbol{X}}_4)^{T}\boldsymbol{S}^{-1}\left[\boldsymbol{x} - \tfrac{1}{2}(\bar{\boldsymbol{X}}_2 + \bar{\boldsymbol{X}}_4)\right] \quad (14)$$

$$h_{34}(x) = (\bar{\boldsymbol{X}}_3 - \bar{\boldsymbol{X}}_4)^{T}\boldsymbol{S}^{-1}\left[\boldsymbol{x} - \tfrac{1}{2}(\bar{\boldsymbol{X}}_3 + \bar{\boldsymbol{X}}_4)\right] \quad (15)$$

And consequently, a set of allocation rules would emerge, thus, allocate a future observation to group

$$D_1 \text{ if } h_{12}(\mathbf{x}) > 0\,, h_{13}(\mathbf{x}) > 0 \text{ and } h_{14}(\mathbf{x}) > 0 \quad (16)$$

$$D_2 \text{ if } h_{12}(\mathbf{x}) > 0\,, h_{23}(\mathbf{x}) > 0 \text{ and } h_{24}(\mathbf{x}) > 0 \quad (17)$$

$$D_3 \text{ if } h_{13}(\mathbf{x}) > 0\,, h_{32}(\mathbf{x}) > 0 \text{ and } h_{34}(\mathbf{x}) > 0 \quad (18)$$

$$D_4 \text{ if } h_{14}(\mathbf{x}) > 0\,, h_{24}(\mathbf{x}) > 0 \text{ and } h_{34}(\mathbf{x}) > 0 \quad (19)$$

Since, allocation in this case involves a simultaneous consideration of more than two groups, it is imperative that the rule be consequently more than one for each possible allocation. From the above derivations and presentations, it is undoubtedly clear that as the number of group increases, functions and allocation rules increases correspondingly and the direction of the inequality signs would almost be impossible to state. This would naturally make allocation tedious and confusing. Situation could also arise in which the score for an individual to be classified do not satisfy one of the inequalities. It becomes more difficult to handle when the assumption of equal variance-covariance for all groups is violated as the resulting functions and allocation rule will be extremely lengthy thereby becoming almost impossible to evaluate. Generalization of the Fisher's LDF procedure for more than two groups exists in notation, however, the application is somewhat difficult and inconsistent (Tao et al., 2006) and the performance of the methods for several groups is not generally reliable (David, H., 1996). Although use of computer aided calculations and packages like Support Vector Machine (Vapnik, 1998), Pairwise Comparison (Hastie and Tibshirani, 1998), Multi-Class Objective Function (Weston and Watkins, 1998) among others have hidden this. However, its reality for theoretical purposes and consistency cannot be overlooked or undermined. The choice of the direction of the inequality signs in the allocation rules is out of an attempt to ensure that each pair of group combinations is considered on both lower and upper bounds of the optimum values. As Gilbert (1969) and Moore (1973) pointed out, Fisher's LDF is optimal when "two" populations have multivariate normal distributions with equal covariance matrices. It suffices to say that, even when the assumptions are rightly met, use of the Fisher's LDF under conditions of multi-groups may not be optimal.

Again, according to Press and Wilson (1978), classification of an observation into one of several population is discriminate analysis, while relating qualitative variables to other variables through a logistic (cumulative density function) functional form is logistic regression. Although estimates generated from one of these methods are often used in the other. However the conditions for the application of both are not the same. Discriminate function estimators have often been used in logistic regression in both theory and application (Turett et al 1967). Halperin et al (1971) reported that when discriminant function estimators were compared empirically with maximum likelihood estimators for logistic regression problems, they were found to be generally inferior, although not always by substantial amount. The procedure has performed averagely better than the Fisher's LDF in many situations because it is more like a non-parametric method which makes no assumption about the distribution of the data.

4. Conclusion

Performance of Linear Discriminant Function under some non-optimal conditions have been reviewed as it relates to few commonly used statistical methodology for discrimination and classification. A new result of another situation in which some conventional discriminant analysis procedures has failed to provide clear and explicit methodology has also been reported. Though, a theoretical and mathematical derivations and framework

which must be validated is in process to overcome some of the challenges that has limited optimal result in classification with regard to multiple groups, it is however, worthwhile to point attention to this considering its importance to the effort of obtaining a methodology for discriminant analysis that is efficient in every ramification. Thus, we conclude by suggesting that, in order to overcome the limitation posed to discriminant analysis involving more than two groups with regard to the conventional FLDF, each group should have its own corresponding allocation rule and allocation done by considering pairs of groups. Allocation by pairing groups ensures that the underlying principle of LDA is maintained while the procedures for allocation differ. This will be case confronting the practicing statistician who wants to decide whether he can use existing discriminant analysis procedure, or whether some other procedure will give better results.

References

Anderson, T. W. (1951) Classification by Multivariate Analysis. Psychometrika, 16, 631-650

Aitchson, J. and Aitken, C. G. G. (1976) Multivariate binary discrimination by the Kernel method. Biometrika, 63, 413-420

David, H. (1996). Error-Rate Estimation in Multiple-Group Linear Discriminant Analysis. Technometrics, 38, 389-399

Fisher, R. A. (1936). The use of Multiple Measurements in Taxonomic Problems. Ann. Eugen., 7, 179-188

Gilbert, E. S. (1968). On Discrimination using Qualitative Variables. Journal of American. Statistical Association, 63, 1399-1412.

Hastie, T. & Tibshirani, R. (1998). Classification by Pairwise Coupling. In M. I. Jordan, M. J. Kearns, & S. A. Solla (eds.), Advances in Neural Information Processing Systems. The MIT Press.

Halperin, M, Balckweldeer, W. C. and Verter, J. I (1971) Estimation of the Multivariate Logistic Risk Function: A comparison of the Discriminant Function and Maximum likelihood approaches. Journal of chronic Diseases, 24, 125 – 158.

Hills, M (1967) Discrimination and allocation with Discrete Data. Applied Statistics, 16, 237-250.

Krzanowski, W. J. (1975) Discrimination and Classification Using Both Binary and Continuous Variables. Journal of the American Statistical Association, 352, 782-790

Krzanowski, W. J. (1977) The Performance of Fisher's Linear Discriminant Function under Non-Optimal Conditions. Technometrics, 19, 191-200.

Krzanowski, W. J. (1980) Mixtures of Continuous and Categorical Variables in Discriminant Analysis. Biometrics, 36, 493-499

Krzanowski, W. J. (1982) Mixtures of Continuous and Categorical Variables in Discriminant Analysis; A hypothesis testing approach. Biometrics, 38, 991-1002.

Lachembruch, P. A., Sneeringer, C. and Revo, L. T. (1973). Robustness of the Linear and Quadratic Discriminant Functions to certain types of Non-normality. Computational Statistics, 1, 39-56.

Moore, D. H. (1973). Evaluation of five Discrimination procedures for Binary Variables. Journal of American Statistical Association, 68, 399-404.

Olkin, I. and Tate, R. F. (1961) Multivariate Correlation Models with Mixed Discrete and Continuous variables. Annals of Mathematical Statistics, 32, 448-465

Oyeyemi, G. M. and Mbaeyi, G. C. (2013). On Discrimination Procedure with mixtures of Continuous and Categorical Variables. Unpublished M.Sc thesis, Department of Statistics, University of Ilorin, Ilorin, Kwara State.

Press, S. J. and Wilson, S. (1978) Choosing between Logistic Regression and Discriminant analysis. Journal of the American Statistical Association, 73, 699-701

Tao, L., Shenghou, Z. & Mitsunori,O. (2006) Using Discriminant Analysis for Multi-class Classification: An Experimental Investigation. Knowledge and Information Systems, 10, 453-472

Titterington, D. M. (1977) Analysis of Incomplete Multivariate Binary Data by the Kernel Method. Biometrika, 64, 455-460

Truett, J., Cornfield, J., and Kannel, W. (1967) A Multivariate analysis of the risk of Coronary Heart Disease in Framingha. Journal of chronic Diseases, 20, 511-524.

Vapnik, V. N. (1998). Statistical Learning Theory. Wiley, New York.

Wald, A. (1944). On a Statistical Problem arising in the Classification of an Individual into two Groups. Annals of Mathematical Statistics, 15, 145-163

Zhezhel, Yu. N. (1968). The Efficiency of a Linear Discriminant Function for Arbitrary Distributions. Engineering Cybermetics, 6, 107-111.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar