ISSN 2347-1921
Volume 13 Number 4
Journal of Advance in Mathematics

Comparative Linear Classification Splicing

[1]F. Z. Okwonu and [2]N. O. Ogini
[1]Department of Mathematics Faculty of Science, Delta State University, Abraka
[2]Department of Computer Science Faculty of Science, Delta State University, Abraka, Nigeria
fzokwonu_delsu@yahoo.com

Abstract

The conventional Fisher linear classification analysis has been investigated by numerous researchers and this has led to different modification or splicing due to non- robustness when the assumptions are violated and also when the data set contains influential observations. This paper adduced a winsorized procedure to robustify the probability base classification approach. The comparative classification performance of the Fisher linear classification analysis and its spliced versions when the data set are contaminated are investigated. The simulation results revealed that the robust Fisher's approach based on the minimum covariance determinant estimates outperformed the other procedures; a good competitor to this technique is the winsorized probability base classification technique. Though, the robust Fisher's technique using the minimum covariance determinant estimates breakdown for mixture contamination. On a general note, the conventional Fisher's approach and the probability base technique performed comparable.

Introduction

Linear classification techniques such as the Fisher's method have been studied extensively and its modifications have no restriction in multivariate statistics. The performance of the Fisher's technique will be inaccurate and inefficient if the conventional assumptions are violated(Wang 2005). The Fisher's method is basically designed in such a way that maximizing the between group scatter and minimizing the within group scatter allows for maximum separation of the groups. This concept can easily be visualize graphically (Johnson 2007) and allows the unknown sample observation to be accurately assigned to the exact group. Though, as a dimension reduction procedure, the Fisher linear classification rule outperformed the principal component analysis (PCA) and other dimension reduction techniques(Chiang 2004). Even though, the Fisher's approach is robust against the PCA, upon assumption violation, the Fisher's technique still underperformed if the data set contains influential observations. It has been extensively robustified by using the plug-in and data transformation procedures.

Conventionally, linear classification technique such as the Fisher linear classification rely on the training set to build the model and due to upward biased, the training set is not applied to validate the model. In such scenario, the test data set is used to validate the model in other to obtain unbiased result. Generally, linear classification rule is designed such that the unknown observation is accurately assigned to well defined group(Johnson 2007). Though different robust linear and quadratic classification rules have been proposed using the plug-in technique (Croux 2008; Wang 2014; Chen 1994; Hubert 2011a; Crimin 2007). Wang (2004) applied the winsorized approach to robustify the Fisher discriminant analysis. The crux of it is that the winsorized approach renders the influential observations ineffective.

The minimum covariance determinant (*MCD)* procedure is also a highly robust multivariate estimator and has bounded influence function. It is applied to robustify classical sample mean and covariance matrix used as a replacement parameter for classical estimators of multivariate techniques such as principal component analysis, factor analysis and discriminant analysis (Croux 2000; Hawkins 1997). The *MCD* estimators are usually used as initial estimators when considering two stage techniques. The minimum covariance estimator (*MCD*) (Rousseeuw P. J. 1985) were proposed to robustify the sample mean vectors and covariance matrices. The robustified sample mean vectors and covariance matrices are plug-in into the conventional multivariate procedures to obtain robust multivariate techniques including the Fisher's technique. It has been applied to robustify the linear discriminant analysis and the quadratic discriminant analysis by Hubert and Van Driessen(Hubert 2004b). The procedure strictly depends on information gleaned from the half set. The half set does not utilize the entire data set. This procedure is a data cleaning technique that is used as a preprocessing step before being applied to the technique of interest. This procedure requires concentration steps, detail of this robust high breakdown method and its application to classification is contained in (Rousseeuw 1999b).

Robustification of any classical classification techniques arises from poor classification performance if the data set contains influential observations or assumption violation. For instance, the under classification performance of the Fisher linear classification rule is due to estimation errors of the mean vectors and covariance matrices(Pohar 2004). The high misclassification rate of the conventional linear classification procedures is due to instability of the sample mean and covariance matrix (Maronna 2006; Munoz-Pichardo 2011). Wang (2005) applied the M-estimate to transform the sample observations in other to reduce the influence of the multivariate observations thereby allowing the retainance of the in-lier observations.

The objective of linear classification splicing is to modify existing linear classification technique(s) to achieve more effective classification results. This involves introducing different concept to existing procedure to output better performance. In this paper, the objective is to robustify the conventional Fisher and the probability base linear classification rules by applying the winsorized procedure. The procedure is straight forward in the sense that we substitute the winsorized mean for the classical mean, same follows in computing the covariance matrices respectively. In practice, the winsorized mean is insensitive to influential observations and hence is regarded as a robust estimator(Wilcox 2003).

The reminder of this paper is organized as follows. Section Two briefly describes the Fisher linear classification rule. Section Three contains the Fisher approach using the minimum covariance determinant estimates. The probability base

classification rule is described in Section Four. The winsorized probability base classification rule is presented in Section Five. Simulation and analysis is contained in Section Six while conclusion is contained in Section Seven.

Fisher linear classification analysis(FLCA)

The Fisher linear classification analysis (FLCA)(Johnson, 2007) is a dimension reduction approach which encompass separation and classification. This procedure is a linear combination of measured variables that best describe the allocation of individual or observation to well defined groups. The mean, covariance matrice and the pooled covariance matrix are the building block of this technique. This procedure, assumed that the variance covariance matrices are homoscedastic, the data set are Gaussian in nature and the mean vectors are not equal for each group. The coefficient of this procedure is obtained by post-multiplying the inverse of the pooled covariance matrix by the within group mean vector difference (Okwonu, 2016). To allow separation and allocation, the coefficient must be non zero(Rencher 2002). Allocation of observation to well defined group in the case of group allocation is done by comparing the classification score with the cutoff point. To this end, an observation is assigned to group one if the classification score is greater than the cutoff point otherwise, the observation is assigned to group two.

Fisher linear classification based on minimum covariance determinant(FMCD)

The minimum covariance determinant (*MCD*) estimator proposed by Rousseeuw (1985) like the minimum volume ellipsoid estimator minimizes the determinant of the covariance matrix. This robust multivariate estimator of mean vector and covariance matrices are insensitive to outlying observations (Gervini 2003; Rousseeuw 1990).The minimum covariance determinant procedure search for the subset h_i (out of n_i) of the data set whose covariance matrix has the minimum determinant(Hubert 2004b). The sample observations based on the half set h_i are chosen from the multivariate data set to obtain the *MCD* estimates. The half set h_i strictly depends on the arithmetic of the sample size and sample dimension. These robust estimates are computed based on the clean data set selected by the half set. The performance of this estimator depends on the half set. The *MCD* rely only on the half set not the entire data set. The half set h is constrained in the sense that if the half set is too large, the *MCD* estimator is not robust; however, if the half set is too small the estimator tends to underestimate it accuracy. The robust *MCD* estimates of mean vectors and covariance matrices are plug-in into the Fisher's Equations to obtain the robust Fisher linear classification rule(Hubert 2004b). Detail of this procedure is contained in (Croux 1999; Fauconnier 2009; Hubert 2011b, 2008; Pison 2002; Rousseeuw 1999a; Maronna 2006). Although, the minimum covariance determinant is computed based on the *FAST-MCD* algorithm of Rousseeuw and Van Driessen (1999). Detailed description and theorem to compute the concentration steps of the half set is contained in (Rousseeuw 1999a). The classification procedure allows an observation to be assigned to the correct group; in this case, if the classification score is greater than the cutoff point the observation is assigned to group one otherwise the observation is allocated to the second group (Okwonu,2013).

Probability base linear classification technique(PCT)

Conventionally, the coefficient of the Fisher's technique is based on the within group mean vector difference and the inverse of the pooled covariance matrix. The probability base classification technique(Okwonu 2013) only utilize the within group mean vectors difference for the two groups, say, $d = \bar{x}_1 - \bar{x}_2$, and the sum of the within group mean vectors is given as $\hat{d} = \bar{x}_1 + \bar{x}_2$. To formulate the coefficient of this procedure, the following are obtained;

$$\begin{aligned} \partial &= |d|, \\ \tilde{d} &= 1 + \sqrt{|\hat{d}|}, \\ \beta &= d^2 / \tilde{d}, \\ \varepsilon &= 1 - |\beta^2|. \end{aligned} \tag{1}$$

Based on the definitions in Equation (1), the following is obtained

$$w = e^{\beta} + e^{\beta^2/\varepsilon} + p_i, \tag{2}$$

where $p_i = N_i / N$, is the within group prior probabilities, p_i satisfy the following conditions, $p_i > 0$, and $\sum_{i=1}^{g=2} p_i = 1$. N_i is the sample size for each group, N is the total sample size for the two groups and $p = \sum_{i=1}^{2} p_i$, is the total probability. The component of the decision rules are described as follows;

$$z = \left(\frac{w}{S_{pooled}^{-1}} \right)' x = u'x, \qquad u = \frac{w}{S_{pooled}^{-1}}. \tag{3}$$

Equation (3) describes the classification score z and Equation (4) is the decision point $\bar{z}$,

$$\bar{z} = \frac{\hat{d}}{2} u', \tag{4}$$

Based on the above analysis, an observation in group one is correctly assigned to group one if the classification score is less than the decision point otherwise, the observation is assigned to group two.

Winsorized probability base classification technique(WPCT)

The conventional probability base classification rule discussed previously, utilized the conventional sample mean which may be susceptible to influential observations. In this section, we proposed its modification by using the winsorized mean. That is, instead of applying the classical sample mean and covariance matrices computed from the data set, the winsorized technique is applied to winsorized the data set and hence computes the winsorized mean and covariance matrix. These robust estimates are plug-in in place of the classical mean and covariance matrices respectively. The winsorized mean and covariance are applied to build the model. This technique is straight forward since is strictly based on plug-in approach. Like the previous procedures, the test data is used to validate the model. Allocation of observation is done in the same manner as in the previous methods. The objective is to investigate if the winsorized approach is robust over the conventional method.

Simulation and analysis

In this section, the comparative performance of the above techniques including the robust Fisher's technique based on the minimum covariance determinant (MCD) (Rousseeuw 1999a; Hubert 2004a) are investigated. To investigate the comparative performance of these techniques, the data set are generated based on contaminated normal model $(1-\varepsilon)N_d(0,1_d)+\varepsilon N_d(\mu,\sigma I_d)$ in which majority of the data set come from the normal distribution while the rest are obtained from a contaminated normal distribution. This is done by varying the mean vectors and the variance. The level and proportion of contamination strictly depends on the mean vectors,variance and ε. we intend to mimic the contamination procedure in(Khan 2007). In all, the sample size N_i for all group are equal, where d is the dimension of the sample size and $i(i=1,2)$ is the number of groups, respectively. The robustness of this methods are determined by comparing the mean probability of correct classification with the mean of the optimal probability of correct classification computed based on the data set generated from the normal distribution. The performance analyses are shown in the following figures. The data set for Fig. 1 below is generated based on the contaminated normal model that is asymmetric. That is shift normal contamination $(1-\varepsilon)N_d(0,1)+\varepsilon N_d(20,1).$

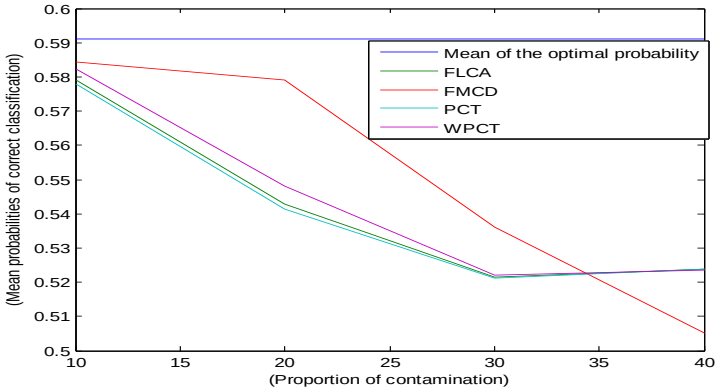

Fig. 1 classification performance based on shift contamination

ISSN 2347-1921
Volume 13 Number 4
Journal of Advance in Mathematics

The data set for Fig. 2 below is obtained based symmetric normal contamination $(1-\varepsilon)N_d(0,1)+\varepsilon N_d(0,20)$. The sample size is equal for all groups with corresponding three dimension.

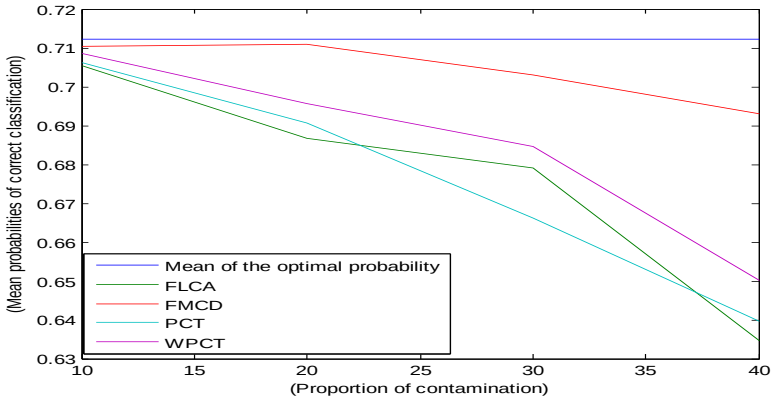

Fig.2 classification performance based on symmetric contamination

Fig.3 below is obtained from data set generated from asymmetric, shift normal contamination $(1-\varepsilon)N_d(0,1)+\varepsilon N_d(50,1)$.

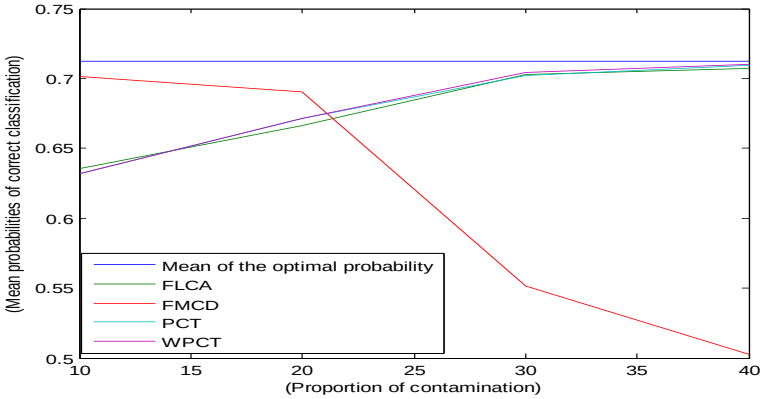

Fig. 3 classification performance based on asymmetric contamination

Fig. 4 below is obtained based on mixture contamination in which the mean and variance are equal,say $(1-\varepsilon)N_d(0,1)+\varepsilon N_d(10,10)$.

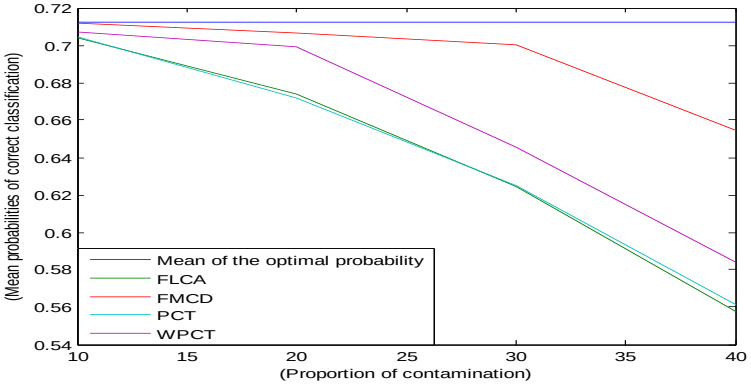

Fig.4 classification performance based on mixture contamination

Conclusion

The winsorized probability base classification rule requires large sample size to compute its estimates; the conventional probability base approach is advantageous in this aspect. For the shift contamination with small proportion of contamination, the FMCD approach outperformed the other techniques but as the proportion of contamination increases, the other procedures outperformed the FMCD. For the asymmetric contamination, the other classification techniques outperformed the FMCD, while the WPCT method outperformed the other techniques for mixture contamination except the FMCD. The comparative analysis revealed that the different techniques performed differently depending on the contamination model the data set are derived.

References

1. Chen, Z. Y., and Muirhead, R. J., 1994, A comparison of robust linear discriminant procedures using projection pursuit methods, *Multivariate analysis and its applications* 24, 163-176.
2. Chiang, L. H., Kotanchek, M. E., and Kordon, A. K., 2004, Fault diagnosis based on Fisher discriminant analysis and support vector machines, *Computer and Chemical Engineer* 28, 1389-1401.
3. Crimin, K., McKean, J. W., and Sheather, S. J., 2007, Discriminant procedures based on efficient robust discriminant coordinates, *journal of nonparametric statistics* 9, 199-213.
4. Croux, C., and Haesbroeck, G.,, 1999, Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, *Journal of multivariate analysis* 71, 161-190.
5. Croux, C., and Haesbroeck, G., , 2000, Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies, *Biometrika* 87, 603-618.
6. Croux, C., Filzmoser, P., and Joossens, K., 2008, Classification efficiencies for robust linear discriminant analysis, *Statistica Sinica* 18, 581-599.
7. Fauconnier, G., and Haesbroeck, G.,, 2009, Outliers detection with minimum covariance determinant estimator in practice, *Statistical methodology* 6, 363-379.
8. Gervini, D., 2003, A robust and efficient adaptive reweighted estimator of multivariate location and scatter, *Journal of multivariate analysis* 84, 116-144.
9. Hawkins, D. M., and Mclachlan, G. J.,, 1997, High breakdown linear discriminant analysis, *Journal of the American Statistical Association* 92, 136-143.
10. Hubert, M., and Van Driessen, K., 2004a, Fast and robust discriminant analysis, *Computational Statistics and Data Analysis* 45, 301-320.
11. Hubert, M., and Van Driessen, K.,, 2004b, Fast and robust discriminant analysis, *Computational Statistics and Data Analysis* 45, 301-320.
12. Hubert, M., Rousseeuw, P. J., and Van Aelst, S.,, 2008, High breakdown robust multivariate methods, *Statistical Science* 23, 92-119.
13. Hubert, M., Rousseeuw, P. J., and Verdonck, T., 2011a, A deterministic algorithm for the *MCD*, *Citeseerx.ist.psu.edu/viewdoc/summary?*, 1-26.
14. Hubert, M., Rousseeuw, P. J., and Verdonck, T.,. 2011b, A deterministic algorithm for the MCD.
15. Johnson, R. A., and Wichern, D. W., 2007, Applied multivariate statistical analysis (Pearson Prentice Hall, Upper Saddle River.
16. Khan, J. A., Van Aelst, S. and Zamar, H. R., 2007, Robust linear model selection based on least angle regression, *Journal of the American Statistiscal Association* 102, 1289-1299.
17. Maronna, R., Martin, R. D., and Yohai, V. J.,, 2006, Robust statistics: Theory and methods (John Wiley, New York.
18. Munoz-Pichardo, J. M., Enguix-Gonzalez, A., Munoz -Garcia, J., and Moreno-Rebollo, J.L.,, 2011, Influence analysis on discriminant coordinates, *Communications in statistics-simulation and computation* 40, 793-807.
19. Okwonu, F. Z., and Othman, A. R., 2013, Probability base classification technique: A preliminary study for two groups, *Journal of Mathematical Theory and Model* 3, 40-46.
20. Okwonu, F. Z., 2016. Supervised difference linear classification techniques for two group's problem. Nigerian Journal of Science and Environment,Vol.13 (1), 111-116.
21. Okwonu, F.Z., 2013. Comparison of several robust unbiased linear classification techniques for two groups. Unpublished manuscript, USM.
22. Pison, G., Van Aelst, S., and Willems, G.,, 2002, Small sample corrections fot LTS and MCD, *Metrika* 55, 111-123.

23. Pohar, M., Blas, M., and Turk. S., 2004, Comparison of logistic regression and linear discriminant analysis: A simulation study, *Metodoloski zvezki* 1, 143-161.
24. Rencher, A. C., 2002, A methods of multivariate analysis (A John Wiley & Sons, Inc.
25. Rousseeuw P. J., 1985, Multivariate Estimators With High Breakdown Point, *Mathematical Statistics and its Applications* B, 283-297.
26. Rousseeuw, P. J., and Van Driessen, K., 1999a, A fast algorithm for the minimum covariance determinant estimator, *Technometrics* 41, 212-223.
27. Rousseeuw, P. J., and Van Driessen, K.,, 1999b, A fast algorithm for the minimum covariance determinant estimator, *Technometrics* 41, 212-223.
28. Rousseeuw, P. J., and Van Zomeren, B. C.,, 1990, Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association* 85, 633-651.
29. Wang, D., and Romagnoli, J. A., 2005, A robust discriminate analysis method for process fault diagnosis, *European Symposium on Computed Aided Process Engineering-15*, 1-6.
30. Wang, Y., Zhang, Y.,Yi, J., Qu, H., and Miu, J., 2014, A robust probability classifier based on the modified x^2 - distance, *Mathematical Problems in Engineering* 2014, 1-11.
31. Wilcox, R. R., and Keselman, H. J., 2003, Modern robust data analysis method: Measures of central tendency, *Psychological Methods* 8, 254-274.