Robust MLCR Linear Classification Technique: An Application to Classify Aede Albopictus Mosquito

Friday Zinzendorf Okwonu[1,2] and Abdul Rahman Othman[2]

[1]Department of Mathematics and Computer Science, Delta State University, Abraka, Delta State, Nigeria

[2] Center for Mathematical Sciences, Universiti Sains Malaysia, 11800, Pulau Penang, Malaysia

Abstract

The classical Fisher linear classification analysis is a well known linear classification procedure. This technique tends to minimize the misclassification error if the data set comes from a multivariate normal distribution. On the other hand, if the data set is contaminated, the misclassification error of this approach tends to increase. Relying on the classification performance of the classical Fisher's technique for contaminated data set, robust Fisher linear classification analysis based on the minimum covariance determinant estimators was proposed. The objective of this procedure is to obtain maximum classification rate when the data set contain influential observations. This procedure only depends on the number of sample observations selected by the half set. The performance of the robust Fisher's procedure strictly depends on the half set. Considering the classification performance of the classical and robust Fisher's techniques, a robust *M* linear classification rule that utilizes all the entire data set is proposed to compare the classification performance of the above methods.

Keywords: *Classification, Mean Probability, Robust.*

1. Introduction

The Fisher linear classification analysis (FLCA) was introduced by Fisher (1936) when he applied it to study the Iris data set for two groups. This technique was developed when the sample size is greater than the sample dimension for each group. Its basic assumptions are homoscedasticity of the covariance matrices and normality of the data set. A comparable classification rule to the Fisher linear classification analysis and the Wald-Anderson classification rule was subsequently proposed [1, 2].

The classical multivariate techniques including the FLCA was proposed based on the sample mean vectors and covariance matrices. The mean vectors and covariance matrices are the building blocks of most classical multivariate techniques but are susceptible to influential observations [3-11]. However, the classical multivariate techniques performs optimally if the mean vectors and covariance matrices are computed from normally distributed data [12, 13]. Based on the shortcoming of the mean vectors and covariance matrices, robust techniques were proposed to robustify the mean vectors and covariance matrices. In this paper, we consider robust Fisher technique based on the minimum covariance determinant estimators (MCD). The *MCD* is a robust multivariate outlier identification and rejection technique that depends strictly on the number of sample observations selected by the half set [14-17]. It is applied to robustify the mean vectors and covariance matrices used in developing the robust Fisher linear classification analysis[17]. The application of this robust procedure was generally accepted when the *FAST-MCD* algorithm of Rousseeuw (1999) was introduced [18].

In general, irrespective of the drawbacks relating to the minimum covariance determinant, it still the method of choice to obtain robust estimates and this is due largely to the availability of the *FAST-MCD* algorithm in most statistical packages. In this paper, we consider robust high breakdown linear classification rule that does not downweight the influential observations. This procedure is based on the within group median. The covariance matrices of this procedure is not pooled rather the covariance matrix is summed to obtain the separation parameter. These procedures are applied to classify the laboratory reared *Aedes Albopictus* mosquitoes as male or female using body size (wing length) measurement.

The organization of this paper begins with the Fisher linear classification analysis in Section 2. Section 3 contains the Fisher linear classification analysis based on the minimum covariance determinant estimators. The robust *M* linear classification rule is described in section 4. Simulation and conclusion are contained in Sections 5 and 6.

2. Fisher Linear Classification Analysis

The mean vectors and covariance matrices are computed from the training samples. These parameters are applied to develop the FLCA. Based on the information provided by the FLCA via the training or validation samples, the objective is to classify an observation as belonging to one of the two populations. The Fisher linear classification rule[19] for two groups problem is defined mathematically as follows,

$$\mathbb{Q} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{\text{pooled}}\mathbf{x} = \mathbf{q}'\mathbf{x}, \quad (1)$$

$$\bar{\mathbb{Q}} = \frac{(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{\text{pooled}}. \quad (2)$$

Where Eq.1 is the Fisher linear classification score $\mathbb{Q}$, $\mathbf{q}$ is the Fisher linear coefficient, $\bar{\mathbf{x}}_i, i = 1, 2,$ is the within group mean vectors, $\mathbf{x}$ is the sample observation, $\mathbf{S}_{pooled}$ is the pooled sample covariance matrices for the two groups and Eq. 2 is the Fisher's comparative average point $\bar{\mathbb{Q}}$.

The computation of the Fisher linear coefficient $\mathbf{q}$ is possible if the group mean vectors are unequal and the sample size for each group is greater than the sample size dimension. This condition is vital to enable separation and classification feasible. The classification score allow visual inspection of the separation between known groups of observations [20]. The separations between the group mean vectors can also be visualized by inspection. This characteristic of the Fisher linear classification analysis is very vital for classification purposes when the population have common covariance matrices. The Fisher linear classification rule is obtained by comparing the classification score with the classification average point. Allocation of an observation to either of the groups based on the Fisher linear classification rule can be described as follows; an observation $\mathbf{x}_1$ is assign to group one Ω_1 if the classification score $\mathbb{Q}$ is greater than or equal to the classification average point $\bar{\mathbb{Q}}$ otherwise the observation $\mathbf{x}_1$ is assign to group two Ω_2 if the classification score $\mathbb{Q}$ is less than the classification average point $\bar{\mathbb{Q}}$.

3. Fisher Linear Classification Analysis Based on the Minimum Covariance Determinant (FMCD)

The *MCD* is not a classification technique rather it is a robust outlier identification and rejection technique. The *MCD* search for the subset h_i (out of N_i) of the data set whose covariance matrix has the minimum determinant[21]. The sample observations based on the half set are chosen from the multivariate data set to obtain the *MCD* estimates of location and scatter. These robust estimates are computed based on the clean data set selected by the half set. The robust *MCD* estimates of location and scatter are plug-in into Eq.1 and Eq.2 to obtain the robust Fisher linear classification rule[21]. This approach can be express mathematically as follows,

$$\mathrm{R}_{\mathbb{Q}} = (\bar{\mathbf{x}}_{mcd1} - \bar{\mathbf{x}}_{mcd2})'\mathbf{S}^{-1}_{mcdpooled}\mathbf{x} = \mathbf{q}'_{mcd}\mathbf{x}, \quad (3)$$

$$\mathrm{R}_{\bar{\mathbb{Q}}} = \frac{(\bar{\mathbf{x}}_{mcd1} + \bar{\mathbf{x}}_{mcd2})}{2}(\bar{\mathbf{x}}_{mcd1} - \bar{\mathbf{x}}_{mcd2})'\mathbf{S}^{-1}_{mcdpooled}. \quad (4)$$

The above equations constitute the robust Fisher's technique based on the estimates computed from the minimum covariance determinant procedure. The parameter $\mathrm{R}_{\mathbb{Q}}$ denote the robust Fisher linear classification score, $\bar{\mathbf{x}}_{mcdi}, i = 1, 2$ are the robust within group mean vectors, $\mathbf{S}_{mcdpooled}$ denote the pooled common sample covariance matrices, $\mathbf{q}_{mcd}$ is the robust linear classification coefficient and $\mathrm{R}_{\bar{\mathbb{Q}}}$ is the robust cutoff point. Detail description and computation of the minimum covariance determinant procedure is contained in [6, 14, 22-26].

The classification procedure is describe as follows; an observation $\mathbf{x}_1$ in group one Ω_1 is classify to group one Ω_1 if the following condition is satisfy, $\mathrm{R}_{\mathbb{Q}} - \mathrm{R}_{\bar{\mathbb{Q}}} \geq 0,$ otherwise the observation $\mathbf{x}_1$ is assign to group two Ω_2 if the following condition hold, $\mathrm{R}_{\mathbb{Q}} - \mathrm{R}_{\bar{\mathbb{Q}}} < 0.$

4. *M* Linear Classification Rule (MLCR)

The objective of this procedure is to obtain stable linear classification coefficient in order to minimize the misclassification error rate. It has been observed that unstable linear classification coefficient allows for high misclassification rate. To achieve the above objective, we modify the conventional linear classification coefficient. In this regard, this approach uses the square root of the summed covariance matrices to obtain the linear

classification coefficient. The components of the *M* linear classification method consist of the within group median $\ddot{\mathbf{x}}_i$ and the inverse of the square root of the summed covariance matrices which are apply as a tool for group separation.

The median is a robust affine equivariant estimator. Like every other robust high breakdown and affine equivariant estimator, the median has low efficiency due to its high breakdown point. As noted by [27], the median has bounded influence function. Stromberge (1997)[28] observed that the probability of the median taking the influential observation as it center is equal to the probability of taking the regular observation as it center. The propose technique is described as follows,

$$\mho = (\ddot{\mathbf{x}}_1 - \ddot{\mathbf{x}}_2)\left(\sqrt{(\sum_{i=1}^{2} \mathrm{S}_i)}\right)^{-1} \mathbf{x} = \varsigma' \mathbf{x}, \qquad (5)$$

$$\mathrm{S}_i = \frac{\sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \ddot{\mathbf{x}}_i)(\mathbf{x}_{ij} - \ddot{\mathbf{x}}_i)'}{(N_i - 1)}, (i = 1, 2).$$

From the above equations, the unbiased sample covariance matrices S_i are computed based on the within group median of the sample observations. Unlike the conventional procedure, in this technique the variation between the within group median is obtain and post multiply by the inverse of the square root of the summed covariance matrices to obtain the coefficient ς, where $\mho$ is the linear classification score. The comparative cutoff point $\bar{\mho}$ is defined as

$$\bar{\mho} = \frac{(\ddot{\mathbf{x}}_1 + \ddot{\mathbf{x}}_2)}{2}\varsigma. \qquad (6)$$

The classification rule is obtained by comparing the classification score $\mho$ with the comparative cutoff point $\bar{\mho}$, say,

$$\mho \geq \bar{\mho}, \qquad (7)$$

Eq.7 implies that an observation in group one Ω_1 is correctly assign to group one Ω_1 otherwise the observation is assign to group two Ω_2 if the following equation is satisfied,

$$\mho < \bar{\mho}. \qquad (8)$$

5. Simulation

The aim of this simulation is to verify if the reared aede *albopictus* mosquitoes were correctly assigned to the groups it belongs based on the wing length measurements obtained. This simulation is performed to classify the laboratory reared *aede albopictus* mosquitoes into two groups (male or female) using the above procedures. According to the laboratory setup and analysis, Group one Ω_1 contains 100 unknown aede *albopictus* mosquitoes and group two Ω_2 contains 200 unknown aede *albopictus* mosquitoes. In each group, 15 adult mosquitoes were randomly selected and the wing length was measured to determine the body size. The preferred predictor variables are x_1 for large wing length and x_2 for small wing length. The point is that we are interested to apply these linear classification procedures to predict if their measurements were accurate. The data set was reshuffled using the uniform distribution $U(0,1)$ and divided into training set (60%) and validation set (40%). Detail of the experimental setup is contained in [29]. The classification mean probability and standard deviation (in bracket) of correct classification are reported in Table 1. The results reported are based on 1000 replications.

Table 1: Mean probability and standard deviation (in bracket) of correct classification: laboratory reared aede albopictus mosquito data

FLCA	FMCD	MLCR
0.9998 (0.0005)	0.9415 (0.0106)	1.0000 (0.0000)

The classification results in Table 1 indicate that MLCR and FLCA have better recognition and classification rate than FMCD. The classification result showed that FLCA and MLCR agreed with the measurement that body size via the wing length can be used to determine the gender of aede *albopictus* mosquito. Thus, FMCD learn towards agreeing to the above conclusion.

The classification results reported in Table 1 are results obtained from the original data set. In the following, the objective is to investigate how the above linear classification techniques perform if influential observations are introduced into the data set. It is interesting to note that the classical Fisher linear classification rule performed well as well as the robust techniques for the above data set. In what follows, we introduced three influential observations into the data set to investigate the performance of these linear classification rules. In group one, we introduced one influential observation and group two, two influential observations, respectively. The aim is to observe if the excellent classification results obtained in Table 1 can be repeated by the various linear classification procedures. Table 2 contains the classification results for this experiment. The classification results from Table 2 indicate that influential observations affect the performance of the linear classification rules. The classification difference tells us how robust the proposed technique is to the classical

procedures compared against. Based on this data set, the MLCR is the best linear classification rule compared to the other linear classification rules with classification difference 0.2412, 0.0485 over FLCA and FMCD. The classification difference of the FMCD over FLCA is 0.1927.

Table 2: Mean probability and standard deviation (in bracket) of correct classification: contaminated laboratory reared aede albopictus mosquito data

FLCA	FMCD	MLCR
0.7248 (0.0095)	0.9175 (0.0125)	0.9660 (0.0081)

The classification results in Table 2 indicate that MLCR has better recognition rate than the other linear classification methods considered. Comparing the classification performance of these techniques for the two Tables, we conclude that the proposed linear classification rule performance better than the other techniques considered for this data set.

6. Conclusions

From the simulations, we observed that the mean probabilities of correct classification for FLCA and MLCR procedures are greater than that of the FMCD procedure for the original data set. However, for the contaminated data set, the mean probability of correct classification for the FMCD approach is greater than that of the FLCA technique. Overall, the MLCR method has better recognition rate than the other linear classification techniques. The analysis revealed that the proposed technique performed better than the known techniques for this data set.

Acknowledgments

This research work was funded through the short term grant of the Universiti Sains Malaysia.

References

[1] A. Kudo, "The classificatory problem viewed as a two-decision problem," *Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics*, vol. 13, pp. 96-125, 1959.

[2] A. Kudo, "The classificatory problem viewed as a two decision problem II," *Memoirs of the Faculty of Science, Kyushu University, Series A*, vol. 14, pp. 63-83, 1960.

[3] P. Filzmoser, and Hron, K., , "Outlier detection for compositional data using robust methods," *Mathematical Geosciences*, vol. 40, pp. 233-248, 2008.

[4] I. Basak, "Robust *M*-estimation in discriminant analysis," *The Indian Journal of Statistics*, vol. 60, pp. 246-268, 1998.

[5] S. J. Devlin, Gnanadesikan, R., and Kettenring, J. R.,, "Robust estimation of dispersion matrices and principal components," *Journal American Statistiscal Association*, vol. 76, pp. 354-362, 1981.

[6] M. Hubert, Rousseeuw, P. J., and Van Aelst, S.,, "High breakdown robust multivariate methods," *Statistical Science*, vol. 23, pp. 92-119, 2008.

[7] J. Jin, and An, J.,, "Robust discriminant analysis and its application to Identify protein coding regions of rice genes," *Mathematical Biosciences*, vol. 232, pp. 96-100, 2011.

[8] S.-J. Kim, Magnani, A., and Boyd, S. P.,, "Robust Fisher discriminant analysis," *Advances in Neural Information Processing System*, vol. 18, pp. 659-666, 2005.

[9] A. M. Pires, and Branco, J. A.,, "Generalization of Fisher's linear discriminant," *www.math.ist.utl.pt/~apires/PDF/APJB_RP96.pdf*, 1996.

[10] E. Roelant, Van Aelst, S., and Williems, G.,, "The minimum weighted covariance determinant estimator," *Metrika*, vol. 70, pp. 177-204, 2009.

[11] G. Wu, Chen, C., and Yan, X.,, "Modified minimum covariance determinant estimator and its application to outlier detection of chemical process data," *Journal of Applied Statistics*, vol. 38, pp. 1007-1020, 2011.

[12] K. Linnet, "On the sensitivity of linear discriminant analysis to sampling variation and analytic errors," *Computers and Biomedical Research*, vol. 21, pp. 158-168, 1988.

[13] Y. Zuo, " Robust location and scatter estimators in multivariate analysis," *WSPC/Trim Size:9in x6in for Review Volume*, pp. 0-31, 2005.

[14] G. Fauconnier, and Haesbroeck, G.,, "Outliers detection with minimum covariance determinant estimator in practice," *Statistical Methodology*, vol. 6, pp. 363-379, 2009.

[15] D. Pena, and Prieto, F. J.,, "Multivariate outlier detection and robust covariance matrix estimation," *Technometrics*, vol. 43, pp. 286-310, 2001.

[16] F. Filzmoser, Jooossens, K., Croux, C., and Leuven, K. U.,, "Multiple group linear discriminant analysis:robustness and error rate," pp. 521-532, 2006.

[17] M. Hubert, and Van Driessen, K.,, "Fast and robust discriminant analysis," *Computational Statistics and Data Analysis*, vol. 45, pp. 301-320, 2004.

[18] M. Hubert, and Debruyne, M.,, "Minimum covariance determinant," *Computational Statistics*, vol. 2, pp. 36-43, 2009.

[19] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179 - 188, 1936.

[20] R. A. Johnson, and Wichern, D. W.,, "Applied multivariate statistical analysis," sixth ed: Prentice Hall, New York, 2007.

[21] M. Hubert, and Van Driessen, K.,, "Fast and robust discriminant analysis," *Computational Statistics and Data Analysis*, vol. 45, pp. 301-320, 2004.

[22] C. Croux, and Haesbroeck, G.,, "Influence function and efficiency of the minimum covariance determinant

scatter matrix estimator," *Journal of Multivariate Analysis*, vol. 71, pp. 161-190, 1999.

[23] M. Hubert, Rousseeuw, P. J., and Verdonck, T.,. "A deterministic algorithm for the MCD," 2011.

[24] G. Pison, Van Aelst, S., and Willems, G.,, "Small sample corrections fot LTS and MCD," *Metrika*, vol. 55, pp. 111-123, 2002.

[25] P. J. Rousseeuw, and Van Driessen, K.,, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, pp. 212-223, 1999.

[26] R. Maronna, Martin, R. D., and Yohai, V. J.,, "Robust statistics: Theory and methods," John Wiley, New York, 2006.

[27] H. P. Lopuhaa, and Rousseeuw, P. J.,, "Breakdown points of affine equivariant estimators of multivariate location and covariance matrices," *The Annals of Statistics*, vol. 19, pp. 229-248, 1991.

[28] A. J. Stromberg, "Robust covariance estimates based on resampling," *Journal of Statistical Planing and Inference*, vol. 57, pp. 321-334, 1997.

[29] F. Z. Okwonu, H. Dieng, A. R. Othman, and O. S. Hui, "Classification of aedes adult mosquitoes in two groups based on Fisher linear discriminant analysis and FZOARO techniques," *Mathematical Theory and Model*, vol. 2, pp. 22-30, 2012.

Friday Zinzendoff Okwonu had his Ph.D in robust statistics with special interest in classification and discrimination. His current interest is biostatistics.

Abdul rahman othman is a professor of robust statistics with versed interest in different area of statistics.