Robust Fisher Linear Classification Technique for Two Groups

F.Z. Okwonu and A.R. Othman

1Department of Mathematics, Faculty of Science, Delta State University, P.M.B. 1, Abraka, Nigeria
2School of Distance Education, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia

Abstract: The sample mean and covariance matrix are the cornerstone of classical multivariate analysis including linear discriminant analysis. Linear discriminant analysis (LDA) depends on the sample mean vectors and covariance matrices computed from different groups of the training samples. Linear discriminant analysis is performed using Fisher’s technique. Classical Fisher linear discriminant analysis (FLDA) is highly susceptible to outliers due to the influence single outlier may impact on the mean vectors. The poor performance of classical FLDA is due to lack of robustness of the classical estimators applied to train the model. In this paper, we propose to robustify Fisher linear discriminant analysis. This approach is developed by applying weight function and compensate constant that performs dual functions. The robustness of these techniques is investigated via numerical simulations using optimal probability of correct classification to determine the classification accuracy and efficacy.

Key words: Classification • Robustness • Hit-Ratio • Misclassification Probability

INTRODUCTION

Fisher linear discriminant analysis (FLDA) is a conventional multivariate technique for dimension reduction and classification. FLDA is a linear combination of observed or measured variables that best describe the separations between known groups of observations. Its basic objective is to classify or predict problems where the dependent variables appear in a qualitative form [1-3]. The classification performance of FLDA will collapse if the training sample contains outlying observations. This occur because Fisher criteria assume the differences between the group means and homoscedasticity of the covariance matrices which lead to overlapped distribution on projection space [4, 5]. Various authors have suggested to replace the classical estimators with robust multivariate estimators, e.g. [6, 7]. Campbell [8] proposed to replace the classical sample mean and covariance matrix with smooth estimators. The authors in the reference therein [9-11] applied minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) to robustify the classical estimators. Rocke et al. [12] and Rousseeuw et al. [13] applied maximum likelihood estimators and M estimators. Several authors say, [14-17] used S-estimators to robustify FLDA. Hubert et al. [18] and Rousseeuw et al. [19] applied FAST MCD estimators of location and shape. In this paper, we propose to robustify FLDA based on weight function and compensate constant which stabilizes the mean vectors, covariance matrices and control the robust discriminant coefficient and the discriminant mean.

This paper is arranged as follows. In section two, we reviewed the classical FLDA procedure. FZOARO classification technique is described in section three. The proposed robust FLDA is presented in section four. Numerical simulations and conclusions are presented in sections five and six.

Fisher Linear Discriminant Analysis: Fisher suggested transforming multivariate observations to univariate observations such that the univariate observations derived from each population is maximally separated. The separation of these univariate observations can be examined by their mean difference [2]. Fisher classification rule maximizes the variation between samples variability to within samples variability. In this section, we describe Fisher linear discriminant analysis for two groups. Consider classifying an observation vector $x_{obs}$ into one of two populations say $\gamma; N_{i}(\mu_{i}, \Sigma_{i}), (i = 1, 2)$ the population mean vectors and covariance matrices are denoted...
as $\mu_n, \Sigma_n$ respectively. Since the population mean vectors and covariance matrices are unknown, the sample estimates is applied. The sample mean vectors, sample covariance matrices and pooled sample covariance matrix are defined as follows.

$$
\bar{x}_{emui} = \frac{1}{n} \sum_{j=1}^{n} x_{obsij}, \\
S_{oroi} = \frac{1}{n} \sum_{j=1}^{n} (x_{obsij} - \bar{x}_{emui})(x_{obsij} - \bar{x}_{emui})', \\
S_{pooled} = \frac{\sum_{i=1}^{g=2} (n_i - 1) S_{oroi}}{\sum_{i=1}^{g=2} n_i - g}.
$$

Using the parameters defined above, Fisher linear discriminant analysis [20] can be started as;

$$
Q = \phi' x_{obsij} = (\bar{x}_{emui} - \bar{x}_{emu2})' S_{pooled}^{-1} x_{obsij},
$$

$$
\bar{Q} = (\bar{x}_{emui} + \bar{x}_{emu2})/2. \phi.
$$

Equations (2.1) and (2.2) are the discriminant score and discriminant mean respectively. The classification rule based on (2.1) and (2.2) can be described as follows: classify $x_{obsi}$ to population one $\gamma_1$ if:

$$
Q \geq \bar{Q}
$$

otherwise allocate $x_{obsi}$ to population two $\gamma_2$ if

$$
Q < \bar{Q}.
$$

Fisher maintained that his technique most adhere strictly to equal variance covariance matrix of the two normal populations. FLDA has been extended to more than two groups [20, 21].

**FZOARO Classification Model:** In this section, we describe a comparable classification and dimension reduction technique proposed in [22, 23]. The proposed technique is developed based on homoscedastic assumption, equal prior probability, the data set are multivariate normal and the sample size is greater than the sample dimension. Applying the definitions of the sample parameters in section two, we state the FZOARO classification model mathematically as follows,

$$
\Omega = \eta_\omega x_{obsij} = (w' \delta) x_{obsij},
$$

$$
\eta_\omega = w' \delta,
$$

$$
w = (\bar{x}_{emu1} - \bar{x}_{emu2})' S_{pooled}^{-1}.
$$

$$
\delta = \sqrt{\frac{\sum S_{pooled}^{-1}}{\chi^2_\alpha}},
$$

$$
h = 3n / 4n,
$$

$$
\Omega = \frac{(\bar{x}_{emu1} + \bar{x}_{emu2})}{2} \eta_\omega.
$$

The parameter $\Omega$ is the discriminant score, $\eta_\omega$ is the FZOARO discriminant coefficient and $\Omega$ is the discriminant mean. The classification procedure is performed by comparing the discriminant score with the discriminant mean. That is, classify $x_{obsi}$ to population one $\sigma_1$ if $\Omega$ is greater than or equal to $\Omega$ or classify $x_{obsi}$ to population two $\sigma_2$ if $\Omega$ is less than $\Omega$.

**Robust FLDA:** Robustness implies reduction in error rates due to classification procedure that does not conform to the assumptions which the conventional model was built. The poor performance of classical FLDA can be attributed to the classical estimators used in training the model. Different authors have diversify different ways to robustify FLDA and the most popular technique used is the plug in approach, see [6, 8-10, 14, 16, 17, 24-27]. The criterion used in this section is based on comparing the Mahalanobis distance (MD) with Chi-square value. This technique applies weight to the sample observations to obtain the weighted sample mean and weighted pooled variance covariance matrix. The method is briefly described as follows,

$$
MD^2 = (x_{obsij} - \bar{x}_{emu1})' S_{pooled}^{-1} (x_{obsij} - \bar{x}_{emu1}),
$$

$$
w_i = \begin{cases} 
1 & \text{if } \quad MD \leq \sqrt{\chi^2_\alpha}, \\
0 & \text{if } \quad MD > \sqrt{\chi^2_\alpha},
\end{cases}
$$

$$
\bar{x}_i = \frac{\sum_{j=1}^{n_i} w_j x_{ij}}{\sum_{j=1}^{n_i} w_j},
$$

$$
\Omega = \frac{(\bar{x}_{emu1} + \bar{x}_{emu2})}{2} \eta_\omega.
$$
Equations (4.1), (4.2) and (4.3) are the weight function, weighted within group means and weighted within group variance covariance matrices respectively. Equation (4.3) is pooled to obtain the pooled covariance matrix. A compensate constant $\epsilon_{mu}$ is added to the inverse of the pooled covariance matrix to compensate for the zero lost weight. The robust FLDA score (4.5) and its discriminant mean (4.6) are described as follows;

$$\Omega_i = \Xi' x_{sbi} = \left( \rho_i' \delta \right) x_{sbi}, \quad (4.5)$$

$$\Xi' = \rho_i \delta, \quad \rho_i = (\bar{x}_{emu1} - \bar{x}_{emu2})' S_{pooled}^{-1}, \quad (4.6)$$

The classification rule based on equations (4.5) and (4.6) is described as follows; classify $x_{sbi}$ to population one $P_1$ if $\Omega_i$ is greater than or equal to $\tilde{\Omega}$ otherwise allocate $x_{sbi}$ to population two $P_2$ if $\Omega_i$ is less than $\tilde{\Omega}$. In addition to the assumptions of FZOARO, the new approach concurs with the conventional assumptions of classical Fisher linear discriminant analysis.

**Simulation:** The numerical simulation is designed to compare the comparative classification accuracy of the conventional FLDA, FZOARO and the proposed robust FLDA. Each group has sample sizes of 100 generated data set and majority of the data set is generated from the normal distribution and the remaining from the contaminated distribution, say 5%, 10% and 15% respectively. The simulation is designed using symmetric and asymmetric contamination model. The sample observations used in this experiment is divided into training sample (60%) and validation (40%). The sample observations are randomly selected in each case.

The experiment is run 100 times and the mean and standard deviation () of correct classification is reported. Table 1 gives the mean and standard deviation for these techniques for symmetric contaminated normal data. FZOARO perform similar as classical FLDA, however, the robust approach outperformed the FLDA and FZOARO.

Table 2 illustrates the classification rate of the FLDA, FZOARO and the robust FLDA based on asymmetric contaminated normal data. This experiment shows that FZOARO outperformed FLDA and RFLDA for 5% contamination and RFLDA outperformed FLDA. For 10% and 15% RFLDA outperformed the other techniques and FLDA and FZOARO performs similar. Comparing the performance of the various techniques considered based on the optimal classification rate; we conclude that RFLDA outperformed other techniques.

**CONCLUSION**

A new procedure for the robustification of Fisher linear discriminant analysis has been proposed. Numerical simulation was conducted to compare the classification capability of the proposed robust approach with classical Fisher linear discriminant analysis and FZOARO methods. The result reported after taking the average run of 100 runs indicate that the robust approach outperformed the classical FLDA and FZOARO for both contamination model.
REFERENCES