GENERALIZATION OF FISHER'S LINEAR DISCRIMINANT

Ana M. Pires and João A. Branco
Instituto Superior Técnico, Departamento de Matemática
Av. Rovisco Pais, 1096 Lisboa Codex
Tel: ++351 (0)1 8417053 and ++351 (0)1 8417051, Fax: ++351 (0)1 8499242
email: anpires@math.ist.utl.pt and jbranco@math.ist.utl.pt

Abstract: In this paper a generalization of Fisher's linear discriminant is proposed. With this new procedure it is possible to estimate linear discriminant functions which are not affected by outlying observations. The proposed method and the classical method are compared by applying both to real and simulated data sets. The generalized approach has shown advantages over the classical one.

Keywords: Discriminant analysis, Fisher's linear discriminant, robust procedures.

1. INTRODUCTION

Fisher's linear discriminant (Fisher, 1936) is very popular among users of discriminant analysis. Some of the reasons for this are its simplicity and unnecessity of strict assumptions. However it has optimality properties only if the underlying distributions of the groups are multivariate normal. It is also easy to verify that the discriminant rule obtained can be very harmed by only a small number of outlying observations. Outliers are very hard to detect in multivariate data sets and even when they are detected simply discarding them is not the most efficient way of handling the situation. Therefore the need for robust procedures that can accommodate the outliers and are not strongly affected by them.

In this paper we propose a generalization of Fisher's linear discriminant which leads easily to a very robust procedure.

2. DESCRIPTION OF THE METHOD

For two groups with locations $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and common dispersion matrix $\boldsymbol{\Sigma}$, Fisher's separation criterion looks for the vector $\boldsymbol{\alpha}$ (with dimension m, the number of features or variables), which maximizes the ratio

$$\frac{\left[\boldsymbol{\alpha}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^2}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{\alpha}}. \tag{1}$$

The solution is well known and easy to obtain by standard algebra

$$\boldsymbol{\alpha} \propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

A cutoff point, $-\alpha_0$, is then determined in an appropriate way. For equal costs of misclassification, $C(i|j)$, with $i \neq j = 1, 2$, and equal *a priori* probabilities of group membership, π_i, $\alpha_0 = -\boldsymbol{\alpha}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Fisher's classification rule for a new observation $\mathbf{x}$ of unknown origin is

$$\begin{array}{l} \text{classify in group 1 if } \boldsymbol{\alpha}^T\mathbf{x} + \alpha_0 > 0 \\ \text{classify in group 2 if } \boldsymbol{\alpha}^T\mathbf{x} + \alpha_0 < 0 \end{array}.$$

When the population parameters are unknown, as it is usually the case, they are estimated by their training samples counterparts, $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and $\mathbf{S}$. Then (1) is equivalent to

$$\frac{(n_1 + n_2 - 2)\left[\text{ave}(\boldsymbol{\alpha}^T\mathbf{x}_{1j}) - \text{ave}(\boldsymbol{\alpha}^T\mathbf{x}_{2j})\right]^2}{(n_1 - 1)\text{var}(\boldsymbol{\alpha}^T\mathbf{x}_{1j}) + (n_2 - 1)\text{var}(\boldsymbol{\alpha}^T\mathbf{x}_{2j})}, \tag{2}$$

where ave and var denote the sample mean and sample variance operators, applied to the one-dimensional samples of projected observations, $\boldsymbol{\alpha}^T\mathbf{x}_{ij}$, $i = 1, 2$; $j = 1, \ldots, n_i$. The reason for the non-robustness of Fisher's linear discriminant function (ldf) lies in the use of the sample means and variances, which can be affected by one sufficiently large point.

A straightforward generalization of (2) is to allow for general univariate estimators of location (T) and dispersion (S):

$$I(\boldsymbol{\alpha}) = \frac{\left[T(\boldsymbol{\alpha}^T\mathbf{x}_{1j}) - T(\boldsymbol{\alpha}^T\mathbf{x}_{2j})\right]^2}{a_1 S^2(\boldsymbol{\alpha}^T\mathbf{x}_{1j}) + a_2 S^2(\boldsymbol{\alpha}^T\mathbf{x}_{2j})} \tag{3}$$

and then find $\hat{\boldsymbol{\alpha}}$ such that $I(\hat{\boldsymbol{\alpha}})$ is the maximum of $I(\boldsymbol{\alpha}) : \mathbb{R}^m \to \mathbb{R}_0^+$, subject to the constrain

$\|\boldsymbol{\alpha}\| = 1$. If T and S^2 are the sample mean and variance Fisher's ldf is obtained. When T and S are robust estimators the ldf inherits their robustness properties. In particular, a high breakdown point that does not depend on m is attainable (the breakdown point is the maximum proportion of misbehaving observations in the sample that an estimator can accommodate before it gives completely arbitrary estimates; see Hampel *et al*, 1986, for a formal definition).

The use of general coefficients a_i in the denominator of (3) also allows for more flexibility. $a_i = (n_i - 1)/(n_1 + n_2 - 2)$ is appropriate when the group covariance matrices are similar, while $a_i = 1/2$ is a good choice when the dispersions of the groups are different, in the sense that an approximate best linear discriminant function is obtained.

Several possibilities are available for T and S, for instance the pair T =median, S =MAD (median absolute deviation of the median), or members of the general class of M-estimators. Among these our preference goes to the simultaneous M-estimators with Huber type weights (see for instance Hoaglin *et al*, 1992), because of their good robustness and regularity properties. With these estimators an explicit solution is no longer available, therefore a numerical algorithm has to be implemented in order to find the direction $\hat{\boldsymbol{\alpha}}$ that maximizes $I(\boldsymbol{\alpha})$ (see Pires, 1995).

The cutoff point is determined using the same type of estimators. If the dispersions of the groups are assumed equal, $C(1|2) = C(2|1)$ and $\pi_1 = \pi_2$, then

$$\hat{\alpha}_0 = -\frac{T(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{1j}) + T(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{2j})}{2}. \qquad (4)$$

If the assumption of equal dispersions is not valid then we propose that $\hat{\alpha}_0$ be estimated by the solution of the second degree equation

$$\frac{\left[\hat{\alpha}_0 + T(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{2j})\right]^2}{S^2(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{2j})} - \frac{\left[\hat{\alpha}_0 + T(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{1j})\right]^2}{S^2(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{1j})} = = \log \frac{S^2(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{1j})}{S^2(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{2j})} \qquad (5)$$

which minimizes

$$\pi_1[1 - \Phi(y_1)] + \pi_2[1 - \Phi(y_2)],$$

with

$$y_1 = \frac{T(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{1j}) + \hat{\alpha}_0}{S(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{1j})}$$

and

$$y_2 = \frac{-T(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{2j}) - \hat{\alpha}_0}{S(\hat{\boldsymbol{\alpha}}^T \mathbf{x}_{2j})}.$$

Expressions (4) and (5) can be easily modified in order to include the general case of unequal costs or *a priori* probabilities (see Pires, 1995).

3. EXAMPLES

To compare the proposed method with the classical method we have applied both to several real and simulated data sets. The results obtained are presented in the next examples.

Example 1: Let us consider the most favorable situation to Fisher's ldf, that is gaussian data. $n_1 = n_2 = 30$ observations were generated from two trivariate normal distributions with identity covariance matrices, $\mathcal{N}_3(\boldsymbol{\mu}_i, \mathbf{I})$, the first group with location $\boldsymbol{\mu}_1 = (0, 0, 0)^T$ and the second with $\boldsymbol{\mu}_2 = (1.5, 1.5, 1.5)^T$ (under these conditions the optimum error rate is $e_{opt} = \Phi(-1.5\sqrt{3}/2) \simeq 9.697\%$). From the training sample three linear discriminant functions were obtained: the classical and the generalized for two values of the tunning constant of the M-estimators (1.645 and 1.96). The actual error rate was then evaluated using the theoretical distribution:

Method	e_{act}
Classical	9.715%
Gen(1.96)	9.760%
Gen(1.645)	10.299%

Although the classical method is better, as antecipated, the results of the generalized approach are very good, especially for the higher tunning constant (this is expected since the M-estimators converge to the classical estimators as their tunning constant increases).

Example 2: It is also important to compare the methods when the data are contaminated. We considered two well separated groups but where one of them has 10% outlying points. The data for the first group consists of 90 observations generated from $\mathcal{N}_2(\mathbf{0}, \mathbf{I})$, plus 10 observations on the point $(10, 10)$. The second group consists of 100 observations from $\mathcal{N}_2(\boldsymbol{\mu}_2, \mathbf{I})$, with $\boldsymbol{\mu}_2 = (4, 0)^T$. The contamination scheme used for the first group intends to model, in a simplified way, the occurrence of gross errors or undesired sampling from a different population. The theoretical situation shall not take the outliers into account, therefore the optimum error rate

is $e_{opt} = \Phi(-2) \simeq 2.275\%$. The methods used were the same that in Example 1 and led to

Method	e_{act}
Classical	6.802%
Gen(1.96)	2.630%
Gen(1.645)	2.574%

The results for the generalized method are now much better than the classical and were not affected by the contamination in the training sample.

Example 3: This real data set was studied by Hermans and Habbema (1975) and concerns the detection of hemophilia carriers based on two variables (log Factor-VIII activity, x_1, and log Factor-VIII like antigen, x_2). The training sample consists of observations on 30 women known to be non-carriers and on 22 known carriers, and is represented in Figure 1. In this case the theoretical distributions are unknown, therefore the actual error rate had to be estimated. The bootstrap procedure (Efron, 1983) was used. The following results were obtained:

Method	e_{boot}
Classical	5.285%
Gen*(1.96)	2.745%
Gen*(1.645)	2.732%

* Assuming unequal variances

The results with the new method are better. This can partially be justified by the unequal variance assumption.

Example 4: The data used in this example is described in Macieira-Coelho *et al.* (1990). Those authors studied the possibility of using discriminant analysis to predict, as a screening tool, the existance of choronary heart disease, based on four clinical variables and on five variables obtained in a stress test. The training sample consists of observations on 30 "normal" and on 83 "sick" patients (the true state can be accurately determined by choronariography, an expensive and risky exam that is not advisable as a routine procedure). From the nine original variables only seven were selected for the final analysis (various selection procedures are described in Pires, 1995). The results for the bootstrap error rate were

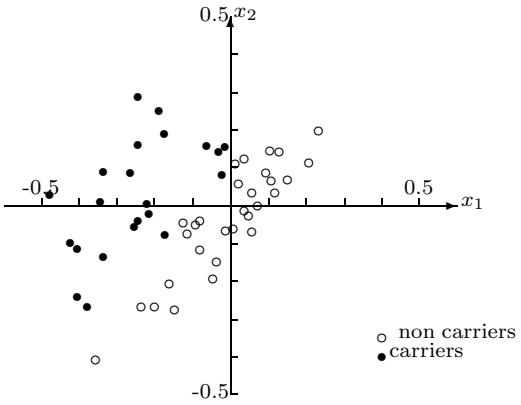

Figure 1: Data of Example 3.

Method	e_{boot}
Classical	16.1%
Gen*(1.645)	17.1%
Gen*(1.96)	14.0%

* Assuming unequal variances

Although the three error rates are quite high, it is seen that the second generalized approach led to some improvement. It is possible that a linear discriminant is not the most suitable method for this problem, however with the high number of variables and small number of observations it turned out to be the only feasible.

As seen in this example, and also in Example 1, the choice of the tunning constant may be important. From our practice we suggest that at least the two values, 1.645 and 1.96, be used and that the one yielding the smaller error rate be choosen.

4. CONCLUSIONS

The results from the previous examples show that both methods lead to similar misclassification error rates when the data are well behaved (that is, approximately gaussian) but that the generalized approach is better (in the same sense) in the presence of outliers.

It is evident that the new procedure is robust against contamination of the training samples and that it can be safely used otherwise, therefore we strongly recommend it, provided that a linear method is adequate.

References

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-

validation. *Journal of the American Statistical Association* **78**. 316–331.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Hermans, J. and Habbema, J. D. F. (1975). Comparison of five methods to estimate posterior probabilities. *EDV in Medizin und Biologie* **6**, 14–19.

Hoaglin, D., Mosteller, F. and Tukey, J. W. (1992). *Análise Exploratória de Dados, Técnicas Robustas — Um Guia*. Edições Salamandra, Lisboa. (Portuguese translation of: Hoaglin, D., Mosteller, F. and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.)

Macieira-Coelho, E., Oliveira, M. F. and Amaral-Turkman, M. A. (1990). Diagnóstico de cardiopatia isquémica no doente ambulatório: análise multivariada de dados clínicos e electrocardiográficos. *Acta Médica Portuguesa* **3**, 277–282.

Pires, A. M. (1995). *Análise Discriminante: Novos Métodos Robustos de Estimação*. Tese de Doutoramento. IST, Universidade Técnica de Lisboa, Lisboa.