

Effect of Prior Probabilities on the Classificatory Performance of Parametric and Mathematical Programming Approaches to the Two-Group Discriminant Problem

CONSTANTINE LOUCOPOULOS

loucopoc@esumail.emporia.edu

Box 4023, School of Business, Emporia State University, Emporia, KS 66801, USA

Abstract. A mixed-integer programming model (MIP) incorporating prior probabilities for the two-group discriminant problem is presented. Its classificatory performance is compared against that of Fisher's linear discriminant function (LDF) and Smith's quadratic discriminant function (QDF) for simulated data from normal and nonnormal populations for different settings of the prior probabilities of group membership. The proposed model is shown to outperform both LDF and QDF for most settings of the prior probabilities when the data are generated from nonnormal populations but underperforms the parametric models for data generated from normal populations.

Keywords: Discriminant Analysis, Classification, Mathematical Programming, Simulation

1. Introduction

Mathematical programming approaches to discriminant analysis have attracted considerable research interest in recent years. Simulation studies by Freed and Glover [5], Joachimsthaler and Stam [7], Stam and Jones [12], Hosseini and Armacost [6] and Loucopoulos and Pavur [10], have shown that these mathematical programming approaches are viable alternatives to Fisher's [2] linear discriminant function (LDF) and Smith's [11] quadratic discriminant function (QDF). An appealing characteristic of mathematical programming (MP) approaches is that they do not rely on the assumption of multivariate normality, nor do they impose any conditions on the covariance structures for optimal classificatory performance. In contrast, both LDF and QDF assume multivariate normality, with equal or unequal covariance structures respectively.

Despite a plethora of proposed mathematical programming models for the two-group discriminant problem by Freed and Glover [3], [4], Choo and Wedley [1], Koehler and Erenguc [8] and Lam, Choo and Moy [9], the effect of prior probabilities on the classificatory performance of MP models has not received any research interest. This paper proposes a mathematical programming model incorporating the effect of prior probabilities and compares its holdout classificatory performance against that of LDF and QDF for various settings of the prior probabilities of group membership. The proposed model is presented in the next section. The simulation experiment for the comparison of classificatory performance is described in Sec-

tion 3, with simulation results analyzed in Section 4 and conclusions presented in Section 5.

2. An MIP Model Incorporating Prior Probabilities for the Two-Group Problem

In this section, a modification to the MIP model for the two-group discriminant problem is proposed. The modification involves the incorporation of prior probabilities into the objective function and the elimination of the risk of unacceptable solutions with the inclusion of appropriate constraints. This mixed-integer programming model is presented below.

Notation:

a_k is the weight assigned to attribute variable X_k
($k = 1, 2, \dots, p$)

$X_k^{(i)}$ is the value of variable X_k for observation i
($i = 1, 2, \dots, n$)

$I_i = \begin{cases} 1 & \text{if observation } i \text{ is misclassified} \\ 0 & \text{otherwise} \end{cases}$

π_j is the prior probability of membership in group G_j ($j = 1, 2$)

c is the cutoff value for group G_1

ϵ is the width of the gap separating groups G_1 and G_2

M is the maximum deviation of a misclassified observation from the cutoff value of its group

Thus, a_k ($k = 1, 2, \dots, p$), I_i ($i = 1, 2, \dots, n$) and c are decision variables whose values are to be determined by the model, whereas π_j ($j = 1, 2$), M and ϵ are parameters.

Formulation:

$$\min \pi_1 \sum_{i \in G_1} I_i + (1 - \pi_1) \sum_{i \in G_2} I_i$$

s. t.

$$\sum_{k=1}^p a_k X_k^{(i)} - M I_i \leq c \quad i \in G_1$$

$$\sum_{k=1}^p a_k X_k^{(i)} + M I_i \geq c + \epsilon \quad i \in G_2$$

$$\sum_{i \in G_1} I_i \leq n_1 - 1$$

$$\sum_{i \in G_2} I_i \leq n_2 - 1$$

where $a_k (k = 1, 2, \dots, p)$ and c are sign-unrestricted variables.

The objective of this formulation is the minimization of the weighted sum of misclassifications with the prior probabilities of group membership being the weights. In this formulation, a discriminant score $\sum_{k=1}^p a_k X_k^{(i)}$ is computed for each observation. According to the first constraint, an observation $i \in G_1$ will be correctly classified, if its discriminant score $\sum_{k=1}^p a_k X_k^{(i)}$ does not exceed c . Otherwise it is misclassified. However, its discriminant score cannot exceed c by more than M , where M is a preset large positive constant.

According to the second constraint, an observation $i \in G_2$ will be correctly classified if its discriminant score $\sum_{k=1}^p a_k X_k^{(i)}$ exceeds $c + \epsilon$, where ϵ is a preset small positive constant. Otherwise, the observation is misclassified. If $i \in G_2$ is misclassified, the value of its discriminant score cannot fall below $c + \epsilon - M$. The purpose of the gap of width ϵ between the two groups is to enhance group separation.

The last two constraints guarantee that, whatever the values of the prior probabilities or the attribute variables, an unacceptable solution with $a_1 = a_2 = \dots = a_p = 0$ is not feasible. In this case, all the observations would be classified into the same group.

3. Simulation Experiment

The holdout sample classificatory performance of the proposed model was compared against that of Fisher's linear discriminant function (LDF) and Smith's quadratic discriminant function (QDF) using data generated from bivariate normal, contaminated normal and exponential populations. The different configurations included

Table 1. Configurations used in the simulation study

Distribution	Group Location Parameters	Covariance Structures	Cfg
Normal	$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\boldsymbol{\mu}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = I$	N_1
		$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & .75 \\ .75 & 1 \end{bmatrix}$ $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & -.75 \\ -.75 & 1 \end{bmatrix}$	N_2
		$\boldsymbol{\Sigma}_1 = .25I$ $\boldsymbol{\Sigma}_2 = 4I$	N_3
Contaminated Normal	$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\boldsymbol{\mu}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ $\boldsymbol{\mu}_1^{(c)} = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$ $\boldsymbol{\mu}_2^{(c)} = \begin{bmatrix} -6 \\ -6 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = I$	C_1
		$\boldsymbol{\Sigma}_1^{(c)} = \boldsymbol{\Sigma}_2^{(c)} = .25I$	
		$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\boldsymbol{\mu}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ $\boldsymbol{\mu}_1^{(c)} = \boldsymbol{\mu}_2^{(c)} = \begin{bmatrix} -8 \\ -8 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = I$ $\boldsymbol{\Sigma}_1^{(c)} = \boldsymbol{\Sigma}_2^{(c)} = .16I$
	$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\boldsymbol{\mu}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ $\boldsymbol{\mu}_1^{(c)} = \begin{bmatrix} -8 \\ -8 \end{bmatrix}$ $\boldsymbol{\mu}_2^{(c)} = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$ $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & -.5 \\ -.5 & 1 \end{bmatrix}$	C_3
		$\boldsymbol{\Sigma}_1^{(c)} = \begin{bmatrix} .25 & .125 \\ .125 & .25 \end{bmatrix}$ $\boldsymbol{\Sigma}_2^{(c)} = \begin{bmatrix} .25 & -.125 \\ -.125 & .25 \end{bmatrix}$	
Exponential	$\boldsymbol{a}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\boldsymbol{a}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = 4I$ i.e., $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 = \begin{bmatrix} .5 \\ .5 \end{bmatrix}$	E_1
		$\boldsymbol{\Sigma}_1 = I$ $\boldsymbol{\Sigma}_2 = 16I$ i.e., $\boldsymbol{\lambda}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\boldsymbol{\lambda}_2 = \begin{bmatrix} .25 \\ .25 \end{bmatrix}$	E_2
		$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 0 \\ 0 & 16 \end{bmatrix}$ i.e., $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 = \begin{bmatrix} .5 \\ .25 \end{bmatrix}$	E_3

in this simulation study are presented in Table 1. The prior probabilities π_i of membership in group G_i were assigned values .20, .35, .50, .65 and .80, whereas the values of the parameters M and ϵ in the MIP model were set at 100 and .001, respectively. Such values of the parameters M and ϵ are consistent with the practice employed in previous simulation studies on the classificatory performance of

mathematical programming approaches to the discriminant problem, calling for the assignment of a large value to M and a small value to ϵ . Training samples of size 100 (50 per group) were simulated. Holdout samples of size 1000 were generated with the number of observations from group G_i being $1000\pi_i$ ($i = 1, 2$), where π_i represents the prior probability of membership in group G_i . Each experimental condition was replicated 100 times. The simulation study was carried out using SAS 6.11 on a RISC 6000/58H computer.

In configurations N_1 , N_2 and N_3 , the data were simulated from normal populations with equal and unequal covariance structures. In configurations C_1 , C_2 and C_3 , the data were generated from contaminated normal populations with a contaminating fraction of .10. It should be noted that $\mu_i^{(c)}$ and $\Sigma_i^{(c)}$ refer to the mean and covariance structure, respectively, of the contaminant component of group G_i ($i = 1, 2$). In configurations E_1 , E_2 and E_3 , the data were generated from exponential populations with starting points a_i and density function:

$$f(x) = \begin{cases} \lambda e^{-\lambda(x-a)} & x \geq a \\ 0 & \text{otherwise} \end{cases}$$

4. Simulation Results

The percentage misclassification rates of the different models in the holdout sample are presented in Tables 2. Under experimental conditions optimal for Fisher's linear discriminant function (configuration N_1), the proposed MIP model yielded higher mean misclassification rates than either LDF or QDF in the holdout sample. Under experimental conditions optimal for QDF (configurations N_2 and N_3), the proposed model underperformed QDF for all settings of the prior probabilities, but outperformed LDF for certain settings of the prior probabilities.

When the data are generated from contaminated normal populations (configurations C_1 , C_2 and C_3), the MIP model had lower average misclassification rates than both LDF and QDF in the holdout sample. This was true for all values assigned to the prior probabilities π_i .

When the data are generated from exponential populations (configurations E_1 , E_2 and E_3), the MIP model outperformed both LDF and QDF for $\pi_1=.35$, $\pi_1=.50$ and $\pi_1=.65$. However, for $\pi_1=.20$ and $\pi_1=.80$ the results were mixed.

5. Conclusions

This paper examines the effect of prior probabilities on the classificatory performance of a proposed MIP model as well as the standard parametric procedures

Table 2. Holdout misclassification rates (%)

Cfg	Method	Prior Probabilities				
		$\pi_1 = .20$	$\pi_1 = .35$	$\pi_1 = .50$	$\pi_1 = .65$	$\pi_1 = .80$
N_1	MIP	7.212	8.624	9.006	8.571	7.301
	LDF	6.074	7.700	8.271	7.718	6.055
	QDF	6.163	7.828	8.380	7.864	6.202
N_2	MIP	4.534	5.663	6.486	7.163	6.839
	LDF	4.772	6.406	7.306	7.527	6.971
	QDF	2.687	3.943	4.887	5.366	5.160
N_3	MIP	14.587	13.524	11.308	8.776	5.956
	LDF	14.775	13.376	12.539	11.330	8.541
	QDF	6.296	6.361	5.649	4.563	3.025
C_1	MIP	17.005	17.749	18.031	17.733	16.803
	LDF	21.196	37.196	36.799	37.355	21.271
	QDF	21.816	35.091	37.947	35.893	22.426
C_2	MIP	15.162	14.289	12.953	11.202	8.625
	LDF	23.661	35.997	26.633	28.244	19.771
	QDF	21.388	28.839	28.755	28.283	21.389
C_3	MIP	5.347	6.565	7.351	7.777	6.847
	LDF	16.123	15.870	10.629	18.136	17.264
	QDF	13.682	12.620	14.940	19.018	14.995
E_1	MIP	8.773	13.979	18.971	22.981	21.095
	LDF	10.228	15.011	23.404	25.106	19.552
	QDF	10.577	15.987	22.745	24.990	19.897
E_2	MIP	2.437	3.580	4.391	4.881	4.539
	LDF	2.846	6.452	8.900	9.241	7.275
	QDF	3.408	5.009	5.773	5.898	5.282
E_3	MIP	9.397	15.329	21.020	26.607	22.438
	LDF	13.855	17.905	28.668	29.438	21.213
	QDF	14.160	20.489	28.629	29.791	21.786

(LDF and QDF). It is shown that, regardless of the values assigned to the prior probabilities, the proposed MIP model will yield higher misclassification rates in the holdout sample when the experimental conditions are optimal for the parametric procedures. It is also shown that for data generated from contaminated normal populations, the proposed model outperforms both LDF and QDF, regardless of the values assigned to the prior probabilities. For data generated from exponential populations, the MIP model outperformed the other two models when the prior probabilities of membership in group G_1 ($i = 1, 2$) was $.35$, $.50$ or $.65$. However for $\pi_1 = .20$ and $\pi_1 = .80$, the results of the simulation study were inconclusive for data generated from exponential populations.

Because of the numerous possibilities in terms of data configurations, prior probabilities and sample sizes, it may be inappropriate to draw generalized conclusions about the classificatory performance of the proposed model. Further research should focus on the relative performance of the proposed MIP model under different experimental conditions.

References

1. E. U. Choo and W. C. Wedley. Optimal criterion weights in repetitive multicriteria decision making. *Journal of the Operational Research Society*, 36:983–992, 1985.
2. R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.
3. N. Freed and F. Glover. A linear programming approach to the discriminant problem. *Decision Sciences*, 12:68–74, 1981.
4. N. Freed and F. Glover. Simple but powerful goal programming models for the discriminant problem. *European Journal of Operational Research*, 7:44–60, 1981.
5. Evaluating alternative linear programming models to solve the two–group discriminant problem. *Decision Sciences*, 17:151–162, 1986.
6. J. C. Hosseini and R. L. Armacost. The two–group discriminant problem with equal group means vectors: An experimental evaluation of six linear/nonlinear programming formulations. *European Journal of Operational Research*, 77:241–252, 1994.
7. E. Joachimsthaler and A. Stam. Four approaches to the classification problem in discriminant analysis: An experimental study. *Decision Sciences*, 19:322–333, 1988.
8. G. J. Koehler and S. Erenguc. Minimizing misclassifications in linear discriminant analysis. *Decision Sciences*, 21:63–85, 1990.
9. K. F. Lam, E. U. Choo and J. W. Moy. Minimizing deviations from the group mean: A new linear programming approach for the two–group classification problem. *European Journal of Operational Research*, 88:358–367, 1996.
10. C. Loucopoulos and R. Pavur. Experimental evaluation of the classificatory performance of mathematical programming approaches to the three–group discriminant problem: The case of small samples. *Annals of Operations Research* (forthcoming).
11. C. A. B. Smith. Some examples of discrimination. *Annals of Eugenics*, 13:272–282, 1947.
12. A. Stam and D. G. Jones. Classification performance of mathematical programming techniques in discriminant analysis: Results for small and medium sample sizes. *Managerial and Decision Economics*, 11:243–253, 1990.