O P E R A T I O N S R E S E A R C H A N D D E C I S I O N S
No. 1 2020
DOI: 10.37190/ord200107

ON THE BINARY CLASSIFICATION PROBLEM IN DISCRIMINANT ANALYSIS USING LINEAR PROGRAMMING METHODS

MICHAEL O. OLUSOLA*, SIDNEY I. ONYEAGU

Nnamdi Azikiwe University, PMB 5025 Awka, Nigeria

This paper is centred on a binary classification problem in which it is desired to assign a new object with multivariate features to one of two distinct populations as based on historical sets of samples from two populations. A linear discriminant analysis framework has been proposed, called the minimised sum of deviations by proportion (MSDP) to model the binary classification problem. In the MSDP formulation, the sum of the proportion of exterior deviations is minimised subject to the group separation constraints, the normalisation constraint, the upper bound constraints on proportions of exterior deviations and the sign unrestriction vis-à-vis the non-negativity constraints. The two-phase method in linear programming is adopted as a solution technique to generate the discriminant function. The decision rule on group-membership prediction is constructed using the apparent error rate. The performance of the MSDP has been compared with some existing linear discriminant models using a previously published dataset on road casualties. The MSDP model was more promising and well suited for the imbalanced dataset on road casualties.

Keywords: *binary classification, discriminant analysis, error rate, hit rate, linear programming*

1. Introduction

Discriminant analysis techniques involve the use of observations of known class membership to generate functions that separate these observations into the specified classes optimally in terms of a technique-dependent separation criterion [8, 17]. Discriminant analysis methods are used to study the difference between two or more groups as based on one or more attributes and to classify new observations into a group to which they are likely to belong [4]. Allocation of the new observations to a group is a predictive aspect of discriminant analysis. This predictive aspect of discriminant analysis is

*Corresponding author, email address olusolamo@gmail.com
Received 13 November 2019, accepted 20 April 2020

often referred to as classification analysis. The relevance of discriminant analysis in diverse and substantive areas, such as management science, artificial intelligence, criminology, health care, etc., is well known in the literature [7, 18]. Several models have been developed for this purpose [1, 3, 17].

This paper focuses on the classification problem for two groups. Instances, where the practical importance of the two-group problem has found application, include the development of an email program to classify emails as either legitimate or spam, the student admissions problem in which prospective students are grouped into those that are likely to succeed in a course of study or likely to fail, and the classification problem by police of burglaries into solvable and unsolvable cases. The two-group linear discriminant classifier is a function of the form $(\mathbf{w}, c)$: $R^p \to \{1, 2\}$, where R^p is the p-dimensional Euclidean space, $\mathbf{w}$ is the discriminant coefficient, and c is the cut-off value. The classifier is determined from a set of observations whose group membership is known and it partitions the p-dimensional Euclidean space R^p into two regions: a closed half-space $\{\mathbf{x} : \mathbf{xw} \le c\}$, and an open half-space $\{\mathbf{x} : \mathbf{xw} > c\}$ [4]. The set of observations used to generate the discriminant function is called the training sample.

Abramovich and Pensky [2] consider the large p- small n-type of the multi-class classification problem. In this kind of problem, the dimensionality of the parameter space p by far exceeds the sample size n for objects with a large number of classes. The study implements feature selection by a thresholding technique and classification is carried out based on the minimal Mahalanobis distance. The conditions on the effects of significant features and bounds for distances between classes required for successful feature selection and classification are derived. The findings reveal that having a larger number of classes could aid feature selection and improve classification accuracy.

Gaynanova and Wang [8] consider a binary classification problem wherein n mutually independent observations (X_1, Y_1), ..., (X_n, Y_n) from a random pair (X, Y) taking values in $R^p \times \{1, 2\}$ were studied. The goal was to formulate a rule that could assign one of the two labels in $\{1, 2\}$ to a new data point $X \in R^p$ and to determine the subset of p variables that influences the rule. The work identifies the drawbacks of the linear discriminant analysis (LDA) and the quadratic discriminant analysis (QDA). For the LDA, it was said that the assumption of equal covariance matrices (i.e., $\Sigma_1 = \Sigma_2$) of the two groups is unlikely to be satisfied in practice and that the performance of the linear rule is suboptimal, while the QDA performs poorly when p is large as the estimation of the precision matrices Σ_1^{-1} and Σ_2^{-1} is extremely challenging when $p > n$. For large p, the task of assigning $X \in R^p$ to one of the labels in $\{1, 2\}$ is a high-dimensional binary classification problem. A solution to this problem using a variable selection technique to reduce the dimension of the original data is attempted and subsequently applied QDA on the reduced space. The performance of the method is measured by misclassification error rates and then compared with the existing methods.

This study is aimed at developing a linear programming discriminant analysis model that minimises the sum of the proportion of deviations from a reference hyperplane for the two-group discriminant problem. The study is structured on the linear programming (LP) discriminant analysis for the binary classification problem. Our choice for the linear programming approach is the sequel to its advantages [7], viz. LP models are both distribution-free (no assumptions on the distribution of the populations) and free from parametric assumption (no assumption on the covariance matrices), the objective of minimising exterior deviations are easily captured and LP models less sensitive to outliers since the models are based on linear metrics. The linear programming approach to discriminant analysis uses a given dataset to construct a discriminant function. This kind of problem has received considerable attention in the literature [12]. This paper concentrates on a distribution-free approach to discriminant analysis and represents the classification problem as a linear programming problem wherein the proportion of exterior deviation is minimised subject to certain constraints. An exterior deviation is a deviation from the hyperplane of a point improperly classified.

In LP discriminant analysis, the assumption on the equality of the covariance matrices of the group (as in Fisher's method) is relaxed. This is because the discriminant classifier does not depend on the covariance of the population groups. The basic form of LP discriminant analysis models is

$$\text{minimise } f(\mathbf{w}, c) \text{ subject to } \mathbf{Xw} \leq (>)\ c\mathbf{1},\ \mathbf{Yw} > (\leq) c\mathbf{1},\ \mathbf{w} \neq \mathbf{0}$$

where $\mathbf{1}$ is a column vector of 1's that is conformable to the dimension of the product $\mathbf{Xw}$ or $\mathbf{Yw}$. In many instances, the constraint space is closed by introducing the relaxation equal or higher than ($\geq$) in place of higher than ($>$).

The use of LP models for discriminant analysis is not new [4]. The earlier LP methods were largely based on the objective of minimising exterior deviations, maximising internal deviations or both [20], or deriving a discriminant function using a mixed integer programming (MIP) approach [8, 15]. The deviation-based LP models may lead to unbounded solutions and the MIP models are characterised by excessive computational procedures. The MIP models utilise either the branch-and-bound algorithm or the cutting plane algorithm, following an optimal solution obtained without the integral requirements taken into consideration. The computational difficulties of the MIP are linked to the binary variable that must be associated with each training sample observation.

The first two LP formulations in the literature are the minimise the maximum deviation (MMD) model and the minimise the sum of deviations (MSD) model [19]. In the MMD model, an unbounded solution indicates a perfect separation of the two groups. The MSD is the most widely used objective in LP discriminant analysis [5]. Note that if the two groups are linearly separable, the MSD is zero and the discriminant function will be the separating hyperplane. Stam and Jones [19] report that the MMD approach

is found to have a classification power inferior to the MSD and that the MSD has a greater promise to minimise the rate of misclassification under data situations where the more traditional (statistical) approaches are less effective.

Freed and Glover [7] provide a detailed study of the MMD and the MSD models using the normalisation

$$\sum w_j + c = 10$$

The obtained results establish the MSD model as a promising alternative to the conventional linear discriminant techniques, e.g., Fisher's method.

Liu and Maloney [16] develop LP models to solve the two-group discriminant problem for two cases, where the groups are linearly separable and where they are non-separable. Lam and Moy [13] develop a new LP model to solve the multigroup (more than two) classification problem. The model aggregates information contained in the multigroup problem to simultaneously determine the cut-off values for the different classification functions. Gochet et al. [11] introduce a novel problem formulation for the multigroup LP classification problem and show that the new formulation is capable of producing good classification results which can compete with both Fisher's parametric method and the nonparametric *k*-nearest neighbourhood method.

Our concern in this paper is the linear discriminant analysis of the LP kind. Among the LP discriminant methods, the MSD has been identified to be the most competing alternative to the Fisher discriminant procedure [7, 14]. However, MSD has some limitations. For the linear discriminant analysis of two groups, which is of interest in this paper, the basic MSD is biased to the dominant population group when the population sizes are not equal and may yield a solution which is unbounded even when the normalisation constraint in [7] is imposed. This paper is designed to fix these limitations of the MSD. The literature also suggests the extension of the basic MSD as a future research direction [7]. Rather than using the magnitude of exterior deviations, this paper proposes the use of the proportion of exterior deviations and then introduces the size of the individual group into the objective function of the basic MSD. The resulting model is the minimised sum of deviation by proportion (MSDP) model. The objective of the MSDP is to minimise the sum of the individual proportion of exterior deviation scaled up by the size of the other group. The modified objective function in this paper is due to the assertion by Glover [10].

2. Method

Consider a *p*-characteristic classification problem with two sets of samples **X** and **Y** of size N_1 and N_2, representing two distinct populations 1 and 2, respectively. More precisely, **X** is an $N_1 \times p$ matrix whose *i*th row is the *i*th observation vector from group 1,

$i = 1, 2, ..., N_1$, and $\mathbf{Y}$ is an $N_2 \times p$ matrix whose *l*th row is the *l*th observation vector from group 2, $l = 1, 2, ..., N_2$. Let $\mathbf{X} = (\mathbf{x}_i)$ be a data matrix for population 1, $i = 1, 2, ..., N_1$, with $\mathbf{x}_i$ being a $1 \times p$ vector of the measurements of individual *i* from population 1, and $\mathbf{Y} = (\mathbf{y}_l)$ be a data matrix for population 2, $l = 1, 2, ..., N_2$, with $\mathbf{y}_l$ being a $1 \times p$ vector of the measurements of individual *l* from population 2. To be more precise, $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ and $\mathbf{y}_l = (y_{l1}, y_{l2}, ..., y_{lp})$, where p is the number of characteristics (or features or variables) of interest. This study develops an LP model for discriminant analysis using the proportion of exterior deviations. Exterior deviations are undesirable. We define the proportion of exterior deviations as:

$$p_i = \begin{cases} > 0 & \text{if } \mathbf{x}_i \text{ is misclassified} \\ 0 & \text{otherwise} \end{cases}, \; i = 1, 2, ..., m$$

and

$$q_l = \begin{cases} > 0 & \text{if } \mathbf{y}_l \text{ is misclassified} \\ 0 & \text{otherwise} \end{cases}, \; l = 1, 2, ..., n$$

where m and n are the size of the sample from population group 1 and the size of the sample from population group 2, respectively, used in the training sample, with $m \le N_1$ and $n \le N_2$, $0 \le \sum_{i=1}^{m} p_i \le 1$, $0 \le \sum_{l=1}^{n} q_l \le 1$. For a large data size, $m < N_1$ and $n < N_2$. The remaining data sets $N_1 - m$ and $N_2 - n$ constitute the holdout sample.

Moreover, population groups are usually characterised by an unbalanced dataset. A classification model developed from an unbalanced dataset may be unduly influenced by the observations in the dominant class. When there are fewer observations in a group, say G1, than in another group, say G2, the resultant discriminant function from the mathematical programming method could be biased in favour of the majority group, G2 [5]. The approach in this paper is to deal with unbalanced data by taking into account the consequences of their misclassification arising from exterior deviations. This is achieved by incorporating the size of each group directly into the objective function of the LP model by making m copies of $\sum_{l=1}^{n} q_l$ and n copies of $\sum_{i=1}^{m} p_i$. The resulting representation does not enlarge the model formulation as there are no additional variables or constraints that have been created in dealing with the imbalance. We define the objective function for the LP model as

$$z = n\sum_{i=1}^{m} p_i + m\sum_{l=1}^{n} q_l$$

This objective function is in line with [10]. Let $\mathbf{w}' = (w_1,\ w_2,\ ...,\ w_p),\ \ p \geq 2,$ be an $1 \times p$ vector of the variable coefficients in the discriminant function. The linear discriminant function $(\mathbf{w},\ c)\colon R^p \to \{1,\ 2\}$, defined by a cut-off value, c, and the coefficient vector, $\mathbf{w}$, is to be generated to separate the set of observations into two groups. We modify the discriminant rule $\mathbf{Xw} \leq c\mathbf{1},\ \ \mathbf{Yw} \geq c\mathbf{1}$, by introducing the deviation terms θp_i and θq_l to the right-hand side as follows:

- for group 1

$$\mathbf{x}_i\mathbf{w} \leq c + \theta p_i,\ i = 1,\ 2,\ ...,\ m \leq N_1$$

- for group 2

$$\mathbf{y}_l\mathbf{w} \geq c - \theta q_l,\ l = 1,\ 2,\ ...,\ n \leq N_2$$

The additional terms θp_i and θq_l model the magnitude of the exterior deviations in the classification scheme. That is the magnitude by which the data points $\mathbf{x}_i$ and $\mathbf{y}_l$ lie outside their targeted half-spaces. The value $\theta > 0$ is a predetermined positive number. The introduction of θ in the new LP discriminant model evades the trauma associated with the problem of the non-separable dataset and unbounded solution in the MSD. Since $\mathbf{w} \neq \mathbf{0},$ the entries in $\mathbf{w}' = (w_1,\ w_2,\ ...,\ w_p)$ can either be positive, negative or zero, but not all zero. The usual LP approach to deal with such variable coefficients is to treat them as variables that are unrestricted in sign. In this regard, the variable coefficients in the discriminant function $w_k,\ k = 1,\ 2,\ ...,\ p,$ is represented by a pair of non-negative variables as $w_k = w_k^+ - w_k^-,\ \ w_k^+ \geq 0,\ w_k^- \geq 0.$ There are three possible cases for w_k: (i) $w_k > 0$ if $w_k^+ > w_k^-$, (ii) $w_k < 0$ if $w_k^+ < w_k^-$, or (iii) $w_k = 0$ if $w_k^+ = w_k^-$. Since the case $\mathbf{w} = \mathbf{0}$ is undesirable, there is a need to introduce a normalisation constraint to prevent the solution without discriminating power. We constrain the absolute values w_k and c to sum to a constant, that is

$$\sum_{k=1}^{p} (w_k^+ - w_k^-) + c^+ - c^- = s$$

where s is a non-zero constant. This normalisation constraint is in line with [6]. From the totality of the foregoing, we propose the following LP discriminant analysis model,

which we call the minimised sum of deviation by proportion (MSDP), to find a discriminant function that minimises the proportion of exterior deviations subject to certain constraints as follows:

Minimise

$$z = n\sum_{i=1}^{m} p_i + m\sum_{l=1}^{n} q_l$$

subject to:

- the group separation constraints

$$\sum_{k=1}^{p} w_k x_{ik} - c - \theta p_i \le 0, \qquad \sum_{k=1}^{p} w_k y_{lk} - c + \theta q_l \ge 0$$

- the sign unrestricted expression

$$w_k = w_k^+ - w_k^-, \qquad c = c^+ - c^-$$

w_k, c are unrestricted in sign,

- the normalisation constraint

$$\sum_{k=1}^{p} (w_k^+ - w_k^-) + c^+ - c^- = s$$

s is a non-zero constant

- the upper bound constraints on proportions of exterior deviations

$$\sum_{i=1}^{m} p_i \le 1, \qquad \sum_{l=1}^{n} q_l \le 1$$

- the non-negativity constraint

$$p_i,\ q_l,\ w_k^+,\ w_k^-,\ c^+,\ c^- \ge 0,$$

$i = 1,\ 2,\ ...,\ m \le N_1,\ \ l = 1,\ 2,\ ...,\ n \le N_2,\ \ k = 1,\ 2,\ ...,\ p.$

This LP model contains $(m + n + 3)$ functional constraints and $(m + n + 2p + 2)$ decision variables. The number of decision variables is arrived at by taking the sum of the $(m+n)$ proportion of exterior deviations, p_i, q_l, and the $2(p+1)$ non-negative

variables w_k^+, w_k^-, c^+, c^-. Since the number of decision variables is more than the number of functional constraints, the LP model has an undetermined set of linear functional constrains. A basic solution to the LP model can be obtained by setting the following number of variables equal to zero

$$(m+n+2p+2)-(m+n+3)=2p-1$$

and then solving the reduced system which contains $(m+n+3)$ basic variables. The inclusion of θ in the constraints creates flexibility in the use of the LP model as θ, which is predetermined by the user, is the maximum exterior deviation that can be tolerated. It is worthy of note that the solution obtained for the cut-off value *c*, and the coefficient vector **w**, would not be unique. This is because there are no objective means to choose the non-zero constant *s*, the predetermined value θ, as well as the size of the training sample.

The solution to the LP model is necessary to implement the model. The structure of the LP model can be solved using the big M method or the two-phase method. In either method, the constraints of the original problem are revised by introducing artificial variables as needed to obtain an initial basic feasible solution for the artificial problem. This paper adopts the two-phase method. The reason for this is that the feasibility conditions are quickly determined at the first phase without waiting until optimality is established as in the case of the big M method. The LP model is indeed complex and requires excessive computational efforts. Consequently, a computer program is required to obtain a model solution. This paper utilises the MATLAB package as a computing device.

Suppose $\mathbf{w}^*$, $\mathbf{w}^* \neq 0$, and c^* are the solution obtained by implementing the MSDP on a training sample. Then, the discriminant function is expressed as $\mathbf{x}\mathbf{w}^* = c^*$. We propose the following group-membership discriminant rule. Let $\mathbf{x}$ be an object with *p*-characteristics that is to be classified into either group 1 or group 2. Then, we choose a subset of the training sample and compute the apparent error rate ξ. If $\xi \to 0$ the object $\mathbf{x}$ is classified as belonging to group 1 if $\mathbf{x}\mathbf{w}^* \leq c^*$ and group 2 if $\mathbf{x}\mathbf{w}^* > c^*$. If $\xi \to 1$, the classification rule is reversed, that is, the object $\mathbf{x}$ is classified as belonging to group 2 if $\mathbf{x}\mathbf{w}^* \leq c^*$ and group 1 if $\mathbf{x}\mathbf{w}^* > c^*$. If $\xi \to 0.5$, determine a new discriminant classifier by adjusting the predetermined value θ, and/or the cases of the training sample.

In brief, the MSDP model proposed in this paper involves the following steps.

Step 0. Construct the data matrix, $\mathbf{\Omega} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$, with $S(\mathbf{\Omega})$ as the corresponding set of sample labels of the cases in $\mathbf{\Omega}$, where $\mathbf{X}$ is a matrix of measurements whose *i*th row

is the *i*th observation vector from group 1 and $\mathbf{Y}$ is a matrix of measurements whose *l*th row is the *l*th observation vector from group 2.

Step 1. Partition the data matrix ($\mathbf{\Omega}$) into samples: training sample and holdout sample. Set the size of the training sample, T_e, so that $H_e = S(\mathbf{\Omega})/T_e$ is the corresponding holdout sample.

Step 2. Assume

$\mathbf{x}_i\mathbf{w} \le c + \theta p_i$ for each *i* for group 1,

$\mathbf{y}_l\mathbf{w} \ge c - \theta q_l$ for each *l* for group 2.

Step 3. Use the training sample to construct the MSDP discriminant function based on step 2 as follows:

A. The representative sample of an approximate proportion of the group size concerning T_e is selected from each population group by randomization without replacement.

B. Use T_e to construct a linear programming model (i.e., the MSDP).

C. Revise the constraints of the MSDP by introducing artificial variables as needed to obtain an initial basic feasible solution for the artificial problem of the LP model.

D. Choose a value for the allowable maximum exterior deviation. Then, apply the two-phase simplex method.

E. If the solution is unbounded, repeat Step 3D otherwise go to Step 3F.

F. Use the solution in Step 3D to construct the discriminant function.

Step 4. Compute the apparent error rate from the discriminant function in Step 3.

Step 5. If the apparent error rate is small then use the assumed rule in Step 2 to classify the holdout sample. Otherwise, reverse the classification rule

Step 6. Classify the observations in $H_e = S(\mathbf{\Omega})/T_e$ and compute the hit rate.

3. Numerical illustration

We use an unbalanced dataset in R software that describes road casualties monthly. The R dataset is provided as an example dataset by R, e.g., R i386 3.4.0 software. The data consist of 192 cases of time series data of monthly totals of road casualties before and after the introduction of seat belt law, with 169 cases in one class and 23 cases in the other class. Each observation consists of seven variables, according to the law in effect on a seat belt. The variables are car drivers killed, drivers killed or seriously injured, front-seat passengers killed or seriously injured, rear-seat passengers killed or seriously injured, distance driven in kilometres, petrol price index, number of van (light goods vehicle) drivers and the law in effect. The dataset is accessed in R by

```
>data(package=”datasets”)
>Seatbelts
```

The data can be used to classify the casualty rates, according to whether a seat belt law is in effect for a month or not. The seat belt law makes the wearing of seat belts by front-seat occupants of cars and light goods vehicles compulsory. We use discriminant analysis to predict the law in effect using drivers killed, front, rear, kms, petrol price and van-killed as predictors.

We perform discriminant analysis on the R data using approximately 20% of the cases in each group to form the development sample. The size of the development sample is thirty-nine, and this consists of the first thirty-four cases before the government intervention and the first five cases after the seatbelt law was in effect.

We apply three discriminant methods: the MSD, the linear discriminant analysis in Minitab version 17 (Minitab LDA), and the MSDP. The MSD was implemented in MATLAB, and the results obtained were an optimal solution that is unbounded. This could not be used for discrimination.

We obtain the coefficients and the constant values of the discriminant function for the Minitab LDA as shown in Table 1. For the holdout sample, we obtain a hit rate of 88.24% with a false positive rate (1-specificity) of 100% for the Minitab LDA. The Minitab LDA could discriminate the training sample, but not the holdout sample.

Table 1. Coefficients and constant values of the Minitab LDA

Parameter	0	1
Constant	–1255	–2117
Drivers killed	1	1
Drivers_1	0	0
Front	0	–1
Rear	–1	–1
Kms	0	0
Petrol price	19 427	25 070
Van killed	–3	–8

Applying the MSDP to the R dataset at $\theta = 100$, we obtain the following discriminant function:

$$3.1540x_1 - 0.1964x_2 - 0.1841x_3 + 0.5207x_4 - 0.0001x_5 \\ -152.8394x_6 - 0.0963x_7 = 150.64$$

This discriminant function produces a hit rate of 82.35% with a false positive rate of 55.56%. This false-positive rate may be attributed to the undue influence of the dominant cases of no seatbelt law in effect. Among the three discriminant methods, the MSDP is preferred. This is because the Minitab LDA was insensitive to the period the law was in effect as the false positive rate was 100% for the holdout sample which

means it is extremely biased in favour of no government intervention. The results from the Minitab LDA on the holdout sample indicate that the seatbelt law does not affect at all road casualties. One may erroneously conclude that the Minitab LDA is superior to the MSDP because it has a hit rate of 88.24% which is the proportion of the cases of no seatbelt law in effect in the development sample. Since the Minitab LDA could assign all cases in the holdout sample to the group 0, it has limited practical value for the R dataset. To this end, the MSDP classifier is more appropriate for this classification problem with an imbalance dataset.

4. Conclusion

The MSDP, just like other distribution-free techniques, allows the sets of observations to adjudge themselves as based on the discriminant rule without forcing a distribution on them. This model involves two stages: it obtains a set of characteristic weights via the two-phase method and decides on the decision rule for group-membership predictions using a subset of the training sample. This model has the advantage of circumventing unbounded solutions whenever they are encountered by adjusting the user-defined maximum exterior deviation θ. The utility of our LP model is not in doubt as it is built upon minimising incorrect classification of observations arising from exterior deviations. The Minitab LDA, the MSD and the MSDP, were used to examine a dataset in R. The MSDP appeared more promising and well suited in dealing with the problem of imbalanced data, unlike the Minitab LDA which was less effective. This study contributes to the literature by (i) developing an MSDP model that relies on proportion with a user-specified level for maximum tolerance of exterior deviation, (ii) showcasing the performance of the MSDP model for imbalanced datasets and, (iii) using apparent error rate as a basis for classification rule. Further comparison of the MSDP with other algorithms based on linear separation is suggested for future research.

References

[1] ADEGBOYE O.S., *The optimal classification rule for exponential populations*, Austr. J. Stat., 1993, 35 (2), 185–194.

[2] ABRAMOVICH F., PENSKY M., *Classification with many classes: challenges and pluses*, J. Mult. Anal., 2019, 174. Available at https://doi.org/ 10.1016/j.jmva.2019.104536

[3] AWOGBEMI C.A., ONYEAGU S.I., *Distribution of errors of misclassification for the linear discriminant function (a case of Edgeworth series non-normal distribution)*, Math. Theory Model., 2018, 8 (4), 30–44.

[4] ERENGUC S.S., KOEHLER G.J., *Survey of mathematical programming models and experimental results for linear discriminant analysis*, Manage. Dec. Econ., 1990, 11 (4), 215–225.

[5] FALANGIS K., *Mathematical programming models for classification problems with applications to credit scoring*, PhD Thesis, The University of Edinburgh, Edinburgh 2013.

[6] FALANGIS K., GLEN J.J., *Heuristics for feature selection in mathematical programming discriminant analysis models*, J. Oper. Res. Soc., 2010, 61, 804–812.
[7] FREED N., GLOVER F., *Evaluating alternative linear programming models to solve the two-group discriminant problem*, Dec. Sci., 1986, 17, 151–162.
[8] GAYNANOVA I., WANG T., *Sparse quadratic classification rules via linear dimension reduction*, J. Multiv. Anal., 2019, 169, 278–299.
[9] GLEN J.J., *Classification accuracy in discriminant analysis: a mixed integer programming approach*, J. Oper. Res. Soc., 2001, 52 (3), 328–339.
[10] GLOVER F., *Improved linear programming models for discriminant analysis*, Dec. Sci., 1990, 21, 771–785.
[11] GOCHET W., STAM A., SRINIVASAN V., CHEN S., *Multigroup discriminant analysis using linear programming*, Oper. Res., 1997, 45 (2), 213–225.
[12] KOEHLER G.J., *Considerations for mathematical programming models in discriminant analysis*, Manage. Dec. Econ., 1990, 11 (4), 227–234.
[13] LAM K.F., MOY J.W., *Improved linear programming formulations for the multi-group discriminant problem*, J. Oper. Res. Soc., 1996, 47 (12), 1526–1529.
[14] LAM K.F., CHOO E.U., MOY J.W., *Minimizing deviations from the group mean: a new linear programming approach for the two-group classification problem*, Eur. J. Oper. Res., 1996, 88, 358–367.
[15] LIITTSCHWAGER J.M., WANG C., *Integer programming solution of a classification problem*, Manage. Sci., 1978, 24 (14), 1515–1525.
[16] LIU Y.-H., MALONEY J., *Discriminant analysis and linear programming*, Int. J. Math. Edu. Sci. Technol., 1997, 28 (2), 207–210.
[17] MAKINDE O.S., *On misclassification probabilities of linear and quadratic classifiers*, Afr. Stat., 2016, 11 (1), 943–953.
[18] RENCHER A.C., *Method of Multivariate Analysis* (2nd Ed.), Wiley, New York 2002.
[19] STAM A., JONES D.G., *Classification performance of mathematical programming techniques in discriminant analysis. Results for small and medium sample sizes*, Manage. Dec. Econ., 1990, 11 (4), 243–253.
[20] ZIARI H.A., LEATHAM D.J., ELLINGER P.N., *Development of statistical discriminant mathematical programming model via resampling estimation techniques*, Am. J. Agr. Econ., 1997, 79 (4), 1352–1362.