Hindawi Publishing Corporation
Advances in Operations Research
Volume 2009, Article ID 252989, 34 pages
doi:10.1155/2009/252989

Research Article

Mathematical Programming Approaches to Classification Problems

Soulef Smaoui,[1] Habib Chabchoub,[1] and Belaid Aouni[2]

[1] *Unité de Recherche en Gestion Industrielle et Aide à la Décision, Faculté des Sciences Economiques et de Gestion, Sfax, Tunisia*
[2] *Decision Aid Research Group, School of Commerce and Administration, Faculty of Management, Laurentian University, Sudbury, ON, Canada P3E2C6*

Correspondence should be addressed to Belaid Aouni, baouni@laurentian.ca

Received 22 March 2009; Revised 31 August 2009; Accepted 19 October 2009

Recommended by Mahyar Amouzegar

Discriminant Analysis (DA) is widely applied in many fields. Some recent researches raise the fact that standard DA assumptions, such as a normal distribution of data and equality of the variance-covariance matrices, are not always satisfied. A Mathematical Programming approach (MP) has been frequently used in DA and can be considered a valuable alternative to the classical models of DA. The MP approach provides more flexibility for the process of analysis. The aim of this paper is to address a comparative study in which we analyze the performance of three statistical and some MP methods using linear and nonlinear discriminant functions in two-group classification problems. New classification procedures will be adapted to context of nonlinear discriminant functions. Different applications are used to compare these methods including the Support Vector Machines- (SVMs-) based approach. The findings of this study will be useful in assisting decision-makers to choose the most appropriate model for their decision-making situation.

1. Introduction

Discriminant Analysis (DA) is widely applied in many fields such as social sciences, finance and marketing. The purpose of DA is to study the difference between two or more mutually exclusive groups and to classify this new observation into an appropriate group. The popular method used in DA is a statistical approach. The pioneer of these methods is Fisher [1] who proposed a parametric method introducing linear discriminant functions for two-group classification problems. Somewhat later, Smith [2] introduced a quadratic discriminant function, which along with other discriminant analyses, such as logit and probit, has received a good deal of attention over the past several decades. Some recent researches raise the fact that standard assumptions of DA, such as the normality of the data distribution and the

equality of the variance-covariance matrices, are not always verified. The MP approach has also been widely used in DA and it can be considered a valuable alternative to the classical models of DA. The aim of these MP models is either to minimize the violations (distance between the misclassified observations and the cutoff value) or to minimize the number of misclassified observations. They require no assumptions about the population's distribution and provide more flexibility for the analysis by introducing new constraints, such as those of normalization, or by including weighted deviations in the objective functions including higher weightings for misclassified observation deviations and lower weightings for correctly classified observation deviations. However, special difficulties and even anomalous results restrict the performance of these MP models [3]. These difficulties may be classified under the headings of "degeneracy" and "stability" [4, 5]. The solutions can be classed as degenerate if the analysis presents unbounded solutions in which improvement of the objective function is unconstrained. Similarly, the results can be classed as unstable if, for example, they depend on the position of the data in relation to the origin. A solution would be deemed unacceptable in a situation where all of the coefficients of the discriminant function were equal to zero, thus leading to of all the observations being incorrectly classified in the same group [6, 7]. To overcome these problems, different normalization constraints have been identified and variants of MP formulations for classification problems have been proposed [4, 8–11].

For any given discriminant problem, the choice of an appropriate method for analyzing the data is not always an easy task. Several studies comparing statistical and MP approaches have been carried out by a number of researchers. A number of comparative studies using both statistical and MP approaches have been performed on real data [12–14] and most of them use linear discriminant functions. Recently, new MP formulations have been developed based on nonlinear functions which may produce better classification performance than can be obtained from a linear classifier. Nonlinear discriminant functions can be generated from MP methods by transforming the variables [15], by forming dichotomous categorical variables from the original variables [16], based on piecewise-linear functions [17] and on kernel transformations that attempt to render the data linearly separable, or by using Multihyperplanes formulations [18].

The aim of this paper is, thus, to conduct a comparative study in which we analyze the performance of three statistical methods: (1) the Linear Discriminant Function method (LDF), (2) the Logistic function (LG), and (3) Quadratic Discriminant Function (QDF) along with five MP methods based on linear discriminant functions: the MSD model, the Ragsdale and Stam [19] (RS) model, the model of Lam et al. [12] (LPM), the Lam and Moy [10] model MC, and the MCA model [20]. These methods will be compared to the second-order MSD model [15], the popular SVM-based approach, the piecewise-linear models, and the Multihyperplanes models. New classification procedures adapted to the last models based on nonlinear discriminant functions will be proposed. Different applications in the financial and medicine domains are used to compare the different models. We will examine the conditions under which these various approaches give similar or different results.

In this paper, we report on the results of the different approaches cited above. The paper is organized as follows: first, we discuss the standard MP discriminant models, followed by a presentation of MP discriminant models based on nonlinear functions. Then, we develop new classification model based on piecewise-nonlinear functions and hypersurfaces. Next, we present the datasets used in the analysis process. Finally, we compare the performance of the classical and the different MP models including the SVM-based approach and draw our conclusions.

2. The MP Methods

In general, DA is applied to two or multiple groups. In this paper, we discuss the case of discrimination with two groups. Consider a classification problem with k attributes. Let X_h be an $(n_h\ x\ k)$ matrix representing the attributing scores of a known sample of n_h objects from the group $h(G_h\ h = 1,2)$. Hence, x_{ij} is the value of the jth attribute for the ith object, a_j is the weight assigned to the jth attribute in the linear combination which identifies the hyperplane, and $\sum_h(k\ x\ k)$ is the variance-covariance matrices of group h.

2.1. The Linear MP Models for Classification Problem

In this section, we will present seven MP formulations for classification problem. These formulations assume that all group $G_1(G_2)$ cases are below (above) the cutoff score c. This score defines the hyperplane which allows the two groups to be separated as follows:

$$\sum_{j=1}^{k} x_{ij}a_j \le c \quad \forall i \in G_1, \tag{2.1}$$

$$\sum_{j=1}^{k} x_{ij}a_j > c \quad \forall i \in G_2 \tag{2.2}$$

(a_j and c are free), with c: the cutoff value or the threshold.

The MP models provide unbounded, unacceptable solutions and are not invariant to a shift of origin. To remove these weaknesses, different normalization constraints are proposed: (N1) $\sum_{j=1}^{k} a_j + c = 1$; (N2) $\sum_{j=1}^{k} a_j = c$ [4]; (N3) the normalization constant ± 1, that is, $c = \pm 1$, by defining binary variables $c^+$and c^- such as $c = c^+ - c^-$ with $c^+ + c^- = 1$; and (N4) the normalization for invariance under origin shift [11]. In the normalization (N4), the free variables a_j are represented in terms of two nonnegative variables (a_j^+ and a_j^-) such as $a_j = a_j^+ - a_j^-$ and constraining the absolute values of the $a_j(j = 1,2,\ldots,k)$ to sum to a constant as follows:

$$\sum_{j=1}^{k}\left(a_j^+ + a_j^-\right) = 1. \tag{2.3}$$

By using the normalization (N4), two binary variables ζ_j^+ and ζ_j^- will be introduced in the models in order to exclude the occurrence of both $a_j^+ > 0$ and $a_j^- > 0$ [11]. The definition of ζ_j^+ and ζ_j^- requires the following constraints:

$$\varepsilon\zeta_j^+ \le a_j^+ \le \zeta_j^+, \quad \varepsilon\zeta_j^- \le a_j^- \le \zeta_j^- \quad j = 1,\ldots,k, \tag{2.4}$$

$$\begin{gathered} \zeta_j^+ + \zeta_j^- \le 1 \quad j = 1,\ldots,k, \\ \zeta_j^+ = \frac{0}{1}, \quad \zeta_j^- = \frac{0}{1}, \quad a_j^+ \ge 0, \quad a_j^- \ge 0. \end{gathered} \tag{2.5}$$

The classification rule will assign the observation x_0 into group G_1 if $\sum_{j=1}^{k} x_{0j}a_j \le c$ and into group G_2, otherwise.

2.1.1. MSD Model (Minimize the Sum of Deviations)

The problem can be expressed as follows:

$$\text{minimize} \quad \sum_i d_i \tag{2.6}$$

subject to

$$\sum_{j=1}^{k} x_{ij} a_j \leq c + d_i \quad \forall i \in G_1, \tag{2.6a}$$

$$\sum_{j=1}^{k} x_{ij} a_j > c - d_i \quad \forall i \in G_2 \tag{2.6b}$$

(a_j and c are free and $d_i \geq 0$ for all i), where d_i is the external deviation from the hyperplane for observation i.

The objective takes zero value if the two groups can be separated by the hyperplane. It is necessary to introduce one of the normalization constraints cited above to avoid unacceptable solutions that assign zero weights to all discriminant coefficients.

2.1.2. Ragsdale and Stam Two-Stage Model (RS) [19]

$$\text{minimize} \quad \sum_i d_i \tag{2.7}$$

subject to

$$\sum_{j=1}^{k} x_{ij} a_j - d_i \leq c_1 \quad \forall i \in G_1, \tag{2.7a}$$

$$\sum_{j=1}^{k} x_{ij} a_j + d_i > c_2 \quad \forall i \in G_2 \tag{2.7b}$$

(a_j are free and $d_i \geq 0$ for all i), where c_1 and c_2 are two predetermined constants with $c_1 < c_2$. The values chosen by Ragsdale and Stam are $c_1 = 0$ and $c_2 = 1$. Two methods were proposed to determine the cutoff value. The first method is to choose the cutoff value equal to $(c_1+c_2)/2$. The second method requires the resolution of another LP problem which minimizes only the observation deviations whose classification scores lie between c_1 and c_2. The observations that have classification scores below c_1 or above c_2 are assumed to be correctly classified. The advantage of this latter method is to exclude any observation with classification scores on the wrong side of the hyperplane. However, for simplicity, we use the first method in our empirical study. Moreover, we will solve the model by considering c_1 and c_2 decision variables by adding the constraints:

$$c_2 - c_1 = 1. \tag{2.7c}$$

2.1.3. Lam et al. Method [9, 12]

This model abbreviated as LPC is defined by the following.

$$\text{minimize} \quad \sum_{i\in G_1}\left(d_i^+ + d_i^-\right) + \sum_{i\in G_2}\left(e_i^+ + e_i^-\right) \tag{2.8}$$

subject to

$$\sum_{j=1}^{k} x_{ij}a_j + d_i^- - d_i^+ = c_1 \quad \forall i \in G_1, \tag{2.8a}$$

$$\sum_{j=1}^{k} x_{ij}a_j + e_i^- - e_i^+ = c_2 \quad \forall i \in G_2, \tag{2.8b}$$

$$c_2 - c_1 \geq 1 \text{ (normalization constraint)} \tag{2.8c}$$

(c_1, c_2, a_j are free), where d_i^+ and d_i^- are, respectively, the external and the internal deviations from the discriminant axis to observation i in group 1.

e_i^+ and e_i^- are, respectively, the internal and the external deviations from the discriminant axis to observation i in group 2.

c_1 and c_2 are defined as decision variables and can have different significant definitions.

A particular case of this model is that of Lee and Ord [21] which is based on minimal absolute deviations with $c_1 = 0$ and $c_2 = 1$.

A new formulation of LPC is to choose c_h ($h = 1, 2$) as the mean group of the classification scores of the group h, as follows (LPM):

$$\text{minimize} \quad \sum_{i\in G_1}\left(d_i^+ + d_i^-\right) + \sum_{i\in G_2}\left(e_i^+ + e_i^-\right) \tag{2.9}$$

subject to

$$\sum_{j=1}^{k}\left(x_{ij} - \mu_{1j}\right)\ a_j + d_i^- - d_i^+ = 0 \quad \forall i \in G_1, \tag{2.9a}$$

$$\sum_{j=1}^{k}\left(x_{ij} - \mu_{2j}\right)\ a_j + e_i^- - e_i^+ = 0 \quad \forall i \in G_2, \tag{2.9b}$$

$$\sum_{j=1}^{k}\left(\mu_{2j} - \mu_{1j}\right)a_j \geq 1, \tag{2.9c}$$

with $\mu_{hj} = \sum_{r\in Gh} x_{rj}/n_h$ as the mean of all x_{rj} through the group G_h and n_h is the number of observations in the group G_h.

The objective of the LPM model is to minimize the total deviations of the classification scores from their group mean scores in order to obtain the attribute weights a_j which are considered to be more stable than those of the other LP approaches. The weighting obtained from the resolution of LPM will be utilized to compute the classification scores of all the objects. Lam et al. [12] have proposed two formulations to determine the cutoff value c. One of these formulations consists of minimizing the sum of the deviations from the cutoff value c (LP2).

The linear programing model LP2 is illustrated as follows:

$$\text{minimize} \quad \sum_{i=1}^{n} d_i \tag{2.10}$$

subject to

$$\sum_{j=1}^{k} x_{ij} a_j - d_i \leq c \quad \forall i \in G_1, \tag{2.10a}$$

$$\sum_{j=1}^{k} x_{ij} a_j + d_i \geq c \quad \forall i \in G_2 \tag{2.10b}$$

(c is free, and $d_i \geq 0$).

2.1.4. Combined Method [10]

This method combines several discriminant methods to predict the classification of the new observations. This method is divided into two stages: *the first stage* consists of choosing several discriminant models. Each method is then applied independently. The results from the application of each method provide a classification score for each observation. The group having the higher group-mean classification score is denoted as G_H and the one having the lower group-mean classification score is denoted as G_L. *The second stage* consists of calculating the partial weights of the observations using the scores obtained in the first stage. For group G_H, the partial weight t_{ri} of the ith observation obtained from solving the r method ($r = 1, \ldots, R$ where R is the number of methods utilized) is calculated as the difference between the observation's classification scores and the group-minimum classification score divided by the difference between the maximum and the minimum classification scores:

$$t_{ri} = \frac{\sum_{j=1}^{k} x_{ij} a_{rj} - \text{Min}\left(\sum_{j=1}^{k} x_{ij} a_{rj},\ i \in G_H\right)}{\text{Max}\left(\sum_{j=1}^{k} x_{ij} a_{rj},\ i \in G_H\right) - \text{Min}\left(\sum_{j=1}^{k} x_{ij} a_{rj},\ i \in G_H\right)} \quad i \in G_H. \tag{2.11}$$

The largest partial weight is equal to 1 and the smallest partial weight is equal to zero.

The same calculations are used for each observation of the group G_L, but in this case, the partial weight of each observation is equal to one minus the obtained value in the

calculation. Thus, in this group, the observations with the smallest classification scores are the observations with the greatest likelihood of belonging to this group:

$$t_{ri} = 1 - \frac{\sum_{j=1}^{k} x_{ij}a_{rj} - \text{Min}\left(\sum_{j=1}^{k} x_{ij}a_{rj},\ i \in G_L\right)}{\text{Max}\left(\sum_{j=1}^{k} x_{ij}a_{rj},\ i \in G_L\right) - \text{Min}\left(\sum_{j=1}^{k} x_{ij}a_{rj},\ i \in G_L\right)} \quad i \in G_L. \tag{2.12}$$

The same procedure is repeated for all the discrimination methods used in the combined method. The final combined weight w_i is the sum of all the partial weights obtained. The final combined weights of all observations are used as the weighting for the objective function of the LP model in the second stage. A larger combined weight for one observation indicates that there is little chance that this observation has been misclassified. For each combined weight, the authors add a small positive constant ε in order to ensure that all the observations are entered in the classification model, even for those observations with the smallest partial weights obtained by all the discriminant methods.

The LP formulation which combines the results of different discriminant methods is the following weighting *MSD* (*W-MSD*) model:

$$\text{minimize} \quad \sum_{i} wid_i \tag{2.13}$$

subject to

$$\sum_{j=1}^{k} x_{ij}a_j \leq c + d_i \quad \forall i \in G_1, \tag{2.13a}$$

$$\sum_{j=1}^{k} x_{ij}a_j \geq c - d_i \quad \forall i \in G_2, \tag{2.13b}$$

$$\sum_{j=1}^{k} a_j + c = 1 \tag{2.13c}$$

(a_j and c are free for all j and $d_i \geq 0$ for all i). The advantage of this model is its ability to weight the observations. Other formulations are also possible, for example, the weighting RS model (W-RS).

In our empirical study, the three methods LDF, MSD, and LPM are combined in order to form the combined method MC1. Methods LDF, RS, and LPM are combined in order to form the combined method MC2. Other combined methods are also possible.

2.1.5. The MCA Model [22]

$$\text{maximize} \quad \sum_{h=1}^{2} \sum_{i=1}^{n_h} \beta_{hi} \tag{2.14}$$

subject to

$$\sum_{j=1}^{k} x_{ij}a_j - c + (M+\Delta)\beta_{1i} \le M \quad i \in G_1, \tag{2.14a}$$

$$\sum_{j=1}^{k} x_{ij}a_j - c - (M+\Delta)\beta_{2i} \ge -M \quad i \in G_2 \tag{2.14b}$$

(a_j, c are free, $\beta_{hi} = 0, 1, h = 1, 2$, $i = 1, \ldots, n_h$), with $\beta_{hi} = 1$ if the observation is classified correctly, Δ,$\Delta > 0$ is very small, and $M, M > 0$ is large. The model must be normalized to prevent trivial solutions.

2.1.6. The MIP EDEA-DA Model (MIP EDEA-DA) [23]

Two stages characterize this model:

First stage (classification and identification of misclassified observations) is to

$$\text{minimize } d \tag{2.15}$$

subject to

$$\sum_{j=1}^{k} x_{ij}\left(a_j^+ - a_j^-\right) - c - d \le 0 \quad i \in G_1, \tag{2.15a}$$

$$\sum_{j=1}^{k} x_{ij}\left(a_j^+ - a_j^-\right) - c + d \ge 0 \quad i \in G_2, \tag{2.15b}$$

$$\sum_{j=1}^{k}\left(a_j^+ + a_j^-\right) = 1, \tag{2.15c}$$

$$\varepsilon\zeta_j^+ \le a_j^+ \le \zeta_j^+, \quad \varepsilon\zeta_j^- \le a_j^- \le \zeta_j^- \quad j = 1, \ldots, k, \tag{2.15d}$$

$$\zeta_j^+ + \zeta_j^- \le 1 \quad j = 1, \ldots, k, \tag{2.15e}$$

$$\sum_{j=1}^{k}\left(\zeta_j^+ + \zeta_j^-\right) = k, \tag{2.15f}$$

($\zeta_j^+ = 0/1$, $\zeta_j^- = 0/1$, d and c are free), with $a_j^* = (a_j^{+*} - a_j^{-*})$ with c* and d* being the optimal solution of the model (2.15). There are two cases.

(i) If $d^* < 0$, then there is no misclassified observations and all the observations are classed in either group 1 or group 2 by $\sum_j x_{ij}a_j^* = c^*$. We stop the procedure at this stage.

(ii) If $d^* > 0$, then there are misclassified observations and then comes stage 2 after classifying the observations in these appropriate ensembles (E_1, E_2).

The classification rule is

if $\sum_{j=1}^{k} a_j^* x_{ij} < c^* + d^*$, then $i \in G_1$ $(= E_1)$,

if $\sum_{j=1}^{k} a_j^* x_{ij} > c^* - d^*$, then $i \in G_2$ $(= E_2)$,

if $c^* - d^* \leq \sum_{j=1}^{k} a_j^* x_{ij} \leq c^* + d^*$, then the appropriate group of observation i is determined by the second stage.

Second stage (classification) is to

$$\text{minimize} \quad \sum_{i \in C_1} r_i + \sum_{i \in C_2} r_i \tag{2.16}$$

subject to

$$\sum_{j=1}^{k} x_{ij}\left(a_j^+ - a_j^-\right) - c - Mr_i \leq -\varepsilon \quad i \in C_1, \tag{2.16a}$$

$$\sum_{j=1}^{k} x_{ij}\left(a_j^+ - a_j^-\right) - c + Mr_i \geq 0 \quad i \in C_2, \tag{2.16b}$$

$$\sum_{j=1}^{k}\left(a_j^+ + a_j^-\right) = 1, \tag{2.16c}$$

$$\varepsilon\zeta_j^+ \leq a_j^+ \leq \zeta_j^+ , \quad \varepsilon\zeta_j^- \leq a_j^- \leq \zeta_j^- \quad j = 1, \ldots, k, \tag{2.16d}$$

$$\zeta_j^+ + \zeta_j^- \leq 1 \quad j = 1, \ldots, k, \tag{2.16e}$$

$$\sum_{j=1}^{k}\left(\zeta_j^+ + \zeta_j^-\right) = k, \tag{2.16f}$$

where $(\zeta_j^+ = 0/1,\ \zeta_j^- = 0/1,\ a_j^+ \geq 0,\ a_j^- \geq 0,\ r_i = 0/1,$ and c is free), with $C_1 = G_1 - E_1,\ C_2 = G_2 - E_2$.

The classification rule is

if $\sum_{j=1}^{k} a_j^* x_{ij} \leq c^* - \varepsilon$, then $i \in G_1$,

if $\sum_{j=1}^{k} a_j^* x_{ij} > c^*$, then $i \in G_2$.

The advantage of this model is to minimize the number of misclassified observations. However, the performance of the model depends on the choice of numbers M and ε which are subjectively determined by the searcher and depends also on the choice of the computer science used for resolving the model.

2.2. The Nonlinear MP Models

2.2.1. The Second-Order MSD Formulation [15]

The form of the second-order MSD model is to

$$\text{minimize} \quad \sum_{i\in G_1} d_{1i}^{+} + \sum_{i\in G_2} d_{2i}^{-} \tag{2.17}$$

subject to

$$\sum_{j=1}^{k} x_{ij}a_{jL} + \sum_{j=1}^{k} x_{ij}^{2}a_{jQ} + \sum_{h\neq m} x_{ih}x_{im}a_{hm} + d_{1i}^{-} - d_{1i}^{+} \leq c \quad \forall i \in G_1, \tag{2.17a}$$

$$\sum_{j=1}^{k} x_{ij}a_{jL} + \sum_{j=1}^{k} x_{ij}^{2}a_{jQ} + \sum_{h\neq m} x_{ih}x_{im}a_{hm} + d_{2i}^{-} - d_{2i}^{+} \geq c \quad \forall i \in G_2, \tag{2.17b}$$

$$\sum_{j=1}^{k} a_{jL} + \sum_{j=1}^{k} a_{jQ} + \sum_{h\neq m} a_{hm} + c = 1 \tag{2.17c}$$

where $(a_{jL}, a_{jQ}, a_{hm}$ are free, $h,\ j,\ m = 1,\ldots,k, d_{ri}^{+}, d_{ri}^{-} \geq 0,\ r = 1,2,\ i = 1,\ldots,n)$, a_{jL} is the coefficient for the linear terms x_{ij} of attribute j, a_{jQ} are the coefficients for quadratic terms x_{ij}^{2} of attribute j, a_{hm} are the coefficients for the cross-product terms involving attributes h and m, d_{ri}^{+}, d_{ri}^{-} are the external deviations of group r observations, and c is the cutoff value.

The constraint (2.17c) is the normalization constraints which prevent trivial solution. Other normalization constraints are also possible [4, 11]. It is interesting to note that the cross-product terms can be eliminated from the model when the attributes are uncorrelated [15].

In order to reduce the influence of the group size and give more importance to each group deviation costs, we propose the replacement of the objective function (2.17) by the following function (2.17′):

$$(1-\lambda)\frac{\sum_{i=1}^{n_1} \mathrm{d}_{1i}^{+}}{n_1} + \lambda\frac{\sum_{i=1}^{n_2} \mathrm{d}_{2i}^{-}}{n_2}, \tag{2.17′}$$

with $\lambda \in [0,1]$ a constant representing the relative importance of the cost associated with misclassification of the first and the second groups.

The classification rule is

$$\begin{aligned} &\text{if } \sum_{j} x_{ij}a_{jL}^{*} + \sum_{j} x_{ij}^{2}a_{jQ}^{*} + \sum_{h\neq m} x_{ih}x_{im}a_{hm}^{*} \leq c^{*}, \quad \text{then } x_0 \in G_1, \\ &\text{otherwise} \quad x_0 \in G_2. \end{aligned} \tag{2.18}$$

2.2.2. The Piecewise-Linear Models [17]

Recently, two piecewise-linear models are developed by Glen: the MCA and MSD piecewise models. These methods suppose the nonlinearity of the discriminant function. This

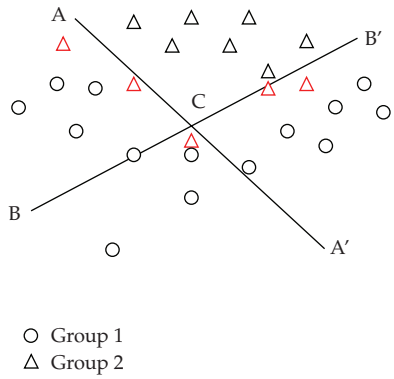

Figure 1: A two-group discriminant problem in two variables.

nonlinearity is approximated by piecewise-linear functions. The concept is illustrated in Figure 1.

In Figure 1, the piecewise-linear functions are ACB′ and BCA′, while the component linear functions are represented by the lines AA′ and BB′. Note that for the piecewise-linear function ACB′, respectively, BCA′, the region of correctly classified group 2 (group 1) is convex. However, the region for correctly classified group 1 (group 2) observations is nonconvex. The optimal of the linear discriminant function is obtained when the two precedent cases are considered separately. The MP must be solved twice: once to constrain all of group 1 elements to a convex region and once to constrain all of group 2 elements to a convex region. Only the second case is considered in developing the following MP models.

(a) The Piecewise-Linear MCA Model [17]

The MCA model for generating a piecewise-linear function in s segment is:

$$\text{maximize} \quad \sum_{h=1}^{2} \sum_{i=1}^{n_h} \beta_{hi} \tag{2.19}$$

subject to

$$\sum_{j=1}^{k} x_{ij}\left(a_{lj}^{+} - a_{lj}^{-}\right) + (M+\varepsilon)\delta_{li} \le c_l + M \quad i \in G_1,\ l = 1,\ldots,s, \tag{2.19a}$$

$$\sum_{j=1}^{k} x_{ij}\left(a_{lj}^{+} - a_{lj}^{-}\right) - (M+\varepsilon)\beta_{2i} \ge c_l - M \quad i \in G_2,\ l = 1,\ldots,s, \tag{2.19b}$$

$$\sum_{l=1}^{s} \delta_{li} - \beta_{1i} \ge 0 \quad i \in G_1, \tag{2.19c}$$

$$\sum_{j=1}^{k}\left(a_{lj}^{+} + a_{lj}^{-}\right) = 1 \quad l = 1,\ldots,s, \tag{2.19d}$$

where c_l is free, $a_{lj}^{+}, a_{lj}^{-} \geq 0$, $l = 1, \ldots, s$, $j = 1, \ldots, k, \beta_{hi} = 0, 1$, $h = 1, 2$, $i = 1, \ldots, n_h$, and $\delta_{li} = 0, 1$, $l = 1, 2, \ldots, s$ $i = 1, \ldots, n_1$, with ε, $\varepsilon > 0$ being a small interval, within which the observations are considered as misclassified, and M is a positive large number,

$\beta_{hi} = 1$ if the observation is correctly classified,

$\delta_{li} = 1$ $(i \in G_1)$, if the group 1 observation is correctly classified by function l on its own.

The correctly classified group 2 observation can be identified by the s constraints of type (2.19b). An observation of group 1 is correctly classified only if it is correctly classified by at least one of the s segments of the piecewise-linear function (constraint (2.19c)).

The classification rule of an observation x_0 is

$$\begin{aligned} &\text{if } \sum_{j=1}^{k} x_{0j} a_{lj}^{*} \leq c_l^{*}, \quad \text{then } x_0 \in G_1, \\ &\text{otherwise} \quad x_0 \in G_2. \end{aligned} \tag{2.20}$$

A similar model must also be constructed for the case in which the nonconvex region is associated with group 2 and the convex region is associated with group 1.

(b) The Piecewise-Linear MSD Model [17]:

$$\text{minimize} \quad \sum_{h=1}^{2} \sum_{i=1}^{n_h} d_{hi}, \tag{2.21}$$

subject to

$$\sum_{j=1}^{k} x_{ij} \left(a_{lj}^{+} - a_{lj}^{-} \right) - e_{li} \leq c_l - \varepsilon \quad i \in G_1,\ l = 1, ..., s, \tag{2.21a}$$

$$\sum_{j=1}^{k} x_{ij} \left(a_{lj}^{+} - a_{lj}^{-} \right) + f_{li} \geq c_l + \varepsilon \quad i \in G_2,\ l = 1, ..., s, \tag{2.21b}$$

$$\sum_{j=1}^{k} \left(a_{lj}^{+} + a_{lj}^{-} \right) = 1 \quad l = 1, \ldots, s, \tag{2.21c}$$

$$d_{2i} - f_{li} \geq 0 \quad i \in G_2,\ l = 1, \ldots, s, \tag{2.21d}$$

$$e_{li} - e_{pi} + U\delta_{li} \leq U \quad i \in G_1,\ l = 1, \ldots, s,\ p = 1, \ldots, s\ (p \neq l), \tag{2.21e}$$

$$d_{1i} - e_{li} + U\delta_{li} \geq -U \quad i \in G_1,\ l = 1, \ldots, s, \tag{2.21f}$$

$$\sum_{l=1}^{s} \delta_{li} = 1 \quad i \in G_1, \tag{2.21g}$$

where c_l is free, $a_{lj}^{+}, a_{lj}^{-} \geq 0$, $(l = 1, \ldots, s,\ j = 1, \ldots, k), d_{hi} = 0,1$ $(h = 1,2,\ i = 1, \ldots, n_h), e_{li} \geq 0, \delta_{li} = 0,1$ $(l = 1, \ldots, s,\ i = 1, \ldots, n_1)$, and $f_{li} \geq 0$ $(l = 1, \ldots, s,\ i = 1, \ldots, n_2)$, with ε, $\varepsilon > 0$ being a small interval and U, $U > 0$ being an upper bound on e_{li}.

e_{li} is the deviation of group 1 observation i from component function l of the piecewise-linear function, where $e_{li} = 0$ if the observation is correctly classified by function l on its own and $e_{li} > 0$ if the observation is misclassified by function l on its own.

f_{li} is the deviation of group 2 observation i from component function l of the piecewise-linear function, where $f_{li} = 0$ if the observation is correctly classified by function l on its own and $f_{li} > 0$ if the observation is misclassified by function l on its own. A group 2 observation is correctly classified if it is classified by each of the s component functions.

d_{2i} is the lower bound on the deviation of group 2 observation i from the s segment piecewise-linear discriminant function, where $d_{2i} = 0$ if the observation is correctly classified and $d_{2i} > 0$ if the observation is misclassified.

The binary variable δ_{li} is introduced in the model to determine d_{1i} by detecting the minimal deviation e_{li}.

The classification rule is the same as that of the piecewise-linear MCA.

The two piecewise-linear MCA and MSD models must be solved twice: once to consider all group 1 observations in convex region and once to consider all group 2 observations in convex region, in order to obtain the best classification. Other models have been developed by Better et al. [18]. These models are more effective for more complex datasets than for the piecewise-linear models and do not require that one of the groups belong to a convex region.

2.2.3. The Multihyperplanes Models [18]

The multihyperplanes models can be interpreted as models identifying many hyperplanes used successively. The objective is to generate tree conditional rules to separate the points. This approach constitutes an innovation in the area of Support Vector Machines (SVMs) in the context of successive perfect separation decision tree. The advantage of this approach is to construct a nonlinear discriminant function without the need for kernel transformation of the data as in SVM. The first model using multihyperplanes is the Successive Perfect Separation decision tree (SPS).

(a) The Successive Perfect Separation Decision Tree (SPS)

The specific structure is developed in the context of SPS decision tree. The decision tree is a tree which results from the application of the SPS procedure. In fact, this procedure permits, at each depth $l < D$, to compel all the observations of either group 1 or group 2 to lie on one side of the hyperplane. Thus, at each depth the tree has one leaf node that terminates the branch that correctly classifies observations in a given group.

In Figure 2, the points represented as circles and triangles must be separate. The PQ, QR, and RS segments of the three hyperplanes separate all the points. We can remark that the circles are correctly classified either by H_1 or by H_2 and H_3. However, the triangles are correctly classified by the tree if it is correctly classified by H_1 and H_2 or by H_1 and H_3. Several tree types are possible. Specific binary variables called "slicing variables" are used to describe the specific structure of the tree. These variables define how the tree is sliced in order to classify an observation correctly.

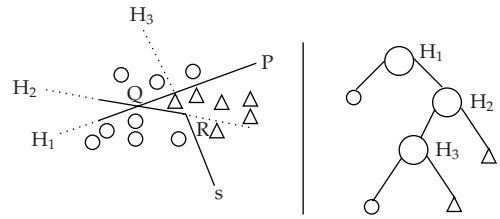

Figure 2: One particular type of SPS tree for $D = 3$.

The specific structure SPS decision tree model is formulated as follows:

$$\begin{aligned} &D = 3, \\ &\text{minimize} \quad \sum_{i=1}^{n} \delta_i^* \end{aligned} \tag{2.22}$$

subject to

$$\sum_{j=1}^{k} x_{ij} a_{jd} - M\delta_{di} \le c_d - \varepsilon \quad i \in G_1,\ d = 1,2,3, \tag{2.22a}$$

$$\sum_{j=1}^{k} x_{ij} a_{jd} + M\delta_{di} \ge c_d + \varepsilon \quad i \in G_2,\ d = 1,2,3, \tag{2.22b}$$

$$M(sl_1 + sl_2) + \delta_i^* \ge \delta_{1i} + \delta_{2i} + \delta_{3i} - 2 \quad i \in G_1, \tag{2.22c}$$

$$M(sl_1 + sl_2) + M\delta_i^* \ge \delta_{1i} + \delta_{2i} + \delta_{3i} \quad i \in G_2, \tag{2.22d}$$

$$M(2 - sl_1 - sl_2) + M\delta_i^* \ge \delta_{1i} + \delta_{2i} + \delta_{3i} \quad i \in G_1, \tag{2.22e}$$

$$M(2 - sl_1 - sl_2) + \delta_i^* \ge \delta_{1i} + \delta_{2i} + \delta_{3i} - 2 \quad i \in G_2, \tag{2.22f}$$

$$M(1 + sl_1 - sl_2) + \delta_i^* \ge \delta_{1i} - M\mu_i \quad i \in G_1, \tag{2.22g}$$

$$M(1 + sl_1 - sl_2) + M\delta_i^* \ge \delta_{2i} + \delta_{3i} - M[1 - \mu_i] \quad i \in G_1, \tag{2.22h}$$

$$M(1 + sl_1 - sl_2) + \delta_i^* \ge \delta_{1i} \quad i \in G_2, \tag{2.22i}$$

$$M(1 + sl_1 - sl_2) + \delta_i^* \ge \delta_{2i} + \delta_{3i} - 1 \quad i \in G_2, \tag{2.22j}$$

$$M(1 - sl_1 + sl_2) + \delta_i^* \ge \delta_{1i} \quad i \in G_1, \tag{2.22k}$$

$$M(1 - sl_1 + sl_2) + \delta_i^* \ge \delta_{2i} + \delta_{3i} - 1 \quad i \in G_1, \tag{2.22l}$$

$$M(1 - sl_1 + sl_2) + \delta_i^* \ge \delta_{1i} - M\mu_i \quad i \in G_2, \tag{2.22m}$$

$$M(1 - sl_1 + sl_2) + M\delta_i^* \ge \delta_{2i} + \delta_{3i} - M[1 - \mu_i] \quad i \in G_2, \tag{2.22n}$$

$$\sum_{j=1}^{k} \sum_{d=1}^{3} a_{jd} = 1, \tag{2.22o}$$

noting that $\delta_i \in \{0,1\}$ $(i \in G_1 \cup G_2)$, $\delta_{di} \in \{0,1\}$ $(i \in G_1 \cup G_2,\ d = 1,2,3)$, $\mu_i \in \{0,1\}$ $(i \in G_1 \cup G_2)$, $sl_t \in \{0,1\}$ $(t = 1,2)$, and a_{jd}, c_d are free $(j = 1,\ldots,k)$, where M is large, while ε is very small constant. Consider the following:

$$\delta_i^* = \begin{cases} 0 & \text{if } i \text{ is correctly classified by the tree,} \\ 1 & \text{otherwise,} \end{cases}$$
$$\delta_{di} = \begin{cases} 0 & \text{if } i \text{ is correctly classified by hyperplane } d, \\ 1 & \text{otherwise.} \end{cases} \tag{2.23}$$

The (2.22c) and (2.22d) constraints represent the type of tree $(0,0)$ and are activated when $sl_1 = 0$ and $sl_2 = 0$. Similarly, the (2.22e) and (2.22f) constraints for tree type $(1,1)$ will only be activated when $sl_1 = 1$ *and* $sl_2 = 1$. However, for the tree types $(0,1)$ and $(1,0)$ corresponding to (2.22g)–(2.22n) constraints, a binary variable μ_i is introduced in order to activate or deactivate either of the constraints relevant to these trees. In fact, when $sl_1 = 0$ *and* $sl_2 = 1$, the (2.22g)–(2.22j) constraints for tree type $(0,1)$ will be activated so that an observation from group 1 will be correctly classified by the tree if it is correctly classified either by the first hyperplane or by both the second and the third hyperplanes. On the other hand, an observation from group 2 is correctly classified by the tree if it is correctly classified either by the hyperplanes 1 and 2 or by the hyperplanes 2 and 3.

This classification is established in the case where $\mu_i = 0$ which permits to activate constraints (2.22g) and (2.22h). The case that corresponds to tree type $(1,0)$ is just a "mirror image" of previous case. However, the model becomes difficult to resolve when the number of possible tree types increases (D large). In fact, as D increases, the number of possible tree types increases and so does the number of constraints. For these reasons, Better et al. [18] developed the following model.

(b) The General Structure SPS Model (GSPS)

$$\text{minimize} \quad \sum_{i=1}^{n} \delta_i[D] \tag{2.24}$$

subject to

$$\sum_{j=1}^{k} x_{ij} a_{jd} - M\left(\sum_{h=1}^{d-1} v_{ih} + \delta_{id}\right) \le c_d - \varepsilon \quad i \in G_1,\ d = 1,\ldots,D, \tag{2.24a}$$

$$\sum_{j=1}^{k} x_{ij} a_{jd} + M\left(\sum_{h=1}^{d-1} v_{ih} + \delta_{id}\right) \ge c_d + \varepsilon \quad i \in G_2,\ d = 1,\ldots,D, \tag{2.24b}$$

$$\mu_d \ge \delta_{id} \quad i \in G_1,\ d = 1,\ldots,D-1, \tag{2.24c}$$

$$1 - \mu_d \ge \delta_{id} \quad i \in G_2,\ d = 1,\ldots,D-1, \tag{2.24d}$$

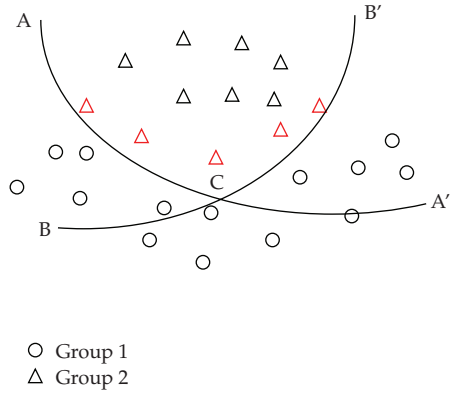

Figure 3: A two-group discriminant problem.

$$v_{id} \leq \mu_d \quad i \in G_1,\ d = 1,\ldots,D-1, \tag{2.24e}$$

$$v_{id} \leq 1 - \mu_d \quad i \in G_2,\ d = 1,\ldots,D-1, \tag{2.24f}$$

$$v_{id} \leq 1 - \delta_{id} \quad i \in G_1 \cup G_2,\ d = 1,\ldots,D-1, \tag{2.24g}$$

$$\sum_{j=1}^{k} \sum_{d=1}^{D} a_{jd} = 1, \tag{2.24h}$$

where $\delta_{di} \in \{0,1\}$ $(i \in G_1 \cup G_2,\ d = 1,\ldots,D)$, $\mu_d \in \{0,1\}$ $(i \in G_1 \cup G_2,\ d = 1,\ldots,d-1)$, $0 \leq v_{id} \leq 1$ $(d = 1,\ldots,D-1)$, and a_{jd}, c_d are free $(j = 1,\ldots,k,\ d = 1,\ldots,D)$. The variables μ_d and v_{id} are not included in the final hyperplane (D). The variable μ_d is defined as

$$\mu_d = \begin{cases} 0 & \text{if all } i \in G_1 \text{ are compelled to lie on one side of hyperplane } d, \\ 1 & \text{if all } i \in G_1 \text{ are compelled to lie on one side of hyperplane } d. \end{cases} \tag{2.25}$$

The constraints (2.24c) and (2.24d) permit to lie all group 1 or group 2 observations on one side of the hyperplane according to μ_d value. In fact, due to constraint (2.24c), if $\mu_d = 0$, all group 1 observations and possibly some group 2 observations lie on one side of the hyperplane d. However, only observations of group 2 will lie on the other side of hyperplane d and so these observations can be correctly classified. Conversely, due to constraint (2.24d), if $\mu_d = 1$, the observations correctly classified by the tree will be those belonging to group 1. The variables v_{i1} permit to identify the correctly classified and misclassified observations of each group from the permanent value 1—δ_{i1}. In fact, in the case where $\mu_1 = 1$, the permanent values δ_{i1} to establish are those of group1 observations such that $\delta_{i1} = 0$, because these particular observations are separate in such manner that we do not need to consider them again. Thus, for these last observations, the fact that $\mu_1 = 1$ and $\delta_{i1} = 0$ forces the v_{i1} to equal 1. If we consider the case to force $v_{i1} = 0$ for group 1 observations, it means that these observations have not yet permanently separated from group 2 observations and one or more hyperplanes are necessary to separate them. Thus, $v_{i1} = 0$ if $\mu_1 = 0$ or $\delta_{i1} = 1$ (verified by the constraints (2.24e) and (2.24g)).

For the empirical study, the SPS and GSPS model will be resolved using the two following normalization constraints:

$$\begin{aligned} &\text{N}'1 \quad \sum_{j=1}^{k}\sum_{d=1}^{D} a_{jd} = 1, \\ &\qquad\qquad\qquad\qquad d = 1,\ldots,D, \\ &\text{N}'2 \quad \sum_{j=1}^{k} a_{jd} = 1. \end{aligned} \tag{2.26}$$

The developed models presented previously are based either on piecewise-linear separation or on the multihyperplanes separation. New models based on piecewise-nonlinear separation and on multihypersurfaces are proposed in the next section.

3. The Proposed Models

In this section different models are proposed. Some use the piecewise-nonlinear separation and the others use the multihypersurfaces.

3.1. The Piecewise-Nonlinear Models (Quadratic Separation)

The piecewise-linear MCA and MSD models are based on piecewise-linear functions. To ameliorate the performance of the models, we propose two models based on piecewise-nonlinear functions. The base concept of these models is illustrated in Figure 3.

The curves AA′ and BB′ represent the piecewise-nonlinear component functions: ACB′ and BCA′. The interpretations are the same as those in Figure 1. However, we can remark that the use of piecewise-nonlinear functions permits to minimize the number of misclassified observations. Based on this idea, we suggest proposing models based on piecewise-nonlinear functions. In these models we suggest to replace the first constraints of piecewise-linear MCA and MSD models by the linear constraints which are nonlinear in terms of the attributes as follows:

$$\sum_{j} x_{ij}a_{ljL} + \sum_{j} x_{ij}^{2}a_{ljQ} + \sum_{h\neq m} x_{ih}x_{im}a_{lhm} + (M+\varepsilon)\delta_{li} \leq c_l + M \quad \forall i \in G_1 \; l = 1,\ldots,s, \tag{3.22a}$$

$$\sum_{j} x_{ij}a_{ljL} + \sum_{j} x_{ij}^{2}a_{ljQ} + \sum_{h\neq m} x_{ih}x_{im}a_{lhm} - (M+\varepsilon)\beta_{2i} \geq c_l - M \quad \forall i \in G_2 \; l = 1,\ldots,s, \tag{3.22b}$$

where a_{ljL}, a_{ljQ}, a_{lhm} *are unrestricted in sign,* $h, j, m = 1,\ldots,k$ *and* $l = 1,\ldots,s$, a_{ljL} are the linear terms of attribute j for the function l, a_{ljQ} are the quadratic terms of attribute j for the function l, a_{lhm} are the cross-product terms of attributes h and m for the function l.

Note that if the attributes are uncorrelated, the cross-product terms can be excluded from the models. Other general nonlinear terms can, also, be included in the models. On the other hand, the normalization constraint is replaced by the following constraint:

$$\sum_{j} a_{ljL} + \sum_{j} a_{ljQ} + \sum_{h\neq m} a_{lhm} = 1 \quad l = 1,\ldots,s. \tag{3.22c}$$

The piecewise-quadratic separation models obtained are the following.

3.1.1. The Piecewise-Quadratic Separation MCA Model (QSMCA)

$$\text{maximize} \quad \sum_{r=1}^{2}\sum_{i=1}^{n_r}\beta_{ri} \tag{3.23}$$

subject to

$$\sum_{j=1}^{k} x_{ij}a_{ljL} + \sum_{j=1}^{k} x_{ij}^2 a_{ljQ} + \sum_{h\neq m} x_{ih}x_{im}a_{lhm} + (M+\varepsilon)\delta_{li} \leq c_l + M \quad \forall i \in G_1 \; l = 1,\ldots,s, \tag{3.23a}$$

$$\sum_{j=1}^{k} x_{ij}a_{ljL} + \sum_{j=1}^{k} x_{ij}^2 a_{ljQ} + \sum_{h\neq m} x_{ih}x_{im}a_{lhm} - (M+\varepsilon)\beta_{2i} \leq c_l + M \quad \forall i \in G_2 \; l = 1,\ldots,s, \tag{3.23b}$$

$$\sum_{l=1}^{s} \delta_{li} - \beta_{1i} \geq 0 \quad i \in G_1, \tag{3.23c}$$

$$\sum_{j=1}^{k} a_{ljL} + \sum_{j=1}^{k} a_{ljQ} + \sum_{h\neq m} a_{lhm} = 1 \quad l = 1,\ldots,s, \tag{3.23d}$$

where c_l, a_{ljL}, a_{ljQ}, a_{lhm} *are unrestricted in sign* ($h, j, m = 1,\ldots,k$ and $l = 1,\ldots,s$), $\beta_{ri} = 0,1$ ($r = 1,2$, $i = 1,\ldots,n_r$), *and* $\delta_{li} = 0,1$ ($l = 1,2,\ldots,s$ $i = 1,\ldots,n_1$). The classification rule of an observation x_0 is

$$\begin{aligned} &\text{if } \sum_{j=1}^{k} x_{0j}a_{ljL}^{*} + \sum_{j=1}^{k} x_{0j}x_{0j}a_{ljQ}^{*2} + \sum_{h\neq m} x_{0h}x_{0m}a_{lhm}^{*} \leq c_l^{*}, \quad \text{then } x_0 \in G_1, \\ &\text{otherwise} \quad x_0 \in G_2. \end{aligned} \tag{3.24}$$

3.1.2. The Piecewise-Quadratic Separation MSD Model (QSMSD)

$$\text{minimize} \quad \sum_{r=1}^{2}\sum_{i=1}^{n_r} d_{ri} \tag{3.25}$$

subject to

$$\sum_{j=1}^{k} x_{ij}a_{ljL} + \sum_{j=1}^{k} x_{ij}^2 a_{ljQ} + \sum_{h\neq m} x_{ih}x_{im}a_{lhm} - e_{li} \leq -c_l - \varepsilon \quad i \in G_1, \; l = 1,\ldots,s, \tag{3.25a}$$

$$\sum_{j=1}^{k} x_{ij}a_{ljL} + \sum_{j=1}^{k} x_{ij}^2 a_{ljQ} + \sum_{h\neq m} x_{ih}x_{im}a_{lhm} + f_{li} \geq c_l + \varepsilon \quad i \in G_2, \; l = 1,\ldots,s, \tag{3.25b}$$

$$\sum_{j=1}^{k} a_{ljL} + \sum_{j=1}^{k} a_{ljQ} + \sum_{h\neq m} a_{lhm} = 1 \quad l = 1,\ldots,s, \tag{3.25c}$$

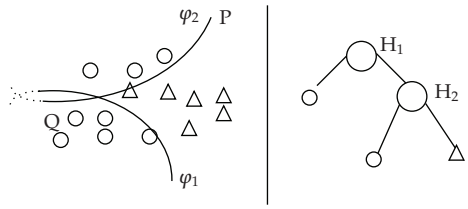

Figure 4: One particular type of QSS tree for $D = 2$.

$$d_{2i} - f_{li} \geq 0 \quad i \in G_2,\ l = 1, \ldots, s, \tag{3.25d}$$

$$e_{li} - e_{pi} + U\delta_{li} \leq U \quad i \in G_1,\ l = 1, \ldots, s,\ p = 1, \ldots, s\ (p \neq l), \tag{3.25e}$$

$$d_{1i} - e_{li} + U\delta_{li} \geq -U \quad i \in G_1,\ l = 1, \ldots, s, \tag{3.25f}$$

$$\sum_{l=1}^{s} \delta_{li} = 1 \quad i \in G_1, \tag{3.25g}$$

where c is free, $a_{lj}^{+}, a_{lj}^{-} \geq 0$ $(l = 1, \ldots, s,\ j = 1, \ldots, k), d_{ri} = 0, 1$ $(r = 1, 2,\ i = 1, \ldots, n_r), e_{li} \geq 0$ $(\delta_{li} = 0, 1\ \ l = 1, \ldots, s,\ i = 1, \ldots, n_1)$, and $f_{li} \geq 0$ $(l = 1, \ldots, s,\ i = 1, \ldots, n_2)$. The interpretation of this model is the same as that of piecewise-linear MSD model. The classification rule is the same as that of QSMCA model.

The construction of piecewise QSMSD and QSMCA models, using the case in which group 1 is in the convex region and the group 2 in the nonconvex region, is also valuable. However, despite the complexity of these models (especially when the datasets are very large), the advantage of piecewise QSMSD and QSMCA models is accelerating the reach of an optimal solution using a reduced number of arcs than segments. But, the disadvantage of the models remains the necessity to resolve twice these models: the case in which group1 is convex and the case in which group 2 is convex. The following quadratic specific structure models could be a way of solving these problems, accelerating the reach of possible solutions to the different models and finding answers to problems of large size.

3.2. The Quadratic Specific Structure Models (QSS)

The quadratic specific structure models are based on the use of nonlinear separation. The following figure illustrates a particular case of the QSS models.

In Figure 4, the points are separated using two curves $\varphi_1$and φ_2. The circle are well classified by φ_1 or by φ_2 and the triangles are well classified by φ_1 and φ_2. As for SPS and GSPS, many tree-specific structures are possible for QSS models. Based on this idea, the quadratic SPS and the quadratic GSPS models are proposed.

3.2.1. The Quadratic SPS Model (QSPS)

Similar to the piecewise QSMSD and QSMCA models, the first constraints of SPS model are replaced by the linear constraints which are nonlinear in terms of the attributes. The QSPS

model is the following:

$$\begin{aligned} &D = 3, \\ &\text{minimize} \quad \sum_{i=1}^{n} \delta_i^* \end{aligned} \tag{3.26}$$

subject to

$$\sum_{j=1}^{k} x_{ij} a_{djL} + \sum_{j=1}^{k} x_{ij}^2 a_{djQ} + \sum_{h \neq m} x_{ih} x_{im} a_{dhm} - M\delta_{di} \leq c_d - \varepsilon \quad i \in G_1,\ d = 1,2,3, \tag{3.26a}$$

$$\sum_{j=1}^{k} x_{ij} a_{ljL} + \sum_{j=1}^{k} x_{ij}^2 a_{ljQ} + \sum_{h \neq m} x_{ih} x_{im} a_{dhm} + M\delta_{di} \geq c_d + \varepsilon \quad i \in G_2,\ d = 1,2,3, \tag{3.26b}$$

$$M(sl_1 + sl_2) + \delta_i^* \geq \delta_{1i} + \delta_{2i} + \delta_{3i} - 2 \quad i \in G_1, \tag{3.26c}$$

$$M(sl_1 + sl_2) + M\delta_i^* \geq \delta_{1i} + \delta_{2i} + \delta_{3i} \quad i \in G_2, \tag{3.26d}$$

$$M(2 - sl_1 - sl_2) + M\delta_i^* \geq \delta_{1i} + \delta_{2i} + \delta_{3i} \quad i \in G_1, \tag{3.26e}$$

$$M(2 - sl_1 - sl_2) + \delta_i^* \geq \delta_{1i} + \delta_{2i} + \delta_{3i} - 2 \quad i \in G_1, \tag{3.26f}$$

$$M(1 + sl_1 - sl_2) + \delta_i^* \geq \delta_{1i} - M\mu_i \quad i \in G_1, \tag{3.26g}$$

$$M(1 + sl_1 - sl_2) + M\delta_i^* \geq \delta_{2i} + \delta_{3i} - M\left[1 - \mu_i\right] \quad i \in G_1, \tag{3.26h}$$

$$M(1 + sl_1 - sl_2) + \delta_i^* \geq \delta_{1i} \quad i \in G_2, \tag{3.26i}$$

$$M(1 + sl_1 - sl_2) + \delta_i^* \geq \delta_{2i} + \delta_{3i} - 1 \quad i \in G_2, \tag{3.26j}$$

$$M(1 - sl_1 + sl_2) + \delta_i^* \geq \delta_{1i} \quad i \in G_1, \tag{3.26k}$$

$$M(1 - sl_1 + sl_2) + \delta_i^* \geq \delta_{2i} + \delta_{3i} - 1 \quad i \in G_1, \tag{3.26l}$$

$$M(1 - sl_1 + sl_2) + \delta_i^* \geq \delta_{1i} - M\mu_i \quad i \in G_2, \tag{3.26m}$$

$$M(1 - sl_1 + sl_2) + M\delta_i^* \geq \delta_{2i} + \delta_{3i} - M\left[1 - \mu_i\right] \quad i \in G_2, \tag{3.26n}$$

$$\sum_{j=1}^{k} a_{djL} + \sum_{j=1}^{k} a_{djQ} + \sum_{h \neq m} a_{dhm} = 1 \quad d = 1, \ldots, D, \tag{3.26o}$$

where $\delta_i^* \in \{0,1\}$ $(i \in G_1 \cup G_2)$, $\delta_{di} \in \{0,1\}$ $(i \in G_1 \cup G_2,\ d = 1,2,3)$, $\mu_i \in \{0,1\}$ $(i \in G_1 \cup G_2)$, $sl_t \in \{0,1\}$ $(t = 1,2)$, and $a_{djL}, a_{djQ}, a_{dhm}, c_d$ are free $(j = 1, \ldots, k)$.

3.2.2. The Quadratic GSPS (QGSPS)

Replacing the first constraints of GSPS model by the linear constraints which are nonlinear in terms of the attributes like those of the QSPS model, we obtain the following QGSPS model:

$$\text{minimize} \quad \sum_{i=1}^{n} \delta_i[D] \tag{3.27}$$

subject to

$$\sum_{j=1}^{k} x_{ij}a_{ljL} + \sum_{j=1}^{k} x_{ij}^2 a_{ljQ} + \sum_{h\neq m} x_{ih}x_{im}a_{dhm} - M\left(\sum_{h=1}^{d-1} v_{ih} + \delta_{id}\right) \leq c_d - \varepsilon \quad i \in G_1,\ d = 1,\ldots,D, \tag{3.27a}$$

$$\sum_{j=1}^{k} x_{ij}a_{ljL} + \sum_{j=1}^{k} x_{ij}^2 a_{ljQ} + \sum_{h\neq m} x_{ih}x_{im}a_{dhm} + M\left(\sum_{h=1}^{d-1} v_{ih} + \delta_{id}\right) \geq c_d + \varepsilon \quad i \in G_2,\ d = 1,\ldots,D, \tag{3.27b}$$

$$\mu_d \geq \delta_{id} \quad i \in G_1,\ d = 1,\ldots,D-1, \tag{3.27c}$$

$$1 - \mu_d \geq \delta_{id} \quad i \in G_2,\ d = 1,\ldots,D-1, \tag{3.27d}$$

$$v_{id} \leq \mu_d \quad i \in G_1,\ d = 1,\ldots,D-1, \tag{3.27e}$$

$$v_{id} \leq 1 - \mu_d \quad i \in G_2,\ d = 1,\ldots,D-1, \tag{3.27f}$$

$$v_{id} \leq 1 - \delta_{id} \quad i \in G_1 \cup G_2,\ d = 1,\ldots,D-1, \tag{3.27g}$$

$$\sum_{j=1}^{k} a_{djL} + \sum_{j=1}^{k} a_{djQ} + \sum_{h\neq m} a_{dhm} = 1 \quad d = 1,\ldots,D, \tag{3.27h}$$

where $\delta_{di} \in \{0,1\}$ $(i \in G_1 \cup G_2,\ d = 1,\ldots,D)$, $\mu_d \in \{0,1\}$ $(i \in G_1 \cup G_2,\ d = 1,\ldots,D-1)$, $0 \leq v_{id} \leq 1$ $(d = 1,\ldots,D-1)$, and $a_{djL}, a_{djQ}, a_{dhm}, c_d$ are free $(j = 1,\ldots,k,\ d = 1,\ldots,D)$.

As mentioned above, the cross-products terms can be excluded from the quadratic models if the attributes are uncorrelated and other types of nonlinear functions are possible.

4. A Comparative Study

4.1. The Datasets

In this study we choose four datasets.

(i) *The first dataset* (D1) is data presented by Johnson and Wichern [24] used by Glen [11] who were trying to apply new approaches to the problem of variable selections using an LP model. This dataset consists of 46 firms (21 bankrupt firms and 25 non-bankrupt firms). The four variables measured were the following financial ratios: cash flow to total debt, net income to total assets, current assets to current liabilities, and current assets to net sales.

Table 1: Normality test and equality of the variance-covariance matrices.

Dataset	Normality	Equality of the variance-covariance matrices ($\sum_1 = \sum_2$)
D1	Unverified	Verified
D2	Unverified	Unverified
D3	Unverified	Verified
D4	Unverified	Unverified

(ii) *The second dataset* (D2) is a Tunisian dataset. The data concerns 62 tumors of breast. Five variables characterize these tumors: four proteins expression scores (EGFR, Her2, Her3, and estrogens) and the size of these tumors in cm. The tumors are divided into two groups according to the SBR grad (grads II and III) which reflects the advancement state of the cancer (source: Centre of Biotechnology of Sfax).

(iii) *The third dataset* is a Japanese dataset (D3). This data contains 100 Japanese banks divided into two groups of 50. Seven financial ratios (return on total assets, equity to total assets, operating costs to profits, return on domestic assets, bad loan ratio, loss ratio on bad loans, and return on equity) characterize this data [25].

(iv) *The fourth dataset* is the Wisconsin Breast Cancer data (D4). This data consist of 683 patients screened for breast cancer divided into two groups: 444 representing a benign case and 139 representing a malignant tumor. Nine attributes characterize this data (clump thickness, uniformity of cell size, uniformity of cell shape, Marginal Adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses) (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/).

The objective is to discriminate between the groups of each dataset using the various methods cited above.

4.2. The Results

Different studies have shown that the reliability of the LDF method depends on the verification of certain hypotheses such as the normality of the data and the equality of the variance-covariance matrices. The results obtained from testing these hypotheses in our datasets are shown in Table 1.

The computer program AMOS 4 is used to verify the normality. To verify the equality of the variance-covariance matrices and to determine the classification rates, the SPSS program is used. According to Table 1, the normality hypothesis is not verified for all datasets, but the equality of variance-covariance matrices is verified for D1 and D3 datasets and not verified for the second and the fourth datasets. The results of the statistical approaches are obtained using SPSS program. The SVM-based approach is solved by the WinSVM package. The experiments are conducted by an Intel (R) Celeron (R) M, processor 1500 MHz in C/C + + environment. Various MPs were solved by CPLEX 10.0 package. For the experiment, we have chosen $M = 1000$ and $\varepsilon = \Delta = 0.005$. Microsoft Excel is used to determine the apparent hit rates (proportion of observations classified correctly), the Leave-One-Out (LOO) hit rates, and the holdout sample hit rates which represents the performance measures to be compared between the models. In fact, in order to evaluate the performance of the different approaches, a Leave-One-Out (LOO) procedure is used for the first three

datasets. The advantage of this procedure is to overcome the problem of the apparent hit rates bias. The LOO hit rate is calculated by omitting each observation in turn from the training sample and using the remaining observations to generate a discriminant function which is then used to classify the omitted observations. Although the computational efficiency of this procedure can be improved in statistical discriminant analysis, it is not practical in MP analysis unless only a relatively number of observations are included. For this reason, the LOO hit rate was not used for the fourth dataset. The performance of the different MP methods using this dataset (D4) is, then, addressed by considering the "split-half" technique. In fact, the important number of observations available in this dataset permits to adopt this latter approach by partitioning the complete observations (386) into training and holdout samples. The training sample of dataset (D4) consisted of a random sample of 73% of the observations in each group, with 340 observations in group1 and 160 in group 2 (500 observations in total). The remaining 27% of the observations (104 group1 observations and 79 group 2 observations) formed the holdout sample. To evaluate the performance of the various approaches, the training sample was used to generate classification models using the different methods and these classification models were then used to determine the holdout sample hit rates. The performance of LDF using this dataset (D4) was also evaluated in the same way.

Furthermore, the "split-half" technique is also employed for the first three datasets in order to evaluate the performance of the SVM-based approach. Similar to the dataset D4, the three datasets D1, D2, and D3 are partitioned into training and holdout samples. The training sample size of the first dataset is equal to 24 observations (11 observations in group 1 and 13 observations in group 2) and its holdout sample size is equal to 22 observations (10 observations in group 1 and 12 observations in group 2). For the second dataset, the training sample contains 45 observations (15 observations in group 1 and 30 observations in group 2). The remaining 17 observations (7 observations in group 1 and 10 observations in group 2) formed the holdout sample. The third dataset is partitioned into 70 observations (35 observations for each group) forming the training sample and 30 observations formed the holdout sample.

In this study, the complete set of observations of each dataset was first used as the training sample giving the apparent hit rate in order to demonstrate the computational feasibility of the different approaches. The use of the "split-half" and LOO procedures permits to allow the performance of classification models generated by the different methods.

4.2.1. The Results of the Linear Programing Model

The results of the correct classification rates (apparent rates) using MCA and MSD methods with various normalization constraints are presented in Table 2.

According to this table, the MCA model performs better than the MSD model for the different normalization constraints used. However, the best classification rates for the MSD model are given by using the constraints (N3) and (N4), except in the case of D1 and D3, where the difference between the three normalization constraints (N2), (N3), and (N4) is not significant. The classification rates of dataset D2 using these constraints are different. This is may be due to the nature of the data, to the fact that the group size is different, or to the fact that the model with (N2) normalization will generate a discriminant function in which the constant term is properly zero, but it will also exclude solutions in which the variable coefficients sum to zero, and rather should be solved with positive and negative normalization constants [4, 11]. However, the performance of the MCA model remains

Table 2: The correct classification rates of the MCA and MSD models (in percentage).

	D1 $n_1 = 21$, $n_2 = 25$, $n = 46$,		D2 $n_1 = 22$, $n_2 = 40$, $n = 62$,		D3 $n_1 = 50$, $n_2 = 50$, $n = 100$,		D4 $n_1 = 444$, $n_2 = 238$, $n = 683$,	
	MCA	MSD	MCA	MSD	MCA	MSD	MCA	MSD
(N1) $\sum_{j=1}^{k} a_j + c = 1$	91,3	84,78	83,87	64,52	96	86	97,2	94,7
	(4)	(7)	(10)	(22)	(4)	(14)	(19)	(36)
(N2) $\sum_{j=1}^{k} a_j = 1$	91,3	89,1	83,87	66,13	96	91	97,2	96,6
	(4)	(5)	(10)	(21)	(4)	(9)	(19)	(23)
(N3) $c \pm 1$	91,3	89,1	83,87	72,6	96	91	97,2	96,6
	(4)	(5)	(10)	(17)	(4)	(9)	(19)	(23)
(N4) invariance under origin shift	91,3	86,95	83,87	75,8	96	91	97,2	96,6
	(4)	(6)	(10)	(15)	(4)	(9)	(19)	(23)

The values between parentheses are the numbers of misclassified observations.

Table 3: LOO hit rates and the holdout hit rates for MCA and MSD models.

	D1	D2	D3	D4
	LOO	LOO	LOO	Holdout hit rate
MSD	89,1 (5)	66,13 (21)	84 (16)	95,6 (8)
MCA	89,1 (5)	70,97 (18)	84 (16)	98,36 (3)

unchanged using the different normalization constraints. For each of the two methods using the normalization constraint (N4), the LOO hit rates for the three datasets D1, D2, and D3 and the holdout sample hit rates for the dataset D4 are presented in Table 3.

From Table 3, we can conclude that the difference between MSD and MCA models is not significant for the first and the third datasets. However, the MCA model performs better than the MSD model for the second and the fourth datasets. Furthermore, the computational time of the MCA models is less than the computational time of the MSD model especially for the fourth dataset. In fact, by using the complete set of observations, the MSD model was solved in less than 7 seconds while the MCA model required less than 4 seconds to obtain the estimated coefficients of the discriminant function. However, for the other datasets, the difference of the solution time between the two models is not significant (less than 2 seconds).

On the other hand, to solve the RS models, two cases are proposed: first c_1 and c_2 take, respectively, the value 0 and 1 (Case 1), and second, the cutoff values c_1 and c_2 are considered decision variables (Case 2). The RS model, for the complete set of observations of the dataset D4, was solved in 3 seconds. The computational time of this model using the other datasets is less than 2 seconds. The apparent and the LOO hit rates for the disriminant function generated by the RS models are shown in Table 4.

The difference between the apparent and LOO hit rates of the RS model in the two cases is well improved particularly for the second and the fourth datasets. For D1 and D2, the difference between the numbers of misclassified observations in the two cases is marginally significant; only one or two misclassified observations are found. However, for D2 and D4, there is a difference. So, when normality and/or equality of the variance-covariance matrices

Table 4: The apparent, the LOO, and the holdout hit rates (in percentage) of the RS model.

	D1		D2		D3		D4	
	Apparent hit rate	LOO	Apparent hit rate	LOO	Apparent hit rate	LOO	Apparent hit rate	Holdout hit rate
$c_1 = 0$ and $c_2 = 1$	89,1 (5)	86,9 (6)	71 (18)	62,9 (23)	94 (6)	83 (17)	96,2 (26)	96,7 (6)
c_1 and c_2 decision variables	91,3 (4)	**89, 1** (5)	79 (13)	**70,97** (18)	96 (4)	**85** (15)	**97,36** (18)	**98,9** (2)

The values in parentheses are the numbers of misclassified observations.

Table 5: The apparent and LOO hit rates (in percentage) of MC1 and MC2 models.

		D1 $n_1 = 21, n_2 = 25$ $n = 46$				D2 $n_1 = 22, n_2 = 40$ $n = 62$				D3 $n_1 = 50, n_2 = 50$ $n = 100$			
		MC1		MC2		MC1		MC2		MC1		MC2	
		Apparent	LOO	Apparent	LOO	Apparent	LOO	Apparent	LOO	Apparent	LOO	Apparent	LOO
2nd step	Weighting	86,9	84,78	86,9	84,78	64,5	62,9	62,9	59,68	93	83	94	85
	MSD	(6)	(7)	(6)	(7)	(22)	(23)	(23)	(25)	(7)	(17)	(6)	(15)
	Weighting	91,3	89,1	91,3	89,1	72,5	66,13	69,35	64,5	94	84	94	85
	RS	(4)	(5)	(4)	(5)	(17)	(21)	(19)	(22)	(6)	(16)	(6)	(15)

are not verified, it would be most appropriate to consider the cutoff values decision variables. The results of the three combined models are given in Table 5.

The MSD weighting model (W-MSD) and the RS weighting model (W-RS) are used in the second stage to solve the MC1 and MC2 models. The results show that the choice of model used in the second stage affects the correctly classified rates. These rates are higher when one uses a *W-RS* model in the second stage. The difference between the models is not very significant for the first and the third datasets when equality of variance-covariance matrices is verified. However, for dataset D2, the MC1 model which combines the LDF, LPM, and RS models performs better than the MC2 model which combines the LDF, RS, and MSD models. In fact, LPM model used in MC1 model has the advantage to force the observations classification scores to cluster around the mean scores of their own groups. The application of the MC1 and MC2 models required a computational effort. In fact, to determine the classification rate, the combined method required to solve each model used in this approach separately. Then, the computational time important is more than 10 seconds for dataset D4, for example. For this reason, the use of such method can not be benefit if the dataset is sufficiently large.

The results of the various models for the four datasets are presented in Table 6.

Table 6 shows that the correctly classified rates (apparent hit rates) obtained by MCA, RS, and MIPEDE-DA are superior to those obtained by the other models especially when the normality and equality of variance-covariance matrices hypotheses are violated. The two combined methods, MC1 and MC2, give similar results for the first dataset. While for the other datasets, the MC1 performs better than MC2. It must be noted that the performance of the combined method can be affected by the choice of the procedures used in this method. Furthermore, the difference between these models is significant especially for dataset D2. In terms of the computational time, we can remark that the resolution of the statistical methods

Table 6: The apparent, the LOO, and the holdout sample hit rates (in percentage) of the different models.

	D1 Nonnormal $\sum_1 = \sum_2$ $n_1 = 21, n_2 = 25$ $n = 46$		D2 Nonnormal $\sum_1 \neq \sum_2$ $n_1 = 22, n_2 = 40$ $n = 62$		D3 Nonnormal $\sum_1 = \sum_2$ $n_1 = 50, n_2 = 50$ $n = 100$		D4 Nonnormal $\sum_1 \neq \sum_2$ $n_1 = 444, n_2 = 238$ $n = 683$	
	Apparent hit rate	LOO	Apparent hit rate	LOO	Apparent hit rate	LOO	Apparent hit rate ($n = 683$)	Holdout hit rate ($n = 183$)
LDF	89,1 (5)	89,1 (5)	74,2 (16)	66,12 (21)	91 (9)	88 (12)	96,3 (25)	99,45 (1)
LG	91,3 (4)	89,1 (5)	74,2 (16)	66,12 (21)	93 (7)	88 (12)	96,9 (21)	98,9 (2)
MSD	89,1 (5)	89,1 (5)	75,8 (15)	66,12 (21)	91 (9)	84 (16)	96,6 (23)	95,6 (8)
RS	91,3 (4)	89,1 (5)	**79**(13)	70,97 (18)	**96**(4)	85 (15)	**97,36**(18)	98,9 (2)
MCA	91,3 (4)	89,1 (5)	83,87 (10)	70,97 (18)	**96**(4)	84 (16)	**97,2**(19)	98,36 (3)
MIPEDEADA	91,3 (4)	89,1 (5)	**85,4**(9)	**75,8**(15)	**96**(4)	**91** (9)	**97,2**(19)	98,9 (2)
LPM	89,1 (5)	89,1 (5)	74,2 (16)	66,12 (21)	93 (7)	84 (16)	96,6 (23)	96,7 (6)
MC1 (LDF, LPM, RS)	91,3 (4)	89,1 (5)	72,5 (17)	66,13 (21)	94 (6)	85 (15)	96,6 (23)	97,27 (5)
MC2 (LDF, MSD, RS)	91,3 (4)	89,1 (5)	69,35 (19)	64,5 (22)	94 (6)	84 (16)	96,3 (25)	96,7 (6)

The values in parentheses are the numbers of misclassified observations.

LDF and LG using the complete dataset is less than one second which is faster than the resolution of the other MP models.

On the other hand, it is important to note that the correct classification rate of the RS model may be changed by selecting the most appropriate cutoff value for c. This cutoff value can be obtained by solving an LP problem in the second stage using a variety of objective functions such as MIP or MSD, instead of simply using the cutoff value equal to $(c_1 + c_2)/2$ [19]. In fact, for the third dataset D3, the apparent hit rate found by Glen [14] using the RS model is equal to 95% which is marginally below the apparent hit rate of 96% found in our study. Effectively, Glen [14] used the 0 and 1 cutoff value in the first stage and the MSD in the second stage of the RS model. Then, we can conclude that RS model can be most performing if the cutoff values are chosen as decision variables and simply using the cutoff value equal to $(c_1+c_2)/2$ in the second stage. Consequently, we do not need to use any binary variables like the case in which MSD or MIP models are applied in the second stage of the RS model. This result is interesting in the sense that the resolution of such model is very easy and does not require much computational time (in general less than 3 seconds). In fact, Glen [14] mentioned that the computational time for the RS model using MIP model in the second stage excluding the process for identifying the misclassified observations of G1 and G2 was lower than the computational time for the MCA model. Indeed, this latter model involves more binary variables than those of the first model (RS). In addition, for the datasets D1 and D3, we remark that the RS, MCA, and MIPEDEA-DA models produce the same apparent hit rates. However, for the second dataset, the MIPEDEA-DA followed by the MCA model performs better than the other approaches. On the other hand, the result obtained by the LOO procedure shows that the MIPEDEA-DA model performs better than the other models for the second and the third datasets, while for the first dataset, the difference between the models

Table 7: Comparison of SPS and GSPS using the N′1 and N′2, normalization constraints.

	D1		D2		D3	
	($n_1 = 21;\ n_2 = 25$)		($n_1 = 22;\ n_2 = 40$)		($n_1 = 50;\ n_2 = 50$)	
	D=2	D=3	D=2	D=3	D=2	D=3
SPSN′1		82,6 (8)				99 (1)
SPSN′2		**100(0)**				**100(0)**
QSPSN′1		**100(0)**		97 (3)		**100(0)**
QSPSN′2		**100(0)**		**100(0)**		**100(0)**
GSPSN′1	58,7 (19)	89,1 (5)	67,7 (20)	100(0)	74 (26)	100(0)
GSPSN′2	100(0)		83,87 (10)	100(0)	99 (1)	100(0)
QGSPSN′1	**100(0)**		96,8 (2)	100(0)	85 (15)	100(0)
QGSPSN′2	**100(0)**		**100(0)**		97 (3)	100(0)

The values in parentheses are the numbers of misclassified observations.

is not significant. In terms of the holdout sample hit rate obtained using the classification models generated from the 73% training sample of dataset D4, the statistical method LDF performs better than the other approaches followed by the LG, the RS, and the MIEDEA-DA models.

4.2.2. The Result of Nonlinear MP Models

(a) Comparison of SPS and GSPS Models Using the Two Normalization Constraints N′1 and N′2

To compare between the two normalization constraints N′1 and N′2, the model SPS and GSPS were solved using the first three datasets (D1, D2, and D3). The results are presented in Table 7.

According to Table 7, the models using the second normalization constraint can perform better than the one using the first normalization constraint. An important result found concerns the SPS models which can not give any solution for the second dataset, while the QSPS models perform very well and especially the one using the second normalization constraint. Furthermore, the GSPSN′2 model performs better than the GSPSN′1 model especially for the first dataset. Thus, compared to the normalization (N′1) used by Better et al. [18], our proposed normalization (N′2) can produce better results. The comparison of the different models developed will be discussed in the following section.

(b) Comparison of Different Models

The results of the different models are presented in Table 8. From Table 8, the nonlinear MP models outperform the classical approaches. This result may be due to the fact that the performance of these latter approaches requires the verification of some standard hypotheses. In fact, the LDF and QDF have the best performance if the data distribution is normal. However, this hypothesis is not verified for these datasets. Although the LG model does not need the verification of such restriction, this model has not provided higher hit rates compared to those of the other approaches especially for the second dataset. On the

Table 8: The apparent classification rates of the different models (in percentage).

	D1		D2		D3		D4	
	($n_1 = 21;\ n_2 = 25$)		($n_1 = 22;\ n_2 = 40$)		($n_1 = 50;\ n_2 = 50$)		($n_1 = 444;\ n_2 = 239$)	
	Nonnormal		Nonnormal		Nonnormal		Nonnormal	
FDL	89,1 (5)		74,2 (16)		91 (9)		96,3 (25)	
LG	91,3 (4)		74,2 (16)		93 (7)		96,9 (21)	
FDQ	76,08 (14)		72,58 (17)		85 (15)		90,8 (63)	
Second order MSD model	93,47 (3)		75,8 (15)		85 (15)		90,8 (63)	
	$S = 2$	$S = 3$	$S = 2$	$S = 3$	$S = 2$	$S = 3$	$S = 2$	$S = 3$
Piecewise MCA	91,3 (4)	**100(0)**	72,5 (17)	**98,39(1)**	99 (1)	**100(0)**		87,55 (85)
Piecewise MSD	97,8 (1)	**97,8(1)**	87,1 (8)	**96,8(2)**	99 (1)	**100(0)**		
Piecewise QSMCA	**100(0)**		**100(0)**		**100(0)**		98,97 (7)	100 (0)
Piecewise QSMSD	97,8 (1)	**100(0)**	**100(0)**		**100(0)**		100 (0)	
	$D = 2$	$D = 3$	$D = 2$	$D = 3$	$D = 2$	$D = 3$	$D = 2$	$D = 3$
SPSN'2		**100(0)**		-		**100(0)**		98,24 (12)
QSPSN'2		**100(0)**		**100(0)**		**100(0)**		98,82 (8)
GSPSN'2	**100(0)**		83,87 (10)	**100(0)**	99 (1)	**100(0)**	98,82 (8)	99,7 (2)
QGSPSN'2	**100(0)**		**100(0)**		97 (3)	**100(0)**	99,85 (1)	100 (0)

The values in parentheses are the numbers of misclassified observations.

other hand, the second-order MSD model, also, performs worse than the other models. Furthermore, the performance of the piecewise QSMCA and QGSPSN'2 models is better than the performance of the piecewise-linear models (MCA and MSD) for the first and second datasets. In fact, the optimal solution is rapidly reached using these models rather than the piecewise-linear approaches (the hit rates are equal 100% on using $S = 2$ and $D = 2$). While, for the second data D2, the piecewise-quadratic models (QSMCA and QSMSD), the multihyperplanes and the multihypersurfaces models perform better than the other approaches. Moreover, the difference between these models and the standard piecewise models is not significant for dataset D3 but we can remark that the piecewise QSMCA and QSMSD can reach optimality rapidly using only $S = 2$. Comparing the nonlinear MP models in terms of computational time, we can remark that the resolution of the QGSPS providing the estimated coefficients of the discriminant function is better than the solution time obtained by GSPS model for all datasets. Using dataset D4, for example, the solution time of the QGSPS with $D = 2$ is equal to 11 seconds (for $D = 3$, the solution time is equal to 21 seconds) while the resolution of the GSPS takes more than 960 seconds. For the other datasets, the solution time of the QGSPS model is less than 3 seconds. On the other hand, employing piecewise models using only the case where the group1 is in the convex region, the optimal solution time is obtained in more than 7 seconds. Otherwise, the time for the resolution of these models would have been approximately double. In fact, using dataset D4 the resolution of piecewise QMCA in the case where G1 is in the convex region, for example, required 8 seconds using

Table 9: Classification rates using the LOO procedure (in percentage).

Models	$S = 2$	$S = 3$
Piecewise-linear MCA*	72,58 (17)	72,58 (17)
Piecewise QSMCA*	85,48 (9)	83,87 (10)
Piecewise-linear MSD*	72.58 (17)	82,26 (11)
Piecewise QSMSD*	79 (13)	87,1 (8)
	$D = 2$	$D = 3$
GSPSN'2	77.4 (14)	82,26 (11)
QGSPSN'2	83.87 (10)	87,1 (8)
QSPSN'2		83,87 (10)

The values in parentheses are the numbers of misclassified observations.
*G2 convex.

three arcs ($s = 3$). However, to obtain the optimal solution, the model must be solved also in the case where G2 is in the convex region and then the computational time will double.

To judge the performance of piecewise-linear MSD and MCA, piecewise QSMSD, QGSPSN'2, GSPSN'2, and QSPSN'2, the LOO (Leave-One-Out) procedure is first used for the dataset D2 which is considered the dataset in which the nonlinear separation is the most adequate. The LOO classification rate was determined by omitting each observation in turn, solving the piecewise models in convex region for group 2. In fact, the optimal solution is attained in this group when convex region and the associated piecewise function are then used to classify the omitted observation. This same procedure (LOO) is applied for the multihypersurfaces and muli-hyperplanes models but without solving twice these models. The classification rates obtained by LOO procedure with $S = 2$, $S = 3$ and $D = 2$, $D = 3$ are presented in Table 9.

The LOO hit rate for each model is, as expected, lower than the apparent hit rate (see Tables 6 and 8). From Table 9, the results show the performance of QGSPSN'2 and QSMSD compared to those of the other models. In fact, the LOO hit rates are equal to 87.1% (54/62) for piecewise QSMSD and QGSPSN'2 while these rates decrease in the other models while using $S = 3$ and $D = 3$. However, for $S = 2$, the piecewise QSMCA model gives the correctly higher classification rate and we can see that the performance of this model decreases if S is augmented to 3 arcs. For the MCA piecewise model, the LOO hit rates are lower than those of the other models. Furthermore, we can remark that there is no improvement in these LOO hit rates when the number of segments increases from two to three. On the other hand, by applying the LOO procedure for the third dataset (D3), we can deduce that the QGSPS and piecewise QSMSD models perform better than the other models with LOO hit rates equal to 96% for QGSPS with $D = 3$ (resp, to 92% for QSMSD with $S = 3$) and suppose that G2 is in the convex region. These LOO hit rates are higher than those obtained using the GSPS (86%) or piecewise MSD models (84%) [17, 18]. Comparing the piecewise QSMSD and the QGSPS models, the latter model is considered to be preferable to the first because it does not need to be resolved twice for the case in which G1 is in the convex region and the case in which G2 is in the convex region. However, for the first dataset (D1), the LOO procedure gives the same classification rates (93%, 48%) for the different models (with D and S equal 3 and G1 in the convex region) except for the GSPS model which has the best classification rate equal to 97, 8%. These results suggest that for this dataset the GSPS model may be the most appropriate. In fact, for this dataset (D1), the use of multihyperplanes is more adequate than

Table 10: The holdout sample hit rate (in percentage) of dataset D4.

Models	$S = 2$	$S = 3$
Piecewise-linear MCA*	91,26 (16)	89,07 (20)
Piecewise QSMCA*	91,26 (16)	99,45 (1)
Piecewise QSMSD*	94,5 (10)	97,27 (5)
	$D = 2$	$D = 3$
GSPSN'2	99,45 (1)	98,36 (3)
QGSPSN'2	95,6 (8)	93,99 (11)
QSPSN'2		100 (0)

The values in parentheses are the numbers of misclassified observations.
*G1 convex.

the use of multihypersurfaces. Then we can conclude that there may be only limited benefits from using nonlinear model when the data requires linear separation (dataset D1).

As mentioned above, due to the important size (683) of the fourth dataset (D4), the procedure used to test the performance of the MP models for this dataset is the split-half technique. The results of the holdout sample hit rate of this dataset are presented in Table 10.

These results are training/holdout specific, but although it would be desirable to determine average holdout hit rate by repeating the calculation with other training/holdout samples, the computational requirement would be excessive. It can be seen from Table 10 that the best holdout hit rate is obtained by the QSPS followed by the GSPS and the piecewise QMCA models. Comparing this holdout hit rate with those obtained by linear MP models, we can remark that the difference is marginally significant, but in terms of computational time, we found that the resolution of the linear MP models using the real dataset takes less than 4 seconds while the resolution of the nonlinear MP models takes more than 8 seconds (dataset D4). Then we can conclude that for this dataset the linear separation is the most adequate and the use of the statistical method LDF is the more appropriate. This result can be also confirmed by the calculation of the LOO hit rate using LDF, QSPS, and GSPS models for dataset D4. In fact, the LDF LOO hit rate is equal to 96, 2% (26 misclassified) which is higher than the LOO hit rate of the QSPS and GSPS which is equal to 94, 5% (37 misclassified) for QSPS and equal to 94, 2% (39 misclassified) for GSPS. This result can be explained by the fact that if the dataset size is sufficiently large, the normality hypothesis is approximately verified and then the statistical method LDF can provide a better performance than that of the other approaches except for the case of the nonlinear separable problems.

In conclusion, the nonlinear MP approaches yield greatly improved classification results over the classical and linear MP approaches especially for the second and the third datasets. These results are due to the fact that normality is violated and the nature of these datasets requires the nonlinear separations.

The statistical and the MP models are also compared to the popular SVM-based approach. In fact, the foundations of this latter approach have been developed by Vapnik in 1995 and are gaining popularity due to many attractive features and promising empirical performance. The result of this approach is presented in the following section.

4.2.3. The Result of SVM-Based Approach

In the experiments, the dot kernel is used to resolve the linear SVM while the most "generic" kernel function employed for the resolution of the nonlinear SVM using the real datasets is

Table 11: The SVM-based approach results.

		D1	D2	D3	D4
Linear SVM	CPU-time average	28 seconds	78 seconds	54 seconds	200 seconds
	Apparent hit rates	91,3 (4)	75,8 (15)	92 (8)	97,2 (19)
	Training apparent hit rates	95,8 (1)	73,3 (12)	87,1 (9)	94,63 (17)
	Holdout sample hit rates	86,36 (3)	64,7 (6)	80 (6)	97,8 (4)
Nonlinear SVM	CPU-time average	29 seconds	53 seconds	51 seconds	330 seconds
	Kernel Function for the complete dataset	radial: C=100 e= 0,01 gamma = 10	radial: c=1000 e=1e-006 gamma = 10	radial: c=1000 e=1e-005 gamma = 2	radial: c=100 e=0,2 Gamma = 5
	Apparent hit rates	100 (0)	100 (0)	100 (0)	100 (0)
	Kernel Function for training dataset	Anova: c=100000 e= 0,01 gamma = 0,1 degree = 2	Anova: c=10 e=0,1 gamma = 0,9 degree = 4	Polynomial: c=100 e=0,2 degree = 2	Polynomial: c=100 e=0 degree = 3
	Training apparent hit rates	100 (0)	100 (0)	100 (0)	100 (0)
	Holdout sample hit rates	86,36 (3)	70,58 (5)	80 (6)	96,17 (7)

"radial basis function". However, the kernel functions providing the best holdout hit rates are chosen and taken in consideration in the analysis.

The results of the SVM-based approach using the different datasets are presented in Table 11. From Table 11 and based on the apparent hit rate, the best result is obtained by the nonlinear SVM for all datasets. However, in terms of computational time and holdout sample hit rates, the linear SVM-based approach performs better than the nonlinear SVM for the first and the fourth datasets. In fact, the holdout sample hit rate for dataset D4, for example, using linear SVM is equal to 97, 8% which is higher than the holdout hit rate obtained by the nonlinear SVM (96, 17%). For datasets D2 and D3, the nonlinear SVM is the most performing. In fact, the resolution of the nonlinear SVM required less time than the resolution of the linear SVM. These results are comparable to those obtained by the MP approaches. However, the advantage of the MP approaches over the SVM-based approach is that the computational time of the first approache is in general lower than the computational time of the latter approach. Furthermore, the MP approaches do not need the use of the kernel function. In fact, despite the simplicity of the SVM- based approach, the performance of this approach depends on the choice of the appropriate kernel function that can be employed for the resolution of classification problem. This required a prior knowledge of the problem especially when the

training data is nonlinearly separable. However, the preferred methods used for selecting a kernel are the bootstrapping and cross-validation. But, this requires a computational effort.

5. Conclusion

In this paper, various linear and nonlinear MP models and three traditional methods, LDF, QDF, and LG, are applied in order to examine the conditions under which the models taken into consideration will give similar or different results. The advantage of this study over other comparative studies lies in our use of different real datasets with different sizes. Different normalization constraints are used in the MP models in order to analyze their performance. The results of the various dataset analyses show that the correctly classification rates obtained by the various MP models are better than those of classical models when standard hypotheses are violated. These results confirm those found in the literature. Besides these results, we note that the correct classification rates are also similar when the differences in size between the groups and the sample size itself are low. Indeed, the performance of the linear MP models improves when the deviation between the size of the groups is significant, when the size of the real dataset is lower, and when the equality of variance-covariance matrices is not verified. On the other hand, the best classification rates are obtained (1) when using the normalization constraint N3 or N4 for the MSD model, (2) when selecting the cutoff values c_1 and c_2 as decision variables for the RS model, (3) when normality and equality of the matrices of variance and covariance are violated, and (4) when using the *W-RS* model in the second stage for the combined model. Comparing the linear MP approaches, the results of this study favor the RS, MCA, and MIPEDEA-DA models when the linear separation is adequate, but the MIPEDEA-DA model gives better results than those of the MCA and RS models when the data requires nonlinear separation. However, when standard hypotheses are violated, we think that the use of RS or MCA models is most practical especially for the linear separation case. In fact, these models are the most simple and do not necessitate any restrictions as for MIPEDEA-DA (higher number of misclassified observations in the first stage of the MIPEDEA-DA models). On the other hand, when the misclassified observation number is important, the use of other nonlinear separation models is recommended.

In this paper, various nonlinear MP models are applied, and new MP models based on piecewise-nonlinear functions and on hypersurfaces are proposed in order to ameliorate the performance of different MP methods. The use of nonlinear discriminant functions gives better performances than linear classifier. In fact, the models based on multihyperplanes and on multihypersurfaces are supposed to be the most favorable compared to piecewise models because they do not require the resolution of the two models (group 1 in convex region and group 2 in convex region). The results are also effective in the sense that they do not require many constraints, many binary variables, and special order sets. On the other hand, the multihypersurfaces models especially QGSPSN'2 model perform better than multihyperplanes models because we can reach the solution faster using a reduced number of hypersurfaces rather than hyperplanes. However, the performance of piecewise QSMCA and QSMSD models is better than the performance of piecewise-linear MSD and MCA models. Furthermore, when normality is violated and when the number of misclassified observations obtained by classical or linear MP models is important, the use of nonlinear MP approaches or the SVM-based approach can be the most appropriate. However, the advantage of MP approaches is that these approaches do not need the determination of the kernel function. In fact, a limited number of known kernel transformations are relied by SVM-based approach to project the original data into very high-dimensional space in order to render it linearly

separable. From these kernel transformations, which is the best for a particular classification problem? To respond to this question, a computational effort should be needed to choose the best one on using the kernel selection methods such as bootstrapping and cross-validation.

For future research, other techniques for choosing the kernel function can be developed and other types of piecewise-nonlinear models and multihypersurfaces models can be used in order to ameliorate the performance of these approaches. For example, in the MP methods, we can build a classification function which, in addition to the usual linear attributes terms (x_{ij}), takes the natural logarithm ($Logx_{ij}$) of the original data. Other complicated nonlinear functions can be used as well. These functions involve solving MP models with nonlinear constraints rather than linear constraints which are nonlinear in terms of the attributes. In addition, we can view the possibility of changing the objective function of multihyperplanes or multihypersurfaces models by minimizing the undesirable deviations of misclassified observations from the different hyperplanes or hypersurfaces rather than minimizing the number of misclassified observations by the hyperplanes or hypersurfaces.

References

[1] R. A. Fisher, "The use of multiple measurements in taxonomy problems," *Annal Eugenics*, vol. 7, pp. 179–188, 1936.

[2] C. A. B. Smith, "Some examples of discrimination," *Annals of Eugenics*, vol. 13, pp. 272–282, 1947.

[3] P. E. Markowski and C. A. Markowski, "Concepts, theory, and techniques: some difficulties and improvements in applying linear programming formulations to the discriminant problem," *Decision Sciences*, vol. 16, no. 3, pp. 237–247, 1985.

[4] N. Freed and F. Glover, "Resolving certain difficulties and improving the classification power of LP discriminant analysis formulations," *Decision Sciences*, vol. 17, pp. 589–595, 1986.

[5] F. Glover, S. Keene, and B. Duea, "A new class of models for the discriminant problem," *Decision Sciences*, vol. 19, no. 2, pp. 269–280, 1988.

[6] G. J. Koehler, "Unacceptable solutions and Hybrid discriminant model," *Decision Sciences*, vol. 20, pp. 844–848, 1989.

[7] G. J. Koehler, "Considerations for mathematical programming models in discriminant analysis," *Managerial and Decision Economics*, vol. 11, no. 4, pp. 227–234, 1990.

[8] S. M. Bajgier and A. V. Hill, "An experimental comparison of statistical and linear programming approach to discriminant problem," *Decision Sciences*, vol. 13, pp. 604–618, 1982.

[9] K. F. Lam and J. W. Moy, "An experimental comparison of some recently developed LP approaches to the discriminant problem," *Computers and Operations Research*, vol. 24, no. 7, pp. 593–599, 1997.

[10] K. F. Lam and J. W. Moy, "Combining discriminant methods in solving classification problems in two-group discriminant analysis," *European Journal of Operational Research*, vol. 138, no. 2, pp. 294–301, 2002.

[11] J. J. Glen, "Integer programming methods for normalisation and variable selection in mathematical programming discriminant analysis models," *Journal of the Operational Research Society*, vol. 50, no. 10, pp. 1043–1053, 1999.

[12] 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*, vol. 88, no. 2, pp. 358–367, 1996.

[13] T. Sueyoshi, "DEA-discriminant analysis: methodological comparison among eight discriminant analysis approaches," *European Journal of Operational Research*, vol. 169, no. 1, pp. 247–272, 2006.

[14] J. J. Glen, "A comparison of standard and two-stage mathematical programming discriminant analysis methods," *European Journal of Operational Research*, vol. 171, no. 2, pp. 496–515, 2006.

[15] A. P. D. Silva and A. Stam, "Second order mathematical programming formulations for discriminant analysis," *European Journal of Operational Research*, vol. 72, no. 1, pp. 4–22, 1994.

[16] J. J. Glen, "Dichotomous categorical variable formation in mathematical programming discriminant analysis models," *Naval Research Logistics*, vol. 51, no. 4, pp. 575–596, 2004.

[17] J. J. Glen, "Mathematical programming models for piecewise-linear discriminant analysis models," *Operational Research Society*, vol. 50, pp. 1043–1053, 2005.

[18] M. Better, F. Glover, and M. Samourani, "Multi-hyperplane formulations for classification and discriminant analysis," 2006, http://www.opttek.com/News/pdfs/Multi%20HyperplaneW%20DSI%20paper%20-%20final.pdf.
[19] C. T. Ragsdale and A. Stam, "Mathematical programming formulations for the discriminant problem: an old dog does new tricks," *Decision Science*, vol. 22, no. 2, pp. 296–307, 1991.
[20] J. J. Glen, "An iterative mixed integer programming method for classification accuracy maximizing discriminant analysis," *Computers & Operations Research*, vol. 30, no. 2, pp. 181–198, 2003.
[21] C. K. Lee and J. K. Ord, "Discriminant analysis using least absolute deviations," *Decision Sciences*, vol. 21, no. 1, pp. 86–96, 1990.
[22] J. J. Glen, "Classification accuracy in discriminant analysis: a mixed integer programming approach," *Journal of the Operational Research Society*, vol. 52, no. 3, pp. 328–339, 2001.
[23] T. Sueyoshi, "Mixed integer programming approach of extended DEA-discriminant analysis," *European Journal of Operational Research*, vol. 152, no. 1, pp. 45–55, 2004.
[24] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, NJ, USA, 2nd edition, 1988.
[25] T. Sueyoshi, "Extended DEA-discriminant analysis," *European Journal of Operational Research*, vol. 131, no. 2, pp. 324–351, 2001.