JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 112, No. 2, pp. 265–293, February 2002 (©2002)

Polyhedral Separability through Successive LP

A. ASTORINO[1] AND M. GAUDIOSO[2]

Communicated by O. L. Mangasarian

Abstract. We address the problem of discriminating between two finite point sets $\mathscr{A}$ and $\mathscr{B}$ in the n-dimensional space by h hyperplanes generating a convex polyhedron. If the intersection of the convex hull of $\mathscr{A}$ with $\mathscr{B}$ is empty, the two sets can be strictly separated (polyhedral separability). We introduce an error function which is piecewise linear, but not convex nor concave, and define a descent procedure based on the iterative solution of the LP descent direction finding subproblems.

Key Words. Classification, separability, machine learning.

1. Introduction

A variety of classical statistical problems can be approached advantageously by optimization tools. Discriminant analysis is one of these fields (Ref. 1). It has many applications in diverse areas such as management science [credit grant decisions (Ref. 2), make or buy policies (Ref. 3), managerial efficiency evaluation (Ref. 4)], medicine (Refs. 5–7), chemistry (Ref. 8), and others].

Classification is the most common function implemented by discriminant analysis methods and algorithms (Ref. 1). A classification process is aimed at assigning each element of a given set to a specific subset. In this paper, we consider classification problems. In particular, given two nonempty disjoint finite points sets $\mathscr{A}$ and $\mathscr{B}$ in $\mathbb{R}^n$ ($\mathscr{A} \cap \mathscr{B} = \varnothing$), consisting of m and k points respectively, that are represented by the columns of the matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times k}$, the objective is to construct a criterion for discriminating between the elements of the two sets.

This problem has been approached by means of mathematical programming. If the convex hulls of the two point sets do not intersect

[1]Researcher, Istituto per la Sistemistica e l'Informatica, ISI CNR, Rende, Cosenza, Italia.
[2]Professor, Dipartimento di Elettronica, Informatica e Sistemistica, Università della Calabria, Rende, Cosenza, Italia.

0022-3239/02/0200-0265/0

$[\text{conv}(\mathscr{A}) \cap \text{conv}(\mathscr{B}) = \varnothing]$, $\mathscr{A}$ and $\mathscr{B}$ are linearly separable (Ref. 9); i.e., there exists a separating hyperplane such that all the points of one set are on one side of the hyperplane and all the points of the other set are on the other side. Linear programming can be used to construct such a hyperplane (Ref. 10).

If the convex hulls of $\mathscr{A}$ and $\mathscr{B}$ intersect, it is possible to obtain a hyperplane (v, γ), with $v \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$, that minimizes some misclassification measure. A possible criterion is to minimize the sum of the distances to the separating hyperplane of the misclassified points (Ref. 11). Previous approaches use distance surrogates that maintain the linearity of the problem to be solved.

In particular, some formulations (Ref. 12) include the objective of minimizing the maximum deviation or the sum of the deviations of the misclassified points from a reference hyperplane, together with weighted variants of these objectives. Other formulations (Ref. 13) maximize the quality of the separation provided by (v, γ). The quality, if negative, measures the largest error connected to (v, γ) with respect to the historical data.

More in general, classification problems are investigated intensively by applying specialized methodologies stemming from diverse areas. In particular, we mention the numerous and interesting applications based on the use of neural networks (Refs. 14–19).

As far as mathematical programming-based approaches are concerned, very popular appears to be the support vector machine (SVM) approach (Refs. 20–30), where the formulation of the classification problem in terms of quadratic programming is aimed at obtaining a good separation margin.

Decision trees are used within approaches such as the global tree optimization (GTO, Refs. 31–36), where classification through nonlinear separation surfaces is pursued. In fact, our method can be considered a member of the GTO approach.

Our method is inspired by a technique, investigated by Bennett and Mangasarian (Ref. 10), which consists in minimizing some norm of the average violations of the following system:

$$A^T v \leq e(\gamma - 1),$$

$$B^T v \geq e(\gamma + 1),$$

that has solution $v \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$ if and only if the two sets $\mathscr{A}$ and $\mathscr{B}$ are linearly separable.

The objective function can be written as the weighted sum of the L_1 norms of the vectors of the classification errors for both the sets $\mathscr{A}$ and $\mathscr{B}$:

$$z^* = \min_{v,\gamma} (1/m) \| \max\{0, A^T v - e\gamma + e\} \|_1 + (1/k) \| \max\{0, -B^T v + e\gamma + e\} \|_1 .$$

The sets $\mathscr{A}$ and $\mathscr{B}$ are linearly separable if and only if $z^* = 0$ and it is proved that the trivial solution $v = 0$ cannot occur.

The previous formulation is equivalent to the following linear program:

$$\begin{aligned} z^* = \min_{v,\gamma,y,z} \quad & e^T y/m + e^T z/k, \\ \text{s.t.} \quad & y \geq A^T v - e\gamma + e, \\ & z \geq -B^T v + e\gamma + e, \\ & y \geq 0, \\ & z \geq 0, \end{aligned}$$

where y_i is nonnegative and represents the error for point $a_i \in \mathscr{A}$ and z_l is nonnegative and represents the error for point $b_l \in \mathscr{B}$.

When the two sets $\mathscr{A}$ and $\mathscr{B}$ cannot be separated by a hyperplane, we can look for nonlinear separating surfaces.

In Ref. 37, strict separation by means of one or more hyperplanes or surfaces (nonlinear manifolds) is investigated. Linear programming plays a role as well.

In Ref. 38, the problem of determining whether or not the two sets $\mathscr{A}$ and $\mathscr{B}$ can be separated by two hyperplanes is approached as a bilinear program, that is processed by a finite iterative linear programming algorithm.

The essence of our method is to obtain a separation with more than one hyperplane, using an error function derived from the previously described approach (Ref. 10).

Whenever the two sets $\mathscr{A}$ and $\mathscr{B}$ cannot be separated by a hyperplane,

$$\text{conv}(A) \cap \text{conv}(B) \neq \emptyset,$$

but the convex hull of $\mathscr{A}$ and the set $\mathscr{B}$ do not intersect,

$$\text{conv}(\mathscr{A}) \cap B = \emptyset,$$

they are h-polyhedrally separable ($h \leq k$, Ref. 39). That is, there exist h hyperplanes such that the points of $\mathscr{A}$ are contained in a convex polyhedron (intersection of h half-spaces) and the points of $\mathscr{B}$ are left outside the polyhedron.

The objective of the paper is to state an algorithm aimed at finding, for a given $h \geq 1$, such set of hyperplanes $\{v^{(j)}, \gamma_j\}$, $j = 1, \ldots, h$, if they exist or to find h hyperplanes that minimize a measure of the classification error. We introduce an error function which is piecewise linear, but not convex nor concave, and we define a descent procedure based on the iterative solution of the LP descent direction finding subproblems.

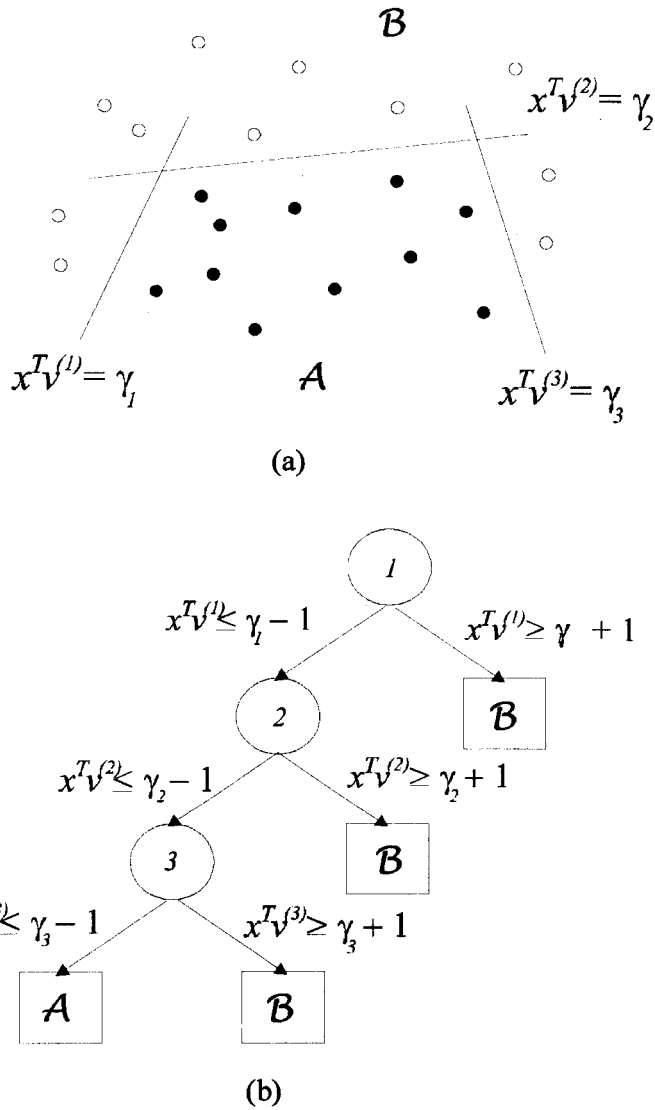

Fig. 1. (a) Polyhedral separability; (b) GTO approach.

Figure 1 shows the possible interpretation of our method in terms of the GTO approach.

The paper is organized as follows. In Section 2, we discuss the concept of *h*-polyhedral separability introduced by Megiddo (Ref. 39); in Section 3,

we give the formulation of the error function and study its characteristics. In Section 4, we describe an algorithm for minimizing the error function. In Section 5, we present an application concerning the follow-up of patients after kidney transplant.

Throughout the paper, we adopt the following notations. The convex hull of a set $\mathscr{X}$ will be denoted by conv($\mathscr{X}$). For a vector x in the n-dimensional real space $\mathbb{R}^n$, x_+ will denote the vector in $\mathbb{R}^n$ with components $(x_+)_i := \max\{x_i, 0\}$, $i = 1, \ldots, n$, and $\max(x)$ will denote the scalar $\max_{1 \le i \le n} x_i \in \mathbb{R}$. For a given matrix A, a_i will denote the ith column. The L_1-norm of x, $\sum_{i=0}^{n} |x_i|$, will be denoted by $\|x\|_1$. A vector of ones in a real space of arbitrary dimension will be denoted by e. A matrix of ones in a real space of arbitrary dimension will be denoted by E.

2. *h*-Polyhedral Separability

We develop here the concept of h-polyhedral separability introduced by Megiddo (Ref. 39) for two nonempty disjoint finite point sets.

Definition 2.1. h-Polyhedral Separability. The sets $\mathscr{A}$ and $\mathscr{B}$ are h-polyhedrally separable if there exists a set of h hyperplanes $\{w^{(j)}, \xi_j\}$, with

$$w^{(j)} \in \mathbb{R}^n, \qquad \xi_j \in \mathbb{R}, \qquad j = 1, \ldots, h,$$

such that

(i) $a_i^T w^{(j)} < \xi_j, \qquad \forall i = 1, \ldots, m, \forall j = 1, \ldots, h,$

(ii) $b_l^T w^{(j)} > \xi_j, \qquad \forall l = 1, \ldots, k,$ for at least one $j \in \{1, \ldots, h\}$.

According to the above definition, $\mathscr{A}$ and $\mathscr{B}$ are h-polyhedral separable whenever $\mathscr{A}$ is contained in a convex polyhedron (intersection of h half-spaces) and the points of $\mathscr{B}$ are left outside the polyhedron (Fig. 2).

The following property holds.

Proposition 2.1. The sets $\mathscr{A}$ and $\mathscr{B}$ are h-polyhedrally separable if and only if there exists a set of h-hyperplanes $\{v^{(j)}, \gamma_j\}$, with

$$v^{(j)} \in \mathbb{R}^n, \qquad \gamma_j \in \mathbb{R}, \qquad j = 1, \ldots, h,$$

such that

(i) $a_i^T v^{(j)} \le \gamma_j - 1, \qquad \forall i = 1, \ldots, m, \forall j = 1, \ldots, h,$

(ii) $\forall l = 1, \ldots, k$, there exists an index $j \in \{1, \ldots, h\}$ such that $b_l^T v^{(j)} \ge \gamma_j + 1$.

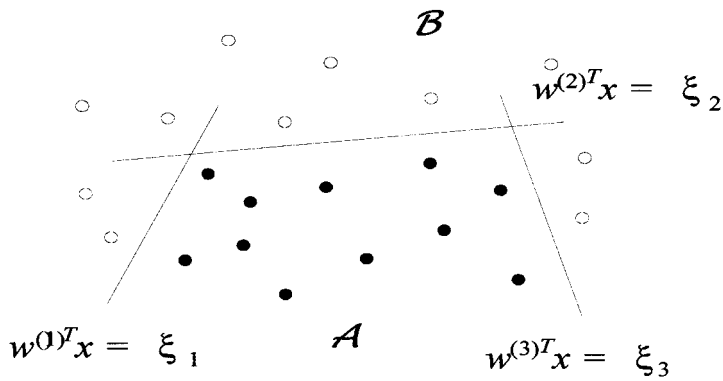

Fig. 2. 3-polyhedral separability.

Proof. Sufficiency is straightforward. As for the necessary condition, we have that, from Definition 2.1, it follows that there exist h-hyperplanes $\{w^{(j)}, \xi_j\}$ such that

$$a_i^T w^{(j)} < \xi_j, \qquad \forall i = 1, \ldots, m, \forall j = 1, \ldots, h;$$

i.e.,

$$-a_i^T w^{(j)} + \xi_j > 0, \qquad \forall i = 1, \ldots, m, \forall j = 1, \ldots, h, \tag{1}$$

and moreover,

$$b_l^T w^{(j)} > \xi_j, \qquad \forall l = 1, \ldots, k, \text{ for at least one } j \in \{1, \ldots, h\}.$$

For all $l = 1, \ldots, k$, let us define $J(l) \subseteq \{1, \ldots, h\}$, with $J(l) \neq \emptyset$, as the set of indices j of the hyperplanes $\{w^{(j)}, \xi_j\}$ such that

$$b_l^T w^{(j)} - \xi_j > 0, \qquad \forall j \in J(l). \tag{2}$$

We define also

$$\delta_1 \triangleq \min_{\substack{i=1,\ldots,m \\ j=1,\ldots,h}} [-a_i^T w^{(j)} + \xi_j] > 0, \tag{3a}$$

$$\delta_2 \triangleq \min_{\substack{l=1,\ldots,k \\ j \in J(l)}} [b_l^T w^{(j)} - \xi_j] > 0, \tag{3b}$$

$$\delta \triangleq \min\{\delta_1, \delta_2\} > 0. \tag{3c}$$

Thus, we have

$$\delta \leq -a_i^T w^{(j)} + \xi_j, \qquad \forall i = 1, \ldots, m, \forall j = 1, \ldots, h, \tag{4a}$$

$$\delta \leq b_l^T w^{(j)} - \xi_j, \qquad \forall i = l, \ldots, k, \forall j \in J(l). \tag{4b}$$

Now, consider a new set of hyperplanes $\{v^{(j)}, \gamma_j\}$, $j = 1, \ldots, h$, defined as follows:

$$v^{(j)} \triangleq w^{(j)}/\delta, \qquad \forall j = 1, \ldots, h, \tag{5a}$$

$$\gamma_j \triangleq \xi_j/\delta, \qquad \forall j = 1, \ldots, h. \tag{5b}$$

From (4)–(5) we have

$$\begin{aligned} a_i^T v^{(j)} &= (1/\delta) a_i^T w^{(j)} \\ &\leq (1/\delta)(\xi_j - \delta) \\ &= \xi_j/\delta - 1 \\ &= \gamma_j - 1, \qquad \forall i = 1, \ldots, m, \forall j = 1, \ldots, h, \\ b_l^T v^{(j)} &= (1/\delta) b_l^T w^{(j)} \\ &\geq (1/\delta)(\xi_j + \delta) \\ &= \xi_j/\delta + 1 \\ &= \gamma_j + 1, \qquad \forall l = 1, \ldots, k, \forall j \in J(l). \end{aligned}$$

□

The above proposition provides an equivalent definition of h-polyhedral separability.

Now, we give a necessary and sufficient condition for separability.

Proposition 2.2. The sets $\mathscr{A}$ and $\mathscr{B}$ are h-polyhedrally separable, for some $h \leq |\mathscr{B}|$, if and only if

$$\text{conv}(\mathscr{A}) \cap \mathscr{B} = \emptyset.$$

Proof. Necessary Condition. From the definition, the sets

$$\mathscr{X} = \{x \mid w^{(j)T} x < \xi_j, j = 1, \ldots, h\} \quad \text{and} \quad \mathscr{B}$$

are disjoint, and the condition follows from the fact that $\mathscr{X}$ is convex and contains $\mathscr{A}$.

Sufficient Condition. The hypothesis ensures that, for all $b_l \in \mathscr{B}$, we have $b_l \notin \text{conv}(\mathscr{A})$. Thus, there exists a hyperplane separating b_l from $\text{conv}(\mathscr{A})$; i.e., there exists $(w^{(j(l))}, \xi_{j(l)})$ such that

$$\begin{aligned} &b_l^T w^{(j(l))} > \xi_{j(l)}, \\ &x^T w^{(j(l))} < \xi_{j(l)}, \qquad \forall x \in \text{conv}(A). \end{aligned}$$

Consequently,

$$a_i^T w^{(j(l))} < \xi_{j(l)}, \qquad \forall i = 1, \ldots, m.$$

Thus, we have $|\mathscr{B}|$ hyperplanes $\{(w^{j(1)}, \xi_{j(l)}), \ldots, (w^{j(k)}, \xi_{j(k)})\}$ such that

$$a_i^T w^{j(l))} < \xi_{j(l)}, \qquad \forall i = 1, \ldots, m, \forall l = 1, \ldots, k,$$

$$b_l^T w^{j(l))} > \xi_{j(l)}, \qquad \forall l = 1, \ldots, k.$$

Of course, some of the $|\mathscr{B}|$ hyperplanes may be redundant, and this corresponds to some hyperplane that separates more than just one point of $\mathscr{B}$ from conv($\mathscr{A}$). □

We observe that the condition ensuring polyhedral separability is obviously weaker than the one for linear separability. Moreover, the role played by the sets $\mathscr{A}$ and $\mathscr{B}$ is not symmetric. This is shown in Fig. 3, where we observe that the set $\mathscr{A}$ is polyhedrally separable from the set $\mathscr{B}$, but that the reverse is not true.

3. Error Function

In this section, we address the problem of finding a polyhedral separation of two sets $\mathscr{A}$ and $\mathscr{B}$ once we have fixed h, the number of hyperplanes.

Given any set of hyperplanes $\{v^{(j)}, \gamma_j\}$, $j = 1, \ldots, h$, we will refer to any point $a_i \in \mathscr{A}$, $i = 1, \ldots, m$, as well classified if the following condition is satisfied:

$$a_i^T v^{(j)} - \gamma_j + 1 \leq 0, \qquad \forall j = 1, \ldots, h,$$

which can be rewritten as

$$\max_{1 \leq j \leq h} [a_i^T v^{(j)} - \gamma_j + 1] \leq 0.$$

On the other hand, we define the classification error for a_i as

$$\max_{1 \leq j \leq h} [0, a_i^T v^{(j)} - \gamma_j + 1]. \tag{6}$$

Of course, to a well-classified point there corresponds zero classification error.

Analogously, a point $b_l \in \mathscr{B}$, $l = 1, \ldots, k$, is well classified by the given set of hyperplanes if the following condition is satisfied:

$$-b_l^T v^{(j)} + \gamma_j + 1 \leq 0$$

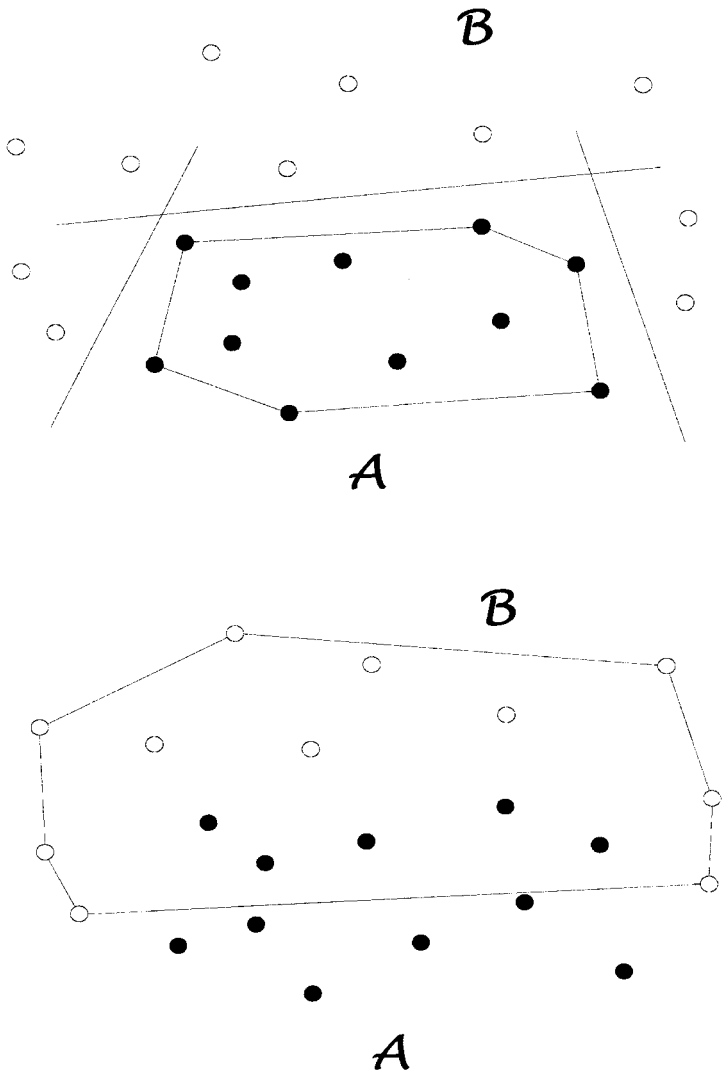

Fig. 3. h-polyhedral separability condition.

for at least one index $j \in \{1, \ldots, h\}$. The above condition can be rewritten as

$$\min_{1 \le j \le h} [-b_l^T v^{(j)} + \gamma_j + 1] \le 0.$$

Thus, the classification error for point b_l is

$$\max\left\{0,\ \min_{1\le j\le h}\left[-b_l^T v^{(j)}+\gamma_j+1\right]\right\}. \tag{7}$$

An (averaged) classification error function can be defined as

$$z([(v^{(j)},\gamma_j)]) \triangleq (1/m)\sum_{i=1}^{m}\left\{\max_{1\le j\le h}\left[a_i^T v^{(j)}-\gamma_j+1\right]_+\right\} + (1/k)\sum_{l=1}^{k}\left\{\min_{1\le j\le h}\left[-b_l^T v^{(j)}+\gamma_j+1\right]\right\}_+. \tag{8}$$

The function z is by definition nonnegative. Minimizing z corresponds to minimize the weighted sum of the L_1 norms of the vectors of classification errors for both the sets $\mathscr{A}$ and $\mathscr{B}$.

Proposition 3.1. The sets $\mathscr{A}$ and $\mathscr{B}$ are h-polyhedrally separable if and only if there exists an optimal set of hyperplanes $\{(v^{(j)},\gamma_j)\}$, $j=1,\ldots,h$, such that the corresponding value z^* of z is equal to 0. In this case,

(i) $v^{(j)}=0$, $\forall j=1,\ldots,h$, cannot be the optimal solution;
(ii) if there exists a nonempty set $\bar{J}\subset\{1,\ldots,h\}$ such that $v^{(j)}=0$, $\forall j\in\bar{J}$, then the sets $\mathscr{A}$ and $\mathscr{B}$ are $(h-|\bar{J}|)$-polyhedrally separable.

Proof. The existence of a set of hyperplanes $\{(v^{(j)},\gamma_j)\}$ such that $z^*=0$ is equivalent to

$$\max_{1\le j\le h}\left[a_i^T v^{(j)}-\gamma_j+1\right]\le 0, \qquad \forall i=1,\ldots,m,$$
$$\min_{1\le j\le h}\left[-b_l^T v^{(j)}+\gamma_j+1\right]\le 0, \quad \forall l=1,\ldots,k.$$

This system can be rewritten as

$$a_i^T v^{(j)}-\gamma_j+1\le 0, \qquad \forall i=1,\ldots,m,\ \forall j=1,\ldots,h,$$
$$-b_l^T v^{(j)}+\gamma_j+1\le 0, \quad \forall l=1,\ldots,k, \text{ for at least one } j\in\{1,\ldots,h\},$$

i.e.,

$$a_i^T v^{(j)}\le \gamma_j-1, \qquad \forall i=1,\ldots,m,\ \forall j=1,\ldots,h,$$
$$b_l^T v^{(j)}\ge \gamma_j+1, \qquad \forall l=1,\ldots,k, \text{ for at least one } j\in\{1,\ldots,h\},$$

which is the definition of h-polyhedral separation between the sets $\mathscr{A}$ and $\mathscr{B}$.

(i) If $v^{(j)} = 0$, $\forall j = 1, \ldots, h$, we would have

$$z^* = \min_{\gamma} \left[\max\left\{0, 1 - \min_{1 \le j \le h} \gamma_j\right\} + \max\left\{0, 1 + \min_{1 \le j \le h} \gamma_j\right\}\right] = 2,$$

that is in contrast to the h-polyhedral separability between $\mathscr{A}$ and $\mathscr{B}$ ($z^* = 0$). In fact,

$$\max\left\{0, 1 - \min_{1 \le j \le h} \gamma_j\right\} = \begin{cases} 0, & \text{if } \min\limits_{1 \le j \le h} \gamma_j \ge 1, \\ 1 - \min\limits_{1 \le j \le h} \gamma_j, & \text{if } \min\limits_{1 \le j \le h} \gamma_j < 1, \end{cases}$$

$$\max\left\{0, 1 + \min_{1 \le j \le h} \gamma_j\right\} = \begin{cases} 0, & \text{if } \min\limits_{1 \le j \le h} \gamma_j \le -1, \\ 1 + \min\limits_{1 \le j \le h} \gamma_j, & \text{if } \min\limits_{1 \le j \le h} \gamma_j > -1; \end{cases}$$

thus, we have that

$$z^* = \min \begin{cases} 1 - \min\limits_{1 \le j \le h} \gamma_j \ge 2, & \text{for } \min\limits_{1 \le j \le h} \gamma_j \le -1, \\ 2, & \text{for } -1 < \min\limits_{1 \le j \le h} \gamma_j < 1, \\ 1 + \min\limits_{1 \le j \le h} \gamma_j \ge 2, & \text{for } \min\limits_{1 \le j \le h} \gamma_j \ge 1; \end{cases}$$

i.e.,

$$z^* = 2.$$

(ii) Now, we suppose that there exists a set $\bar{J} \subset \{1, \ldots, h\} = J$ such that

$$v^{(j)} = 0, \qquad \forall j \in \bar{J}.$$

By hypothesis, we have

$$0 = z^* = \min(1/m) \sum_{i=1}^{m} \max\left\{\max_{j \in J \setminus \bar{J}} [a_i^T v^{(j)} - \gamma_j + 1], \max_{j \in \bar{J}} (-\gamma_j + 1)\right\}_+ + (1/k) \sum_{l=1}^{k} \min\left\{\min_{j \in J \setminus \bar{J}} [-b_l^T v^{(j)} + \gamma_j + 1], \min_{j \in \bar{J}} (\gamma_j + 1)\right\}_+ .$$

It follows that

$$\max_{j \in J \setminus \bar{J}} [a_i^T v^{(j)} - \gamma_j + 1] \le 0,$$

$$\max_{j \in \bar{J}} [-\gamma_j + 1] \le 0,$$

i.e.,

$$a_i^T v^{(j)} - \gamma_j + 1 \le 0, \qquad \forall i = 1, \ldots, m, \forall j \in J \setminus \bar{J}, \tag{9a}$$

$$\gamma_j \ge 1, \qquad \forall j \in \bar{J}, \tag{9b}$$

and

$$\left[\min_{j \in J \setminus \bar{J}} [-b_l^T v^{(j)} + \gamma_j + 1] \le 0 \text{ or } \min_{j \in \bar{J}} (\gamma_j + 1) \le 0 \right].$$

We observe that

$$\min_{j \in \bar{J}} (\gamma_j + 1) \le 0$$

cannot hold. In fact, in this case, we would have

$$\gamma_j \le -1, \qquad \text{for at least one } j \in \bar{J},$$

which is in contrast to

$$\gamma_j \ge 1, \qquad \forall j \in \bar{J},$$

in (9). Thus, we must have

$$\min_{j \in J \setminus \bar{J}} [-b_l^T v^{(j)} + \gamma_j + 1] \le 0, \tag{10}$$

which together with the first part of (9) ensures that the sets $\mathscr{A}$ and $\mathscr{B}$ are $(h - |\bar{J}|)$-polyhedral separable. □

4. Algorithm

Now, we rewrite the problem (8) in compact form. Let $(V, \gamma) \in \mathbb{R}^{h \times (n+1)}$ be the matrix grouping the decision variables. $V \in \mathbb{R}^{h \times n}$ is the matrix whose rows are the vectors $v^{(j)}$, and $\gamma \in \mathbb{R}^h$ is the vector whose jth component is γ_j. In the sequel, $V^{(j)}$ will denote the jth row of V, $j = 1, \ldots, h$; thus,

$$V^{(j)} \triangleq v^{(j)T}.$$

The error function (8) is written as follows:

$$f(V, \gamma) \triangleq f_1(V, \gamma) + f_2(V, \gamma), \tag{11}$$

where

$$f_1(V,\gamma) \triangleq (1/m) \sum_{i=1}^{m} [\max(Va_i - \gamma + e)_+],$$

$$f_2(V,\gamma) \triangleq (1/k) \sum_{l=1}^{k} [\min(-Vb_l + \gamma + e)]_+;$$

it is the sum of two terms taking into account, respectively, the classification errors for $\mathscr{A}$ and $\mathscr{B}$. Functions of this type are investigated in Ref. 40.

Remark 4.1. See Fig. 4. $f(V,\gamma)$ is piecewise affine; $f_1(V,\gamma)$ is convex; $f_2(V,\gamma)$ is not convex nor concave (in fact, it is quasiconcave). In Fig. 4, given two affine functions $p_1(x)$ and $p_2(x)$ of the scalar variable x, we depict in (a) a function of type f_1 and in (b) a function of the type f_2.

4.1. Directional Derivative of the Error Function.

(a) Given a set of m affine functions of the type $p_i(x)$, $x \in \mathbb{R}^n$,

$$p_i(x) = c_i^T x + d_i,\ c_i \in \mathbb{R}^n,\ d_i \in \mathbb{R}, \qquad i = 1, \ldots, m,$$

we can define a max function,

$$p(x) \triangleq \max_{1 \le i \le m} [p_i(x)]_+,$$

which is convex, being the pointwise maximum of convex functions. Now, we define

$$\hat{p}_i(x) \triangleq \max[0, p_i(x)], \qquad \forall i = 1, \ldots, m,$$

which is nondifferentiable on the hyperplane

$$H_i \triangleq \{x \,|\, p_i(x) = c_i^T x + d_i = 0\}.$$

Its subdifferential is

$$\partial \hat{p}_i(x) = \begin{cases} \{0\}, & \text{if } c_i^T x + d_i = p_i(x) < 0, \\ \operatorname{conv}(0, c_i), & \text{if } x \in H_i, \\ \{c_i\}, & \text{if } c_i^T x + d_i = p_i(x) > 0. \end{cases}$$

Thus,

$$\partial p(x) = \operatorname*{conv}_{i \in \bar{I}(x)} \partial \hat{p}_i(x),$$

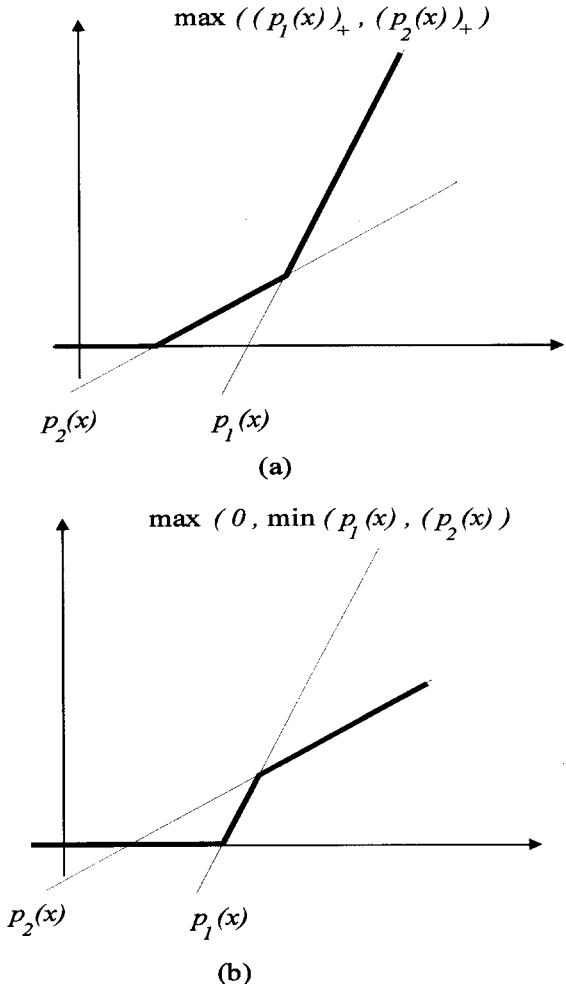

Fig. 4. (a) Function of type f_1; (b) function of the type f_2.

where

$$\bar{I}(x) = \{i | p(x) = \hat{p}_i(x)\}.$$

Now, we consider the directional derivative $p'(x, d)$ of p along the direction d. It holds that

$$p'(x, d) = \max_{g \in \partial p(x)} g^T d,$$

whose expression is the following:

$$p'(x,d)=\begin{cases}0, & \text{if } p(x)=0 \text{ and } I(x)=\varnothing,\\ \max\limits_{i\in I(x)}[(c_i^T d)_+], & \text{if } p(x)=0 \text{ and } I(x)\neq\varnothing,\\ \max\limits_{i\in I(x)} c_i^T d, & \text{if } p(x)>0,\end{cases}$$

where

$$I(x)=\{i\,|\,p(x)=p_i(x)\}.$$

(b) Given a set of affine functions

$$q_l(x)=r_l^T x+s_l, \qquad l=1,\ldots,k,$$

we define

$$q(x)\triangleq\left[\min_{1\le l\le k}(q_l(x))\right]_+.$$

For the directional derivative, we have

$$q'(x,d)=\begin{cases}0, & \text{if } q(x)=0 \text{ and } I(x)=\varnothing,\\ \min\limits_{i\in I(x)}[(r_l^T d)_+], & \text{if } q(x)=0 \text{ and } I(x)\neq\varnothing,\\ \min\limits_{l\in I(x)} r_l^T d, & \text{if } q(x)>0,\end{cases}$$

where now

$$I(x)=\{l\,|\,q(x)=q_l(x)\}.$$

4.2. Finding a Descent Direction. The problem of minimizing the error function f cannot be transformed into a linear program, as it is the case in linear separability. In fact, it is a problem of the min–max–min type, at least as far as function f_2 is concerned.

For a fixed (V,γ), we define

$$\epsilon_j^{(i)}\triangleq V^{(j)}a_i-\gamma_j+1, \qquad \forall i=1,\ldots,m,\ \forall j=1,\ldots,h,$$

$$\epsilon_{\max}^{(i)}\triangleq\max_{1\le j\le h}[\epsilon_j^{(i)}], \qquad \forall i=1,\ldots,m.$$

Of course, when positive, $\epsilon_{\max}^{(i)}$ is the classification error for point $a_i\in\mathscr{A}$.

Analogous definitions can be given for the set $\mathscr{B}$,

$$\varrho_j^{(l)}\triangleq -V^{(j)}b_l+\gamma_j+1, \qquad \forall l=1,\ldots,k,\ \forall j=1,\ldots,h,$$

$$\varrho_{\min}^{(l)}\triangleq\min_{1\le j\le h}[\varrho_j^{(l)}], \qquad \forall l=1,\ldots,k;$$

when positive, $\varrho_{\min}^{(l)}$ is the classification error for point $b_l\in\mathscr{B}$.

Some additional definitions are necessary to describe our approach.

Definition 4.1. The index sets I_a^A and I_a^B of misclassified and borderline points of $\mathscr{A}$ and $\mathscr{B}$ respectively are defined as

$$I_a^A = \{i \,|\, \epsilon_{\max}^{(i)} \geq 0\},$$
$$I_a^B = \{l \,|\, \varrho_{\min}^{(l)} \geq 0\}.$$

We remark that the misclassified and borderline points are often referred to in the literature as the support vectors.

Definition 4.2. For given positive parameters ϵ_A and ϵ_B, the index sets I_e^A and I_e^B of almost misclassified points of $\mathscr{A}$ and $\mathscr{B}$ respectively are defined as

$$I_e^A = \{i \,|\, -\epsilon_A \leq \epsilon_{\max}^{(i)} < 0\},$$
$$I_e^B = \{l \,|\, -\epsilon_B \leq \varrho_{\min}^{(l)} < 0\}.$$

Definition 4.3. For a given $\epsilon > 0$, the index set of the active and almost active hyperplanes for $a_i \in \mathscr{A}$, $i \in I_a^A \cup I_e^A$, is defined as

$$J_i^A = \{s \,|\, \epsilon_{\max}^{(i)} - \epsilon_s^{(i)} \leq \epsilon\}.$$

Definition 4.4. The index set of the active hyperplanes for $b_l \in \mathscr{B}$, $l \in I_a^B \cup I_e^B$, is defined as

$$S_l^B = \{s \,|\, \varrho_s^{(l)} = \varrho_{\min}^{(l)}\}.$$

We describe here a strategy either to conclude that, at any point, an optimality condition is satisfied or to calculate a descent direction for f by solving a linear program or possibly a set of linear programs.

Let $(\bar{V}, \bar{\gamma})$ be any set of decision variables (the current solution in an iterative scheme). We observe that the points a_i with $i \notin I_a^A \cup I_e^A$ and the points b_l with $l \notin I_a^B \cup I_e^B$ do not play any role in determining the local behavior of the objective function.

We recall that our objective function has the form

$$f(V, \gamma) = f_1(V, \gamma) + f_2(V, \gamma).$$

Around $(\bar{V}, \bar{\gamma})$, the function $f_1(V, \gamma)$ coincides with

$$f_1^{(I_a^A \cup I_e^A)}(V, \gamma) \triangleq (1/m) \sum_{i \in I_a^A \cup I_e^A} [\max(V^{(j)} a_i - \gamma_j + e)_+].$$

Any point a_i, $i \notin I_a^A \cup I_e^A$, is characterized by the fact that, at the current solution,

$$\epsilon_{\max}^{(i)} < -\epsilon_A < 0,$$

which implies

$$\bar{V}^{(j)} a_i - \bar{\gamma}_j + 1 < -\epsilon_A, \qquad \forall j = 1, \ldots, h.$$

It is easy to show that, for any possible displacement (Δ, δ), with $\Delta \in \mathbb{R}^{h \times n}$, $\delta \in \mathbb{R}^h$, $\Delta^{(j)}$ the jth row of Δ, and δ_j the jth component of δ, we have that

$$[\bar{V}^{(j)} + \Delta^{(j)}] a_i - (\bar{\gamma}_j + \delta_j) + 1,$$

remains nonpositive $\forall j = 1, \ldots, h$, provided that

$$\|\Delta^{(j)}, \delta_j\| \leq \epsilon_A / M^A,$$

with

$$M^A = \max_{1 \leq i \leq m} \left\| \begin{matrix} a_i \\ -1 \end{matrix} \right\|.$$

As for the function $f_2(V, \gamma)$, we have again that around $(\bar{V}, \bar{\gamma})$, only the points b_l, $l \in I_a^B \cup I_e^B$, play a role. In particular, the points b_l, $l \notin I_a^B \cup I_e^B$, cannot become misclassified provided that

$$\|\Delta^{(j)}, \delta_j\| \leq \epsilon_B / M^B, \qquad \forall j = 1, \ldots, h,$$

where

$$M^B = \max_{1 \leq l \leq k} \left\| \begin{matrix} -b_l \\ 1 \end{matrix} \right\|.$$

Thus, we can assume that, around $(\bar{V}, \bar{\gamma})$, the function $f_2(V, \gamma)$ has the form

$$f_2^{(I_a^B \cup I_e^B)}(V, \gamma) \triangleq (1/k) \sum_{l \in I_a^B \cup I_e^B} \left\{ \min_{s \in S_l^B} [-V^{(s)} b_l + \gamma_s + 1] \right\}_+.$$

Let $l_1, \ldots, l_p, \ldots l_t$, with

$$t = |I_a^B \cup I_e^B|,$$

be the set of indices of the misclassified, borderline, and almost misclassified points of $\mathscr{B}$. For each combination $S^B = \{s_{l_1}, \ldots, s_{l_p}, \ldots, s_{l_t}\}$ of indices of hyperplanes, with

$$S^B \in S_{l_1}^B \times \cdots \times S_{l_p}^B \times \cdots \times S_{l_t}^B,$$

we can define the function

$$f_2^{(I_a^B \cup I_e^B, S^B)}(V, \gamma) \triangleq (1/k) \sum_{l_p \in I_a^B \cup I_e^B} [-V^{(s_{l_p})} b_{l_p} + \gamma_{s_{l_p}} + 1]_+.$$

We observe that

(i) $f_2^{(I_a^B \cup I_e^B, S^B)}(V, \gamma)$ is convex for all possible choices of the combination S^B;

(ii) $f_2(V, \gamma) \leq f_2^{(I_a^B \cup I_e^B, S^B)}(V, \gamma)$ around $(\bar{V}, \bar{\gamma})$ for all possible choices of the combination S^B;

(iii) around $(\bar{V}, \bar{\gamma})$, for each point (V, γ), there exists at least one combination $S^B \in S_{l_1}^B \times \cdots \times S_{l_t}^B$ such that

$$f_2(V, \gamma) = f_2^{(I_a^B \cup I_e^B, S^B)}(V, \gamma).$$

We can state the following theorem.

Theorem 4.1. The current solution $(\bar{V}, \bar{\gamma})$ is a local minimum of $f(V, \gamma)$ if and only if

$$\begin{aligned} \xi^* &= f_1(\bar{V}, \bar{\gamma}) + f_2(\bar{V}, \bar{\gamma}) \\ &= f(\bar{V}, \bar{\gamma}), \end{aligned} \tag{12}$$

where

$$\xi^* = \min_{S^B \in S_{l_1}^B \times \cdots \times S_{l_t}^B} \xi^{S^B},$$

with ξ^{S^B} defined as

$$\xi^{S^B} \triangleq \min_{V, \gamma} [f_1^{(I_a^A \cup I_e^A)}(V, \gamma) + f_2^{(I_a^B \cup I_e^B, S^B)}(V, \gamma)]. \tag{13}$$

Proof. If we define

$$\hat{f}(V, \gamma, S^B) \triangleq f_1^{(I_a^A \cup I_e^A)}(V, \gamma) + f_2^{(I_a^B \cup I_e^B, S^B)}(V, \gamma),$$

the necessary condition follows from the fact that $\hat{f}(V, \gamma, S^B)$ is convex and, around $(\bar{V}, \bar{\gamma})$,

$$\begin{aligned} \hat{f}(V, \gamma, S^B) &\geq f(V, \gamma) \\ &\geq f(\bar{V}, \bar{\gamma}) \\ &= \hat{f}(\bar{V}, \bar{\gamma}, S^B), \qquad \forall S^B. \end{aligned}$$

Thus,

$$\begin{aligned} \xi^{S^B} &= f(\bar{V}, \bar{\gamma}), \qquad \forall S^B, \\ \xi^* &= f(\bar{V}, \bar{\gamma}). \end{aligned}$$

As for the sufficient condition, let us suppose that (12) holds and the current solution $(\bar{V}, \bar{\gamma})$ is not a local minimum of $f(V, \gamma)$. It follows that there exists $(\hat{V}, \hat{\gamma})$ around $(\bar{V}, \bar{\gamma})$ such that

$$f(\hat{V}, \hat{\gamma}) < f(\bar{V}, \bar{\gamma}).$$

By the definition of $f_2^{(I_a^B \cup I_e^B, S^B)}(V, \gamma)$, there exists at least one combination $\bar{S}^B \in S_{l_1}^B \times \cdots \times S_{l_t}^B$ such that

$$f_2(\hat{V}, \hat{\gamma}) = f_2^{(I_a^B \cup I_e^B, \bar{S}^B)}(\hat{V}, \hat{\gamma}).$$

Thus, we have that

$$\begin{aligned}
\xi^* &\leq \xi^{\bar{S}^B} \\
&\leq f_1(\hat{V}, \hat{\gamma}) + f^{(I_a^B \cup I_e^B, \bar{S}^B)}(\hat{V}, \hat{\gamma}) \\
&= f(\hat{V}, \hat{\gamma}) \\
&< f(\bar{V}, \bar{\gamma}),
\end{aligned}$$

which contradicts (12). □

In the case $\xi^* = f(\bar{V}, \bar{\gamma})$, the above theorem provides a stopping criterion in view of an iterative descent procedure. In the case $\xi^* < f(\bar{V}, \bar{\gamma})$, let us focus on any combination S^{*B} such that $\xi^{S^{*B}} < f(\bar{V}, \bar{\gamma})$.

Let (V^*, γ^*) be the optimal solution to the problem (13) defined by S^{*B}, and let

$$(\Delta^*, \delta^*) = (V^* - \bar{V}, \gamma^* - \bar{\gamma}).$$

Of course, taking into account the convexity of $\hat{f}(V, \gamma, S^{*B})$, the observation (ii), and the fact that $\hat{f}(\bar{V}, \bar{\gamma}, S^{*B}) = f(\bar{V}, \bar{\gamma})$, we have that (Δ^*, δ^*) provides us with a descent direction for the objective function $f(V, \gamma)$.

Let us define the predicted reduction,

$$\bar{\sigma} \triangleq \hat{f}(V^*, \gamma^*, S^{*B}) - f(\bar{V}, \bar{\gamma});$$

taking into account that $\hat{f}$ majorizes f in a neighborhood of radius at least

$$\bar{\epsilon} \triangleq \min\{\epsilon_A / M^A, \epsilon_B / M^B\}$$

(here, we have adopted L_∞ as an appropriate matrix norm), it is easy to conclude that the reduction of f that can be achieved along the descent direction (Δ^*, δ^*) adopting a stepsize $0 < \alpha \leq 1$ is at least equal to

$$\bar{\sigma}\bar{\epsilon} / \|\Delta^*, \delta^*\|,$$

which implies that, introducing into the statement of problem (13) an appropriate limitation on the norm of (V, γ) and thus of (Δ, δ), we come out with

the possibility of obtaining along the direction (Δ^*, δ^*) a guaranteed fraction of the predicted reduction $\bar{\sigma}$.

For a given combination $\bar{S}^B = (s_{l_1}, s_{l_2}, \ldots, s_{l_p}, \ldots, s_{l_t})$, the calculation of ξ^{S^B} reduces to the solution of the following linear program:

$$\xi^{S^B} = \min \left[(1/m) \sum_{i \in I_a^A \cup I_e^A} \vartheta_i + (1/k) \sum_{l_p \in I_a^B \cup I_e^B} \mu_{l_p} \right], \tag{14a}$$

$$\text{s.t.} \quad \vartheta_i \geq \Delta^{(j)} a_i - \delta_j + \epsilon_j^{(i)}, \qquad \forall i \in I_a^A \cup I_e^A, \forall j = 1, \ldots, h, \tag{14b}$$

$$\vartheta_i \geq 0, \qquad \forall i \in I_a^A \cup I_e^A, \tag{14c}$$

$$\mu_{l_p} \geq -\Delta^{(s_{l_p})} b_{l_p} + \delta_{s_{l_p}} + \rho_{s_{l_p}}^{(l_p)}, \quad \forall l_p \in I_a^B \cup I_e^B, \tag{14d}$$

$$\mu_{l_p} \geq 0, \qquad \forall l_p \in I_a^B \cup I_e^B, \tag{14e}$$

$$-E \leq \Delta \leq E, \tag{14f}$$

$$-e \leq \delta \leq e, \tag{14g}$$

where, according to the above observations, we have added appropriate box constraints on (Δ, δ) by means of E, a matrix of ones of appropriate dimensions, and e, a vector of ones of appropriate dimensions.

Now, we are ready to state our descent algorithm.

Algorithm 4.1.

Input Parameters. h number of hyperplanes; σ stopping tolerance, ϵ_A, ϵ_B; $(V^{(1)}, \gamma^{(1)})$ an initial estimate of the decision variables.

We describe the main iteration, dropping for sake of clarity the iteration index.

Main Iteration. Let $(\bar{V}, \bar{\gamma})$ be the current solution. While there exists a combination not yet considered, execute the following steps:

Step 1. Select a combination $\bar{S}^B$ not yet considered.

Step 2. Calculate $\xi^{\bar{S}^B}$ and (Δ^*, δ^*) by solving problem (14).

Step 3. If

$$\bar{\sigma} = \xi^{\bar{S}^B} - f(\bar{V}, \bar{\gamma}) < -\sigma,$$

calculate a descent stepsize along the direction (Δ^*, δ^*) and exit from main iteration.

If all combinations have been considered, stop (approximate optimality has been achieved).

The descent stepsize is calculated as the maximal one which ensures that no previously well-classified point becomes a borderline point. Thus,

we calculate α_1 and α_2, the maximal stepsize referred to sets $\mathscr{A}$ and $\mathscr{B}$ respectively and then we fix the stepsize

$$\alpha = \min\{\alpha_1, \alpha_2, 1\},$$

where

(i) $$0 < \alpha_1 = \min_{i \notin I_a^A \cup I_e^A,\, j \in \hat{J}_i^A} \{ -\epsilon_j^{(i)} / [\Delta^{(j)*} a_i - \delta_j^*] \},$$

with

$$\hat{J}_i^A = \{ j \,|\, \Delta^{(j)*} a_i - \delta_j^* > 0 \};$$

(ii) $$0 < \alpha_2 = \min_{l \notin I_a^B \cup I_e^B} \alpha_l,$$

where if

$$J_l^B = \{ j \,|\, \rho_j^{(l)} < 0 \text{ and } [-\Delta^{(j)*} b_l + \delta_j^*] > 0 \} \neq \emptyset,$$

then

$$\alpha_l = \max_{j \in J_l^B} \{ -\varrho_j^{(l)} / [-\Delta^{(j)*} b_l + \delta_j^*] \};$$

else, set $\alpha_l = 1$.

Theorem 4.2. For a fixed σ, Algorithm 4.1 terminates in a finite number of steps.

Proof. Finiteness of the algorithm follows from the facts that being nonnegative, the objective function is bounded from below and that, at each descent iteration, we obtain a reduction of the objective function which is at least equal to $\bar{\sigma}\bar{\epsilon} / \|\Delta^*, \delta^*\|$. Thus, we can only have a finite number of descents. □

The setting of the input parameters deserves some discussion. The parameters ϵ_A and ϵ_B have impact on the number of points that are considered almost misclassified. Therefore, the bigger ϵ_A and ϵ_B are, the larger is the size of the LP (14). On the other hand, to adopt relatively big values for ϵ_A and ϵ_B prevents the risk of having small stepsize at Step 3 of the algorithm and consequently small reduction in the objective function.

A possible way to set the stopping tolerance σ is to fix it at a given fraction of the value of the optimal error function at the solution of the linear separation method (Ref. 10).

We observe that the algorithm is aimed at finding a local minimum of the objective function, which is not convex.

The existence of local minima may constitute a relevant drawback, since it may happen that the algorithm stops at a point $(\bar{V}, \bar{\gamma})$ which is an approximate local minimum with an objective function value $f(\bar{V}, \bar{\gamma})$ far away from zero. Of course, this fact does not ensure that $\mathscr{A}$ and $\mathscr{B}$ are not h-polyhedrally separable, since such a conclusion would be drawn only in the case the local minimum were a global one too.

Thus, in our implementation, we have added a procedure that gives the possibility to escape from a local minimum, once it has been detected by the algorithm. This objective is pursued by modifying the definition of S^B, by accommodating also the indices of the almost active hyperplanes. In particular, for each misclassified, borderline, or almost misclassified point b_l of $\mathscr{B}$, we consider now S_l^B defined as

$$S_l^B \triangleq \{s \,|\, \rho_s^{(l)} - \rho_{\min}^{(l)} \leq \epsilon\},$$

for a given value of the threshold parameter $\epsilon > 0$.

5. Examples and Numerical Tests

Before describing the numerical tests performed to validate the approach, we present three simple explanatory examples of separation in $\mathbb{R}^2$.

In Figs. 5–7, three different pairs of sets $\mathscr{A}$ and $\mathscr{B}$ are depicted together with the hyperplanes obtained using our approach. In particular, in Fig. 5, we have considered the case $h = 1$ and $h = 2$ ($\mathscr{A}$ and $\mathscr{B}$ are not linearly separable, but they are 2-polyhedrally separable). In Fig. 6, $\mathscr{A}$ and $\mathscr{B}$ are not linearly separable; however they are 3-polyhedrally separable. Finally, in Fig. 7, $\mathscr{A}$ and $\mathscr{B}$ are not linearly separable and it is not even possible to find a nontrivial hyperplane minimizing the classification error according to Ref. 10; in fact, in this case, $\mathscr{A}$ and $\mathscr{B}$ share the same barycenter. Nevertheless, $\mathscr{A}$ and $\mathscr{B}$ are 4-polyhedrally separable.

Our algorithm has been implemented in C on an IBM-Netfinity Pentium II 350 MHz, using CPLEX as the LP solver. All runs of our h-polyhedral separation algorithm have been performed setting

$$\epsilon_A = 0.001, \qquad \epsilon_B = 0.001, \qquad \epsilon = 0.001.$$

The parameter h has been set equal to 2; only for the kidney transplant dataset, it has been set equal to 4. Moreover, we have assumed a constant value $\sigma = 0.001$ of the stopping tolerance.

The datasets used have been extracted from the UC-Irvine Repository (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/) and are commonly used for performance evaluation of classification algorithms. In particular,

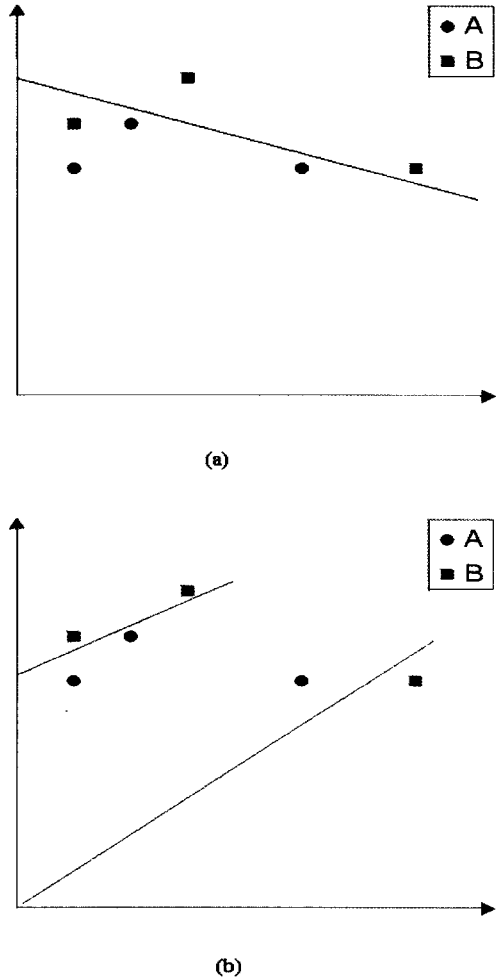

Fig. 5. (a) Case $h = 1$; (b) Case $h = 2$.

the datasets considered are the Wisconsin Breast Cancer Diagnosis (WBCD), the Wisconsin Breast Cancer Prognosis (WBCP), the Cleveland Heart Disease (Heart), the BUPA Liver Disorders (Liver), the Pima Indians Diabets (Diabets), and the United States Congressional Voting Records (Votes).

Tenfold cross-validation (a widely used protocol), which consists in splitting the dataset of interest into ten equally sized pieces, has been adopted. For each dataset, the method has been run ten times, and each

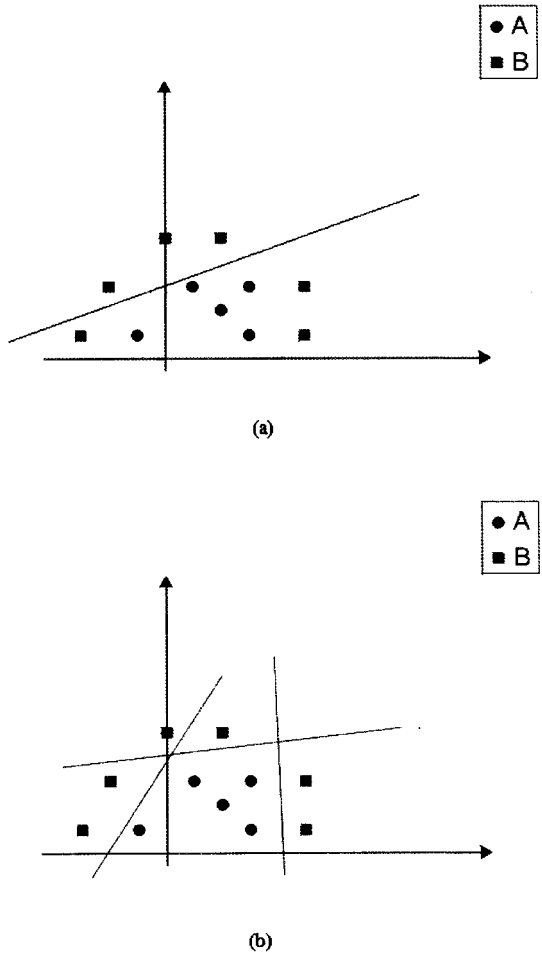

Fig. 6. (a) Case $h = 1$; (b) Case $h = 3$.

time nine pieces have been used as the training set and the remaining one as the testing set. The accuracy is defined as the ratio between the number of well-classified points of both $\mathscr{A}$ and $\mathscr{B}$ and the dataset size. The average accuracy on both the training and testing sets has been calculated for each dataset.

A number of algorithms and the related results drawn from the literature have been considered for comparison purposes. In particular, we have taken into consideration the robust linear programming (RLP, Ref. 10),

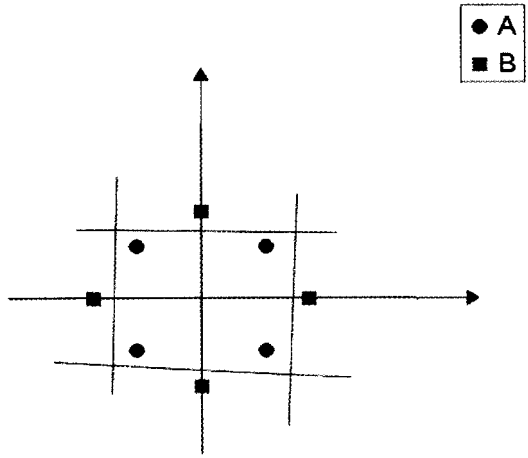

Fig. 7. Case $h = 4$.

the hybrid misclassification minimization (HMM, Ref. 41), the parametric misclassification minimization (PMM, Refs. 42–43), a single support vector machine (SVM, Ref. 35), the global tree optimization (GTO, Ref. 35), and GTO with SVM (GTO/SVM, Ref. 35).

The results obtained by our method are reported in Table 1 together with those reported in the above quoted papers for each considered method. At the testing level, our method has provided comparable results.

As for WBCP, we cannot compare our results with those described in (Refs. 41, 35), where a smaller dataset has been considered. By "Linear Separation" we denote our implementation of the method (Ref. 10).

Besides the tests on standard datasets, we describe here an application to patient follow-up after kidney transplant. In this pathology, the patient can evolve toward different states. The problem is to forecast such evolution on the basis of clinical data obtained after the surgery.

The different possible after-transplant evolutions are: (a) full success; (b) more or less serious intoxications; (c) rejection. The problem is to obtain the classification of the patients on the basis of four parameters (citrate, trimethylamine *N*-oxide, creatine, hippurate), measured at early post-surgical stage.

The dataset of clinical data, provided by the Reparto di Nefrologia, Ospedale Civile dell'Annunziata di Cosenza, consists of 71 samples, 21 for the rejection (set $\mathscr{A}$) and 50 for the other evolutions (set $\mathscr{B}$). In Table 2, we give the results of the leave-one-out validation of the two methods tested (Linear Separation and 4-Polyhedral Separation).

Table. 1. Comparison of training and testing correctness on standard datasets.

Dataset	m k n	Method	Average training set correctness	Average testing set correctness	Number of LP solved (average)	Running time (sec)
		RLP	97.73	97.21		
		HMM	97.87	97.36		
	239	PMM	98.57	96.47		
WBCD	443	SVM	—	97.20		
	9	GTO	—	95.70		
		GTO/SVM	—	96.60		
		2-Polyhedral Sep.	98.40	96.92	26	4
		RLP	84.47	83.51		
		HMM	87.50	82.84		
	81	PMM	91.43	82.16		
Heart	216	SVM	—	81.50		
	13	GTO	—	82.50		
		GTO/SVM	—	82.50		
		2-Polyhedral Sep.	88.33	83.77	10	1
		RLP	68.99	66.93		
		HMM	72.21	66.64		
	145	PMM	74.85	68.37		
Liver	200	SVM	—	60.60		
	6	GTO	—	68.10		
		GTO/SVM	—	68.70		
		2-Polyhedral Sep.	74.66	69.95	11	2
		RLP	76.77	76.00		
		HMM	78.42	75.89		
	268	PMM	80.55	76.67		
Diabets	500	SVM	—	77.60		
	8	GTO	—	76.80		
		GTO/SVM	—	76.80		
		2-Polyhedral Sep.	76.48	75.53	23	6
		RLP	97.45	95.63		
		HMM	98.03	95.62		
	168	PMM	98.82	94.01		
Votes	267	SVM	—	92.60		
	16	GTO	—	92.40		
		GTO/SVM	—	94.00		
		2-Polyhedral Sep.	98.37	94.43	16	2
	46					
WBCP	148	Linear Sep.	70.57	57.76	22	5
	32	2-Polyhedral Sep.	81.69	61.35	23	6

Table 2. Training and testing set correctness on the kidney transplant dataset, leave-one-out validation.

Method	Average training set correctness	Average testing set correctness	Number of LP solved (average)	Running time (sec)
Linear separation	64.8	60.6	3	1
4-Polyhedral Separation	73.6	71.8	3	1

6. Conclusions

We have presented a classification method based on the construction of a piecewise linear error function. The results obtained on a number of benchmark datasets appear competitive with those described in the literature, even with those obtained by methods based on margin maximization. Of course, application of our method, which requires a significantly bigger numerical effort with respect to methods such as linear separation, appears justified only whenever such methods have a poor performance.

References

1. Elder, J., iv, and Pregibon, D., *A Statistical Perspective on Knowledge Discovery in Databases*, Advances in Knowledge Discovery and Data Mining, Edited by U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI/MIT Press, pp. 83–115, 1996.
2. Street, W. N., and Mangasarian, O. L., *Improved Generalization via Tolerant Training*, Journal of Optimization Theory and Applications, Vol. 96, pp. 259–279, 1998.
3. Srinivasan, V., and Kim, Y. H., *Credit Granting: A Comparative Analysis of Classification Procedures*, Journal of Finance, Vol. 42, pp. 665–683, 1987.
4. Charnes, A., Cooper, W. W., and Rhodes, E., *Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through*, Management Science, Vol. 27, pp. 668–687, 1981.
5. Mangasarian, O. L., Setiono, R., and Wolberg, W. H., *Pattern Recognition via Linear Programming: Theory and Application to Medical Diagnosis*, Large-Scale Numerical Optimization, Edited by T. F. Coleman and Y. Li, SIAM, Philadelphia, Pennsylvania, pp. 22–31, 1990.
6. Mangasarian, O. L., and Wolberg, W. H., *Cancer Diagnosis via Linear Programming*, SIAM News, Vol. 23, pp. 1–18, 1990.
7. Mangasarian, O. L., Street, W. N., and Wolberg, W. H., *Breast Cancer Diagnosis and Prognosis via Linear Programming*, Operations Research, Vol. 43, pp. 570–577, 1995.
8. Jurs, P. C., *Pattern Recognition Used to Investigate Multivariate Data in Analytical Chemistry*, Science, Vol. 232, pp. 1219–1224, 1986.

9. MANGASARIAN, O. L., *Linear and Nonlinear Separation of Patterns by Linear Programming*, Operations Research, Vol. 13, pp. 444–452, 1965.
10. BENNETT, K. P., and MANGASARIAN, O. L., *Robust Linear Programming Discrimination of Two Linearly Inseparable Sets*, Optimization Methods and Software, Vol. 1, pp. 23–34, 1992.
11. MANGASARIAN, O. L., *Arbitrary-Norm Separating Plane*, Operations Research Letters, Vol. 24, pp. 15–23, 1999.
12. GLOVER, F., *Improved Linear Programming Models for Discriminant Analysis*, Decision Sciences, Vol. 21, pp. 771–785, 1990.
13. GRINOLD, R. C., *Mathematical Programming Methods of Pattern Classification*, Management Science, Vol. 19, pp. 272–289, 1972.
14. BENNETT, K. P., and MANGASARIAN, O. L., *Neural Network Training via Linear Programming*, Advances in Optimization and Parallel Computing, Edited by P. M. Pardalos, North Holland, Amsterdam, Holland, pp. 56–67, 1992.
15. FIESLER, E., and BEALE, R., *Handbook of Neural Computation*, Institute of Physics Publishing and Oxford University Press, Oxford, England, 1997.
16. HAYKIN, S., *Neural Networks: A Comprehensive Foundation*, Macmillan Publishing Company, 1994.
17. HERTZ, J., KROGH, A., and PALMER, R. G., *Introduction to the Theory of Neural Computation*, Addison–Wesley, Redwood City, California, 1991.
18. MANGASARIAN, O. L., *Mathematical Programming in Neural Networks*, ORSA Journal on Computing, Vol. 5, pp. 349–360, 1995.
19. RIPLEY, B. D., *Pattern Recognition and Neural Network*, Cambridge University Press, Cambridge, Massachussets, 1996.
20. BENNETT, K. P., *On Mathematical Programming Methods and Support Vector Machines*, Advances in Kernel Methods: Support Vector Machines, Edited by A. Schoelkopf, C. Burges, and A. Smola, MIT Press, Cambridge, Massachusetts, 1999.
21. BRADLEY, P. S., and MANGASARIAN, O. L., *Feature Selection via Concave Minimization and Support Vector Machines*, Machine Learning, Proceedings of the 15th International Conference (ICML '98), Edited by J. Shavlik, Morgan Kaufmann, San Francisco, California, pp. 82–90, 1998.
22. BRADLEY, P. S., and MANGASARIAN, O. L., *Massive Data Discrimination via Linear Support Vector Machines*, Optimization Methods and Software, Vol. 13, pp. 1–10, 2000.
23. BURGES, C. J. C., *A Tutorial on Support Vector Machines for Pattern Recognition*, Data Mining and Knowledge Discovery, Vol. 2, pp. 121–167, 1998.
24. JOACHIMS, T., *Making Large-Scale SVM Learning Practical*, Advances in Kernel Methods: Support Vector Learning, Edited by B. Schoelkopf, C. Burges, and A. Smola, MIT Press, Cambridge, Massachusetts, 1999.
25. MANGASARIAN, O. L., and MUSICANT, D. R., *Successive Overrelaxation for Support Vector Machines*, IEEE Transactions on Neural Networks, Vol. 10, pp. 1032–1037, 1999.
26. MANGASARIAN, O. L., and MUSICANT, D. R., *Data Discrimination via Nonlinear Generalized Support Vector Machines*, Applications and Algorithms of Complementarity, Edited by M. C. Ferris, O. L. Mangasarian, and J. S. Pang, Kluwer Academic Publishers, Dordrecht, Holland, 2000.

27. OSUNA, E., FREUND, R., and GIROSI, F., *Support Vector Machines: Training and Applications*, Memorandum, MIT Artificial Intelligence Laboratory, 1996.
28. OSUNA, E., FREUND, R., and GIROSI, F., *An Improved Training Algorithm for Support Vector Machines*, Proceedings of IEEE NNSP'97, Amelia Island, Florida, pp. 24–26, 1997.
29. PLATT, J., *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,* Advances in Kernel Methods: Support Vector Learning, Edited by B. Schoelkopf, C. Burges, and A. Smola, MIT Press, Cambridge, Massachusetts, pp. 185–208, 1999.
30. VAPNIK, V. N., *The Nature of Statistical Learning Theory*, Springer, New York, NY, 1995.
31. BENNETT, K. P., *Decision Tree Construction via Linear Programming*, Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference, Utica, Illinois, pp. 97–101, 1992.
32. BENNETT, K. P., and MANGASARIAN, O. L., *Bilinear Separation of Two Sets in n-Space*, Computational Optimization and Applications, Vol. 2, pp. 207–227, 1993.
33. BENNETT, K. P., *Global Tree Optimization: A Nongreedy Decision-Tree Algorithm,* Computing Science and Statistics, Vol. 26, pp. 156–160, 1994.
34. BENNETT, K. P., and BLUE, J., *Optimal Decision Trees*, Mathematics Report 96-214, Rensselaer Polytechnic Institute, Troy, NY, 1996.
35. BENNETT, K. P., and BLUE, J., *A Support Vector Machine Approach to Decision Trees*, Mathematics Reports 97-100, Rensselaer Polytechnic Institute, Troy, NY, 1997.
36. BENNETT, K. P., and BREDENSTEINER, E., *Feature Minimization within Decision Trees*, Computational Optimization and Applications, Vol. 10, pp. 111–126, 1998.
37. MANGASARIAN, O. L., *Multisurface Method of Pattern Separation*, IEEE Transactions on Information Theory, Vol. 14, pp. 801–807, 1968.
38. BENNETT, K. P., and MANGASARIAN, O. L., *Bilinear Separation of Two Sets in n-Space*, Computational Optimization and Applications, Vol. 2, pp. 207–227, 1993.
39. MEGIDDO, N., *On the Complexity of Polyhedral Separability*, Discrete and Computational Geometry, Vol. 3, pp. 325–337, 1988.
40. TARDELLA, F., *Piecewise Concavity and Discrete Approaches to Continuous Minmax Problems*, Approximation and Complexity in Numerical Optimization: Continuous and Discrete Problems, Edited by P. M. Pardalos, Kluwer Academic Publishers, Dordrecht, Holland, 1999.
41. CHEN, C., and MANGASARIAN, O. L., *Hybrid Misclassification Minimization*, Advances in Computational Mathematics, Vol. 5, pp. 127–136, 1996.
42. BENNETT, K. P., and BREDENSTEINER, E. J., *A Parametric Optimization Method for Machine Learning*, INFORMS Journal on Computing, Vol. 9, pp. 311–318, 1997.
43. MANGASARIAN, O. L., *Misclassification Minimization*, Journal of Global Optimization, Vol. 5, pp. 309–323, 1994.