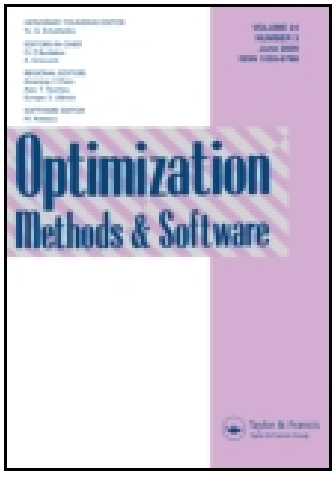

Optimization Methods and Software
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/goms20

Max-min separability

Adil M. Bagirov
[a] CIAO, School of Information Technology and Mathematical Sciences , The University of Ballarat , Victoria, 3353, Australia
Published online: 31 Jan 2007.

To cite this article: Adil M. Bagirov (2005) Max-min separability, Optimization Methods and Software, 20:2-3, 277-296, DOI: 10.1080/10556780512331318263

To link to this article: http://dx.doi.org/10.1080/10556780512331318263

Optimization Methods and Software
Vol. 20, No. 2–3, April–June 2005, 277–296

Taylor & Francis
Taylor & Francis Group

Max–min separability

ADIL M. BAGIROV*

CIAO, School of Information Technology and Mathematical Sciences,
The University of Ballarat, Victoria, 3353, Australia

(Received 8 February 2003; in final form 3 December 2003)

We consider the problem of discriminating two finite point sets in the n-dimensional space by a finite number of hyperplanes generating a piecewise linear function. If the intersection of these sets is empty, then they can be strictly separated by a max–min of linear functions. An error function is introduced. This function is nonconvex piecewise linear. We discuss an algorithm for its minimization. The results of numerical experiments using some real-world datasets are presented, which show the effectiveness of the proposed approach.

Keywords: Classification; Separability; Nonconvex optimization; Nonsmooth optimization

1. Introduction

The problems of supervised data classification arise in many areas including management science, medicine, and chemistry (see [1–3]). The aim of supervised data classification is to establish rules for the classification of some observations assuming that the classes of data are known. To find these rules, known training subsets of the given classes are used. During the last decades, many algorithms have been proposed and studied to solve data classification problems. These algorithms are mainly based on statistical, machine learning, and neural network approaches (see, for example, [4–7]).

One of the promising approaches to data classification problems is based on mathematical programing techniques. There are two main approaches to apply mathematical programing techniques for solving supervised data classification problems. The first approach is an outer approach and is based on the separation of the given training sets by means of a certain, not necessarily linear, function (see [8–19]).

The second approach is an inner approach. In this approach, the given training sets are approximated by cluster centers. The new data vectors are assigned to the closest cluster and correspondingly to the set that contains this cluster (see [20,21]).

*Corresponding author. Email: a.bagirov@ballarat.edu.au

Optimization Methods and Software
ISSN 1055-6788 print/ISSN 1029-4937 online © 2005 Taylor & Francis Group Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/10556780512331318263

In this article we develop an algorithm based on an outer approach. We will consider the problem of the separation of two sets, that is, we assume that the dataset under consideration contains two classes. Thus, we are given two nonempty finite point sets A and B in $\mathbb{R}^n$. If the convex hull of these sets do not intersect, that is, $\text{co}\, A \cap \text{co}\, B = \emptyset$, then they are linearly separable and there exists a hyperplane that separates these two sets. Linear programing techniques can be used to construct such a hyperplane. If the convex hulls of A and B intersect, then linear programing techniques can be applied to obtain a hyperplane that minimizes some misclassification measure. Algorithms based on such an approach are developed in refs. [10,11,15,16].

In the paper [8], the concept of polyhedral separability was introduced. In this article, consider the case when $\text{co}\, A \cap B = \emptyset$. The set A is approximated by a polyhedral set. It is proved that the sets A and B are h-polyhedrally separable for some $h \leq |B|$, where $|B|$ is the cardinality of the set B. Thus, in this case, the sets A and B can be separated by a certain piecewise linear function. The authors introduce an error function that is nonconvex piecewise linear function. An algorithm for minimizing this function is proposed. The problem of the calculation of the descent direction in this algorithm is reduced to a certain linear programing problem.

In this article we introduce the notion of a max–min separability, which can be considered as a generalization of the polyhedral separability. If the sets A and B are disjoint, then they are max–min separable. We describe an error function for this case and discuss an algorithm for its minimization.

It should be noted that the most general nonsmooth classifiers have been introduced in the paper [22]. Classifiers based on linear, polyhedral and max–min separability can be considered as particular cases of those classifiers.

Some numerical experiments using real-world datasets have been carried out. We present their results. These results show that classification algorithms based on the notion of the max–min separability give better and sometimes considerably better results than those based on the linear and polyhedral separability.

The structure of this article is as follows. Section 2 provides some preliminaries. The definition and some results related to the max–min separability are given in section 3. The error function is described and studied in section 4. An algorithm for minimizing the error function is discussed in section 5. Results of numerical experiments are presented in section 6. Section 7 concludes the article.

2. Preliminaries

2.1 *Linear separability*

Let A and B be given sets containing m and p n-dimensional vectors, respectively:

$$A = \{a^1, \ldots, a^m\}, \quad a^i \in \mathbb{R}^n, \quad i = 1, \ldots, m,$$
$$B = \{b^1, \ldots, b^p\}, \quad b^j \in \mathbb{R}^n, \quad j = 1, \ldots, p.$$

An algorithm for finding a hyperplane $\{x, y\}$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^1$ separating these two sets is described in ref. [11]. This hyperplane is the solution to the following problem:

$$\begin{aligned} &\text{minimize} \quad f(x, y) \\ &\text{subject to} \quad (x, y) \in \mathbb{R}^{n+1}, \end{aligned} \tag{1}$$

where

$$f(x, y) = \frac{1}{m}\sum_{i=1}^{m}\max\left(0, \langle x, a^i\rangle - y + 1\right) + \frac{1}{p}\sum_{j=1}^{p}\max\left(0, -\langle x, b^j\rangle + y + 1\right),$$

where $\langle\cdot,\cdot\rangle$ stands for a scalar product in $\mathbb{R}^n$. It is shown in ref. [11] that the problem (1) is equivalent to the following linear program:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{m}\sum_{i=1}^{m} t_i + \frac{1}{p}\sum_{j=1}^{p} z_j \\ \text{subject to} \quad & t_i \geq \langle x, a^i\rangle - y + 1, \quad i = 1, \ldots, m, \\ & z_j \geq -\langle x, b^j\rangle + y + 1, \quad j = 1, \ldots, p, \\ & t \geq 0, z \geq 0, \end{aligned}$$

where t_i is nonnegative and represents the error for the point $a^i \in A$, and z_j is nonnegative and represents the error for the point $b^j \in B$.

The sets A and B are linearly separable if and only if $f^* = f(x^*, y_*) = 0$, where (x^*, y_*) is the solution to the problem (1). It is proved that the trivial solution $x = 0$ cannot occur.

2.2 *Polyhedral separability*

The concept of h-polyhedral separability was developed in ref. [8]. The sets A and B are h-polyhedrally separable if there exists a set of h hyperplanes $\{x^i, y_i\}$, with

$$x^i \in \mathbb{R}^n, \quad y_i \in \mathbb{R}^1, \quad i = 1, \ldots, h$$

such that

(i) for any $j = 1, \ldots, m$ and $i = 1, \ldots, h$

$$\langle x^i, a^j\rangle - y_i < 0;$$

(ii) for any $k = 1, \ldots, p$ there exists at least one $i \in \{1, \ldots, h\}$ so that

$$\langle x^i, b^k\rangle - y_i > 0.$$

It is proved in ref. [8] that the sets A and B are h-polyhedrally separable, for some $h \leq p$ if and only if

$$\operatorname{co} A \bigcap B = \emptyset.$$

The problem of polyhedral separability of the sets A and B is reduced to the following problem:

$$\begin{aligned} \text{minimize} \quad & f(x, y) \\ \text{subject to} \quad & (x, y) \in \mathbb{R}^{(n+1)\times h}, \end{aligned} \tag{2}$$

where

$$\begin{aligned} f(x, y) = \frac{1}{m}\sum_{j=1}^{m}\max\left[0, \max_{1\leq i\leq h}\left\{\langle x^i, a^j\rangle - y_i + 1\right\}\right] \\ + \frac{1}{p}\sum_{k=1}^{p}\max\left[0, \min_{1\leq i\leq h}\left\{-\langle x^i, b^k\rangle + y_i + 1\right\}\right] \end{aligned}$$

is an error function. Note that this function is a nonconvex piecewise linear function. It is proved that $x^i = 0,\ i = 1, \ldots, h$ cannot be the optimal solution. Let $\{\bar{x}^i, \bar{y}_i\}, i = 1, \ldots, h$, be a global solution to the problem (2). The sets A and B are h-polyhedrally separable if and only if $f(\bar{x}, \bar{y}) = 0$. If there exists a nonempty set $\bar{I} \subset \{1, \ldots, h\}$ such that $x^i = 0,\ i \in \bar{I}$, then the sets A and B are $(h - |\bar{I}|)$-polyhedrally separable. In ref. [8], an algorithm for solving problem (2) is developed. The calculation of the descent direction at each iteration of this algorithm is reduced to a certain linear programing problem.

3. Max–min separability

In this section we develop the concept of max–min separability. Let $H = \{h_1, \ldots, h_l\}$, where $h_j = \{x^j, y_j\},\ j = 1, \ldots, l$, with $x^j \in \mathbb{R}^n,\ y_j \in \mathbb{R}^1$, be a finite set of hyperplanes. Let $J = \{1. \ldots, l\}$. Consider any partition of this set $J^r = \{J_1, \ldots, J_r\}$ such that

$$J_k \neq \emptyset,\ \ k = 1, \ldots, r, \qquad J_k \bigcap J_j = \emptyset, \quad \bigcup_{k=1}^{r} J_k = J.$$

Let $I = \{1, \ldots, r\}$. A particular partition $J^r = \{J_1, \ldots, J_r\}$ of the set J defines the following max–min-type function:

$$\varphi(z) = \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, z \rangle - y_j \right\}, \quad z \in \mathbb{R}^n. \tag{3}$$

Let $A, B \subset \mathbb{R}^n$ be given disjoint sets, that is, $A \cap B = \emptyset$.

DEFINITION 1 *The sets A and B are max–min separable if there exist a finite number of hyperplanes* $\{x^j, y_j\}$ *with* $x^j \in \mathbb{R}^n,\ y_j \in \mathbb{R}^1,\ j \in J = \{1, \ldots, l\}$ *and a partition* $J^r = \{J_1, \ldots, J_r\}$ *of the set J such that*

(i) *for all* $i \in I$ *and* $a \in A$

$$\min_{j \in J_i} \left\{ \langle x^j, a \rangle - y_j \right\} < 0;$$

(ii) *for any* $b \in B$, *there exists at least one* $i \in I$ *such that*

$$\min_{j \in J_i} \left\{ \langle x^j, b \rangle - y_j \right\} > 0.$$

Remark 1 It follows from Definition 1 that if the sets A and B are max–min separable, then $\varphi(a) < 0$ for any $a \in A$ and $\varphi(b) > 0$ for any $b \in B$, where the function φ is defined by equation (3). Thus, the sets A and B can be separated by a function represented as a max–min of linear functions.

Remark 2 Linear and polyhedral separability can be considered as particular cases of the max–min separability. If $I = \{1\}$ and $J_1 = \{1\}$, then we have the linear separability and if $I = \{1, \ldots, h\}$ and $J_i = \{i\},\ i \in I$, we obtain h-polyhedral separability.

PROPOSITION 1 *The sets A and B are max–min separable if and only if there exists a set of hyperplanes* $\{x^j, y_j\}$ *with* $x^j \in \mathbb{R}^n,\ y_j \in \mathbb{R}^1,\ j \in J$, *and a partition* $J^r = \{J_1, \ldots, J_r\}$ *of the set J such that*

(i)

$$\min_{j\in J_i}\left\{\langle x^j, a\rangle - y_j\right\} \le -1 \quad \text{for all } i \in I \text{ and } a \in A;$$

(ii) *for any $b \in B$, there exists at least one $i \in I$ such that*

$$\min_{j\in J_i}\left\{\langle x^j, b\rangle - y_j\right\} \ge 1.$$

Proof Sufficiency is straightforward.

Necessity: As A and B are max–min separable there exists a set of hyperplanes $\{\bar{x}^j, \bar{y}_j\}$ with $\bar{x}^j \in \mathbb{R}^n,\ \bar{y}_j \in \mathbb{R}^1,\ j \in J$, a partition J^r of the set J and numbers $\delta_1 > 0,\ \delta_2 > 0$ such that

$$\max_{a\in A}\max_{i\in I}\min_{j\in J_i}\left\{\langle \bar{x}^j, a\rangle - \bar{y}_j\right\} = -\delta_1$$

and

$$\min_{b\in B}\max_{i\in I}\min_{j\in J_i}\left\{\langle \bar{x}^j, b\rangle - \bar{y}_j\right\} = \delta_2.$$

We put $\delta = \min\{\delta_1, \delta_2\} > 0$. Then, we have

$$\max_{i\in I}\min_{j\in J_i}\left\{\langle \bar{x}^j, a\rangle - \bar{y}_j\right\} \le -\delta, \quad \forall a \in A, \tag{4}$$

$$\max_{i\in I}\min_{j\in J_i}\left\{\langle \bar{x}^j, b\rangle - \bar{y}_j\right\} \ge \delta, \quad \forall b \in B. \tag{5}$$

We consider the new set of hyperplanes $\{x^j, y_j\}$ with $x^j \in \mathbb{R}^n$, $y_j \in \mathbb{R}^1$, $j \in J$, defined as follows:

$$x^j = \frac{\bar{x}^j}{\delta}, \quad j \in J,$$

$$y^j = \frac{\bar{y}^j}{\delta}, \quad j \in J.$$

Then, it follows from equations (4) and (5) that

$$\max_{i\in I}\min_{j\in J_i}\left\{\langle x^j, a\rangle - y_j\right\} \le -1, \quad \forall a \in A,$$

$$\max_{i\in I}\min_{j\in J_i}\left\{\langle x^j, b\rangle - y_j\right\} \ge 1, \quad \forall b \in B,$$

which completes the proof. ■

PROPOSITION 2 *The sets A and B are max–min separable if and only if there exists a piecewise linear function separating them.*

Proof As max–min of linear functions is a piecewise linear function, the necessity is straightforward.

Sufficiency: It is known that any piecewise linear function can be represented as a max–min of linear functions of the form (3) (see [23]). Then, we get that there exists max–min of linear functions that separates the sets A and B which in its turn means that these sets are max–min separable. ■

Now, we consider particular cases where max–min separability takes place.

PROPOSITION 3 *Assume that the set A can be represented as a union of sets* $A_i,\ i = 1, \ldots, q$:

$$A = \bigcup_{i=1}^{q} A_i,$$

and for any $i = 1, \ldots, q$

$$B \bigcap \operatorname{co} A_i = \emptyset. \tag{6}$$

Then, the sets A and B are max–min separable.

Proof It follows from equation (6) that $b \notin \operatorname{co} A_i$ for all $b \in B$ and $i \in \{1, \ldots, q\}$. Then, for each $b \in B$ and $i \in \{1, \ldots, q\}$, there exists a hyperplane $\{x^i(b),\ y_i(b)\}$ separating b from the set $\operatorname{co} A_i$, that is

$$\langle x^i(b), b\rangle - y_i(b) > 0,$$

$$\langle x^i(b), a\rangle - y_i(b) < 0, \quad \forall a \in \operatorname{co} A_i,\ \ i = 1, \ldots, q.$$

Then, we have

$$\min_{i=1,\ldots,q} \left\{\langle x^i(b), b\rangle - y_i(b)\right\} > 0$$

and

$$\min_{i=1,\ldots,q} \left\{\langle x^i(b), a\rangle - y_i(b)\right\} < 0, \quad \forall a \in A.$$

Thus, we obtain that for any $b^j \in B,\ j = 1, \ldots, p$, there exists a set of q hyperplanes $\{x^i(b^j),\ y_i(b^j)\},\ i = 1, \ldots, q$, such that

$$\min_{i=1,\ldots,q} \left\{\langle x^i(b^j), b^j\rangle - y_i(b^j)\right\} > 0 \tag{7}$$

and

$$\min_{i=1,\ldots,q} \left\{\langle x^i(b^j), a\rangle - y_i(b^j)\right\} < 0, \quad \forall a \in A. \tag{8}$$

Consequently, we have pq hyperplanes

$$\left\{x^i(b^j),\ y_i(b^j)\right\}, \quad i = 1, \ldots, q,\ \ j = 1, \ldots, p.$$

The set of these hyperplanes can be rewritten as follows:

$$H = \{h_1, \ldots, h_l\}, \quad h_{i+(j-1)q} = \left\{x^i(b^j),\ y_i(b^j)\right\},\ \ i = 1, \ldots, q,\ \ j = 1, \ldots, p,$$
$$l = pq.$$

Let $J = \{1, \ldots, l\},\ I = \{1, \ldots, p\}$ and

$$\bar{x}^{i+(j-1)q} = x^i(b^j), \quad \bar{y}_{i+(j-1)q} = y_i(b^j), \quad i = 1, \ldots, q,\ \ j = 1, \ldots, p.$$

Consider the following partition of the set J:

$$J^p = \left\{J_1, \ldots, J_p\right\}, \quad J_k = \{(k-1)q + 1, \ldots, kq\},\ \ k = 1, \ldots, p.$$

It follows from equations (7) and (8) that for all $k \in I$ and $a \in A$

$$\min_{j \in J_k} \left\{\langle \bar{x}^j, a\rangle - \bar{y}_j\right\} < 0$$

and for any $b \in B$, there exists at least one $k \in I$ such that

$$\min_{j \in J_k} \left\{\langle \bar{x}^j, b\rangle - \bar{y}_j\right\} > 0,$$

which means that the sets A and B are max–min separable. ■

COROLLARY 1 *The sets A and B are max–min separable if and only if they are disjoint: $A \cap B = \emptyset$.*

Proof (Necessity is straightforward.)

Sufficiency: The set A can be represented as a union of its own points. As the sets A and B are disjoint, the condition (6) is satisfied. Then, the proof of this corollary follows from Proposition 3. ■

Remark 3 For many cases, the number of hyper-planes necessary for the max–min separation of the sets A and B are significantly less than pq, that is many hyperplanes are redundant and this corresponds to some hyperplane that separates more than just one point of B from the sets $\operatorname{co} A_i,\ i = 1, \ldots, q$. The result from Proposition 4 confirms it.

PROPOSITION 4 *Assume that the set A can be represented as a union of sets $A_i,\ i = 1, \ldots, q$, and the set B as a union of sets $B_j,\ j = 1, \ldots, d$, such that*

$$A = \bigcup_{i=1}^{q} A_i, \quad B = \bigcup_{j=1}^{d} B_j$$

and

$$\operatorname{co} A_i \bigcap \operatorname{co} B_j = \emptyset \quad \text{for all } i = 1, \ldots, q, \quad j = 1, \ldots, d. \tag{9}$$

Then, the sets A and B are max–min separable with no more than $q \cdot d$ hyperplanes.

Proof Let $i \in \{1, \ldots, q\}$ and $j \in \{1, \ldots, d\}$ be any fixed indices. As $\operatorname{co} A_i \bigcap \operatorname{co} B_j = \emptyset$, there exists a hyperplane $\{x^{ij}, y_{ij}\}$ with $x^{ij} \in \mathbb{R}^n,\ y_{ij} \in \mathbb{R}^1$ such that

$$\langle x^{ij}, a\rangle - y_{ij} < 0 \quad \forall a \in \operatorname{co} A_i$$

and

$$\langle x^{ij}, b\rangle - y_{ij} > 0 \quad \forall b \in \operatorname{co} B_j.$$

Consequently, for any $j \in \{1, \ldots, d\}$ there exists a set of hyperplanes $\{x^{ij}, y_{ij}\},\ i = 1, \ldots, q$, such that

$$\min_{i=1,\ldots,q} \langle x^{ij}, b\rangle - y_{ij} > 0, \quad \forall b \in B_j \tag{10}$$

and

$$\min_{i=1,\ldots,q} \langle x^{ij}, a\rangle - y_{ij} < 0, \quad \forall a \in A. \tag{11}$$

Thus, we get a system of $l = dq$ hyperplanes:

$$H = \{h_1, \ldots, h_l\},$$

where $h_{i+(j-1)q} = \left\{x^{ij}, y_{ij}\right\}, \quad i = 1, \ldots, q, \quad j = 1, \ldots, d.$ Let $J = \{1, \ldots, l\}, \quad I = \{1, \ldots, d\}$, and

$$\bar{x}^{i+(j-1)q} = x^{ij}, \quad \bar{y}_{i+(j-1)q} = y_{ij}, \quad i = 1, \ldots, q, \quad j = 1, \ldots, d.$$

Consider the following partition of the set J:

$$J^d = \{J_1, \ldots, J_d\}, \quad J_k = \{(k-1)q + 1, \ldots, kq\}, \quad k = 1, \ldots, d.$$

It follows from equations (10) and (11) that for all $k \in I$ and $a \in A$

$$\min_{j \in J_k} \left\{ \langle \bar{x}^j, a \rangle - \bar{y}_j \right\} < 0$$

and for any $b \in B$, there exists at least one $k \in I$ such that

$$\min_{j \in J_k} \left\{ \langle \bar{x}^j, b \rangle - \bar{y}_j \right\} > 0,$$

that is, the sets A and B are max–min separable with at most $q \cdot d$ hyperplanes. ■

Remark 4 One can expect that the numbers q and d are not large in many situations. So, in these cases, the number of hyperplanes necessary for the max–min separation of the sets under consideration is not large.

4. Error function

In this section, we introduce an error function and establish one more criterion for max–min separability.

Given any set of hyperplanes $\{x^j, y_j\}$, $j \in J = \{1, \ldots, l\}$, with $x^j \in \mathbb{R}^n$, $y_j \in \mathbb{R}^1$, and a partition $J^r = \{J_1, \ldots, J_r\}$ of the set J, we say that a point $a \in A$ is well classified if the following condition is satisfied:

$$\max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, a \rangle - y_j \right\} + 1 \leq 0.$$

Thus, we can define the classification error for a point $a \in A$ as follows:

$$\max \left[0, \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \right]. \tag{12}$$

To a well-classified point this error is zero.

Analogously, a point $b \in B$ is said to be well classified if the following condition is satisfied:

$$\min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^j, b \rangle + y_j \right\} + 1 \leq 0.$$

Then, the classification error for a point $b \in B$ can be written as

$$\max \left[0, \min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \right]. \tag{13}$$

Thus, an averaged classification error function can be defined as

$$f(x, y) = \left(\frac{1}{m} \right) \sum_{k=1}^{m} \max \left[0, \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, a^k \rangle - y_j + 1 \right\} \right] + \left(\frac{1}{p} \right) \sum_{t=1}^{p} \max \left[0, \min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^j, b^t \rangle + y_j + 1 \right\} \right], \tag{14}$$

where $x = (x^1, \ldots, x^l) \in \mathbb{R}^{l \times n}$ and $y = (y_1, \ldots, y_l) \in \mathbb{R}^l$. It is clear that $f(x, y) \geq 0$ for all $x \in \mathbb{R}^{l \times n}$ and $y \in \mathbb{R}^l$.

Then, the problem of max–min separability is reduced to the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad f(x, y) \\ &\text{subject to} \quad (x, y) \in \mathbb{R}^{(n+1)\times l}. \end{aligned} \tag{15}$$

PROPOSITION 5 *The sets A and B are max–min separable if and only if there exist a set of hyperplanes* $\{x^j, y_j\}$, $j \in J = \{1, \ldots, l\}$, *and a partition* $J^r = \{J_1, \ldots, J_r\}$ *of the set J such that* $f(x, y) = 0$.

Proof Necessity: Assume that the sets A and B are max–min separable. Then, it follows from Proposition 1 that there exists a set of hyperplanes $\{x^j, y_j\}$, $j \in J$, and a partition $J^r = \{J_1, \ldots, J_r\}$ of the set J such that

$$\min_{j\in J_i}\{\langle x^j, a\rangle - y_j\} \le -1, \quad \forall a \in A, \ \ i \in I = \{1, \ldots, r\} \tag{16}$$

and for any $b \in B$, there exists at least one $t \in I$ such that

$$\min_{j\in J_t}\{\langle x^j, b\rangle - y_j\} \ge 1. \tag{17}$$

Consequently, we have

$$\max_{i\in I} \min_{j\in J_i}\{\langle x^j, a\rangle - y_j + 1\} \le 0, \quad \forall a \in A,$$

$$\min_{i\in I} \max_{j\in J_i}\{-\langle x^j, b\rangle + y_j + 1\} \le 0, \quad \forall b \in B.$$

Then, from the definition of the error function, we obtain that $f(x, y) = 0$.

Sufficiency: Assume that there exists a set of hyperplanes $\{x^j, y_j\}$, $j \in J = \{1, \ldots, l\}$, and a partition $J^r = \{J_1, \ldots, J_r\}$ of the set J such that $f(x, y) = 0$. Then, from the definition of the error function f, we immediately get that the inequalities (16) and (17) are satisfied; that is, the sets A and B are max–min separable. ■

PROPOSITION 6 *Assume that the sets A and B are max–min separable with a set of hyperplanes* $\{x^j, y_j\}$, $j \in J = \{1, \ldots, l\}$, *and a partition* $J^r = \{J_1, \ldots, J_r\}$ *of the set J. Then*

(i) $x^j = 0$, $j \in J$, *cannot be an optimal solution to the problem* (15);
(ii) *if for any* $t \in I$, *there exists at least one* $b \in B$ *such that*

$$\max_{j\in J_t}\left\{-\langle x^j, b\rangle + y_j + 1\right\} = \min_{i\in I}\max_{j\in J_i}\left\{-\langle x^j, b\rangle + y_j + 1\right\}, \tag{18}$$

and if there exists $\tilde{J} = \{\tilde{J}_1, \ldots, \tilde{J}_r\}$ *such that* $\tilde{J}_t \subset J_t$, $\forall t \in I$, $\tilde{J}_t$ *is nonempty at least for one* $t \in I$ *and* $x^j = 0$ *for any* $j \in \tilde{J}_t$, $t \in I$, *then the sets A and B are max–min separable with a set of hyperplanes* $\{x^j, y_j\}$, $j \in J^0$ *and a partition* $\bar{J} = \{\bar{J}_1, \ldots, \bar{J}_r\}$ *of the set* J^0 *where*

$$\bar{J}_t = J_t \setminus \tilde{J}_t, \ \ t \in I \quad \text{and} \quad J^0 = \bigcup_{i=1}^{r} \bar{J}_i.$$

Proof

(i) As the sets A and B are max–min separable, we get from Proposition 5 that $f(x, y) = 0$. If $x^j = 0,\ j \in J$, then it follows from equation (14) that for any $y \in \mathbb{R}^l$

$$f(0, y) = \left(\frac{1}{m}\right)\sum_{k=1}^{m} \max\left[0, \max_{i\in I}\min_{j\in J_i}\{-y_j + 1\}\right] + \left(\frac{1}{p}\right)\sum_{t=1}^{p} \max\left[0, \min_{i\in I}\max_{j\in J_i}\{y_j + 1\}\right].$$

We denote

$$R = \max_{i\in I}\min_{j\in J_i}\{-y_j\}.$$

Then, we have

$$\min_{i\in I}\max_{j\in J_i} y_j = -\max_{i\in I}\min_{j\in J_i}\{-y_j\} = -R.$$

Thus

$$f(0, y) = \max[0, R + 1] + \max[0, -R + 1].$$

It is clear that

$$\max[0, R + 1] + \max[0, -R + 1] = \begin{cases} -R + 1 & \text{if } R \le -1, \\ 2 & \text{if } -1 < R < 1, \\ R + 1 & \text{if } R \ge 1. \end{cases}$$

Thus, for any $y \in \mathbb{R}^l$

$$f(0, y) \ge 2.$$

On the other side, $f(x, y) = 0$ for the optimal solution (x, y), that is, $x^j = 0,\ j \in J$, cannot be the optimal solution.

(ii) Consider the following sets:

$$I^1 = \{i \in I: \bar{J}_i \ne \emptyset\}, \quad I^2 = \{i \in I : \tilde{J}_i \ne \emptyset\}, \quad I^3 = I^1 \bigcap I^2.$$

It is clear that $\tilde{J}_i = \emptyset$ for any $i \in I^1 \setminus I^3$ and $\bar{J}_i = \emptyset$ for any $i \in I^2 \setminus I^3$.

It follows from the definition of the error function that

$$0 = f(x, y) = \frac{1}{m}\sum_{k=1}^{m} \max\left[0, \max_{i\in I}\min_{j\in J_i}\left\{\langle x^j, a^k\rangle - y_j + 1\right\}\right] + \frac{1}{p}\sum_{t=1}^{p} \max\left[0, \min_{i\in I}\max_{j\in J_i}\left\{-\langle x^j, b^t\rangle + y_j + 1\right\}\right].$$

As the function f is nonnegative we obtain

$$\max_{i\in I}\min_{j\in J_i}\left\{\langle x^j, a\rangle - y_j + 1\right\} \le 0, \quad \forall a \in A, \tag{19}$$

$$\min_{i\in I}\max_{j\in J_i}\left\{-\langle x^j, b\rangle + y_j + 1\right\} \le 0, \quad \forall b \in B. \tag{20}$$

It follows from equations (18) and (20) that for any $i \in I^2$ there exists a point $b \in B$ such that

$$\max_{j \in J_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \leq 0. \tag{21}$$

If $i \in I^3 \subset I^2$, then we have

$$0 \geq \max_{j \in J_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} = \max \left\{ \max_{j \in \bar{J}_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\}, \max_{i \in \bar{J}_i} \left\{ y_j + 1 \right\} \right\},$$

which means that

$$\max_{j \in \bar{J}_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \leq 0 \tag{22}$$

and

$$\max_{j \in \bar{J}_i} \left\{ y_j + 1 \right\} \leq 0. \tag{23}$$

If $i \in I^2 \setminus I^3$ then from equation (21) we obtain

$$0 \geq \max_{j \in J_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} = \max_{j \in \bar{J}_i} \left\{ y_j + 1 \right\}.$$

Thus, we get that for all $i \in I^2$ the inequality (23) is true. Equation (23) can be rewritten as follows:

$$\max_{j \in \bar{J}_i} y_j \leq -1, \quad \forall i \in I^2. \tag{24}$$

Consequently, for any $i \in I^2$

$$\min_{j \in \bar{J}_i} \left\{ -y_j + 1 \right\} = -\max_{j \in \bar{J}_i} y_j + 1 \geq 2. \tag{25}$$

It follows from equation (19) that for any $i \in I$ and $a \in A$

$$\min_{j \in J_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \leq 0. \tag{26}$$

Then, for any $i \in I^3$, we have

$$0 \geq \min_{j \in J_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} = \min \left\{ \min_{j \in \bar{J}_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\}, \min_{j \in \bar{J}_i} \left\{ -y_j + 1 \right\} \right\}.$$

Taking into account equation (25), we get that for any $i \in I^3$ and $a \in A$

$$\min_{j \in \bar{J}_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \leq 0. \tag{27}$$

If $i \in I^2 \setminus I^3$, then it follows from equation (26) that

$$\min_{j \in \bar{J}_i} \left\{ -y_j + 1 \right\} \leq 0,$$

which contradicts equation (25). Thus, we obtain that $I^2 \setminus I^3 \neq \emptyset$ cannot occur, $I^2 \subset I^1$ and $I^3 = I^2$. It is clear that $\bar{J}_i = J_i$ for any $i \in I^1 \setminus I^2$. Then, it follows from equation (19)

that for any $i \in I^1 \setminus I^2$ and $a \in A$

$$\min_{j \in \bar{J}_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \leq 0. \tag{28}$$

From equations (27) and (28), we can conclude that for any $i \in I$ and $a \in A$

$$\min_{j \in \bar{J}_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \leq 0. \tag{29}$$

It follows from equation (20) that for any $b \in B$ there exists at least one $i \in I$ such that

$$\max_{j \in J_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \leq 0.$$

Then, from expression

$$\max_{j \in J_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} = \max \left\{ \max_{j \in \bar{J}_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\}, \max_{i \in \bar{J}_i} \left\{ y_j + 1 \right\} \right\},$$

we get that for any $b \in B$, there exists at least one $i \in I$ such that

$$\max_{j \in \bar{J}_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \leq 0. \tag{30}$$

Thus it follows from equations (29) and (30) that the sets A and B are max–min separable with the set of hyperplanes $\{x^j, y_j\}$, $j \in J^0$, and a partition $\bar{J}$ of the set J^0. ■

5. Minimization of the error function

In this section, we discuss an algorithm for solving problem (15). The objective function f in this problem has the following form:

$$f(x, y) = f_1(x, y) + f_2(x, y),$$

where

$$f_1(x, y) = \frac{1}{m} \sum_{k=1}^{m} \max \left[0, \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, a^k \rangle - y_j + 1 \right\} \right], \tag{31}$$

$$f_2(x, y) = \frac{1}{p} \sum_{t=1}^{p} \max \left[0, \min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^j, b^t \rangle + y_j + 1 \right\} \right]. \tag{32}$$

The problem (15) is a global optimization problem. However, the number of variables in this problem is large and the global optimization methods cannot be directly applied to solve it. Therefore, we will discuss algorithms for finding local minima of the function f.

The functions f_1 and f_2 are nonconvex piecewise linear. These functions are Lipschitz continuous and consequently subdifferentiable in sense of Clarke [24,25]. Moreover, both functions are semismooth (for the definition of the semismooth functions, see [26]. Therefore,

the function f is also subdifferentiable. The function f_1 contains the following max–min functions:

$$\varphi_{1k}(x, y) = \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, a^k \rangle - y_j + 1 \right\}, \quad k = 1, \ldots, m$$

and the function f_2 contains the following min-max functions:

$$\varphi_{2t}(x, y) = \min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^j, b^t \rangle + y_j + 1 \right\}, \quad t = 1, \ldots, p.$$

The differential properties of max–min functions are studied, for example, in refs. [27,28]. The functions $\varphi_{1k},\ k = 1, \ldots, m$, and $\varphi_{2t},\ t = 1, \ldots, p$, are not regular (for the definition of the regular functions, see [24]. Then, the functions, $f_1,\ f_2$, and consequently the function f are also not regular. Therefore, the calculation of subgradients of the function f is a difficult task. This implies that the methods of nonsmooth optimization, which use subgradients at each iteration, seem not to be effective for solving the problem (15).

In the paper [29], optimization problems with twice continuously differentiable objective functions and max–min constraints were considered and these problems were converted into problems with smooth objective and constraint functions. However, this approach cannot be applied to the problem (15).

Direct search methods of optimization can be used for solving the problem (15). Among such methods, we mention here two widely used methods: Powell's method [30], which is based on a quadratic approximation of the objective function, and Nelder–Mead's simplex method [31]. As was mentioned in ref. [30], Powell's method performs well when the number of variables is less than 20. For the simplex method this number is even smaller. Moreover, both methods are effective when the objective function is smooth. However, in the max–min separability problem, the number of variables is $(n + 1) \times l$ which in many cases is greater than 20, and the objective function in this problem is a quite complicated nonsmooth function.

In this article we use the discrete gradient method to solve the problem (15). The description of this method can be found in refs. [32,33] (see also [34]. The discrete gradient method can be considered as a version of the bundle method [35–37]. In this method, subgradients of the objective function are replaced by its discrete gradients.

The discrete gradient method uses only values of the objective function. It should be noted that the calculation of the objective function in the problem (15) can be expensive. We will show that the use of the discrete gradient method allows one to significantly reduce the number of the objective function evaluations. We will not describe the discrete gradient method in this article, however, we will give a definition of the discrete gradient in order to demonstrate that its computation in the problem (15) can be simplified.

5.1 *Definition of the discrete gradient*

Let F be a locally Lipschitz continuous function defined on $\mathbb{R}^n$. Let

$$S_1 = \{g \in \mathbb{R}^n : \|g\| = 1\}, \quad G = \{e \in \mathbb{R}^n : e = (e_1, \ldots, e_n),\ |e_t| = 1,\ t = 1, \ldots, n\},$$

$$P = \{z(\lambda) : z(\lambda) \in \mathbb{R}^1,\ z(\lambda) > 0, \lambda > 0, \lambda^{-1} z(\lambda) \longrightarrow 0, \lambda \longrightarrow 0\},$$

$$U(g, \alpha) = \{u \in \{1, \ldots, n\} : |g_u| \geq \alpha\},$$

where $\alpha \in (0, n^{-1/2}]$ is a fixed number.

Here, S_1 is the unit sphere, G is a set of vertices of the unit hypercube in $\mathbb{R}^n$, and P is a set of univariate positive infinitesimal functions.

We define operators $H_u^t\colon \mathbb{R}^n \to \mathbb{R}^n$ for $u = 1, \ldots, n,\ t = 0, \ldots, n$, by the formula

$$H_u^t g = \begin{cases} (g_1, \ldots, g_t, 0, \ldots, 0) & \text{if } t < u, \\ (g_1, \ldots, g_{u-1}, 0, g_{u+1}, \ldots, g_t, 0, \ldots, 0) & \text{if } t \geq u. \end{cases} \tag{33}$$

We can see that

$$H_u^t g - H_u^{t-1} g = \begin{cases} (0, \ldots, 0, g_t, 0, \ldots, 0) & \text{if } t = 1, \ldots, n,\ \ t \neq u, \\ 0 & \text{if } t = u. \end{cases} \tag{34}$$

Let $e(\beta) = (\beta e_1, \beta^2 e_2, \ldots, \beta^n e_n)$, where $\beta \in (0, 1]$. For $x \in \mathbb{R}^n$, we consider vectors

$$x_u^t(g, e, z, \lambda, \beta) = x + \lambda g - z(\lambda) H_u^t e(\beta), \tag{35}$$

where $g \in S_1,\ e \in G,\ i \in U(g, \alpha),\ z \in P,\ \lambda > 0,\ t = 0, \ldots, n,\ t \neq u$.

It follows from equation (34) that

$$x_u^{t-1}(g, e, z, \lambda, \beta) - x_u^t(g, e, z, \lambda, \beta) = \begin{cases} (0, \ldots, 0, z(\lambda) e_t(\beta), 0, \ldots, 0) & \text{if } t = 1, \ldots, n, t \neq u, \\ 0 & \text{if } t = u. \end{cases} \tag{36}$$

It is clear that $H_u^0 g = 0$ and $x_u^0(g, e, z, \lambda, \beta) = x + \lambda g$ for all $u \in U(g, \alpha)$.

DEFINITION 2 (see [38]) *The discrete gradient of the function F at the point $x \in \mathbb{R}^n$ is the vector $\Gamma^u(x, g, e, z, \lambda, \beta) = (\Gamma_1^u, \ldots, \Gamma_n^u) \in \mathbb{R}^n$, $g \in S_1$, $u \in U(g, \alpha)$, with the following coordinates*:

$$\Gamma_t^u = [z(\lambda) e_t(\beta)]^{-1} \left[F(x_u^{t-1}(g, e, z, \lambda, \beta)) - F(x_u^t(g, e, z, \lambda, \beta)) \right], \quad t = 1, \ldots, n,\ \ t \neq u,$$

$$\Gamma_u^u = (\lambda g_u)^{-1} \left[F(x_u^n(g, e, z, \lambda, \beta)) - F(x) - \sum_{t=1, t \neq u}^{n} \Gamma_t^u (\lambda g_t - z(\lambda) e_t(\beta)) \right].$$

Remark 5 The discrete gradient is defined with respect to a given direction $g \in S_1$ and a given component $u \in U(g, \alpha)$ of g for fixed α. We can see that for the calculation of one discrete gradient, we have to calculate $(n + 1)$ values of a function F, at the point x and at the points $x_u^t(g, e, z, \lambda, \beta),\ t = 0, \ldots, n,\ t \neq u$. For the calculation of another discrete gradient at the same point with respect to any other direction $g^1 \in S_1$, we have to calculate this function n times, because we have already calculated F at the point x.

Now, let us return to the objective function f of the problem (15). This function depends on $(n + 1)l$ variables, where l is the number of hyperplanes. The function f_1 defined in equation (31) contains max–min functions φ_{1k} which in their turn contain the following linear functions:

$$\psi_{1jk}(x, y) = \langle x^j, a^k \rangle - y_j + 1, \quad j \in J_i,\ i \in I.$$

We can see that for each $j \in J_i,\ i \in I$, variables $\{x^j, y_j\}$ appear in the function ψ_{1jk} and these variables do not appear in any other linear function $\psi_{1tk},\ t \in J_i,\ i \in I,\ t \neq j$. In other words,

for each $j \in J_i,\ i \in I$, variables $\{x^j, y_j\}$ appear only in one linear function. For a given $t = 1, \ldots, (n+1)l$, we set

$$q_t = \left\lfloor \frac{t-1}{n+1} \right\rfloor + 1, \quad d_t = t - (q_t - 1)(n+1)$$

where $\lfloor c \rfloor$ stands for the floor of a number c. We define by X the vector of all variables $\{x^j, y_j\}$, $j = 1, \ldots, l$:

$$X = (X_1, X_2, \ldots, X_{(n+1)l}),$$

where

$$X_t = \begin{cases} x_{d_t}^{q_t} & \text{if } 1 \le d_t \le n, \\ y_{q_t} & \text{if } d_t = n+1 \end{cases}$$

It follows from Definition 2 that in order to calculate a discrete gradient of the function f_1, first we have to define the following set of $(n+1)l$ points:

$$X_u^0, \ldots, X_u^{u-1}, X_u^{u+1}, \ldots, X_u^{(n+1)l}, \quad u \in U(g, \alpha), \ \ g \in \mathbb{R}^{(n+1)l}.$$

It follows from equation (36) that the points X_u^{t-1} and X_u^t differ exactly by one coordinate. This coordinate appears only in the linear function ψ_{1q_tk}. It follows from the definition of the operator H_u^t that $X_t^t = X_t^{t-1}$ and, therefore, the points X_t^{t+1} and X_t^t differ exactly by one coordinate. Thus we get

$$\psi_{1jk}(X_u^t) = \psi_{1jk}(X_u^{t-1}) \quad \forall j \ne q_t.$$

The function ψ_{1q_tk} can be calculated at the point X_u^t using the value of this function at the point X_u^{t-1}, $t \ge 1$:

$$\psi_{1q_tk}(X_u^t) = \begin{cases} \psi_{1q_tk}(X_u^{t-1}) - z(\lambda) a_{d_t}^k e_t(\beta) & \text{if } 1 \le d_t \le n, \\ \psi_{1q_tk}(X_u^{t-1}) + z(\lambda) e_t(\beta) & \text{if } d_t = n+1 \end{cases} \tag{37}$$

In order to calculate the function f_1 at the point X_u^t, $t \ge 1$, first we have to calculate the functions ψ_{1q_tk} for all $a^k \in A, \quad k = 1, \ldots, m$, using equation (37). Then, we update the function f_1 using these values and the values of all other linear functions at the point X_u^{t-1} according to equation (31). Thus, we have to apply a full calculation of the function f_1 using the formula (31) only at the point $X_u^0 = X + \lambda g$.

As the functions f_1 and f_2 are very similar in structure, f_2 it can be calculated in the same manner using a formula similar to equation (37).

Thus, for the calculation of each discrete gradient, we have to apply a full calculation of the objective function f only at the point $X_u^0 = X + \lambda g$ directly from equations (31) and (32).

We can conclude that for the calculation of the first discrete gradient at a point X, we calculate the function f at two points: X, and $X_u^0 = X + \lambda g,\ g \in S_1, \lambda > 0$. This function is updated at the point X_u^t, $t \ge 1$, using a simplified scheme. For the calculation of another discrete gradient at the same point X with respect to any other direction $g^1 \in S_1$, we calculate the function f only at the point $X + \lambda g^1$. As the number of variables $(n+1)l$ in the problem (15) can be large, this algorithm allows one to significantly reduce the number of the objective function evaluations during the calculation of a discrete gradient.

On the other hand, the function f_1 contains max–min-type functions and their computation can be simplified using an algorithm proposed in ref. [39]. The function f_2 contains min–max-type functions and a similar algorithm can be used for their calculation.

Results of numerical experiments show that the use of these algorithms allows one to significantly accelerate the computation of the objective function f and its discrete gradients. Because at each iteration of the discrete gradient method a few discrete gradients are calculated, we can say that the discrete method is effective for solving the problem (15).

6. Numerical experiments

In this section, we present results of numerical experiments with some real-world datasets. The datasets used are the Wisconsin Breast Cancer Diagnosis (WBCD), the Wisconsin Breast Cancer Prognosis (WBCP), the Cleveland Heart Disease (heart), the Pima Indians Diabetes (diabetes), the BUPA Liver Disorders (liver), the United States Congressional Voting Records (votes), and the ionosphere. All datasets contain two classes. The description of these datasets can be found in ref. [40].

Our algorithm has been implemented in Lahey Fortran 95 on a Pentium 4 (1.7 GHz). First, we take entire datasets and check their polyhedral or max–min separability considering various number of hyperplanes. Results of numerical experiments are presented in table 1. We use the following notations: m is the number of instances in the first class, p is the number of instances in the second class, n is the number of attributes, h is the number of hyperplanes used for polyhedral separability, r is the cardinality of the set I, and j is the cardinality of the sets $J_i, i \in I$, in the max–min separability. The sets J_i contain the same number of indices for all $i \in I$. In our experiments, we restrict r to 10 and j to 5. The accuracy is defined as the ratio between the number of well-classified points of both A and B and the total number of points in the dataset.

From the results presented in table 1, we can conclude that in none of the datasets, classes are linearly separable. Classes in heart, votes, and WBCP are polyhedrally separable, and in WBCD they are 'almost' polyhedrally separable. We considered different values for h in diabetes and liver datasets and present best results. These results show that classes in these datasets are unlikely to be polyhedrally separable. Classes in WBCD, heart, ionosphere, votes, and WBCP are max–min separable with a presented number of hyperplanes, whereas classes in diabetes and liver datasets are likely to be max–min separable with quite large number of hyperplanes. On the other side, results for these datasets show that the use of max–min separability allows one to achieve significantly better separation.

Then, we applied our algorithm to solve data classification problems in earlier-mentioned datasets. A10-fold cross-validation, which consists in splitting the dataset into 10 equally sized pieces, has been used. Therefore, the algorithm ran 10 times and each time nine pieces have been used as the training set and the remaining one as the test set.The results are presented in tables 2–5. In these tables, we also present results available from the literature and obtained

Table 1. Results of numerical experiments.

			Polyhedral		Max–min	
Database	$m/p/n$	Linear	h	accuracy	$r \times j$	accuracy
WBCD	239/444/9	97.36	7	98.98	5×2	100
Heart	137/160/13	84.19	10	100	2×5	100
Ionosphere	126/225/34	93.73	4	97.44	2×2	100
Votes	168/267/16	96.80	5	100	2×3	100
WBCP	46/148/32	76.80	4	100	3×2	100
Diabetes	268/500/8	76.95	12	80.60	7×4	89.45
Liver	145/200/6	68.41	12	74.20	10×2	87.83

Table 2. Results for WBCD and heart datasets.

Database	Algorithm	Average training set correctness	Average test set correctness
WBCD	RLP	97.73	97.21
	HMM	97.87	97.36
	PMM	98.57	96.47
	SVM	–	97.20
	Linear	97.55	96.72
	2-Polyhedral	98.40	96.92
	5-Polyhedral	98.69	98.21
	10-Polyhedral	98.83	98.66
	2×2-Max–min	98.25	98.06
	5×2-Max–min	98.69	98.51
Heart	Linear	85.30	82.76
	2-Polyhedral	88.21	84.48
	5-Polyhedral	96.49	86.21
	10-Polyhedral	99.93	94.48
	2×2-Max–min	93.28	85.52
	4×2-Max–min	99.89	94.83

by a number of algorithms for comparison purposes. In particular, we include results by the robust linear programing (RLP) [11], the hybrid misclassification minimization (HMM) [15], the parametric misclassification minimization (PMM) [10,16], a single support vector machine (SVM) [9], and the h-polyhedral separability (with $h = 2$) [8] algorithms. We report only the results that have been obtained by using the same number of instances in each classes.

We also report the results obtained for linear separability and h-polyhedral separability (with $h \geq 2$) by the implementation of our algorithm. In our experiments, we restrict r to 5 and the cardinality j of the sets J_i, $i \in I$, to 2.

The results presented in table 2 show that algorithms based on polyhedral and max–min separability achieved good results in both training and test phases for WBCD and heart datasets. For heart dataset, they significantly improved the classification accuracy in comparison with algorithms based on linear programing techniques.

We can see from the results presented in table 3 that for ionosphere dataset the algorithm based on the max–min separability achieve a significantly better result than all other algorithms. For WBCP dataset, results obtained by the algorithms based on polyhedral and max–min

Table 3. Results for ionosphere and WBCP datasets.

Database	Algorithm	Average training set correctness	Average test set correctness
Ionosphere	RLP	94.90	86.09
	HMM	96.56	88.36
	PMM	98.42	87.52
	Linear	94.48	86.18
	2-Polyhedral	98.11	85.00
	4-Polyhedral	99.68	88.53
	2×2-max–min	100	96.47
WBCP	Linear	77.10	65.56
	2-Polyhedral	92.39	71.11
	4-Polyhedral	100	92.22
	2×2-max–min	100	88.33
	4×2-max–min	100	94.44

Table 4. Results for diabetes and liver datasets.

Database	Algorithm	Average training set correctness	Average test set correctness
Diabetes	RLP	76.77	76.00
	HMM	78.42	75.89
	PMM	80.55	76.67
	SVM	–	77.60
	Linear	76.56	75.79
	2-Polyhedral	76.48	75.53
	5-Polyhedral	78.27	73.16
	12-Polyhedral	80.79	74.87
	2×2-max–min	80.77	78.03
	5×2 max–min	83.93	81.18
Liver	RLP	68.99	66.93
	HMM	72.21	66.64
	PMM	74.85	68.37
	SVM	–	60.60
	Linear	68.52	67.06
	2-Polyhedral	74.66	69.95
	5-Polyhedral	72.32	66.47
	10-Polyhedral	74.53	66.77
	2×2-max–min	77.75	73.24
	5×2-max–min	86.59	79.12

Table 5. Results for votes dataset.

Database	Algorithm	Average training set correctness	Average test set correctness
Votes	RLP	97.45	95.63
	HMM	98.03	95.62
	PMM	98.82	94.01
	SVM	–	92.60
	Linear	96.64	95.00
	2-Polyhedral	98.37	94.43
	5-Polyhedral	100	98.10
	2×2-max–min	99.72	94.29
	3×2 max–min	100	97.62

separability are comparable and they are considerably better than those obtained by the linear separability.

Both diabetes and liver datasets have very complicated structure. Results presented in table 4 show that the algorithm based on max–min separability improves accuracy in both training and test phases in comparison with all other algorithms. Such an improvement is significant in the case of liver dataset. These results again confirm that classes in these datasets are unlikely to be polyhedrally separable.

The algorithms based on polyhedral and max–min separability achieved best results for the votes dataset and their results are comparable.

7. Conclusions

In this article we have developed the concept of the max–min separability. If finite points sets A and B are disjoint, then they can be separated by a certain piecewise linear function

presented as a max–min of linear functions. We have introduced the error function for this case and proposed the algorithm for the separation of two sets based on max–min separability. This algorithm has been applied to solve data classification problems in some test datasets. Results from numerical experiment show the effectiveness of this algorithm. However, this algorithm requires a significantly bigger computational effort than algorithms based on linear programing techniques.

One of the important questions when one tries to apply the max–min separability is the number l of hyperplanes needed for the max–min separation of two sets A and B. In this article, we assumed this number to be known. However, in many cases it is not known *a priori*. The development of an algorithm for the estimation of l will be the subject of our further research.

Acknowledgements

The author would like to thank the two anonymous referees whose comments have improved this article. This research was supported by the Australian Research Council.

References

[1] Jurs, P.C., 1986, Pattern recognition used to investigate multivariate chemistry. *Science*, **232**, 1219–1224.
[2] Mangasarian, O.L. and Wolberg, W.H., 1995, Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, **43**, 570–577.
[3] Srinivasan, V. and Kim, Y.H., 1987, Credit granting: comparative analysis of classification procedures. *Journal of Finance*, **42**, 665–683.
[4] Fukunaga, K., 1990, *Introduction to Statistical Pattern Recognition* (2nd edn) (Boston, MA: Academic Press).
[5] McLachlan, G.J., 1992, *Discriminant Analysis and Statistical Pattern Recognition* (New York: John Wiley).
[6] Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (Eds), 1994, *Machine Learning, Neural and Statistical Classification* (London: Ellis Horwood Series in Artificial Intelligence).
[7] Quinlan, J.R., 1993, *C4.5: Programs for Machine Learning* (San Mateo: Morgan Kaufmann).
[8] Astorino, A. and Gaudioso, M., 2002, Polyhedral separability through successive LP. *Journal of Optimization Theory and Applications,* **112**(2), 265–293.
[9] Bennet, K.P. and Blue, J., 1997, A support vector machine approach to decision trees. Mathematics Report 97–100, Rensselaer Polytechnic Institute, Troy, New York.
[10] Bennet, K.P. and Bredersteiner, E.J., 1997, A parametric optimization method for machine learning. *INFORMS Journal on Computing*, **9**, 311–318.
[11] Bennett, K.P. and Mangasarian, O.L., 1992, Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, **1**, 23–34.
[12] Bradley, P.S. and Mangasarian, O.L., 2000, Massive data discrimination via linear support vector machines. *Optimization Methods and Software*, **13**, 1–10.
[13] Bradley, P.S., Fayyad, U.M. and Mangasarian, O.L., 1999, Data mining: overview and optimization opportunities. *INFORMS Journal on Computing*, **11**, 217–238.
[14] Burges, C.J.C., 1998, A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.
[15] Chen, C. and Mangasarian, O.L., 1995, Hybrid misclassification minimization. Mathematical Programming Technical Report, 95-05, University of Wisconsin.
[16] Mangasarian, O.L., 1994, Misclassification minimization. *Journal of Global Optimization*, **5**, 309–323.
[17] Mangasarian, O.L., 1997, Mathematical programming in data mining. *Data Mining and Knowledge Discovery*, **1**, 183–201.
[18] Thorsten, J., 2002, *Learning to Classify Text Using Support Vector Machines* (Dordrecht: Kluwer Academic Publishers).
[19] Vapnik, V.N., 1995, *The Nature of Statistical Learning Theory* (New York: Springer).
[20] Bagirov, A.M., Rubinov, A.M. and Yearwood, J., 2001, Using global optimization to improve classification for medical diagnosis and prognosis. *Topics in Health Information Management*, **22**, 65–74.
[21] Bagirov, A.M., Rubinov, A.M. and Yearwood, J., 2002, A global optimization approach to classification. *Optimization and Engineering*, **3**(2), 129–155.
[22] Demyanov, V.F., Mathematical diagnostics via nonsmooth analysis. *Optimization Methods and Software* (submitted).
[23] Bartels, S.G., Kuntz, L. and Sholtes, S., 1995, Continuous selections of linear functions and nonsmooth critical point theory. *Nonlinear Analysis, TMA*, **24**, 385–407.
[24] Clarke, F.H., 1983, *Optimization and Nonsmooth Analysis* (New York: Wiley-Interscience).

[25] Demyanov, V.F. and Rubinov, A.M., 1995, *Constructive Nonsmooth Analysis* (Frankfurt am Main: Peter Lang).
[26] Mifflin, R., 1977, Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, **15**, 957–972.
[27] Demyanov, A.V., Demyanov, V.F. and Malozemov, V.N., 2002, Minmaxmin problems revisited. *Optimization Methods and Software*, **17**(5), 783–804.
[28] Polak, E., 1997, *Optimization: Algorithms and Consistent Approximations* (New York: Springer Verlag).
[29] Kirjner-Neto, C. and Polak, E., 1998, On the conversion of optimization problems with maxmin constraints to standard optimization problems. *SIAM J. Optimization*, **8**(4) 887–915.
[30] Powell, M.J.D., 2002, UOBYQA: unconstrained optimization by quadratic approximation. *Mathematical Programming*, Series B, **92**(3), 555–582.
[31] Nelder, J.A. and Mead, R., 1965, A simplex method for function minimization. *Computer Journal*, **7**, 308–313.
[32] Bagirov, A.M., 1999, Derivative-free methods for unconstrained nonsmooth optimization and its numerical analysis. *Investigacao Operacional*, **19**, 75–93.
[33] Bagirov, A.M., 1999, Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices. In: A. Eberhard *et al.* (Eds) *Progress in Optimization: Contribution from Australasia* (Dordrecht: Kluwer Academic Publishers).
[34] Bagirov, A.M., 2002, A method for minimization of quasidifferentiable functions. *Optimization Methods and Software*, **17**(1), 31–60.
[35] Hiriart-Urruty, J.-P. and Lemarechal, C., 1993, *Convex Analysis and Minimization Algorithms*, Vols 1 and 2 (Berlin, Heidelberg, New York: Springer Verlag).
[36] Kiwiel, K.C., 1985, *Methods of Descent for Nondifferentiable Optimization*, *Lecture Notes in Mathematics*, Vol. 1133 (Berlin: Springer Verlag).
[37] Makela, M.M. and Neittaanmaki, P., 1992, *Nonsmooth Optimization* (Singapore: World Scientific).
[38] Bagirov, A.M. and Gasanov, A., 1995, A method of approximating a quasidifferential. *Russian Journal of Computational Mathematics and Mathematical Physics,* **35**(4), 403–409.
[39] Evtushenko, Yu.G., 1972, A numerical method for finding best guaranteed estimates. *USSR Journal of Computational Mathematics and Mathematical Physics*, **12**, 109–128.
[40] Murphy, P.M. and Aha, D.W., 1992, UCI repository of machine learning databases. Technical report, Department of Information and Computer Science, University of California, Irvine. Available online at: www.ics.uci.edu/mlearn/MLRepository.html.