Max-min separability

Adil M. Bagirov

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

Abstract

We consider the problem of discriminating two finite points 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.

Key words: Classification, separability, nonconvex optimization, nonsmooth optimization.

1 Introduction

The problems of supervised data classification arise in many areas including management science, medicine, chemistry (see [23, 28, 38]). 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 networks approaches (see, for example, [21, 30, 31, 37]).

One of the promising approaches to data classification problems is based on mathematical programming techniques. There are two main approaches to apply mathematical programming 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 [1, 9, 10, 11, 12, 13, 14, 15, 27, 29, 39, 40]).

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 which contains this cluster (see [6, 7]).

In this paper 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 points 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 which separates these two sets. Linear programming techniques can be used to construct such a hyperplane. If the convex hulls of A and B intersect then linear programming techniques can be applied to obtain a hyperplane which minimizes some misclassification measure. Algorithms based on such an approach are developed in [10, 11, 15, 27].

In the paper [1] the concept of polyhedral separability was introduced. In this paper the case when $\text{co } A \cap B = \emptyset$ was considered. 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 which 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 programming problem.

In this paper 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 the most general nonsmooth classifiers have been introduced in the paper [17]. 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 paper is as follows. Section 2 provides some preliminaries. In Section 3 the definition and some results related to the max-min separability are given. 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 paper.

2 Preliminaries

2.1 Linear separability

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

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

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 [11]. This hyperplane is the solution to the following problem:

$$\text{minimize } f(x, y) \quad \text{subject to } (x, y) \in \mathbb{R}^{n+1} \quad (2.1)$$

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 [11] that the problem (2.1) is equivalent to the following linear program:

$$\text{minimize } \frac{1}{m} \sum_{i=1}^m t_i + \frac{1}{p} \sum_{j=1}^p z_j$$

subject to

$$\begin{aligned} t_i &\geq \langle x, a^i \rangle - y + 1, \quad i = 1, \dots, m, \\ z_j &\geq -\langle x, b^j \rangle + y + 1, \quad j = 1, \dots, p, \\ t &\geq 0, \quad z \geq 0, \end{aligned}$$

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 (2.1). It is proved that the trivial solution $x = 0$ cannot occur.

2.2 Polyhedral separability

The concept of h -polyhedral separability was developed in [1]. The sets A and B are h -polyhedrally separable if there exists a set of h hyperplanes $\{x^i, y_i\}$, with

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

such that

- 1) for any $j = 1, \dots, m$ and $i = 1, \dots, h$

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

- 2) for any $k = 1, \dots, p$ there exists at least one $i \in \{1, \dots, h\}$ so that

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

It is proved in [1] that the sets A and B are h -polyhedrally separable, for some $h \leq p$ if and only if

$$\text{co } A \cap B = \emptyset.$$

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

$$\text{minimize } f(x, y) \quad \text{subject to } (x, y) \in \mathbb{R}^{(n+1) \times h} \quad (2.2)$$

where

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

$$\frac{1}{p} \sum_{k=1}^p \max \left[0, \min_{1 \leq i \leq h} \left\{ -\langle x^i, b^k \rangle + y_i + 1 \right\} \right]$$

is an error function. Note that this function is a nonconvex piecewise linear function. It is proved that $x^i = 0$, $i = 1, \dots, h$ cannot be the optimal solution. Let $\{\bar{x}^i, \bar{y}_i\}$, $i = 1, \dots, h$ be a global solution to the problem (2.2). 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 $\bar{I} \subset \{1, \dots, h\}$ such that $x^i = 0$, $i \in \bar{I}$, then the sets A and B are $(h - |\bar{I}|)$ -polyhedrally separable. In [1] an algorithm for solving problem (2.2) is developed. The calculation of the descent direction at each iteration of this algorithm is reduced to a certain linear programming problem.

3 Max-min separability

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

$$J_k \neq \emptyset, \quad k = 1, \dots, r, \quad J_k \cap J_j = \emptyset, \quad \bigcup_{k=1}^r J_k = J.$$

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

$$\varphi(z) = \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, z \rangle - y_j \right\}, \quad z \in \mathbb{R}^n. \quad (3.1)$$

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

Definition 3.1 *The sets A and B are max-min separable if there exist a finite number of hyperplanes $\{x^j, y_j\}$ with $x^j \in \mathbb{R}^n$, $y_j \in \mathbb{R}^1$, $j \in J = \{1, \dots, l\}$ and a partition $J^r = \{J_1, \dots, J_r\}$ of the set J such that*

1) for all $i \in I$ and $a \in A$

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

2) 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 3.1 It follows from Definition 3.1 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 (3.1). Thus the sets A and B can be separated by a function represented as a max-min of linear functions.

Remark 3.2 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 $I = \{1, \dots, h\}$ and $J_i = \{i\}$, $i \in I$ we obtain h -polyhedral separability.

Proposition 3.1 *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 $J^r = \{J_1, \dots, J_r\}$ of the set J such that*

1)

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

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

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

Proof: Sufficiency is straightforward.

Necessity. Since 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

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

and

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

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

$$\max_{i \in I} \min_{j \in J_i} \{ \langle \bar{x}^j, a \rangle - \bar{y}_j \} \leq -\delta, \quad \forall a \in A, \quad (3.2)$$

$$\max_{i \in I} \min_{j \in J_i} \{ \langle \bar{x}^j, b \rangle - \bar{y}_j \} \geq \delta, \quad \forall b \in B. \quad (3.3)$$

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:

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

$$y_j = \bar{y}_j / \delta, \quad j \in J.$$

Then it follows from (3.2) and (3.3) that

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

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

which completes the proof. $\triangle$

Proposition 3.2 *The sets A and B are max-min separable if and only if there exists a piecewise linear function separating them.*

Proof: Since max-min of linear functions is 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 (3.1) (see [8]). 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. $\triangle$

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

Proposition 3.3 *Assume that the set A can be represented as a union of sets A_i , $i = 1, \dots, q$:*

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

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

$$B \cap \text{co} A_i = \emptyset. \quad (3.4)$$

Then the sets A and B are max-min separable.

Proof: It follows from (3.4) that $b \notin \text{co} A_i$ for all $b \in B$ and $i \in \{1, \dots, q\}$. Then, for each $b \in B$ and $i \in \{1, \dots, q\}$ there exists a hyperplane $\{x^i(b), y_i(b)\}$ separating b from the set $\text{co} A_i$, that is

$$\begin{aligned} \langle x^i(b), b \rangle - y_i(b) &> 0, \\ \langle x^i(b), a \rangle - y_i(b) &< 0, \quad \forall a \in \text{co} A_i, \quad i = 1, \dots, q. \end{aligned}$$

Then we have

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

and

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

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

$$\min_{i=1, \dots, q} \left\{ \langle x^i(b^j), b^j \rangle - y_i(b^j) \right\} > 0 \quad (3.5)$$

and

$$\min_{i=1, \dots, q} \left\{ \langle x^i(b^j), a \rangle - y_i(b^j) \right\} < 0, \quad \forall a \in A. \quad (3.6)$$

Consequently we have pq hyperplanes

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

The set of these hyperplanes can be rewritten as follows:

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

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

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

Consider the following partition of the set J :

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

It follows from (3.5) and (3.6) that for all $k \in I$ and $a \in A$

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

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

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

which means that the sets A and B are max-min separable. $\triangle$

Corollary 3.1 *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. Since the sets A and B are disjoint the condition (3.4) is satisfied. Then the proof of the corollary follows from Proposition 3.3. $\triangle$

Remark 3.3 For many cases the number of hyperlanes 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 $\text{co } A_i$, $i = 1, \dots, q$. The result from the next Proposition 3.4 confirm it.

Proposition 3.4 *Assume that the set A can be represented as a union of sets A_i , $i = 1, \dots, q$ and the set B as a union of sets B_j , $j = 1, \dots, d$ such that*

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

and

$$\text{co } A_i \cap \text{co } B_j = \emptyset \quad \text{for all } i = 1, \dots, q, \quad j = 1, \dots, d. \quad (3.7)$$

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

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

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

and

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

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

$$\min_{i=1, \dots, q} \langle x^{ij}, b \rangle - y_{ij} > 0, \quad \forall b \in B_j \quad (3.8)$$

and

$$\min_{i=1, \dots, q} \langle x^{ij}, a \rangle - y_{ij} < 0, \quad \forall a \in A. \quad (3.9)$$

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

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

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

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

Consider the following partition of the set J :

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

It follows from (3.8) and (3.9) 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. $\triangle$

Remark 3.4 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 $\{x^j, y_j\}$, $j \in J = \{1, \dots, l\}$ with $x^j \in \mathbb{R}^n$, $y_j \in \mathbb{R}^1$ and a partition $J^r = \{J_1, \dots, J_r\}$ of the set J , we say that a point $a \in A$ is well classified if the following condition 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:

$$\max \left[0, \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \right]. \quad (4.1)$$

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

$$\max \left[0, \min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \right]. \quad (4.2)$$

Thus, an averaged classification error function can be defined as

$$\begin{aligned} f(x, y) = & (1/m) \sum_{k=1}^m \max \left[0, \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, a^k \rangle - y_j + 1 \right\} \right] \\ & + (1/p) \sum_{t=1}^p \max \left[0, \min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^j, b^t \rangle + y_j + 1 \right\} \right] \end{aligned} \quad (4.3)$$

where $x = (x^1, \dots, x^l) \in \mathbb{R}^{l \times n}$, $y = (y_1, \dots, y_l) \in \mathbb{R}^l$. 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:

$$\text{minimize } f(x, y) \text{ subject to } (x, y) \in \mathbb{R}^{(n+1) \times l}. \quad (4.4)$$

Proposition 4.1 *The sets A and B are max-min separable if and only if there exist a set of hyperplanes $\{x^j, y_j\}$, $j \in J = \{1, \dots, l\}$ and a partition $J^r = \{J_1, \dots, J_r\}$ 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 3.1 that there exists a set of hyperplanes $\{x^j, y_j\}, j \in J$ and a partition $J^r = \{J_1, \dots, J_r\}$ of the set J such that

$$\min_{j \in J_i} \{\langle x^j, a \rangle - y_j\} \leq -1, \quad \forall a \in A, \quad i \in I = \{1, \dots, r\} \quad (4.5)$$

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

$$\min_{j \in J_t} \{\langle x^j, b \rangle - y_j\} \geq 1. \quad (4.6)$$

Consequently we have

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

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

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

Sufficiency. Assume that there exist a set of hyperplanes $\{x^j, y_j\}, j \in J = \{1, \dots, l\}$ and a partition $J^r = \{J_1, \dots, J_r\}$ 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 (4.5) and (4.6) are satisfied, that is the sets A and B are max-min separable. $\triangle$

Proposition 4.2 Assume that the sets A and B are max-min separable with a set of hyperplanes $\{x^j, y_j\}, j \in J = \{1, \dots, l\}$ and a partition $J^r = \{J_1, \dots, J_r\}$ of the set J . Then

1) $x^j = 0, j \in J$ cannot be an optimal solution to the problem (4.4);

2) if

(a) for any $t \in I$ there exists at least one $b \in B$ such that

$$\max_{j \in J_t} \{-\langle x^j, b \rangle + y_j + 1\} = \min_{i \in I} \max_{j \in J_i} \{-\langle x^j, b \rangle + y_j + 1\}, \quad (4.7)$$

(b) there exists $\tilde{J} = \{\tilde{J}_1, \dots, \tilde{J}_r\}$ 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 $\bar{J} = \{\bar{J}_1, \dots, \bar{J}_r\}$ of the set J^0 where

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

Proof: 1) Since the sets A and B are max-min separable we get from Proposition 4.1 that $f(x, y) = 0$. If $x^j = 0$, $j \in J$ then it follows from (4.3) that for any $y \in \mathbb{R}^l$

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

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

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

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

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

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.

2) Consider the following sets:

$$I^1 = \{i \in I : \bar{J}_i \neq \emptyset\},$$

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

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

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

Since the function f is nonnegative we obtain

$$\max_{i \in I} \min_{j \in J_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \leq 0, \quad \forall a \in A, \quad (4.8)$$

$$\min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \leq 0, \quad \forall b \in B. \quad (4.9)$$

It follows from (4.7) and (4.9) that for any $i \in I^2$ there exists a point $b \in B$ such that

$$\max_{j \in \bar{J}_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \leq 0. \quad (4.10)$$

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

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

which means that

$$\max_{j \in \bar{J}_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \leq 0 \quad (4.11)$$

and

$$\max_{j \in \bar{J}_i} \{y_j + 1\} \leq 0. \quad (4.12)$$

If $i \in I^2 \setminus I^3$ then from (4.10) we obtain

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

Thus we get that for all $i \in I^2$ the inequality (4.12) is true. (4.12) can be rewritten as follows:

$$\max_{j \in \bar{J}_i} y_j \leq -1, \quad \forall i \in I^2. \quad (4.13)$$

Consequently for any $i \in I^2$

$$\min_{j \in \bar{J}_i} \{-y_j + 1\} = -\max_{j \in \bar{J}_i} y_j + 1 \geq 2. \quad (4.14)$$

It follows from (4.8) that for any $i \in I$ and $a \in A$

$$\min_{j \in \bar{J}_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \leq 0. \quad (4.15)$$

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

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

Taking into account (4.14) we get that for any $i \in I^3$ and $a \in A$

$$\min_{j \in \bar{J}_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \leq 0. \quad (4.16)$$

If $i \in I^2 \setminus I^3$ then it follows from (4.15) that

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

which contradicts (4.14). 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 (4.8) that for any $i \in I^1 \setminus I^2$ and $a \in A$

$$\min_{j \in \bar{J}_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \leq 0. \quad (4.17)$$

From (4.16) and (4.17) we can conclude that for any $i \in I$ and $a \in A$

$$\min_{j \in \bar{J}_i} \left\{ \langle x^j, a \rangle - y_j + 1 \right\} \leq 0. \quad (4.18)$$

It follows from (4.9) that for any $b \in B$ there exists at least one $i \in I$

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

Then from expression

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

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

$$\max_{j \in \bar{J}_i} \left\{ -\langle x^j, b \rangle + y_j + 1 \right\} \leq 0. \quad (4.19)$$

Thus it follows from (4.18) and (4.19) 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 . $\triangle$

5 Minimization of the error function

In this section we discuss an algorithm for solving problem (4.4). The objective function f in this problem has the following form:

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

where

$$f_1(x, y) = \frac{1}{m} \sum_{k=1}^m \max \left[0, \max_{i \in I} \min_{j \in \bar{J}_i} \left\{ \langle x^j, a^k \rangle - y_j + 1 \right\} \right], \quad (5.1)$$

$$f_2(x, y) = \frac{1}{p} \sum_{t=1}^p \max \left[0, \min_{i \in I} \max_{j \in \bar{J}_i} \left\{ -\langle x^j, b^t \rangle + y_j + 1 \right\} \right]. \quad (5.2)$$

The problem (4.4) 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 the finding of 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 ([16, 18]). Moreover,

both functions are semismooth (for the definition of the semismooth functions see [32]). Therefore the function f is also subdifferentiable. The function f_1 contains the following max-min functions:

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

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

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

The differential properties of max-min functions are studied, for example, in [19, 35]. The functions φ_{1k} , $k = 1, \dots, m$ and φ_{2t} , $t = 1, \dots, p$ are not regular (for the definition of the regular functions see, [16]). Then the functions f_1 , f_2 and consequently the function f are not regular, too. 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 problem (4.4).

In the paper [24] optimization problems with twice continuously differentiable objective functions and max-min constraints were considered and these problems were converted to the problems with the smooth objective and constraint functions. However, this approach cannot be applied to the problem (4.4).

Direct search methods of optimization can be used for solving problem (4.4). Among such methods we mention here two widely used methods: the Powell's method (see [36]) which based on a quadratic approximation of the objective function and the Nelder-Mead's simplex method [34]. As was mentioned in [36] the 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 a 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 paper we use the discrete gradient method to solve the problem (4.4). The description of this method can be found in [3, 4] (see, also, [5]). The discrete gradient method can be considered as a version of the bundle method ([22, 25, 26]). 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 (4.4) can be expensive. We will show that the use of the discrete gradient method allow to significantly reduce the number of the objective function evaluations. We will not describe the discrete gradient method in this paper, however, we will give a description of the discrete gradient in order to demonstrate that its computation in the problem (4.4) can be simplified.

5.0.1 Definition of the discrete gradient

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

$$S_1 = \{g \in \mathbb{R}^n : \|g\| = 1\}, \quad G = \{e \in \mathbb{R}^n : e = (e_1, \dots, e_n), |e_j| = 1, j = 1, \dots, n\},$$

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

$$I(g, \alpha) = \{i \in \{1, \dots, n\} : |g_i| \geq \alpha\},$$

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 cube in $\mathbb{R}^n$ and P is a set of univariate positive infinitesimal functions.

We define operators $H_i^j : \mathbb{R}^n \rightarrow \mathbb{R}^n$ for $i = 1, \dots, n$, $j = 0, \dots, n$ by the formula

$$H_i^j g = \begin{cases} (g_1, \dots, g_j, 0, \dots, 0) & \text{if } j < i, \\ (g_1, \dots, g_{i-1}, 0, g_{i+1}, \dots, g_j, 0, \dots, 0) & \text{if } j \geq i. \end{cases} \quad (5.3)$$

We can see that

$$H_i^j g - H_i^{j-1} g = \begin{cases} (0, \dots, 0, g_j, 0, \dots, 0) & \text{if } j = 1, \dots, n, j \neq i, \\ 0 & \text{if } j = i. \end{cases} \quad (5.4)$$

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

$$x_i^j(g, e, z, \lambda, \beta) = x + \lambda g - z(\lambda) H_i^j e(\beta), \quad (5.5)$$

where $g \in S_1$, $e \in G$, $i \in I(g, \alpha)$, $z \in P$, $\lambda > 0$, $j = 0, \dots, n$, $j \neq i$.

It follows from (5.4) that

$$x_i^{j-1}(g, e, z, \lambda, \beta) - x_i^j(g, e, z, \lambda, \beta) = \begin{cases} (0, \dots, 0, z(\lambda) e_j(\beta), 0, \dots, 0) & \text{if } j = 1, \dots, n, j \neq i, \\ 0 & \text{if } j = i. \end{cases} \quad (5.6)$$

It is clear that $H_i^0 g = 0$ and $x_i^0(g, e, z, \lambda, \beta) = x + \lambda g$ for all $i \in I(g, \alpha)$.

Definition 5.1 (see [2]) *The discrete gradient of the function f at the point $x \in \mathbb{R}^n$ is the vector $\Gamma^i(x, g, e, z, \lambda, \beta) = (\Gamma_1^i, \dots, \Gamma_n^i) \in \mathbb{R}^n$, $g \in S_1$, $i \in I(g, \alpha)$, with the following coordinates:*

$$\Gamma_j^i = [z(\lambda) e_j(\beta)]^{-1} \left[f(x_i^{j-1}(g, e, z, \lambda, \beta)) - f(x_i^j(g, e, z, \lambda, \beta)) \right], \quad j = 1, \dots, n, j \neq i,$$

$$\Gamma_i^i = (\lambda g_i)^{-1} \left[f(x_i^n(g, e, z, \lambda, \beta)) - f(x) - \sum_{j=1, j \neq i}^n \Gamma_j^i (\lambda g_j - z(\lambda) e_j(\beta)) \right].$$

Remark 5.1 The discrete gradient is defined with respect to a given direction $g \in S_1$. 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 $x_i^j(g, e, z, \lambda, \beta)$, $j = 0, \dots, n$, $j \neq i$. For the calculation of the next 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 (4.4). This function depends on $(n+1)l$ variables where l is the number of hyperplanes. The function f_1 contains max-min functions φ_{1k} which in their turn contain the following linear functions:

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

We can see that for each $j \in J_i, i \in I$ the function ψ_{1jk} has own variables $\{x^j, y_j\}$ and these variables are not 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 $\{x^j, y_j\}$ belong only to one linear function. For a given $i = 1, \dots, (n+1)l$ we set

$$q_i = \left\lfloor \frac{i-1}{n+1} \right\rfloor + 1, \quad d_i = i - (q_i - 1)(n+1)$$

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

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

where

$$X_i = \begin{cases} x_{d_i}^{q_i} & \text{if } 1 \leq d_i \leq n, \\ y_{q_i} & \text{if } d_i = n+1. \end{cases}$$

It follows from the definition of the discrete gradient and (5.6) that for the calculation of $(n+1)l - 1$ coordinates $\Gamma_i^j, j = 1, \dots, (n+1)l, j \neq i, i \in I(g, \alpha)$ of the discrete gradient we have to change one coordinate of the previous point. In particular, for the calculation of $\Gamma_i^j, j \in \{1, \dots, (n+1)l\}, j \neq i$ only j -th coordinate of the previous point $X_i^{j-1}(g, e, z, \lambda, \beta)$ is changed. This variable belongs to one linear function, more exactly to the function $\psi_{1q_j k}$. All other linear functions have the same value as in the point $X_i^{j-1}(g, e, z, \lambda, \beta)$. The function $\psi_{1q_j k}$ can be calculated at the point $X_i^j(g, e, z, \lambda, \beta)$ using the value of this function at the previous point $X_i^{j-1}(g, e, z, \lambda, \beta), j \geq 1$:

$$\psi_{1q_j k}(X_i^j(g, e, z, \lambda, \beta)) = \begin{cases} \psi_{1q_j k}(X_i^{j-1}(g, e, z, \lambda, \beta)) - z(\lambda) a_{d_j}^k e_j(\beta) & \text{if } 1 \leq d_j \leq n, \\ \psi_{1q_j k}(X_i^{j-1}(g, e, z, \lambda, \beta)) + z(\lambda) e_j(\beta) & \text{if } d_j = n+1. \end{cases} \quad (5.7)$$

In order to calculate the function f_1 at the point $X_i^j(g, e, z, \lambda, \beta)$ first we have to calculate the functions $\psi_{1q_j k}$ for all $a^k \in A, k = 1, \dots, m$ using (5.7), then to calculate f_1 using (5.1) and the values of all other linear functions at the previous point. Thus we calculate the function f_1 only at the point $X + \lambda g$ directly from (5.1).

The function f_2 can be calculated in the same manner using a formula similar to (5.7). Thus we calculate the function f only at the point $X + \lambda g$ directly from (5.1) and (5.2).

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 $X_i^0(g, e, z, \lambda, \beta) = X + \lambda g, g \in S_1, \lambda > 0$. This function can be calculated at any other points by the simplified scheme using (5.7). For the calculation of the next 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: $X + \lambda g^1$. Since the number of variables $(n+1)l$ in the problem (4.4) can be large this algorithm allows to significantly reduce the number of the objective function evaluations.

On the other hand the function f_1 contains max-min-type functions and their computation can be simplified using an algorithm proposed in [20]. 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. Since 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 problem (4.4).

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 2 classes. The description of these datasets can be found in [33].

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 notation: m - is the number of instances in the first class, p - is the number of instances in the second class, n - number of attributes, h 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.

Table 1: Results of numerical experiments

Database	$m/p/n$	Linear	Polyhedral		Max-min	
			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

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 to achieve significantly better separation.

Then we applied our algorithm to solve data classification problems in above mentioned datasets. Tenfold cross-validation, which consists in splitting the dataset into ten equally sized pieces, has been used. So the algorithm ran ten 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 by a number of algorithms for comparison purposes. In particular, we include results by the robust linear programming (RLP) [11], the hybrid misclassification minimization (HMM) [15], the parametric misclassification minimization (PMM) [10, 27], a single support vector machine (SVM) [9] and the h -polyhedral separability (with $h = 2$) [1] algorithms. We report only results which have been obtained by using the same number of instances in each classes.

We also report 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.

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

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 programming techniques.

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

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 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 and in the case of liver dataset significantly, accuracy in both training and test phases in comparison with all other algorithms. 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 paper 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

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

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 programming techniques.

Acknowledgements

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

References

- [1] A. Astorino and M. Gaudioso, Polyhedral separability through successive LP, *Journal of Optimization Theory and Applications*, 112(2), 2002, 265-293.
- [2] A.M. Bagirov and A. Gasanov, A method of approximating a quasidifferential, *Russian Journal of Computational Mathematics and Mathematical Physics*, Vol. 35, No. 4, 1995, pp. 403-409.

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

- [3] A.M. Bagirov, Derivative-free methods for unconstrained nonsmooth optimization and its numerical analysis, *Investigacao Operacional*, 19, 1999, 75-93.
- [4] A.M. Bagirov, 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*, Kluwer Academic Publishers, 1999, 147-175.
- [5] A.M. Bagirov, A method for minimization of quasidifferentiable functions, *Optimization Methods and Software*, 17(1), 2002, 31-60.
- [6] A.M. Bagirov, A.M. Rubinov and J. Yearwood, Using global optimization to improve classification for medical diagnosis and prognosis, *Topics in Health Information Management*, 22, 2001, 65-74.
- [7] A.M. Bagirov, A.M. Rubinov and J. Yearwood, A global optimization approach to classification, *Optimization and Engineering*, No. 3, 2002, 129-155.
- [8] S.G. Bartels, L. Kuntz and S. Sholtes, Continuous selections of linear functions and nonsmooth critical point theory, *Nonlinear Analysis, TMA*, 24, 1995, 385-407.
- [9] K.P. Bennet and J. Blue, A support vector machine approach to decision trees, Mathematics Report 97-100, Rensselaer Polytechnic Institute, Troy, New York, 1997.
- [10] K.P. Bennet and E.J. Bredersteiner, A parametric optimization method for machine learning, *INFORMS Journal on Computing*, 9, 1997, 311-318.
- [11] K.P. Bennett and O.L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, *Optimization Methods and Software*, 1, 1992, 23-34.
- [12] P.S. Bradley and O.L. Mangasarian, Massive data discrimination via linear support vector machines, *Optimization Methods and Software*, 13, 2000, 1-10.

- [13] P.S. Bradley, Usama M. Fayyad and O.L. Mangasarian, Data mining: overview and optimization opportunities, *INFORMS Journal on Computing*, 11, 1999, 217-238.
- [14] Burges, C.J.C., A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 2, 1998, 121-167.
- [15] C. Chen and O.L. Mangasarian, Hybrid misclassification minimization, Mathematical Programming Technical Report, 95-05, University of Wisconsin, 1995.
- [16] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
- [17] V.F. Demyanov, Mathematical Diagnostics via nonsmooth analysis, Submitted to: *Optimization Methods and Software*.
- [18] V.F. Demyanov and A.M. Rubinov, *Constructive Nonsmooth Analysis*, Peter Lang, Frankfurt am Main, 1995.
- [19] A.V. Demyanov, V.F. Demyanov and V.N. Malozemov, Minmaxmin problems revisited, *Optimization Methods and Software*, 17(5), 2002, 783-804.
- [20] Yu.G. Evtushenko, A numerical method for finding best guaranteed estimates, *USSR Journal of Computational Mathematics and Mathematical Physics*, 12, 1972, 109-128.
- [21] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. 2nd edition. Academic Press, 1990.
- [22] J.-P. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Vol. 1 and 2, Springer Verlag, Berlin, Heidelberg, New York, 1993.
- [23] P.C. Jurs, Pattern recognition used to investigate multivariate chemistry, *Science*, 232, 1986, 1219-1224.
- [24] C. Kirjner-Neto and E. Polak, On the conversion of optimization problems with max-min constraints to standard optimization problems, *SIAM J. Optimization*, Vol. 8, No. 4, 1998, 887-915.
- [25] K.C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, *Lecture Notes in Mathematics*, 1133, Springer Verlag, Berlin, 1985.
- [26] M.M. Makela and P. Neittaanmaki, *Nonsmooth Optimization*, World Scientific, Singapore, 1992.
- [27] O.L. Mangasarian, Misclassification minimization, *Journal of Global Optimization*, 5, 1994, 309-323.
- [28] O.L. Mangasarian and W.H. Wolberg, Breast cancer diagnosis and prognosis via linear programming, *Operations Research*, 43, 1995, 570-577.

- [29] O.L. Mangasarian, Mathematical programming in data mining, *Data Mining and Knowledge Discovery*, 1, 1997, 183-201.
- [30] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York, John Wiley, 1992.
- [31] D. Michie, D.J. Spiegelhalter and C.C. Taylor (eds.), *Machine Learning, Neural and Statistical Classification*. Ellis Horwood Series in Artificial Intelligence, London, 1994.
- [32] R. Mifflin, Semismooth and semiconvex functions in constrained optimization, *SIAM Journal on Control and Optimization*, 15, 1977, 957-972.
- [33] P.M. Murphy and D.W. Aha, UCI repository of machine learning databases. Technical report, Department of Information and Computer science, University of California, Irvine, 1992. www.ics.uci.edu/mllearn/MLRepository.html.
- [34] J.A. Nelder and R. Mead, A simplex method for function minimization, *Comput. J.*, Vol. 7, pp. 308-313.
- [35] E. Polak, *Optimization: Algorithms and Consistent Approximations*, Springer Verlag, New York, 1997.
- [36] M.J.D. Powell, UOBYQA: unconstrained optimization by quadratic approximation, *Mathematical Programming*, Series B, Vol. 92, No. 3, 2002, pp. 555-582.
- [37] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, 1993.
- [38] V. Srinivasan, and Y.H. Kim, Credit granting: comparative analysis of classification procedures, *Journal of Finance*, Vol. 42, 1987, 665-683.
- [39] J. Thorsten, *Learning to Classify Text Using Support Vector Machines*, Kluwer Academic Publishers, Dordrecht, 2002.
- [40] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.