Max-min separability: incremental approach and application to supervised data classification

Adil M. Bagirov and Dean Webb

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

Bülent Karasözen

Department of Mathematics & Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey.

Abstract

A new algorithm for the computation of a piecewise linear function separating two finite point sets in n-dimensional space is developed and the algorithm is applied to solve supervised data classification problems. The algorithm computes hyperplanes incrementally and it finds as many hyperplanes as necessary to separate two sets with respect to some tolerance. An error function is formulated and an algorithm for its minimization is discussed. We present results of numerical experiments using several UCI test data sets and compare the proposed algorithm with two support vector machine solvers: LIBSVM and SVM_light.

Key words: Separability, nonconvex optimization, nonsmooth optimization, supervised data classification.

1 Introduction

The problem of discriminating between two sets is very important in applied mathematics and in particular, in supervised data classification. If the intersection of the convex hulls of two sets is empty then it follows from convex analysis that these two sets can be separated by an affine function that is using one hyperplane. In this case two sets are linearly separable. However, in many applications two sets are not linearly separable.

Over the last decade various approaches have been proposed to find a piecewise linear function separating two sets. The paper [9] introduces the concept of bilinear separability, where two hyperplanes are used to separate sets. The problem of bilinear separability is reduced to a certain bilinear programming problem.

The paper [1] introduces the concept of polyhedral separability as the generalization of linear separability. In this case one of the sets is approximated by a polyhedral set and

the rest of the space is used to approximate the second set. The authors formulated the error function which can be represented as the sum of nonsmooth convex and nonsmooth nonconvex functions. An algorithm for minimizing the error function is developed where the problem of finding the descent directions is reduced to a linear programming problem.

The use of different kernels in support vector machines allows one to compute different types of nonlinear functions separating two sets [10, 16, 18].

The concept of max-min separability was introduced in [3]. In this approach two sets are separated using a piecewise linear function. Since a continuous piecewise linear function can be represented as a max-min of linear functions such a separability is called max-min separability. Max-min separability is the generalization of linear, bilinear and polyhedral separabilities. It has been proven that any two finite point sets can be separated by a piecewise linear function. The error function in this case is nonconvex nonsmooth. An algorithm for minimizing the error function is developed. Results presented in [6] demonstrate that the algorithm based on max-min separability is effective for solving supervised data classification problems in many large scale data sets.

In [3, 6] it is assumed that the number of hyperplanes and their distribution over minima functions is known a priori. However, in real situations such information is not available. Moreover, the number of variables in an error function increases as the number of hyperplanes increases and as a result the problem of minimization of the error function becomes large scale optimization problem. Finally, the number of local minimizers of the error function increases as the number of hyperplanes and the number of data points increase. The problem of minimization of the error function becomes a very complicated global optimization problem.

In this paper we propose an incremental approach for computation of a piecewise linear function separating two finite point sets. Given a tolerance such an approach allows us to find as many hyperplanes as necessary to separate sets with respect to this tolerance. Also it allows us to find the "near" global minimizer of the error function. We use the discrete gradient method to minimize the error function. The special structure of the error function, such as piecewise partial separability allows us to modify the discrete gradient method.

We present the results of numerical experiments using several UCI test data sets and compare the proposed algorithm with two support vector machine solvers: LIBSVM ([11]) and SVM_light ([16]). We also present the computational results to demonstrate the ability of the algorithm to compute a piecewise linear function separating two sets.

The structure of this paper is as follows. In Section 2 the definition and some results related to the max-min separability are given. The problem of minimization of an error function is discussed in Section 3. An algorithm for solving max-min separability problems is discussed in Section 4. Results of numerical experiments are presented in Section 5. Section 6 concludes the paper.

2 Max-min separability

In this section we describe the concept of max-min separability and introduce an error function (see [3]).

2.1 Definition and properties

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

$$A = \{a^1, \ldots, a^m\},\ a^i \in \mathbb{R}^n,\ i = 1, \ldots, m,$$

$$B = \{b^1, \ldots, b^p\},\ b^j \in \mathbb{R}^n,\ j = 1, \ldots, p.$$

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 $J = \{1, \ldots, l\}$. For a given $1 \leq r \leq l$ consider any partition $J^r = \{J_1, \ldots, J_r\}$ of this set such that

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

From now on we use the following notation for hyperplanes. If the set of hyperplanes is given without partition we use only one index for x and y. If the set of hyperplanes is given with partition we use two indices for x and y.

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

$$\varphi(z) = \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^{ij}, z \rangle - y_{ij} \right\},\ z \in \mathbb{R}^n. \tag{1}$$

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

Definition 1 *([3]) 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 $J^r = \{J_1, \ldots, J_r\},\ I = \{1, \ldots, r\},\ 1 \leq r \leq l$ of the set J such that*

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

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

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

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

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 (1). Thus the sets A and B can be separated by a function represented as a max-min of linear functions. Therefore this kind of separability is called a max-min separability.

Remark 2 Linear and polyhedral separability, introduced in [8] and [1], respectively, 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 the h-polyhedral separability. Moreover, max-min separability is a generalization of the bilinear separability (see [6]).

Proposition 1 *(see [3]). 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 $J^r = \{J_1, \ldots, J_r\},\ I = \{1, \ldots, r\},\ 1 \le r \le l$ of the set J such that*

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

$$\min_{j \in J_i} \left\{ \langle x^{ij}, a \rangle - y_{ij} \right\} \le -1;$$

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

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

Proposition 2 *(see [3]). The sets A and B are max-min separable if and only if there exists a continuous piecewise linear function separating them.*

Remark 3 It follows from Proposition 2 that the notions of max-min and piecewise linear separability are equivalent.

Proposition 3 *(see [3]). The sets A and B are max-min separable if and only if they are disjoint: $A \bigcap B = \emptyset$.*

Remark 4 Proposition 3 means any two disjoint finite point sets are max-min separable.

The next proposition shows that in most cases the number of hyperlanes necessary for the max-min separation of the sets A and B is not too large.

Proposition 4 *(see [3]). 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

$$co\, A_i \bigcap co\, B_j = \emptyset \ \text{ for all } i = 1, \ldots, q,\ j = 1, \ldots, d. \tag{2}$$

Then the number of hyperplanes necessary for the separation of the sets A and B is at most $q \cdot d$.

Remark 5 Proposition 4 demonstrate that in most cases the cardinality $|J_i|$ of all sets $J_i,\ i \in I$ is the same. If the assumptions of Proposition 4 are satisfied then the cardinality of all these sets is either d or q. We will use this fact for the design of an incremental algorithm in Section 4.

2.2 Error function

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 $J^r = \{J_1, \ldots, J_r\},\ \ I = \{1, \ldots, r\},\ 1 \le r \le l$ of the set J, we say that a point $a \in A$ is well separated from the set B if the following condition is satisfied:

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

Then we can define the separation error for a point $a \in A$ as follows:

$$\max \left[0, \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^{ij}, a \rangle - y_{ij} + 1 \right\} \right]. \tag{3}$$

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

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

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

$$\max \left[0, \min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^{ij}, b \rangle + y_{ij} + 1 \right\} \right]. \tag{4}$$

Thus, an averaged error function can be defined as

$$F(x, y) = (1/m) \sum_{k=1}^{m} \max \left[0, \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^{ij}, a^k \rangle - y_{ij} + 1 \right\} \right]$$

$$+ (1/p) \sum_{t=1}^{p} \max \left[0, \min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^{ij}, b^t \rangle + y_{ij} + 1 \right\} \right] \tag{5}$$

where $x = (x^{11}, \ldots, x^{r|J_r|}) \in \mathbb{R}^{l \times n},\ y = (y_{11}, \ldots, y_{r|J_r|}) \in \mathbb{R}^l$. It is clear that $F(x, y) \ge 0$ for all $x \in \mathbb{R}^{l \times n}$ and $y \in \mathbb{R}^l$.

Proposition 5 *([3]). The sets A and B are max-min separable if and only if there exists a set of hyperplanes $\{x^j, y_j\}, j \in J = \{1, \ldots, l\}$ and a partition $J^r = \{J_1, \ldots, J_r\},\ \ I = \{1, \ldots, r\},\ 1 \le r \le l$ of the set J such that $F(x, y) = 0$.*

Proposition 6 *([3]). 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 $J^r = \{J_1, \ldots, J_r\},\ \ I = \{1, \ldots, r\},\ 1 \le r \le l$ of the set J. Then*

1) $x^{ij} = 0,\ j \in J_i,\ i \in I$ cannot be an optimal solution;

2) if

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

$$\max_{j \in J_t} \left\{ -\langle x^{tj}, b \rangle + y_{tj} + 1 \right\} = \min_{i \in I} \max_{j \in J_i} \left\{ -\langle x^{ij}, b \rangle + y_{ij} + 1 \right\}, \tag{6}$$

(b) 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* $x^{tj} = 0$ *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*

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

Remark 6 The error function (5) is nonconvex and if the sets A and B are max-min separable, then the global minimum of this function $F(x^*, y_*) = 0$ and the global minimizer is not unique.

3 Minimization of error function

The problem of the max-min separability is reduced to the following mathematical programming problem:

$$\text{minimize} \quad F(x, y) \quad \text{subject to} \quad (x, y) \in \mathbb{R}^{(n+1)\times l} \tag{7}$$

where the objective function F has the following form:

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

and

$$f_1(x, y) = \frac{1}{m} \sum_{k=1}^{m} \max \left[0, \max_{i \in I} \min_{j \in J_i} \left\{\langle x^{ij}, a^k \rangle - y_{ij} + 1\right\}\right], \tag{8}$$

$$f_2(x, y) = \frac{1}{p} \sum_{t=1}^{p} \max \left[0, \min_{i \in I} \max_{j \in J_i} \left\{-\langle x^{ij}, b^t \rangle + y_{ij} + 1\right\}\right]. \tag{9}$$

Since both functions f_1 and f_2 are represented as a sum of max-min of linear functions they are locally Lipschitz continuous. Numerical methods of nonconvex, nonsmooth optimization can be applied to solve Problem (7). Most of these methods are based on the concept of the Clarke subdifferential. For a locally Lipschitz continuous function f defined on $\mathbb{R}^n$ the Clarke subdifferential can be defined by

$$\partial f(x) = \text{co} \left\{v \in \mathbb{R}^n : \exists (x^k \in D(f), x^k \to x, k \to +\infty) : v = \lim_{k \to +\infty} \nabla f(x^k)\right\},$$

here $D(f)$ denotes the dense set where f is differentiable [12]. For locally Lipschitz continuous functions the set $\partial f(x)$ is non-empty, convex and compact set. This means that the function F from (7) is subdifferentiable.

The generalized directional derivative of f at x in the direction g is defined as

$$f^0(x, g) = \limsup_{y \to x, \alpha \to +0} \alpha^{-1}[f(y + \alpha g) - f(y)].$$

For the locally Lipschitz function f the generalized directional derivative exists and $f^0(x,g) = \max\{\langle v, g\rangle : v \in \partial f(x)\}$. If the function f is directionally differentiable at x then $f^0(x,g) \geq f'(x,g)$ for all $g \in \mathbb{R}^n$, where $f'(x,g)$ is a derivative of the function f at the point x in the direction g:

$$f'(x,g) = \lim_{\alpha \to +0} \alpha^{-1}[f(x+\alpha g) - f(x)].$$

f is called a regular function at $x \in \mathbb{R}^n$, if it is directionally differentiable at this point and $f'(x,g) = f^0(x,g)$ for all $g \in \mathbb{R}^n$. The function f is non-regular at x if $f^0(x,g) > f'(x,g)$ for at least one $g \in \mathbb{R}^n$. Since in general max-min-type functions are not regular, the function F from (7) is a non-regular function. Calculus for non-regular functions exists as inclusions and this calculus cannot be used to compute subgradients of the function F. In the next subsection we consider a scheme for estimating subgradients of the function F. First, we show that the function F is quasidifferentiable and semismooth.

A function $f : \mathbb{R}^n \to \mathbb{R}^1$ is called semismooth at $x \in \mathbb{R}^n$, if it is locally Lipschitz continuous at x and for every $g \in \mathbb{R}^n$, the limit

$$\lim_{g' \to g, \alpha \to +0, v \in \partial f(x+\alpha g')} \langle v, g\rangle$$

exists [14]. The semismooth function f is directionally differentiable and

$$f'(x,g) = \lim_{g' \to g, \alpha \to +0, v \in \partial f(x+\alpha g')} \langle v, g\rangle.$$

Proposition 7 *The objective function F in (7) is semismooth.*

Proof: The proof follows from the facts that linear functions are semismooth, minimum of semismooth functions is semismooth, maximum of semismooth functions is semismooth and finally, the sum of semismooth functions is again semismooth [14]. Both functions f_1 and f_2 are represented as a sum of max-min of linear functions and therefore they are semismooth and consequently the function F is semismooth. $\triangle$

A function f is called quasidifferentiable at a point x if it is locally Lipschitz continuous, directionally differentiable at this point and there exist convex, compact sets $\underline{\partial} f(x)$ and $\overline{\partial} f(x)$ such that:

$$f'(x,g) = \max_{u \in \underline{\partial} f(x)} \langle u, g\rangle + \min_{v \in \overline{\partial} f(x)} \langle v, g\rangle.$$

The set $\underline{\partial} f(x)$ is called a subdifferential, the set $\overline{\partial} f(x)$ is called a superdifferential and the pair $[\underline{\partial} f(x), \overline{\partial} f(x)]$ is called a quasidifferential of the function f at a point x [13].

Proposition 8 *The objective function F in (7) is quasidifferentiable and its subdifferential and superdifferential are polytopes.*

Proof: In order to prove quasidifferentiability of F it is sufficient to show that functions f_1 and f_2 can be represented as a difference of two convex functions. We will show it only for the function f_1. For a given $k \in \{1, \ldots, m\}$ we introduce

$$\psi_{ki}(x^{ij}, y_{ij}) = \langle x^{ij}, a^k\rangle - y_{ij} + 1, \ j \in J_i, \ i \in I,$$

$$\varphi_{ki}(x,y) = \min_{j\in J_i} \psi_{ki}(x^{ij}, y_{ij}), i \in I,$$

$$\eta_k(x,y) = \max_{i\in I} \min_{j\in J_i} \psi_{ki}(x^{ij}, y_{ij}),$$

$$\bar{\eta}_k(x,y) = \max\{0, \eta_k(x,y)\}.$$

Following [17] we get

$$\varphi_{ki}(x,y) = \varphi^1_{ki}(x,y) - \varphi^2_{ki}(x,y),$$

where

$$\varphi^1_{ki}(x,y) = \sum_{j\in J_i} \psi_{ki}(x^{ij}, y_{ij}),$$

$$\varphi^2_{ki}(x,y) = \max_{j\in J_i} \sum_{t\in J_i, t\neq j} \psi_{kt}(x^{ij}, y_{ij}),\ i \in I.$$

Both functions φ^1_{ki} and φ^2_{ki} are convex piecewise linear functions and their subdifferentials are polytopes. Then we have

$$\eta_k(x,y) = \eta^1_k(x,y) - \eta^2_k(x,y),$$

where

$$\eta^1_k(x,y) = \max_{i\in I} \left(\varphi^1_{ki}(x,y) + \sum_{t\in I, t\neq i} \varphi^2_{kt}(x,y) \right)$$

$$\eta^2_k(x,y) = \sum_{i\in I} \varphi^2_{ki}(x,y).$$

Again both functions η^1_k and η^2_k are convex piecewise linear functions and their subdifferentials are polytopes. Finally, we can write

$$\bar{\eta}_k(x,y) = \bar{\eta}^1_k(x,y) - \bar{\eta}^2_k(x,y),$$

where

$$\bar{\eta}^1_k(x,y) = \max\left(\eta^1_k(x,y), \eta^2_k(x,y)\right),$$

$$\bar{\eta}^2_k(x,y) = \eta^2_k(x,y).$$

Thus, the function f_1 is represented as a difference of two convex functions and its sub and superdifferential are polytopes. $\triangle$

It is clear that the function F is piecewise linear.

Next we will describe an algorithm to approximate subgradients of the function F. This algorithm is introduced in [4, 5]. All necessary proofs can also be found in these papers.

We consider a function f defined on $\mathbb{R}^n$ and assume that this function is quasidifferentiable. We also assume that both sets $\underline{\partial} f(x)$ and $\overline{\partial} f(x)$ are polytopes at any $x \in \mathbb{R}^n$ that is at a point $x \in \mathbb{R}^n$ there exist sets

$$C = \{c^1, \ldots, c^l\},\ \ c^i \in \mathbb{R}^n,\ \ i = 1, \ldots, l, l \geq 1$$

and

$$D = \{d^1, \ldots, d^q\}, \quad d^j \in \mathbb{R}^n, \quad j = 1, \ldots, q, q \geq 1$$

such that

$$\underline{\partial} f(x) = \operatorname{co} C, \quad \overline{\partial} f(x) = \operatorname{co} D.$$

We denote by $\mathcal{F}$ the class of all semismooth, quasidifferentiable functions whose subdifferential and superdifferential are polytopes at any x. The objective function F in (7) belongs to this class.

Let $G = \{e \in \mathbb{R}^n : e = (e_1, \ldots, e_n), |e_j| = 1, \ j = 1, \ldots, n\}$ be a set of all vertices of the unit hypercube in $\mathbb{R}^n$. We take $e \in G$ and consider the sequence of n vectors $e^j = e^j(\alpha), \ j = 1, \ldots, n$ with $\alpha \in (0, 1]$:

$$\begin{array}{rcl} e^1 & = & (\alpha e_1, 0, \ldots, 0), \\ e^2 & = & (\alpha e_1, \alpha^2 e_2, 0, \ldots, 0), \\ \ldots & = & \ldots\ldots\ldots \\ e^n & = & (\alpha e_1, \alpha^2 e_2, \ldots, \alpha^n e_n). \end{array}$$

Let $\lambda > 0$ be a given number. Consider the following points

$$x^0 = x, \quad x^j = x^0 + \lambda e^j(\alpha), \quad j = 1, \ldots, n.$$

It is clear that

$$x^j = x^{j-1} + (0, \ldots, 0, \lambda \alpha^j e_j, 0, \ldots, 0), \quad j = 1, \ldots, n.$$

Let $v = v(\alpha, \lambda) \in \mathbb{R}^n$ be a vector with the following coordinates:

$$v_j = (\lambda \alpha^j e_j)^{-1} \left[f(x^j) - f(x^{j-1}) \right], \quad j = 1, \ldots, n. \tag{10}$$

For any fixed $e \in G$ and $\alpha \in (0, 1]$ we introduce the set:

$$V(e, \alpha) = \left\{ w \in \mathbb{R}^n : \ \exists (\lambda_k \to +0, \quad k \to +\infty), \quad w = \lim_{k \to +\infty} v(\alpha, \lambda_k) \right\}.$$

Proposition 9 *[4, 5]. Assume that $f \in \mathcal{F}$. Then there exists $\alpha_0 > 0$ such that*

$$V(e, \alpha) \subset \partial f(x), \ \forall \ \alpha \in (0, \alpha_0].$$

Remark 7 It follows from Proposition 9 that in order to approximate subgradients of quasidifferentiable functions one can choose a vector $e \in G$, sufficiently small $\alpha > 0, \ \lambda > 0$ and apply (10) to compute a vector $v(\alpha, \lambda)$. This vector is an approximation to a subgradient.

This algorithm allows one to find subgradients. The notion of a discrete gradient can be applied to approximate subsets of the subdifferential and these subsets can be used to find descent directions. The discrete gradient is introduced in [2]. We recall here its definition.

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

$$S_1 = \{g \in \mathbb{R}^n : \|g\| = 1\},$$

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

Here S_1 is the unit sphere and P is the set of univariate positive infinitesimal functions. We take any $g \in S_1,\ e \in G$ and a number $\alpha \in (0, 1]$. Then we define $|g_i| = \max\{|g_k|,\ k = 1, \ldots, n\}$ and the sequence of n vectors $e^j(\alpha),\ j = 1, \ldots, n$ as above. For given $x \in \mathbb{R}^n$ and $z \in P$ consider a sequence of $n + 1$ points:

$$\begin{array}{lll} x^0 = & x+ & \lambda g, \\ x^1 = & x^0+ & z(\lambda)e^1(\alpha), \\ \ldots = & \ldots & \ldots \\ x^n = & x^0+ & z(\lambda)e^n(\alpha). \end{array}$$

Definition 2 *The discrete gradient of the function f at the point $x \in \mathbb{R}^n$ is the vector $\Gamma^i(x, g, e, z, \lambda, \alpha) = (\Gamma^i_1, \ldots, \Gamma^i_n) \in \mathbb{R}^n, g \in S_1$ with the following coordinates:*

$$\Gamma^i_j = [z(\lambda)\alpha^j e_j)]^{-1} \left[f(x^j) - f(x^{j-1}) \right], \quad j = 1, \ldots, n, \quad j \neq i,$$

$$\Gamma^i_i = (\lambda g_i)^{-1} \left[f(x + \lambda g) - f(x) - \lambda \sum_{j=1, j \neq i}^{n} \Gamma^i_j g_j \right].$$

Remark 8 One can see that the discrete gradient is defined with respect to a given direction $g \in S_1$ and in order to compute it, first we define points $x^0, \ldots, x^n$ and compute the values of the function f at these points that is we compute $n+2$ values of this function including the point x.

For a given $\alpha > 0$ we define the following set:

$$B(x, \alpha) = \{v \in \mathbb{R}^n : \ \exists (g \in S_1,\ e \in G,\ z_k \in P,\ z_k \to +0,\ \lambda_k \to +0,\ k \to +\infty),$$
$$v = \lim_{k \to +\infty} \Gamma^i(x, g, e, z_k, \lambda_k, \alpha)\}. \tag{11}$$

Proposition 10 *Assume that $f \in \mathcal{F}$. Then there exists $\alpha_0 > 0$ such that*

$$co\, B(x, \alpha) \subset \partial f(x),\ \forall\ \alpha \in (0, \alpha_0].$$

Remark 9 Since the objective function F in (7) is piecewise linear we use a simplified scheme described in [7] to compute its discrete gradients. This scheme allows us to use only two instead of $n + 2$ evaluations of the function F to compute one discrete gradient.

In this section we described an algorithm to approximate subgradients of the objective F in (7). Results of this section demonstrate that the discrete gradient method described in [4, 5] can be used to minimize the error function F.

4 An incremental algorithm

The number of hyperplanes l necessary to separate two sets is not known a priori. In this section we suggest an algorithm for the computation of a piecewise linear function separating two sets and this algorithm computes hyperplanes incrementally. It computes as many hyperplanes as necessary for separating the sets with respect to a given tolerance.

Following Proposition 4 we assume that the sets $J_i,\ i \in I$ have the same cardinality. Let $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ be tolerances.

Algorithm 1 An incremental algorithm

Step 0. (Initialization) Select any starting point $(x^1, y_1),\ x^1 \in \mathbb{R}^n,\ y_1 \in \mathbb{R}^1$. Set $X^1 = (x^1, y_1),\ I_1 = \{1\},\ J_1^1 = \{1\},\ f_1 = f(x^1, y_1),\ r_1 = |I| = 1,\ d_1 = |J_1| = 1$, the number of hyperplanes $l = 1$ and $k = 1$.

Step 1. (Computation of a piecewise linear function) Solve Problem (7) starting from the point $X^k \in \mathbb{R}^{(n+1)\times l}$. Let $X^{k,*}$ be a solution to this problem, F_k^* is the corresponding objective function value, $f_{1,k}^*$ and $f_{2,k}^*$ are the values of functions f_1 and f_2, respectively.

Step 2. (The first stopping criterion) If $f_{1,k}^* \le \varepsilon_1$ and $f_{2,k}^* \le \varepsilon_1$ then stop. $X^{k,*}$ is a final solution.

Step 3. (The second stopping criterion) If $k \ge 2$,

$$f_{1,k-1}^* - f_{1,k}^* \le \varepsilon_2$$

and

$$f_{2,k-1}^* - f_{2,k}^* \le \varepsilon_2$$

then stop. $X^{k,*}$ is a final solution.

Step 4. (Adding new hyperplanes)

1. If $f_{1,k}^* > \varepsilon_1$, then set $d_{k+1} = d_k + 1,\ J_i^{k+1} = J_i^k \bigcup \{d_{k+1}\}$ for all $i \in I_k$. Set $x^{ij} = x^{i,j-1,*},\ i \in I_k,\ j = d_{k+1}$.

2. If $f_{2,k}^* > \varepsilon_1$, then set $r_{k+1} = r_k + 1,\ I_{k+1} = I_k \bigcup \{r_{k+1}\},\ J_{r_{k+1}}^k = J_{r_k}^k$. Set $x^{ij} = x^{i-1,j,*},\ i = r_{k+1},\ j \in J_{r_k}^k$.

Step 5. (New starting point) Set $X^{k+1} = \left(x^{ij},\ j \in J_i^{k+1},\ i \in I_{k+1}\right)$, $k = k + 1$ and go to Step 1.

Explanations to Algorithm 1. The algorithm starts by computing one hyperplane to separate sets (Steps 0 and 1). There are two different stopping criteria in this algorithm. The stopping criterion in Step 2 means that the computed piecewise linear function separates two sets with the tolerance $\varepsilon_1 > 0$. The stopping criterion in Step 3 implies that adding new hyperplanes the algorithm cannot significantly decrease the value of the error

function. This may happen when a large number of hyperlanes are needed to separate sets. Such a criterion allows one to avoid problems with overfitting in supervised data classification problems. However, this stopping criterion does not mean that a piecewise linear function separating two sets has been computed. Step 4 provides rules for adding new hyperplanes and defining their normal vectors. These vectors are defined to guarantee a decrease of the error function in the next iteration compared to the current iteration. Step 5 defines a starting point for the minimization of the error function for the next iteration. Since the problem of minimization of the error function is a global optimization problem, such a strategy allows us to find the "near" global solution.

5 Results of numerical experiments

In this section we present the results of numerical experiments. We consider two different applications of the proposed algorithm. In the first application we apply the algorithm to find a piecewise linear function separating two sets. The aim is to demonstrate the ability of the algorithm to find such piecewise linear functions. Then we apply the algorithm to solve the supervised data classification problems. We also use two SVM solvers to compare results. The SVM solvers used in this paper are: LIBSVM and SVM_light. LIBSVM can be found on: http://www.csie.ntu.edu.tw/ cjlin/libsvm. More detailed information is given in [11]. SVM_light can be found on: http://svmlight.joachims.org and the detailed information on it is given in [16].

5.1 Data sets

The brief description of the data sets used in our numerical experiments is given in Table 1. A more detailed description can be found in [15]. We consider data sets with only continuous attributes. Table 1 contains information about the number of instances in both training and test sets, as well as the number of attributes (including class attribute) and classes.

5.2 Computation of piecewise linear functions

In this subsection we apply the proposed incremental algorithm to separate two finite point sets. We compare the max-min and the linear separability results. These results are presented in Table 2. Accuracy of the separation is defined as a ratio of the number of well-classified points to the total number of data points (in %). We also give the number of hyperplanes in the case of the max-min separability (for the linear separability it is one).

For data sets with two classes we compute only one piecewise linear function, however for data sets with more than two classes we use one vs. the rest strategy and compute a piecewise linear function separating one class from all others. As a result the number of the piecewise linear functions coincides with the number of classes. In Table 2 we present the number of hyperplanes for each of the classes for the data sets with more than two classes.

Table 1: Brief description of data sets

Data sets	(train,test)	No. of attributes	No. of classes
Shuttle control (SH)	(43500,14500)	10	7
Letter recognition (LET)	(15000,5000)	17	26
Landsat satellite image (LSI)	(4435,2000)	37	6
Pen-based recognition of handwritten (PD)	(7494,3498)	17	10
Page blocks (PB)	(4000,1473)	11	5
Optical recognition (OD)	(3823, 1797)	65	10
Spambase (SB)	(3682,919)	58	2
Image segmentation (SEG)	(1848,462)	20	7
Yeast (YEAST)	(1191, 293)	9	10
Vehicle (VEH)	(679,167)	19	4
German credit (GC)	(999,1)	25	2
Australian credit (AC)	(689,1)	15	2
Breast cancer (BC)	(682,1)	10	2

Results presented in the first and the third columns of Table 2 demonstrate that the use of max-min separability allows one to significantly improve the results obtained by using only one hyperplane that is using the linear separability. The use of max-min separability provides a much clearer picture on the structure of a data set. Some of these data sets, for example the yeast data set, have a quite complex structure, however the breast cancer data set has one large cluster for each class and only a few points around them. The number of hyperplanes necessary for separating two sets corresponds to the number $q \times d$ from Proposition 4 and this number contains information on the structure of classes in a data set.

5.3 Supervised data classification via max-min separability

In this subsection we apply the proposed incremental algorithm to solve supervised data classification problems. Computational results are presented in Tables 3 and 4. We also present results using LIBSVM (C-SVM and nu-SVM) and SVM_light with different kernels. Dash lines in the tables show that either an algorithm failed to solve a classification problem on a data set or a result on this data set is not available.

The code of the incremental max-min separability algorithm was written in Lahey Fortran 95. All computations (including LIBSVM and SVM_light) were carried out on a Pentium IV Intel(R) with CPU 1.83 GHz and 1.00 GB of RAM.

We restricted the number of hyperplanes in the incremental algorithm by 10 and the training set was divided into two parts. The first part was used to compute piecewise linear functions with different number of hyperplanes, whereas the second part was used to choose

Table 2: Results on the separation of sets

Data sets	Max-min sep.		Lin. separation
	Accuracy	No.of hyper.	Accuracy
PD	99.86	(9,12,12,4,3,4,8,8,6,6)	93.30
PB	95.12	(25,9,8,16,3)	91.23
SEG	99.35	(2,8,1,1,2,16,12)	91.17
YEAST	73.85	(49,49,36,6,20,12,49,24,6,1)	52.36
VEH	99.53	(16,25,4,8)	80.73
GC	97.00	25	73.60
AC	97.10	36	86.09
BC	100.00	12	97.51

a piecewise linear function with the best performance. For German credit, Australian credit and Breast Cancer data sets the leave-one-out estimate was used, whereas for all other data sets one training and one test set were used.

Table 3: Results for supervised data classification problems

Algorithms	SH	LET	LSI	PD	PB	SEG
Max-min separability	99.83	90.24	87.55	97.03	85.54	95.24
Linear separability	88.55	65.74	82.85	90.02	82.89	92.21
C-SVM-linear	97.24	84.30	85.75	95.63	87.64	95.02
C-SVM-polynomial	82.32	37.58	66.70	96.37	87.44	90.69
C-SVM-radial	97.61	82.24	85.35	97.83	87.17	93.94
C-SVM-sigmoid	96.04	77.24	83.30	89.62	88.93	90.69
nu-SVM-linear	-	77.06	79.25	82.88	-	92.21
nu-SVM-polynomial	-	77.52	76.50	88.85	-	91.34
nu-SVM-radial	-	78.98	80.00	84.59	-	92.64
nu-SVM-sigmoid	-	76.10	79.35	84.59	-	92.21
SVM_light-linear	89.62	55.18	75.80	85.68	84.86	-
SVM_light-polynomial	79.16	77.26	-	96.43	-	-
SVM_light-radial	79.54	78.56	-	-	84.79	-
SVM_light-sigmoid	68.25	-	-	-	84.52	-

One can draw the following conclusions from the results presented in Tables 3 and 4.

1. The incremental algorithm based on max-min separability produces either the best results or results comparable with the best ones. These results confirm that the algorithm is reliable and an effective algorithm for solving supervised data classification problems.

Table 4: Results for supervised data classification problems (cont.)

Algorithms	AC	GC	BC	SPAM	VEH	Yeast
Max-min separability	85.36	72.70	96.93	92.06	81.44	57.34
Linear separability	85.94	73.00	96.78	92.93	76.05	47.78
C-SVM-linear	85.51	77.18	96.93	91.64	76.05	57.34
C-SVM-polynomial	85.36	74.27	97.36	87.92	47.31	39.93
C-SVM-radial	85.51	76.18	97.07	87.05	71.86	55.97
C-SVM-sigmoid	85.80	76.68	96.78	78.45	62.27	51.19
nu-SVM-linear	85.51	75.38	95.31	88.25	71.26	-
nu-SVM-polynomial	85.51	75.58	96.49	88.36	75.45	-
nu-SVM-radial	85.51	77.28	96.63	88.57	75.45	-
nu-SVM-sigmoid	85.51	72.87	94.58	89.23	73.65	-
SVM_light-linear	85.51	78.00	97.22	-	-	-
SVM_light-polynomial	87.39	84.00	97.36	-	-	-
SVM_light-radial	91.59	72.50	92.09	-	-	-
SVM_light-sigmoid	57.10	64.30	-	-	-	-

2. A comparison of the results obtained by the piecewise linear and the linear separability for multi-class data sets clearly demonstrates that the use of the piecewise linear separability significantly improves results by the linear separability.

3. Among SVM algorithms C-SVM-linear produced very good results for data sets considered in this paper. The choice of kernels strongly depends on data sets. For example, C-SVM-polynomial gives a very good result for the Page blocks data set, however it fails for the Land Satellite Image data set.

4. The incremental algorithm based on max-min separability is especially effective for solving supervised data classification problems in multi-class large scale data sets.

5. The algorithm based on max-min separability is much more time consuming than any other algorithm considered in this paper.

6 Conclusions and further work

In this paper we have developed a new algorithm for the computation of a piecewise linear function separating two finite point sets. This algorithms computes hyperplanes incrementally and it computes as many hyperplanes as necessary with respect to a given tolerance.

The error function in max-min separability is nonconvex and nonsmooth. The discrete gradient method from [4, 5] is applied to minimize it. A simplified scheme is applied to compute discrete gradients. An incremental approach allows one to find a good starting

point for the next iteration of the algorithm and as a result the algorithm finds either global or the "near" global solution.

The computational results presented in this paper demonstrate that the incremental algorithm is a good alternative to the very powerful SVM algorithms. However it requires more CPU time. Different approaches can be used to significantly reduce CPU time. One such approach is to modify the discrete gradient method for minimization of the error function exploiting its piecewise separability. Another approach can be the use of only those data points which contribute to the error function. Since we use an incremental approach the number of such data points will be reduced as the number of hyperplanes increase. These all will be the subject of our future research.

Acknowledgements

Dr. Adil Bagirov is the recipient of an Australian Research Council Australian Research Fellowship (Project number: DP 0666061). Dr. Adil Bagirov and Prof. Bülent Karasözen are thankful for the support of TUBITAK(Turkish Scientific and Technical Research Council) and the Australian Mathematical Sciences Institute which initiated this current work by their supporting mutual visits.

References

[1] Astorino, A. and Gaudioso, M. (2002), Polyhedral separability through successive LP, *Journal of Optimization Theory and Applications,* 112(2), pp. 265-293.

[2] 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*, Kluwer Academic Publishers, pp. 147-175.

[3] Bagirov, A.M. (2005), Max-min separability, *Optimization Methods and Software*, 20(2-3), pp. 271-290.

[4] Bagirov, A.M., Ghosh, M. and Webb, D. (2006) A derivative-free method for linearly constrained nonsmooth optimization, *Journal of Industrial and Management Optimization,* 2(3): pp. 319-338.

[5] Bagirov, A.M., Karasözen, B. and Sezer, M. (2007) Discrete gradient method: a derivative free method for nonsmooth optimization, *Journal of Optimization Theory and Applications,* accepted for publication.

[6] Bagirov A.M. and Ugon, J. (2005) Supervised data classification via max-min separability, In: A.M. Rubinov and V. Jeyakumar (eds), *Trends in Continuous Optimization, Applied Optimization,* Vol. 99, Springer, Dordrecht, pp. 175-208

[7] Bagirov, A.M. and Ugon, J. (2006) Piecewise partially separable functions and a derivative-free method for large-scale nonsmooth optimization. *Journal of Global Optimization* 35: pp.163-195.

[8] Bennet, K.P.and Mangasarian, O.L. (1992), Robust linear programming discrimination of two linearly inseparable sets, *Optimization Methods and Software*, 1, pp. 23-34.

[9] Bennet, K.P. and Mangasarian, O.L. (1993), Bilinear separation of two sets in n-space, *Computational Optimization and Applications*, 2(3), pp. 207-227.

[10] Burges, C.J.C. (1998), A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 2, pp. 121-167.

[11] Chih-Chung Chang and Chih-Jen Lin, (2001) LIBSVM: a library for support vector machines.

[12] Clarke, F.H. (1983), *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York.

[13] Demyanov, V.F. and Rubinov, A.M. (1995) *Constructive Nonsmooth Analysis*, Peter Lang, Frankfurt am Main, 1995.

[14] Mifflin, R. (1977), Semismooth and semiconvex functions in constrained optimization, *SIAM Journal on Control and Optimization,* 15, pp. 957-972.

[15] Newman, D.J., Hettich, S., Blake, C.L. and Merz, C.J. (1998), UCI repository of machine learning databases. Technical report, Department of Information and Computer science, University of California, Irvine, 1992. www.ics.uci.edu/MLRepository.html.

[16] Thorsten, J.(2002), *Learning to Classify Text Using Support Vector Machines*, Kluwer Academic Publishers, Dordrecht, 2002.

[17] Tuy, H. (1999) *Convex Analysis and Global Optimization,* Kluwer Academic Publishers, Dordrecht, Boston.

[18] Vapnik, V.N. (1995), *The Nature of Statistical Learning Theory,* Springer, New York.