CLASSIFICATION PROBLEMS AND NONLINEAR OPTIMIZATION: SEPARATION BY MEANS OF NONLINEAR SURFACES

A. ASTORINO*, A. FUDULI†, M. GAUDIOSO‡, AND E. GORGONE§

Abstract. Classification is a machine learning technique aimed at assigning exactly one from among several classes to each sample in a population, the sample being identified as a vector of numerical values of a certain number of attributes. The classification process works as follows. Starting from a dataset of samples with known class membership, a mathematical tool is developed (the classifier) which is able to predict class membership for any newly incoming sample.

Construction of a classifier requires in general solution of a numerical optimization problem to detect an "optimal" separation surface in the space of the samples. Most of the classification methods (e.g. the classic Support Vector Machine one) are based on the use of separating hyperplanes.

In this paper, instead, we focus on some recent classification techniques which use nonlinear separation surfaces. We survey several approaches of both the supervised and the semisupervised type.

Key words. Classification, Machine Learning, Nonlinear Optimization, Nonsmooth Optimization

1. Introduction.

The objective of classification, a relevant chapter of machine learning, is to categorize data into different classes on the basis of their similarities. The field has become more and more relevant as possible applications touch a quite wide range of practical problems in areas such as text and web classification, object recognition in machine vision, gene expression profile analysis, DNA and protein analysis, medical diagnosis, customer profiling etc.

We focus on the binary classification framework where we are given two sets of data expressed in the form of two sets of points in any given n-dimensional space. Each point (sample) is representative of an individual characterized by n quantitative parameters (features), which are assumed to be real numbers and, consequently, each individual is completely represented by a vector in $\mathbb{R}^n$. The objective of any binary classification model is to construct a criterion (the so-called classifier) for discriminating between the two sets, in order to categorize any new unlabeled point as a point belonging exactly to one of them. The ability to correctly classify a new point is called the "generalization capability".

Classification deals, consequently, with separation of sets in finite dimensional spaces by means of appropriate separation surfaces. Separation of sets has been a very important field of interest in mathematics, we just recall here the celebrated Hahn-Banach theorem. In particular, separability of two sets by means of a hyperplane is a property of the geometry of the sets connected to the notion of convex set. Thus linear separability of sets [31, 32, 16] has been at the basis of the most popular approach to

*Istituto di Calcolo e Reti ad Alte Prestazioni-C.N.R., 87036 Rende (CS), Italia. E-mail: astorino@icar.cnr.it

†Dipartimento di Matematica e Informatica, Università della Calabria, 87036 Rende (CS), Italia. E-mail: antonio.fuduli@unical.it

‡Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia. E-mail: gaudioso@dimesunical.it

§Faculté des Sciences, Université Libre de Bruxelles, B-1050 Bruxelles, Belgium. E-mail: enrico.gorgone@ulb.ac.be

classification, the Support Vector Machine (SVM) model [41, 22], where one looks for a hyperplane separating the two given sample sets. Unfortunately data sets coming from practical applications are very rarely linearly separable, consequently construction of a classifier requires the choice of a hyperplane which minimizes an error function penalizing point misclassification. For this reason, in the last years many researchers coming from the mathematical programming community entered the machine learning field, providing optimization tools aimed at minimizing such an error function.

In particular the SVM approach consists in calculating a hyperplane which both minimizes the classification error and maximizes the width of the no-man-land between the two sets, in the sense that the separation margin (the width of the strip around the separating hyperplane where no point of the two sets falls) is maximized. This objective can be achieved by solving a structured quadratic programming problem and many efforts have been devoted in last years to design efficient algorithms for such typically large scale optimization problems [27, 35]. Margin maximization is aimed at improving generalization capability of the classifier, that is its ability to classify correctly the new incoming samples.

Parallel to the development of SVM methods, the use of nonlinear separating surfaces in the dataset space, instead of hyperplanes, has received in recent years some attention. The objective of the paper is to survey such proposals, mainly in terms of the numerical optimization algorithms which are required to construct the surfaces. In particular we will consider the following approaches:

- Polyhedral separation
- Ellipsoidal separation
- Spherical separation
- Conical separation

All of them are characterized by the fact that the separation process takes place in the original input space and does not require mapping to higher dimension spaces.

Most of the classification methods we consider in this paper are machine learning techniques of the supervised type (clustering, instead, is a typical unsupervised technique). In our framework, a classifier is constructed by assuming that the class membership of all samples is known.

More recently semisupervised approaches have been intensively studied as well. They stem from the observation that in many applications both labeled and unlabeled samples are available; consequently the objective is to let both these two families of data enter into classifier construction. The main advantage is that quite often in practical cases most of the samples are unlabeled and then it appears appealing to exploit the entire available information.

The semisupervised approach has been successfully used in classification algorithms based on separation by means of a hyperplane (see for example [28, 29, 18, 2]). Some attempts to use the semisupervised approach also in the framework of nonlinear surfaces are in [3, 4]. In this paper we will survey also some semisupervised methods in such area.

We now introduce more formally the definitions and notations in use throughout the paper. Let

$$\mathcal{X} = \{x_1, \dots, x_p\}$$

be a set of samples (or points) $x_i \in \mathbb{R}^n$.

In the supervised learning, we assume that, corresponding to any point x_i of $\mathcal{X}$, a label y_i is given. The case $y_i \in \mathbb{R}$ is known as "regression", while, when the label y_i takes values in a discrete finite set, we come out with a classification problem. A particular case of the latter is the binary classification, where, for each i, the label y_i can assume only two possible values, +1 or -1, representing in fact for each sample its class membership. The objective of the supervised learning is to predict the label of any new sample only on the basis of the information of the labeled points (the training set).

As for the semisupervised learning, we assume that the set $\mathcal{X}$ is partitioned into two sets

$$\mathcal{X}_L = \{(x_i, y_i) \mid x_i \in \mathbb{R}^n,\ y_i \in \{+1, -1\}, \quad i = 1, \dots, p\},$$

and

$$\mathcal{X}_U = \{x_i \mid x_i \in \mathbb{R}^n, \quad i = p+1, \dots, p+k\},$$

which are the sets of the labeled and unlabeled points, respectively, the latter being those points whose class membership is unknown. Then the classifier is obtained on the basis of the information coming from both $\mathcal{X}_U$ and $\mathcal{X}_L$. This may be appealing especially when the number of unlabeled data is much more bigger than that of the labeled ones and when to obtain the labels on the data is very costly: this is the case, for example, of webpage classification and of speech recognition.

In setting the various classification models, we partition X_L into two nonempty sets:

$$\mathcal{X}_+ := \{(x_i, y_i) \mid x_i \in \mathbb{R}^n,\ y_i = +1, \quad i = 1, \dots, m\}$$

and

$$\mathcal{X}_- := \{(x_i, y_i) \mid x_i \in \mathbb{R}^n,\ y_i = -1, \quad i = m+1, \dots, p\}.$$

Throughout the paper we use the following notation. We denote by $\| \cdot \|$ the Euclidean norm in $\mathbb{R}^n$ and by $a^T b$ the inner product of the vectors a and b. Moreover, in correspondence to a finite set of points $\mathcal{A}$ in $\mathbb{R}^n$, we indicate by $|\mathcal{A}|$ its cardinality and by $\mathrm{conv}(\mathcal{A})$ its convex hull.

The paper is organized as follows. In section 2 we review some different approaches to supervised binary classification based on the use of nonlinear surfaces. In particular, we describe the polyhedral separation in 2.1, ellipsoidal separation in 2.2, spherical separation in 2.3 and conical separation in 2.4. In section 3 are presented the semisupervised counterpart of two binary classification approaches based, respectively, on polyhedral separation in 3.1 and on spherical separation in 3.2.

2. Supervised approaches. In this section we survey some approaches to binary classification in the supervised setting based on the use of nonlinear separating surfaces. They share the common characteristic that the optimization problem to be solved in order to find such surface is no longer a quadratic programming problem (as in the SVM approach) and it is often of the nonsmooth optimization type [7].

In next subsections we consider separation by means of polyhedral, ellipsoidal, spherical and conical surfaces in $\mathbb{R}^n$.

2.1. Polyhedral separation. The concept of polyhedral separability has been introduced in [33] and developed more recently in [8, 14, 26, 34]. In particular we assume that the sets $\mathcal{X}_+$ and $\mathcal{X}_-$ are polyhedrally separable if one set is contained in a polyhedron and the points of the other one are left outside it. In the sequel we briefly recall the approach used in [8], starting from the following definition.

Definition 2.1. *The set $\mathcal{X}_+$ is h-*polyhedrally separable *from $\mathcal{X}_-$ if and only if there exists a set of h hyperplanes $\{w^{(j)}, b_j\}$, with $w^{(j)} \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$, $j = 1, \dots, h$, such that*

- *$x_i^T w^{(j)} + b_j \leq -1$, for all $i = 1, \dots, m$ and for all $j = 1, \dots, h$;*
- *for all $i = m+1, \dots, p$, there exists an index $j \in \{1, \dots, h\}$ such that $x_i^T w^{(j)} + b_j \geq 1$.*

On the basis of previous definition, a polyhedron separating $\mathcal{X}_+$ from $\mathcal{X}_-$ is computed as the intersection of h halfspaces associated to h hyperplanes.

The following property has been proved in [8].

Theorem 2.2. *The set $\mathcal{X}_+$ is h-polyhedrally separable from $\mathcal{X}_-$, with $h \leq |\mathcal{X}_-|$, if and only if*

$$\operatorname{conv}(\mathcal{X}_+) \cap \mathcal{X}_- = \emptyset.$$

We remark that the role played by the two sets is not symmetric: in fact separating $\mathcal{X}_+$ from $\mathcal{X}_-$ is not the same as separating $\mathcal{X}_-$ from $\mathcal{X}_+$.

Once h has been fixed, a polyhedral separation of $\mathcal{X}_+$ and $\mathcal{X}_-$ can be obtained by minimizing the following error function:

$$\begin{aligned} z_P(\mathbf{w}, \mathbf{b}) \quad := \quad & \frac{1}{|\mathcal{X}_+|} \sum_{i=1}^{m} \max_{j=1,\dots,h} \{0, y_i(x_i^T w^{(j)} + b_j) + 1\} + \\ & \frac{1}{|\mathcal{X}_-|} \sum_{i=m+1}^{p} \max\{0, \min_{j=1,\dots,h} [y_i(x_i^T w^{(j)} + b_j) + 1]\}, \end{aligned} \tag{2.1}$$

where $\mathbf{w}^T := [w^{(1)T}, w^{(2)T}, \dots, w^{(h)T}]$ and $\mathbf{b}^T := [b_1, b_2, \dots, b_h]$. Note that the weights $\frac{1}{|\mathcal{X}_+|}$ and $\frac{1}{|\mathcal{X}_-|}$ have been introduced to account for unbalanced cardinalities of the two sets $\mathcal{X}_+$ and $\mathcal{X}_-$.

Separation is achieved whenever a solution $(\mathbf{w}, \mathbf{b})$ is found such that $z_P(\mathbf{w}, \mathbf{b}) = 0$. In general we minimize function (2.1) and stay content with any optimal solution $(\mathbf{w}^*, \mathbf{b}^*)$, even when $z_P(\mathbf{w}^*, \mathbf{b}^*) > 0$.

Function z_P is nondifferentiable; moreover it is nonconvex and nonconcave, but it is quasi-concave. Its minimization corresponds to a min-max-min type problem and then, differently from the linear separability case, it cannot be transformed into a linear program. Embedding some ideas coming from nonsmooth optimization, in [8] an algorithm where, at each iteration, the descent direction is obtained by solving a sequence of linear programs was proposed.

Some variants of the above model are described in the literature. In [34] a mixed binary programming framework is introduced and a polyhedral classifier based on the trade-off between accuracy and generalization potential is discussed. In particular misclassification measure, separation margin and number of active attributes are simultaneously considered.

Polyhedral conic functions are adopted in [26] to construct separation surfaces by solving a sequence of linear programs. At first stage a polyhedral conic function is obtained to partition the whole input space into two parts such that as many as possible points of $\mathcal{X}_+$ are inside the corresponding sublevel set and all points of $\mathcal{X}_-$ remain outside. The points of $\mathcal{X}_+$ correctly classified are now excluded from further consideration and the process is repeated on the remaining ones, thus generating a new separating function. The process continues until all points in $\mathcal{X}_+$ have been located.

Finally maxmin separation has been introduced in [13], where two sets are separated by means of a continuous piecewise linear function. In fact polyhedral separation can be considered a special case of the maxmin one. Polyhedral conic functions and maxmin classification have been jointly adopted in [14].

2.2. Ellipsoidal separation. A nonlinear classifier based on ellipsoidal separation has been developed in [9]. Other methods available in the literature are aimed at calculating the minimal volume ellipsoid enclosing all the points of the set $\mathcal{X}_+$ (see for example [36] and [37]).

Now we focus on the approach proposed in [9]. It consists in finding an ellipsoid enclosing all points of $\mathcal{X}_+$ and no point of $\mathcal{X}_-$. More specifically, the set $\mathcal{X}_+$ is *ellipsoidally separable* from $\mathcal{X}_-$ if and only if there exists an ellipsoid

$$E(Q) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n \mid \ (x-x_0)^T Q (x-x_0) \le 1\},$$

centered in $x_0 \in \mathbb{R}^n$, with $Q \in \mathbb{R}^{n\times n}$ symmetric and positive definite, such that

$$y_i[(x_i-x_0)^T Q (x_i-x_0) - 1] \le 0 \quad i=1,\dots,p.$$

The problem of finding a separation ellipsoid is not always feasible: such an ellipsoid might exist only if the intersection of the convex hull of $\mathcal{X}_+$ with the set $\mathcal{X}_-$ is empty; in fact, differently from the polyhedral separability case, this condition is only necessary but not sufficient.

A possible approach is to minimize an appropriate measure of the classification error, looking for an ellipsoid enclosing as many as possible points of $\mathcal{X}_+$ and as few as possible points of $\mathcal{X}_-$.

If the center x_0 is fixed (it could be selected as any centroid of the set $\mathcal{X}_+$), the ellipsoidal separation of $\mathcal{X}_+$ from $\mathcal{X}_-$ can be obtained by minimizing the following error function:

$$z_E(Q) \stackrel{\triangle}{=} \sum_{i=1}^{p} \max\{0, y_i[(x_i-x_0)^T Q (x_i-x_0) - 1]\}$$

which is nonnegative, convex and piecewise affine, being sum of nonnegative, convex and piecewise affine functions. As a consequence the problem

$$\min_{Q\in\mathbb{R}^{n\times n}, Q\succ 0} \ z_E(Q) \tag{2.2}$$

is a convex nonsmooth minimization problem, since the objective function z_E is convex and the space of the symmetric and positive definite matrices is convex as well [39].

Problem (2.2) can be solved exactly by any semidefinite programming software (for example [40]). In [9] a heuristic iterative descent algorithm of the local search type [42] has been presented. In fact, given a symmetric and positive definite matrix $Q^{(k)}$ (the estimate of an optimal solution available at the iteration k), matrix $Q^{(k+1)}$ is computed as

$$Q^{(k+1)} = Q^{(k)} + \alpha_k a_k a_k^T,$$

where $a_k \in \mathbb{R}^n$ is obtained by solving a quadratic program aimed at finding a descent direction, while $\alpha_k \in \mathbb{R}$ is an appropriate positive stepsize. In passing from $Q^{(k)}$ to $Q^{(k+1)}$, a sufficient decrease condition on the objective function is satisfied.

Once x_0 has been fixed and the corresponding optimal value of Q has been computed, a further improvement of the classification error could be obtained by moving the center of the ellipsoid. This can be done redefining the classification error as function of x_0 and minimizing the following error function:

$$z_{E_0}(x_0) \overset{\triangle}{=} \sum_{i=1}^{p} \max\{0, y_i[(x_i - x_0)^T Q(x_i - x_0) - 1]\},$$

which is nonsmooth and nonconvex. In [9] the minimization of z_{E_0} is performed by calculating, at the current point, the steepest descent direction. However, the overall proposed approach is based on the alternation between the minimization of z_E, keeping x_0 fixed, and that of z_{E_0}, keeping Q fixed.

An alternative approach, where several ellipsoids are adopted for classification purposes, is described in [23].

2.3. Spherical separation. A particular case of ellipsoidal separation is the spherical one, obtained taking $Q = (1/R^2)I$, where $R > 0$ is the radius of the separating sphere and I is the identity matrix.

In this case the set $\mathcal{X}_+$ is *spherically separable* from $\mathcal{X}_-$ if and only if there exists a sphere $S(x_0, R)$, centered in $x_0 \in \mathbb{R}^n$ of radius $R \in \mathbb{R}$, such that

$$y_i(\|x_i - x_0\|^2 - R^2) \leq 0 \quad i = 1, \ldots, p. \tag{2.3}$$

In addition, when inequalities (2.3) are strictly satisfied, then the sets $\mathcal{X}_+$ and $\mathcal{X}_-$ are strictly spherically separated, i.e.

$$\begin{array}{ll} \|x_i - x_0\|^2 \leq (R - M)^2 & i = 1, \ldots, m \\ \|x_i - x_0\|^2 \geq (R + M)^2 & i = m+1, \ldots, p, \end{array} \tag{2.4}$$

for some M, the margin, such that $0 < M \leq R$. Setting $q \overset{\triangle}{=} 2RM$ and $r \overset{\triangle}{=} R^2 + M^2$, inequalities (2.4) become:

$$q + y_i(\|x_i - x_0\|^2 - r) \leq 0 \quad i = 1, \ldots, p.$$

In general it is not easy to know in advance whether the two sets are strictly spherically separable; then in [10, 5, 6, 30] a classification error function has been defined in order to find a minimal error separating sphere.

In particular, in [10, 5] a separating sphere has been obtained by minimizing the following objective function:

$$z_{S_1}(x_0, r) \overset{\triangle}{=} r + C\sum_{i=1}^{p} \max\{0, y_i(\|x_i - x_0\|^2 - r)\},$$

with $M = 0$ and $r \geq 0$. Note that in this model it is embedded also the objective of minimizing the volume of the sphere. This is accounted for by adding to the error measure also the radius, of course by means of a positive tradeoff parameter C.

More precisely in [10] an ad hoc algorithm that finds the optimal solution in $O(t \log t)$, with $t = \max\{m, p - m\}$, has been presented for the case where the center x_0 is fixed. It basically consists of two phases: the sorting phase and the cutting phase. In the first one the sample points are sorted according to their distance from the center, while in the second one an optimal cut is found. The adopted simplification is rather drastic, nevertheless a judicious choice of the center (e.g. the barycenter of the set $\mathcal{X}_+$) has allowed to obtain reasonably good separation results at a very low computational cost.

In [5] the spherical separation problem has been tackled without any constraint in the location of the center by means of DCA (DC Algorithm) [38, 1], based on a DC (Difference of Convex) decomposition of the objective function. A similar DCA approach has been proposed in [30] for minimizing the following error function:

$$z_{S_2}(x_0, r) \overset{\triangle}{=} r^2 + C\sum_{i=1}^{p} \max\{0, y_i(\|x_i - x_0\|^2 - r)\}^2,$$

with $M = 0$. The choice of z_{S_2} instead of z_{S_1} is motivated by the fact that using z_{S_2} allows a DC decomposition where all the computations in DCA are explicit.

Margin maximization has been introduced in [6], where the following problem has been defined:

$$\left\{ \begin{array}{ll} \min\limits_{x_0,q,r} & z_{S_3}(x_0, q, r) \\ & 0 \leq q \leq r, \end{array} \right. \tag{2.5}$$

with

$$z_{S_3}(x_0, q, r) \overset{\triangle}{=} C\sum_{i=1}^{p} \max\{0, y_i(\|x_i - x_0\|^2 - r) + q\} - q.$$

Note that the term $-q$ is aimed at maximizing the margin. Moreover, similarly to [10], also in [6] an ad hoc algorithm that finds the optimal solution in $O(t \log t)$ has been designed when in z_{S_3} the center x_0 is fixed.

2.4. Conical separation. A revolution cone in $\mathbb{R}^n$ is a set $\Gamma(x_0, v, s)$ defined as follows:

$$\Gamma(x_0, v, s) \overset{\triangle}{=} \{x \in \mathbb{R}^n : s\,\|x - x_0\| - v^T(x - x_0) = 0\}, \tag{2.6}$$

where $x_0 \in \mathbb{R}^n$ and $v \in \mathbb{S}_n \subset \mathbb{R}^n$ are, respectively, the *apex* and the *axis* of the cone, with $\mathbb{S}_n \overset{\triangle}{=} \{x \in \mathbb{R}^n : \|x\| = 1\}$. The scalar $s \in [0,1]$, called the *aperture coefficient*, provides the half-aperture angle of the cone given by $\arccos s$. Note that in some applications x_0 might be constrained to belong to some set $X_0 \subseteq \mathbb{R}^n$.

The sets $\mathcal{X}_+$ and $\mathcal{X}_-$ are *conically separable* if and only if there exists a revolution cone $\Gamma(x_0, v, s)$ such that

$$y_i[s\|x_i - x_0\| - z^T(x_i - x_0)] \leq 0 \quad i = 1, \dots, p.$$

Observe that, since a hyperplane is a particular instance of a revolution cone (to see this, just take $s = 0$ and x_0 fixed), linear separability implies conic separability, while the converse statement is not true. Note also that minimizing the coefficient s amounts to make the half-aperture angle of the separation cone as large as possible and, in terms of classification correctness, this fact reduces the possibility of false-negative outcome.

The corresponding optimization problem can be written as follows

$$\begin{cases} \min\limits_{x_0, v, s} & s \\ & y_i[s\|x_i - x_0\| - v^T(x_i - x_0)] \leq 0 \quad i = 1, \dots, p \\ & \|v\|^2 = 1 \\ & 0 \leq s \leq 1 \\ & x_0 \in X_0. \end{cases}$$

More specifically a conic classification model has been presented in [11, 12] where the following problem is tackled:

$$\begin{cases} \min\limits_{x_0, v, s} & z_C(x_0, v, s) \\ & \|v\|^2 = 1 \\ & 0 \leq s \leq 1 \\ & x_0 \in X_0, \end{cases}$$

with

$$z_C(x_0, v, s) \overset{\triangle}{=} \gamma s + \sum_{i=1}^{p} \max\{0, y_i[\|x_i - x_0\|s - v^T(x_i - x_0)]\}.$$

Note that in the above model possible misclassification of some samples is considered and, in fact, the parameter $\gamma > 0$ in the objective function z_C expresses the tradeoff between the two conflicting objectives of maximization of the half-aperture angle of the cone and of minimization of the classification error. Moreover it is possible to prove that the constraint $\|v\|^2 = 1$ can be replaced by

$$\|v\|^2 \geq 1. \tag{2.7}$$

In [11] and [12] the two different cases where X_0 is a singleton (in particular $X_0 = \{0\}$) or a polytope have been treated, respectively.

The former case has been treated in turn in two different ways; the first one works by penalizing the norm constraint (2.7) and formulating the resulting problem as a DC (Difference of Convex) program, while in the second one the constraint (2.7) is linearized and a two phase method based on alternating a proximal-point-type minimization and a feasibility restoration step is adopted.

Also for the second case (X_0 is a polytope) two different approaches have been devised. The first one is again based on penalization of the constraint (2.7) and on alternating minimization first with respect to the couple of variables (x_0, v) and then with respect to the scalar variable s; again a DC decomposition scheme is adopted as far as minimization in the variables (x_0, v) is performed. The second approach is motivated by the fact that, once x_0 is fixed, all the machinery developed for the case when X_0 is a singleton can be easily adapted. Thus the method is based on alternating minimization with respect to (v, s) and x_0.

3. Semisupervised approaches. As previously mentioned, quite often in practical applications class membership is not known in advance for all points in the dataset. Semisupervised classification is a valid approach to treat such case: in fact the samples are partitioned into labeled and unlabeled ones and they all enter into the classification process in different ways.

In this section we describe two semisupervised models presented in [3, 4] characterized by a nonconvex nondifferentiable objective function. In both of them (see also [2] and [17]) the method described in [24] is adopted as a solver capable to cope with both nonconvexity and nonsmoothness.

3.1. Polyhedral separation. We present in this subsection an approach embedding the h-polyhedral separability into the semisupervised framework (see [3]).

We describe first how the semisupervised approach works when applied to linear separability. In this case the TSVM (Transductive SVM) model proposed in [19] and tackled in [2] deals with the problem of finding a hyperplane away from both the labeled and unlabeled points; it is formulated as follows:

$$\min_{w \in \mathbb{R}^n,\ b \in \mathbb{R}} z(w, b),$$

where

$$\begin{aligned} z(w,b) \stackrel{\triangle}{=} \quad & \frac{1}{2}\|w\|^2 + C_1 \sum_{i=1}^{p} \max\{0, 1 - y_i(w^T x_i + b)\} + \\ & C_2 \sum_{i=p+1}^{p+k} \max\{0, 1 - |w^T x_i + b|\}, \end{aligned}$$

with $C_1, C_2 > 0$.

Function z is the sum of three terms. The minimization of the first term corresponds to maximizing the margin (which is the main characterization of any SVM model), the second one is aimed at minimizing the misclassification errors of the labeled points and the last term takes into account the minimization of the number of unlabeled points in the margin zone. Parameters C_1 and C_2 tune the trade-off between these different objectives. Function z is nondifferentiable and, due to the last term involving the unlabeled points, it is also nonconvex. In particular it is piecewise affine, with the number of "pieces" depending on the number of samples, and, when $C_2 = 0$ (in this case the unlabeled points are not taken in consideration), it reduces to the standard SVM error function.

The above problem has been tackled in different ways. In [20] the authors apply a standard gradient descent method to a smooth approximation of the objective function.

In other papers the nonconvex nonsmooth nature of the objective function is faced without resorting to any smoothing. In particular, in [21] the problem is reformulated as a DC (Difference of Convex) one, while in [2] the bundle type method [24] previously cited is directly applied.

Now we describe how such approach can be extended to the h-polyhedral separability framework [3]. Using the same definitions and notations as in (2.1), we consider the following model:

$$\min_{\mathbf{w}\in\mathbb{R}^{n\cdot h},\ \mathbf{b}\in\mathbb{R}^h} \tilde{z}_P(\mathbf{w},\mathbf{b}), \tag{3.1}$$

where

$$\begin{aligned} \tilde{z}_P(\mathbf{w},\mathbf{b}) := \quad & \frac{1}{2}\sum_{j=1}^{h}\|w^{(j)}\|^2 + \\ & C_1\sum_{i=1}^{m}\max_{j=1,\dots,h}\{0, y_i(x_i^T w^{(j)} + b_j) + 1\} + \\ & C_1\sum_{i=m+1}^{p}\max\{0, \min_{j=1,\dots,h}[y_i(x_i^T w^{(j)} + b_j) + 1]\} + \\ & C_2\sum_{j=1}^{h}\sum_{i=p+1}^{p+k}\max\{0, 1 - |w^{(j)T}x_i + b_j|\}. \end{aligned} \tag{3.2}$$

Function $\tilde{z}_P$ is an extension of the TSVM error function z to the h-polyhedral separability. In particular the terms corresponding to the minimization of the misclassification errors of labeled points are treated taking into account the different role played by $\mathcal{X}_+$ and $\mathcal{X}_-$. Note that $\tilde{z}_P$ is nondifferentiable and nonconvex.

The above model belongs to the large margin classifiers class (see for example [15, 20, 21, 25, 28]). In fact a relevant role in $\tilde{z}_P$ is played by the first term, whose minimization corresponds to maximize h separation margins, one for each hyperplane. We recall that, in the SVM technique, the margin is the distance between two parallel hyperplanes, each of them supporting one class. This concept is easily extended to the polyhedral separation by considering h margin areas, comprised between h different couples of supporting hyperplanes (see the example in Fig. 3.1).

Finally note that, when $h > 1$ and $C_2 = 0$, we obtain the SVM version, while, for $h = 1$, function $\tilde{z}_P$ reduces to function z.

3.2. Spherical separation. In [4] spherical separation of two disjoint finite sets of points has been embedded into the semisupervised framework. In particular, an extension of the margin spherical model (2.5) has been proposed by introducing an additional term involving the unlabeled data.

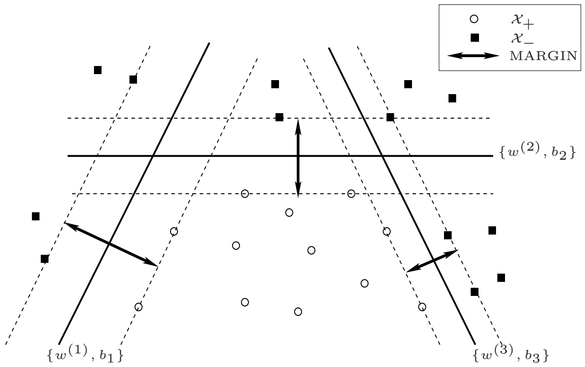

FIG. 3.1. *: The h-polyhedral separation margin for $h = 3$.*

The basic idea is to exploit information coming from unlabeled samples by using the same criterion which is at the basis of the TSVM model, i.e. to minimize the number of unlabeled points in the margin zone. In particular, a point $x_i \in \mathcal{X}_U$ is in the margin area if

$$\|x_i - x_0\|^2 < (R + M)^2 \quad \text{and} \quad \|x_i - x_0\|^2 > (R - M)^2,$$

i.e.

$$\min\{(R + M)^2 - \|x_i - x_0\|^2, \|x_i - x_0\|^2 - (R - M)^2\} > 0.$$

Taking into account the definitions of q and r given in subsection 2.3, we rewrite the above formula as

$$\min\{q + r - \|x_i - x_0\|^2, \|x_i - x_0\|^2 - r + q\} > 0,$$

for all points $x_i \in \mathcal{X}_U$, $i = p + 1, \dots, p + k$, and thus the following merit function is defined:

$$\begin{aligned} &\tilde{z}_{S_3}(x_0, r, q) \overset{\triangle}{=} C_1 \sum_{i=1}^{p} \max\{0, y_i(\|x_i - x_0\|^2 - r) + q\} + \\ &C_2 \sum_{i=p+1}^{p+k} \max\{0, \min[q + r - \|x_i - x_0\|^2, \|x_i - x_0\|^2 - r + q]\} - q, \end{aligned}$$

where $C_1 > 0$ and $C_2 > 0$ are weighting parameters providing the trade-off between the labeled and unlabeled terms. As a consequence, the resulting transductive model is the following:

$$\left\{ \begin{array}{ll} \displaystyle\min_{x_0, r, q} & \tilde{z}_{S_3}(x_0, r, q) \\ & 0 \leq q \leq r, \end{array} \right.$$

whose objective function is still nonsmooth and nonconvex.

4. Conclusions. We have presented some optimization-based techniques for solving binary classification problems which can be expressed in terms of separating two sets in a finite dimension space by means of nonlinear surfaces. We have focused on both the supervised and the semisupervised framework.

We have reviewed several possible choices about the particular surface to be adopted by considering polyhedral, ellipsoidal, spherical and conical separation.

REFERENCES

[1] L. T. H. An and P. D. Tao, *The DC (Difference of Convex Functions) programming and DCA revisited with DC models of real world nonconvex optimization problems*, Annals of Operations Research, 133 (2005), pp. 23–46.

[2] A. Astorino and A. Fuduli, *Nonsmooth optimization techniques for semisupervised classification*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (2007), pp. 2135–2142.

[3] A. Astorino and A. Fuduli, *Support vector machine polyhedral separability in semisupervised learning*, Journal of Optimization Theory and Applications, DOI: 10.1007/s10957-013-0458-6 (2013). to appear.

[4] ———, *Semisupervised spherical separation*, Applied Mathematical Modeling, (2014). submitted to.

[5] A. Astorino, A. Fuduli, and M. Gaudioso, *DC models for spherical separation*, Journal of Global Optimization, 48 (2010), pp. 657–669.

[6] ———, *Margin maximization in spherical separation*, Computational Optimization and Applications, 53 (2012), pp. 301–322.

[7] A. Astorino, A. Fuduli, and E. Gorgone, *Non-smoothness in classification problems*, Optimization Methods and Software, 23 (2008), pp. 675–688.

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

[9] ———, *Ellipsoidal separation for classification problems*, Optimization Methods and Software, 20 (2005), pp. 261–270.

[10] A. Astorino and M. Gaudioso, *A fixed-center spherical separation algorithm with kernel transformations for classification problems*, Computational Management Science, 6 (2009), pp. 357–372.

[11] A. Astorino, M. Gaudioso, and A. Seeger, *Conic separation of finite sets. I. The homogeneous case*, Journal of Convex Analysis, 21 (2014), pp. 1–28.

[12] ———, *Conic separation of finite sets. II. The Non-Homogeneous case*, Journal of Convex Analysis, 21 (2014), pp. 819–831.

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

[14] A. M. Bagirov, J. Ugon, D. Webb, G. Ozturk, and R. Kasimbeyli, *A novel piecewise linear classifier based on polyhedral conic and max-min separabilities*, TOP, 21 (2013), pp. 3–24.

[15] K. P. Bennett and A. Demiriz, *Semi-supervised support vector machines*, in Advances in Neural Information Processing Systems, M. S. Kearns, S. A. Solla, and D. A. Cohn, eds., vol. 12, MIT Press, Cambridge, MA, 1998, pp. 368–374.

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

[17] C. Bergeron, G. Moore, J. Zaretzki, C. Breneman, and K. Bennett, *Fast bundle algorithm for multiple instance learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34 (2012), pp. 1068 – 1079.

[18] O. Chapelle, V. Sindhwani, and S. S. Keerthi, *Optimization techniques for semi-supervised support vector machines*, Journal of Machine Learning Research, 9 (2008), pp. 203–233.

[19] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, *Choosing multiple parameters for support vector machines*, Machine Learning, 46 (2002), pp. 131–159.

[20] O. Chapelle and A. Zien, *Semi-supervised classification by low density separation*, in Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, 2005,

pp. 57–64.

[21] R. Collobert, F. Sinz, J. Weston, and L. Bottou, *Large scale transductive SVMs*, Journal of Machine Learning Research, 7 (2006), pp. 1687–1712.

[22] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[23] A. V. Demyanov and M. Gaudioso, *An approach to classification based on separation of sets by means of several ellipsoids*, Optimization, 54 (2005), pp. 579–593.

[24] A. Fuduli, M. Gaudioso, and G. Giallombardo, *Minimizing nonconvex nonsmooth functions via cutting planes and proximity control*, SIAM Journal on Optimization, 14 (2004), pp. 743–756.

[25] G. Fung and O. L. Mangasarian, *Semi-supervised support vector machines for unlabeled data classification*, Optimization Methods and Software, 15 (2001), pp. 29–44.

[26] R. N. Gasimov and G. Ozturk, *Separation via polyhedral conic functions*, Optimization Methods and Software, 21 (2006), pp. 527–540.

[27] T. Joachims, *Making large-scale SVM learning practical*, in Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press, 1999.

[28] ———, *Transductive inference for text classification using support vector machines*, in International Conference on Machine Learning, 1999, pp. 200–209.

[29] ———, *Transductive learning via spectral graph partitioning*, in Proceedings of the International Conference on Machine Learning, 2003, pp. 290–297.

[30] H. A. Le Thi, H. M. Le, T. Pham Dinh, and N. Van Huynh, *Binary classification via spherical separator by DC programming and DCA*, Journal of Global Optimization, 56 (2013), pp. 1393–1407.

[31] O. L. Mangasarian, *Linear and nonlinear separation of patterns by linear programming*, Operations Research, 13 (1965), pp. 444–452.

[32] ———, *Multisurface method for pattern separation*, IEEE Transactions on Information Theory, IT-14 (1968), pp. 801–807.

[33] N. Megiddo, *On the complexity of polyhedral separability*, Discrete and Computational Geometry, 3 (1988), pp. 325–337.

[34] C. Orsenigo and C. Vercellis, *Accurately learning from few examples with a polyhedral classifier*, Computational Optimization and Applications, 38 (2007), pp. 235–247.

[35] L. Palagi and M. Sciandrone, *On the convergence of a modified version of svmlight algorithm*, Optimization Methods and Software, 20 (2005), pp. 311–328.

[36] J. B. Rosen, *Pattern separation by convex programming*, Journal of Mathematical Analysis and Applications, 10 (1965), pp. 123–134.

[37] P. Sun and R. M. Freund, *Computation of minimum-volume covering ellipsoids*, Operations Research, 52 (2004), pp. 690–706.

[38] P. D. Tao and L. T. H. An, *A D.C. optimization algorithm for solving the trust-region subproblem*, SIAM J. Con. Opt., 8 (1998), pp. 476–505.

[39] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Review, 38 (1996), pp. 49–95.

[40] L. Vandenberghe, S. Boyd, and B. Alkire, http://www.ee.ucla.edu/˜vandenbe/sp.html.

[41] V. Vapnik, *The nature of the statistical learning theory*, Springer Verlag, New York, 1995.

[42] M. Yagiura and T. Ibaraki, *Local search*, in Handbook of Applied Optimization, P. M. Pardalos and M. G. C. Resende, eds., Oxford University Press, 2002, pp. 104–123.