Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Spherical separation and kernel transformations for classification problems

A. Astorino, M. Gaudioso

RT-ICAR-CS-03-16 **Dicembre 2003**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.na.icar.cnr.it*
– Sezione di Palermo, Viale delle Scienze, 90128 Palermo, URL: *www.pa.icar.cnr.it*

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Spherical separation and kernel transformations for classification problems

A. Astorino1[1], M. Gaudioso[2]

Rapporto Tecnico N.:
RT-ICAR-CS-03-16

Data:
Dicembre 2003

---

[1] Istituto di Calcolo e Reti ad Alte Prestazioni, ICAR-CNR, Sede di Cosenza, Via P. Bucci 41C, 87036 Rende (CS)
[2] Università degli Studi della Calabria, DEIS, Via P. Bucci 41C, Rende (CS)

Spherical separation and kernel transformations for classification problems*

A. Astorino§ M. Gaudioso**

Abstract

We state the problem of the optimal separation, via a sphere, of two discrete point sets in a finite dimensional Euclidean space. If the center of the sphere is fixed the problem reduces to an LP problem solvable in $O(p \log p)$ time, where p is the dataset size.

The approach is suitable for use in connection with kernel transformations of the type adopted in the SVM (Support Vector Machine) approach.

Finally we present the numerical results obtained by running our method on some standard test problems drawn from the binary classification literature.

Keywords: Classification, Separability, Kernel Methods, Support Vector Machine.

1 Introduction

Pattern analysis plays a central role in many modern artificial intelligence and computer science problems. The task is to detect regularities that characterize the data coming from a particular source, and the final objective is to design a system capable to make predictions about new data coming from the same source. Pattern analysis may lead to state a number of different problems such as classification, regression, cluster analysis, feature extraction, etc..

*This research has been partially supported by the Italian "Ministero dell'Istruzione, dell'Università e della Ricerca Scientifica", under FIRB project *Large Scale Nonlinear Optimization* (RBNE01WBBB).

§Istituto di Calcolo e Reti ad Alte Prestazioni-C.N.R., c/o D.E.I.S.- Università della Calabria, 87036 Rende (CS), Italia. E-mail: astorino@icar.cnr.it

**Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia. E-mail: gaudioso@deis.unical.it

In particular in this paper we consider binary classification problems, where the objective is to construct a criterion for discriminating between two classes of objects represented by the training set. The problem consists in finding an appropriate surface in $\mathbb{R}^n$ separating two discrete point sets.

Several approaches for binary classification have been proposed. Among the others we mention the pioneering contributions by Ben Rosen [10] and Mangasarian [6]. The kernel method approach, leading to the introduction of the Support Vector Machine (SVM), [12], [13], [4] has been considered a real breakthrough in this area. The basic ideas of SVMs for binary classification problems are to map the data into a higher dimensional space (the feature space) and to separate the two transformed sets by means of one hyperplane. Such a transformation allows to obtain general nonlinear separation surfaces in the original input space (see [12], [13] and [4] for an extensive treatment of the subject).

It is possible, however, to look for nonlinear separation surfaces directly in the input space, (this is the case of polyhedral separation and ellipsoidal separation [1], [2]), or even in the feature space, as we do in present paper.

We suppose that two nonempty and disjoint finite sets of points in the n-dimensional space $\mathbb{R}^n$, say $\mathcal{A} = \{a_1, \dots, a_m\}$ and $\mathcal{B} = \{b_1, \dots, b_k\}$ are given, and we refer to $\mathbb{R}^n$ as to the *input space*. We assume also that the two sets are both non redundant, in the sense that each of them is made up by distinct points.

Our objective is to find, in the input space or in the feature space, a minimal volume sphere separating the set $\mathcal{A}$ from the set $\mathcal{B}$ (i.e. a sphere enclosing all points of $\mathcal{A}$ and no points of $\mathcal{B}$). If the center is fixed, this objective is pursued by solving a Linear Program (LP). The same algorithm can be applied in the feature space, acting on the two transformed sets.

Since we are interested in real-world problems, we must be able to handle very large datasets. Hence, it is not sufficient for an algorithm to work well on small examples, and we require that its performance should scale to large datasets. Our Linear Programming model fits with this need, as it is solvable in $O(p \log p)$ time, where $p = \max\{m, k\}$.

In our opinion spherical separation is a particularly promising approach, if compared with the linear separation approach. In general, if the center is not fixed, the two approaches are equally "parsimonious", as the number of parameters to be selected is in both cases equal to $(n + 1)$ (the center and the radius of the sphere in the former, and the normal and the translation parameter of the hyperplane in the latter). On the other hand, from a heuristic point of view, a judicious choice of the center of a possible separating sphere seems more intuitive than that of the normal to a candidate separating hyperplane. Moreover linear separation can be considered a special case of spherical separation when the distance of the center of the sphere from the data set goes to infinity.

We remark finally that kernel transformation can be adopted in a rather straightforward way in the context of spherical separation.

The paper is organized as follows. In section 2 we discuss the concept of spherical separation. In section 3 we state the problem, assuming that the center of the sphere is given. In section 4 we describe a method for solving the problem. In section 5 we generalize our method by introducing kernel functions. The results of some numerical experiments are finally described in section 6.

Throughout the paper we adopt the following notations. We denote by $\|.\|$ the Euclidean norm in $\mathbb{R}^n$ and by a^Tb the inner product of the vectors a and b. The convex hull of a set $\mathcal{X}$ will be denoted by $conv(\mathcal{X})$; the sphere of center x_0 and radius R will be denoted by $S(x_0, R)$.

2 Separation by a sphere

The spherical separation of a set $\mathcal{A}$ from a set $\mathcal{B}$ consists in finding a minimal volume sphere enclosing all points of $\mathcal{A}$ and no points of $\mathcal{B}$. We remark that the role of the two sets $\mathcal{A}$ and $\mathcal{B}$ is not symmetric, as it may happen that $\mathcal{A}$ is separable from $\mathcal{B}$ but the reverse is not true. In fact a necessary (but not sufficient) condition for the existence of a separating sphere is that the intersection of $conv(\mathcal{A})$ and $\mathcal{B}$ is empty. Since this problem is not always feasible, we resort to the objective of minimizing a function combining the original objective of minimizing the volume with an appropriate measure of the classification error. Such an approach is inspired [11] by the need of obtaining a sphere enclosing as many as possible points of $\mathcal{A}$ and as few as possible points of $\mathcal{B}$, and also of reducing the effect produced by possible outliers.

A sphere centered in $x_0 \in \mathbb{R}^n$ with radius $R \in \mathbb{R}$ is defined as

$$S(x_0, R) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n \mid \ (x - x_0)^T(x - x_0) \leq R^2\}.$$

The sets $\mathcal{A}$ and $\mathcal{B}$ are defined to be not strictly spherically separated by $S(x_0, R)$ if

$$(a_i - x_0)^T(a_i - x_0) \leq R^2$$

for all points $a_i \in \mathcal{A}$ $(i = 1, ..., m)$ and

$$(b_l - x_0)^T(b_l - x_0) \geq R^2$$

for all points $b_l \in \mathcal{B}$ $(l = 1, ..., k)$.

The two sets $\mathcal{A}$ and $\mathcal{B}$ are spherically separable if there exists a not strictly separating sphere. The following property holds.

Proposition 2.1 *If the sets $\mathcal{A}$ and $\mathcal{B}$ are strictly linearly separable then they are spherically separable.*

Proof By hypothesis there exists a separating hyperplane

$$H = \{x \in \mathbb{R}^n \mid w^T x = \gamma\}$$

for some $w \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$ such that

$$a_i \in H^- \quad \forall i = 1, ..., m$$

and

$$b_l \in H^+ \quad \forall l = 1, ..., k,$$

with $H^- = \{x \in \mathbb{R}^n \mid w^T x < \gamma\}$ and $H^+ = \{x \in \mathbb{R}^n \mid w^T x > \gamma\}$. We denote by $\bar{H}^-$ and $\bar{H}^+$ the closed halfspaces $\{x \in \mathbb{R}^n \mid w^T x \leq \gamma\}$ and $\{x \in \mathbb{R}^n \mid w^T x \geq \gamma\}$, respectively .

Let x_0 be any point in H^- and x_P its projection onto H (and, also, onto $\bar{H}^+$).The expression of x_P is the following

$$x_P = x_0 - \frac{w^T x_0 - \gamma}{\|w\|^2} w$$

and, consequently, the distance $d(x_0, H)$ of x_0 from H is

$$d(x_0, H) = \|x_0 - x_P\| = \frac{\gamma - w^T x_0}{\|w\|}.$$

We remark that the sphere $S(x_0, \|x_0 - x_P\|)$ is contained in $\bar{H}^-$.

From the definition of x_P we have :

$$\|x_0 - b_l\| > \|x_0 - x_P\| \quad \forall l = 1, ..., k.$$

If, in addition,

$$\|x_0 - a_i\| \leq \|x_0 - x_P\| \quad \forall i = 1, ..., m$$

then $S(x_0, \|x_0 - x_P\|)$ separates the set $\mathcal{A}$ from $\mathcal{B}$, otherwise the condition

$$\|x_0 - a_i\| > \|x_0 - x_P\|$$

holds for at least one point $a_i \in \mathcal{A}$.

Consider now the half-line

$$x(\alpha) = x_0 + \alpha(x_0 - x_P),$$

for $\alpha \geq 0$. The point x_P is still the projection of $x(\alpha)$ on H (and, also, onto $\bar{H}^+$), moreover

$$d(x(\alpha), H) = \|x(\alpha) - x_P\| = (1+\alpha)\|x_0 - x_P\|$$

and the sphere $S(x(\alpha), \|x(\alpha)-x_P\|)$ is still contained in $\bar{H}^-$. If, in addition, the point $x(\alpha)$ satisfies the condition

$$\|x(\alpha) - a_i\| \leq \|x(\alpha) - x_P\| \quad \forall i = 1, ..., m \tag{1}$$

then the sphere $S(x(\alpha), \|x(\alpha) - x_P\|)$ contains all the points of $\mathcal{A}$ and separates the set $\mathcal{A}$ from the set $\mathcal{B}$.

On the other hand, from the definition of $x(\alpha)$, it follows that

$$\|x(\alpha) - a_i\|^2 = \|x_0 - a_i\|^2 + \alpha^2\|x_0 - x_P\|^2 + 2\alpha(x_0 - a_i)^T(x_0 - x_P) \tag{2}$$

and

$$\begin{array}{l} \|x(\alpha) - x_P\|^2 = (1+\alpha)^2\|x_0 - x_P\|^2 = \\ \|x_0 - x_P\|^2 + \alpha^2\|x_0 - x_P\| + 2\alpha(x_0 - x_P)^T(x_0 - x_P). \end{array} \tag{3}$$

Comparing (2) and (3), we observe that (1) is satisfied if

$$\begin{array}{l} 2\alpha[(x_0 - x_P)^T(x_0 - x_P) - (x_0 - a_i)^T(x_0 - x_P)] \geq \\ \|x_0 - a_i\|^2 - \|x_0 - x_P\|^2 > 0 \ \forall i = 1, ..., m. \end{array}$$

The r.h.s. of the above formula can be rewritten, by taking into account the expression of x_P, as

$$\begin{array}{l} 2\alpha(x_0 - x_P)^T(a_i - x_P) = \\[2ex] \dfrac{2\alpha(w^T x_0 - \gamma)}{\|w\|^2} w^T(a_i - x_P) = \\[2ex] 2\alpha(w^T x_0 - \gamma)(w^T a_i - \gamma) > 0 \ \forall i = 1, ..., m. \end{array}$$

Thus the condition (1) is verified whenever

$$\alpha \geq \bar{\alpha} \stackrel{\triangle}{=} \max_{1 \leq i \leq m} \frac{\|x_0 - a_i\|^2 - \|x_0 - x_P\|^2}{2(w^T x_0 - \gamma)(w^T a_i - \gamma)} > 0$$

$\square$

3 The problem

We state our problem by assuming that the center x_0 of the sphere is given (any centroid either for the set $\mathcal{A}$ or for the set $\mathcal{B}$ or even for the set $\mathcal{A} \bigcup \mathcal{B}$ can be selected). Moreover we assume, without loss of generality, that x_0 does not coincide with any point either of $\mathcal{A}$ or of $\mathcal{B}$.

According to our definitions any sphere $S(x_0, R)$ separates $\mathcal{A}$ from $\mathcal{B}$ provided

$$\begin{array}{ll} (a_i - x_0)^T (a_i - x_0) \le R^2 & \forall i = 1, ..., m \\ (b_l - x_0)^T (b_l - x_0) \ge R^2 & \forall l = 1, ..., k. \end{array}$$

Consequently we define the classification error associated to $S(x_0, R)$ for any point $a_i \in \mathcal{A}$ and for any point $b_l \in \mathcal{B}$, respectively, as:

$$\begin{array}{ll} \xi_i = \max\{0, (a_i - x_0)^T (a_i - x_0) - R^2\} & \forall i = 1, ..., m \\ \mu_l = \max\{0, R^2 - (b_l - x_0)^T (b_l - x_0)\} & \forall l = 1, ..., k. \end{array}$$

The problem of minimizing both the volume of the sphere and the classification error is defined as follows:

$$\begin{array}{lll} \min\limits_{R,\xi,\mu} & R^2 + C \left(\sum\limits_{i=1}^{m} \xi_i + \sum\limits_{l=1}^{k} \mu_l \right) & \\ \text{s.t} & R^2 - (a_i - x_0)^T (a_i - x_0) + \xi_i \ge 0 & \forall i = 1, ..., m \\ & (b_l - x_0)^T (b_l - x_0) - R^2 + \mu_l \ge 0 & \forall l = 1, ..., k \\ & \xi_i \ge 0 & \forall i = 1, ..., m \\ & \mu_l \ge 0 & \forall l = 1, ..., k \end{array} \tag{4}$$

where the positive constant C states the relative importance of the two objectives .

The problem above has a MIP (Mixed Integer Programming) counterpart in case we consider the objective of minimizing the number of misclassified points instead of that of minimizing the classification error:

$$\begin{array}{lll} \min\limits_{R,\xi,\mu,u,v} & R^2 + C \left(\sum\limits_{i=1}^{m} u_i + \sum\limits_{l=1}^{k} v_l \right) & \\ \text{s.t} & R^2 - (a_i - x_0)^T (a_i - x_0) + \xi_i \ge 0 & \forall i = 1, ..., m \\ & (b_l - x_0)^T (b_l - x_0) - R^2 + \mu_l \ge 0 & \forall l = 1, ..., k \\ & 0 \le \xi_i \le M u_i & \forall i = 1, ..., m \\ & 0 \le \mu_l \le M v_l & \forall l = 1, ..., k \\ & u_i = \{0, 1\} & \forall i = 1, ..., m \\ & v_l = \{0, 1\} & \forall l = 1, ..., k \end{array}$$

where M is a sufficiently large positive constant.

By introducing the change of variable

$$z = R^2, \quad z \geq 0 \tag{5}$$

and by defining:

$$\begin{aligned} c_i &\stackrel{\triangle}{=} (a_i - x_0)^T (a_i - x_0) \geq 0 \quad \forall i = 1, ..., m \\ d_l &\stackrel{\triangle}{=} (b_l - x_0)^T (b_l - x_0) \geq 0 \quad \forall l = 1, ..., k \end{aligned}$$

the problem (4) becomes:

$$\begin{array}{lll} f_P = & \min\limits_{z,\xi,\mu} & z + C \left(\sum\limits_{i=1}^{m} \xi_i + \sum\limits_{l=1}^{k} \mu_l \right) \\ & \text{s.t.} & z - c_i + \xi_i \geq 0 \qquad\qquad \forall i = 1, ..., m \\ & & d_l - z + \mu_l \geq 0 \qquad\qquad \forall l = 1, ..., k \\ & & z \geq 0 \\ & & \xi_i \geq 0 \qquad\qquad\qquad\quad \forall i = 1, ..., m \\ & & \mu_l \geq 0 \qquad\qquad\qquad\quad \forall l = 1, ..., k \end{array} \tag{6}$$

that is a Linear Programming problem, whose dual is the following:

$$\begin{array}{lll} f_D = & \max\limits_{\alpha,\beta} & \sum\limits_{i=1}^{m} c_i \alpha_i - \sum\limits_{l=1}^{k} d_l \beta_l \\ & \text{s.t.} & \sum\limits_{i=1}^{m} \alpha_i - \sum\limits_{l=1}^{k} \beta_l \leq 1 \\ & & 0 \leq \alpha_i \leq C \qquad\qquad \forall i = 1, ..., m \\ & & 0 \leq \beta_l \leq C \qquad\qquad \forall l = 1, ..., k \end{array} \tag{7}$$

We observe that both the primal and the dual problems are feasible and in particular the solution

$$\begin{aligned} \alpha_i = 0 \quad \forall i = 1, ..., m \\ \beta_l = 0 \quad \forall l = 1, ..., k \end{aligned}$$

is dual feasible with objective function value equal to zero. The complementary slackness conditions for problems (6) and (7) are the following:

$$\begin{gathered} \left\langle \begin{array}{c} z \left(\sum_{i=1}^{m} \alpha_i - \sum_{l=1}^{k} \beta_l - 1 \right) = 0 \\ \xi_i \left(C - \alpha_i \right) = 0 \quad \forall i = l, ..., m \\ \mu_l \left(C - \beta_l \right) = 0 \quad \forall l = l, ..., k \end{array} \right\rangle \\ \\ \left\langle \begin{array}{c} \alpha_i \left(z - c_i + \xi_i \right) = 0 \quad \forall i = l, ..., m \\ \beta_l \left(-z + d_l + \mu_l \right) = 0 \quad \forall l = l, ..., k \end{array} \right\rangle \end{gathered} \tag{8}$$

In the sequel we indicate by α and β the vectors whose components are, respectively, the α_i's, $i = 1, \dots, m$ and the β_l's, $l = 1, \dots, k$. Moreover we indicate by ξ and μ the vectors whose components are, respectively, the ξ_i's, $i = 1, \dots, m$ and the μ_l's, $l = 1, \dots, k$.

Proposition 3.1 *The following properties hold for z^*, the optimal value of the variable z at any optimal solution for the problem (6):*

i) If $C < \dfrac{1}{m}$ then $z^* = 0$;

ii) If $C > \dfrac{1}{m}$ then $z^* > 0$.

Proof To prove i) it is sufficient to observe that, in case $C < \dfrac{1}{m}$, no dual feasible solution satisfying by equality the constraint $\sum_{i=1}^m \alpha_i - \sum_{l=1}^k \beta_l \leq 1$ exists. The thesis follows by taking into account the complementary slackness condition $z\left(\sum_{i=1}^m \alpha_i - \sum_{l=1}^k \beta_l - 1\right) = 0$.

As for the proof of ii), suppose $C > \dfrac{1}{m}$ and assume by contradiction that (z^*, ξ^*, μ^*) is an optimal solution for (6) with $z^* = 0$. Then it follows that (ξ^*, μ^*) solves the problem

$$\begin{array}{lll} \min\limits_{\xi,\mu} & C\left(\sum\limits_{i=1}^m \xi_i + \sum\limits_{l=1}^k \mu_l\right) & \\ \text{s.t.} & -c_i + \xi_i \geq 0 & \forall i = 1, ..., m \\ & d_l + \mu_l \geq 0 & \forall l = 1, ..., k \\ & \xi_i \geq 0 & \forall i = 1, ..., m \\ & \mu_l \geq 0 & \forall l = 1, ..., k. \end{array} \tag{9}$$

Since, by hypothesis, $c_i > 0$ $\forall i = 1, ..., m$ and $d_l > 0$ $\forall l = 1, ..., k$, it follows that

$$\begin{array}{ll} \xi_i^* = c_i & \forall i = 1, ..., m \\ \mu_l^* = 0 & \forall l = 1, ..., k \end{array}$$

and

$$f_P = C \sum_{i=1}^m c_i.$$

Now consider the feasible solution $(\bar{z}, \bar{\xi}, \bar{\mu})$ to (6) obtained by setting:

$$\bar{z} = \min\{\min_{1 \leq i \leq m} c_i, \min_{1 \leq l \leq k} d_l\} > 0$$

and by calculating $(\bar{\xi}, \bar{\mu})$ as the optimal solution to:

$$\begin{array}{lll}
\min\limits_{\xi,\mu} & \bar{z} + C\left(\sum\limits_{i=1}^{m} \xi_i + \sum\limits_{l=1}^{k} \mu_l\right) & \\
\text{s.t.} & \xi_i \geq c_i - \bar{z} & \forall i = 1, ..., m \\
& \mu_l \geq -d_l + \bar{z} & \forall l = 1, ..., k \\
& \xi_i \geq 0 & \forall i = 1, ..., m \\
& \mu_l \geq 0 & \forall l = 1, ..., k.
\end{array} \tag{10}$$

The optimal values $\bar{\xi}$ and $\bar{\mu}$ are the following:

$$\begin{array}{ll}
\bar{\xi_i} = c_i - \bar{z} & \forall i = 1, ..., m \\
\bar{\mu_l} = 0 & \forall l = 1, ..., k.
\end{array}$$

Consequently the value associated to the feasible solution $(\bar{z}, \bar{\xi}, \bar{\mu})$ is

$$\bar{z} + C\sum_{i=1}^{m}(c_i - \bar{z}) = C\sum_{i=1}^{m} c_i - (m \cdot C - 1)\,\bar{z} < C\sum_{i=1}^{m} c_i = f_P$$

which contradicts the optimality of (z^*, ξ^*, μ^*). □

We remark that since we are not interested in finding trivial (zero radius) spheres, the only interesting choice is to set $C > \dfrac{1}{m}$. In this case, from the previous proposition, taking into account complementary slackness, the constraint $\sum_{i=1}^{m} \alpha_i - \sum_{l=1}^{k} \beta_l \leq 1$ is satisfied by equality at the optimum of the dual problem (7).

Thus we will consider problem (7) in the form

$$\begin{array}{llll}
f_D = & \max\limits_{\alpha,\beta} & \sum\limits_{i=1}^{m} c_i\alpha_i - \sum\limits_{l=1}^{k} d_l\beta_l & \\
& \text{s.t.} & \sum\limits_{i=1}^{m} \alpha_i - \sum\limits_{l=1}^{k} \beta_l = 1 & \\
& & 0 \leq \alpha_i \leq C & \forall i = 1, ..., m \\
& & 0 \leq \beta_l \leq C & \forall l = 1, ..., k.
\end{array} \tag{11}$$

For sake of completeness we remark that, in case $C = \dfrac{1}{m}$, the optimal value z^* can assume any value in the closed interval $[0, \min\{\min\limits_{1\leq i\leq m} c_i, \min\limits_{1\leq l\leq k} d_l\}]$.

4 The Algorithm

Problem (11) is a Linear Program characterized by only one equality constraint and by the presence of lower and upper bounds on all variables.

It is well known that there exists an optimal solution to (11) with at most one variable belonging to the interior of the interval $[0, C]$ (we will refer to such solution as to an optimal basic solution).

We assume, without loss of generality, that the points of the two sets $\mathcal{A}$ and $\mathcal{B}$ are numbered so that:

$$c_1 \geq c_2 \geq \cdots \geq c_m > 0$$

and

$$0 < d_1 \leq d_2 \leq \cdots \leq d_k.$$

Now we describe an algorithm that finds the optimal solution to the dual problem (11) for $C > 1/m$.

Algorithm 1
(Case $C \leq 1$)

Initialization

Set

- $r \stackrel{\triangle}{=} \left\lfloor \frac{1}{C} \right\rfloor$ *(Remark: $m \geq r + 1$)*
- $\bar{p} \stackrel{\triangle}{=} \min\{k, m - r\} \geq 1$
- $\alpha_i = 0 \quad \forall i = 1, ..., m$
- $\beta_l = 0 \quad \forall l = 1, ..., k$

Step 1. *Set* $\alpha_i = C \quad \forall i = 1, ..., r$

If $(c_{r+1} \leq d_1)$ *Set* $\alpha_{r+1} = 1 - Cr$ *and STOP [exit (a): the basic variable is* α_{r+1}*].*
(Remark: If $C = 1$, then $\alpha_{r+1} = 0$ and the solution is a degenerate basic feasible solution).

Endif

If $(c_{r+i} > d_i \quad \forall i = 1, ..., \bar{p})$

If $(\bar{p} > 1)$ *Set* $\alpha_{r+i} = C \quad \forall i = 1, ..., (\bar{p} - 1)$

Endif

Set

- $\alpha_{r+\bar{p}} = 1 - Cr$
- $\beta_l = C \quad \forall l = 1, ..., (\bar{p} - 1)$

and STOP [exit (b): the basic variable is $\alpha_{r+\bar{p}}$].

Endif

Step 2. *Find p^*, the smallest index i, $2 \leq i \leq \bar{p}$ such that $c_{r+i} \leq d_i$ (Remark: Step 2 cannot be entered if $\bar{p} = 1$. Calculation of the index p^* is well posed since the algorithm has not stopped at step 1).*

If $(c_{r+p^*} \geq d_{p^*-1})$
Set

- $\alpha_{r+i} = C \quad \forall i = 1, ..., (p^* - 1)$
- $\alpha_{r+p^*} = 1 - Cr$
- $\beta_l = C \quad \forall l = 1, ..., (p^* - 1)$

and STOP [exit (c): the basic variable is α_{r+p^*}].

Else *Set* $\alpha_{r+i} = C \quad \forall i = 1, ..., (p^* - 1)$

If $(p^* > 2)$ *Set* $\beta_l = C \quad \forall l = 1, ..., (p^* - 2)$

Endif

Set $\beta_{p^*-1} = C(r + 1) - 1$ *and STOP [exit (d): the basic variable is β_{p^*-1}].*

Endif

(Case $C > 1$**)**

Initialization

Set

- $\bar{p} \stackrel{\triangle}{=} \min\{k, m\} \geq 1$
- $\alpha_i = 0 \quad \forall i = 1, ..., m$
- $\beta_l = 0 \quad \forall l = 1, ..., k$

Step 1. *!!!!!!*

If $(c_1 \leq d_1)$ *Set* $\alpha_1 = 1$ *and STOP [exit (a): the basic variable is α_1].*

Endif

If $(c_i > d_i \quad \forall i = 1, ..., \bar{p})$

If $(\bar{p} > 1)$ *Set* $\alpha_i = C \quad \forall i = 1, ..., (\bar{p} - 1)$

Endif

Set

- $\alpha_{\bar{p}} = 1$
- $\beta_l = C \quad \forall l = 1, ..., (\bar{p} - 1)$

and STOP [exit (b): the basic variable is $\alpha_{\bar{p}}$].

Endif

Step 2. *Find p^*, the smallest index i, $2 \leq i \leq \bar{p}$ such that $c_i \leq d_i$ (Remark: Step 2 cannot be entered if $\bar{p} = 1$. Calculation of the index p^* is well posed since the algorithm has not stopped at step 1).*

If $(c_{p^*} \geq d_{p^*-1})$

Set

- $\alpha_i = C \quad \forall i = 1, ..., (p^* - 1)$
- $\alpha_{p^*} = 1$
- $\beta_l = C \quad \forall l = 1, ..., (p^* - 1)$

and STOP [exit (c): the basic variable is α_{p^*}].

Else *Set $\alpha_i = C \quad \forall i = 1, ..., p^*$*

If $(p^* > 2)$ *Set $\beta_l = C \quad \forall l = 1, ..., (p^* - 2)$*

Endif

Set $\beta_{p^*-1} = C - 1$ and STOP [exit (d): the basic variable is β_{p^*-1}].

Endif

Remark. The solution provided by the algorithm is invariant with respect to C for all $C > 1$.

Remark. The preliminary sorting of the c_i's and of the d_l's is required. It can be executed in $O(plogp)$ time, where $p = \max(m, k)$. The algorithm runs in $O(p)$ time.

Theorem 4.1 *The algorithm (1) finds an optimal solution to problem (11).*

Proof We assume that $\bar{\alpha} \geq 0$ and $\bar{\beta} \geq 0$ are those obtained on exit from the algorithm.

We prove the property for the case $C \leq 1$, as the treatment for the case $C > 1$ is analogous.

It is immediate to verify that, corresponding to all possible exits, the constraint

$$\sum_{i=1}^{m} \bar{\alpha}_i - \sum_{l=1}^{k} \bar{\beta}_l = 1 \tag{12}$$

is satisfied by construction.

We denote by $\bar{\alpha}_h$ (exit (a), (b), (c)), or $\bar{\beta}_s$ (exit (d)), the unique basic variable (possibly degenerate) for the appropriate index h or s and we construct a primal solution as follows:

If the basic variable is $\bar{\alpha}_h$ then set $\bar{\xi}_h = 0$, $\bar{z} = c_h$ and

$$\bar{\xi}_i = \begin{cases} 0 & \text{if } \bar{\alpha}_i = 0 \\ c_i - \bar{z} & \text{if } \bar{\alpha}_i = C \end{cases} \quad \text{for } i = 1, ..., m; i \neq h. \tag{13}$$

$$\bar{\mu}_l = \begin{cases} 0 & \text{if } \bar{\beta}_l = 0 \\ \bar{z} - d_l & \text{if } \bar{\beta}_l = C \end{cases} \quad \text{for } l = 1, ..., k. \tag{14}$$

If the basic variable is $\bar{\beta}_s$ then set $\bar{\mu}_s = 0$, $\bar{z} = d_s$ and

$$\bar{\xi}_i = \begin{cases} 0 & \text{if } \bar{\alpha}_i = 0 \\ c_i - \bar{z} & \text{if } \bar{\alpha}_i = C \end{cases} \quad \text{for } i = 1, ..., m. \tag{15}$$

$$\bar{\mu}_l = \begin{cases} 0 & \text{if } \bar{\beta}_l = 0 \\ \bar{z} - d_l & \text{if } \bar{\beta}_l = C \end{cases} \quad \text{for } l = 1, ..., k; l \neq s \tag{16}$$

It easy to verify that the complementary slackness conditions (8) are satisfied as consequence of (12) and of the variable setting (13), (14),(15) and (16) .

To prove the feasibility we need to show first that $(\bar{z}, \bar{\xi}, \bar{\mu})$ are nonnegative. We consider separately the two cases where the basic variable is $\bar{\alpha}_h$ (exits (a),(b),(c)) or $\bar{\beta}_s$ (exit (d)) for some appropriate value of the index h or s respectively.

Consider the case $\bar{\alpha}_h$ is the basic variable. We have $\bar{z} = c_h > 0$ and $\bar{\xi}_i$ is equal either to zero or to $c_i - \bar{z}$, the latter case occurring only in correspondence to an index $i < h$ for which it is, by hypothesis, $c_i \geq c_h = \bar{z}$. On the other hand the nonnegativity of $\bar{\mu}$ follows by observing that whenever it is $\bar{\mu}_l = \bar{z} - d_l$ we have $\bar{\mu}_l = \bar{z} - d_l = c_h - d_l \geq 0$

Consider now the case $\bar{\beta}_s$ is the basic variable. We have $\bar{z} = d_s > 0$ and $\bar{\mu}_l$ is equal either to zero or to $\bar{z} - d_l$, the latter case occurring only in correspondence to an index $l < s$ for which it is by hypothesis $d_l \leq d_s = \bar{z}$. On the other hand the nonnegativity of $\bar{\xi}$ follows by observing that whenever it is $\bar{\xi}_i = c_i - \bar{z} = c_i - d_s$ the condition $c_i - d_s \geq 0$ holds.

Finally, noting that satisfaction of the constraints $\bar{z} - c_i + \bar{\xi}_i \geq 0 \ \ \forall i = 1, ..., m$ and $d_l - \bar{z} + \bar{\mu}_l \geq 0 \ \ \forall l = 1, ..., k$ is ensured by the variable settings and by the initial sorting of the c_i's and of the d_l's, the thesis follows as the solutions $(\bar{z}, \bar{\xi}, \bar{\mu})$ and $(\bar{\alpha}, \bar{\beta})$ are primal and dual feasible respectively and satisfy the complementary slackness conditions. $\square$

Once the the optimal solutions $(\bar{z}, \bar{\xi}, \bar{\mu})$ and $(\bar{\alpha}, \bar{\beta})$ for (6) and (11) respectively have been calculated, recalling the substitution (5), $R^2 = z$, the sphere $S(x_0, \sqrt{\bar{z}})$ can be utilized for classification purposes, in the sense that any new sample point $x \in \mathbb{R}^n$ is classified according to the following rule:

$$\begin{array}{l} x \text{ is a point of the type } \mathcal{A} \text{ if } (x - x_0)^T (x - x_0) < \bar{z} \\ x \text{ is a point of the type } \mathcal{B} \text{ if } (x - x_0)^T (x - x_0) > \bar{z}. \end{array}$$

The point x_0 remains unclassified whenever it is $(x - x_0)^T (x - x_0) = \bar{z}$.

5 Using the kernels

Kernel transformation of the type adopted in SVM can be easily embedded into the spherical separation approach. Our kernel-based approach consists in:

1. mapping the data into a higher dimensional space (the feature space);
2. separating the two transformed sets by means of one sphere.

We consider an embedding map

$$\phi: \ x \in X \subseteq \mathbb{R}^n \ \rightarrow \ \phi(x) \in F \subseteq \mathbb{R}^N,$$

and a kernel function K that for all $x, y \in X$ satisfies

$$K(x, y) = \phi(x)^T \phi(y).$$

We remark that, by using a kernel function K, the inner products in the feature space can be computed without explicitly computing the map ϕ.

The effect of ϕ is to recode our sets $\mathcal{A}$ and $\mathcal{B}$ as

$$\hat{\mathcal{A}} = \{\phi(a_1), ..., \phi(a_m)\} \text{ and } \hat{\mathcal{B}} = \{\phi(b_1), ..., \phi(b_k)\}.$$

Now we proceed looking for a sphere in $\mathbb{R}^N$, centered in $\phi(x_0) \in \mathbb{R}^N$, where $x_0 \in \mathbb{R}^n$ is given, with radius $\hat{R} \in \mathbb{R}$, with the objective of minimizing both the volume and the classification error. We obtain the following problem

$$\begin{array}{rlll} \hat{f}_P = & \min\limits_{\hat{z},\hat{\xi},\hat{\mu}} & \hat{z} + C\left(\sum\limits_{i=1}^{m} \hat{\xi}_i + \sum\limits_{l=1}^{k} \hat{\mu}_l\right) & \\ & \text{s.t.} & \hat{z} - \hat{c}_i + \hat{\xi}_i \geq 0 & \forall i = 1, ..., m \\ & & \hat{d}_l - \hat{z} + \hat{\mu}_l \geq 0 & \forall l = 1, ..., k \\ & & \hat{z} \geq 0 & \\ & & \hat{\xi}_i \geq 0 & \forall i = 1, ..., m \\ & & \hat{\mu}_l \geq 0 & \forall l = 1, ..., k \end{array} \tag{17}$$

where

$\hat{z} = \hat{R}^2$

$\hat{\xi}_i$ is the classification error for the point $\phi(a_i) \in \hat{\mathcal{A}}$

$\hat{\mu}_l$ is the classification error for the point $\phi(b_l) \in \hat{\mathcal{B}}$

and

$$\begin{array}{rl} \hat{c}_i & = (\phi(a_i) - \phi(x_0))^T(\phi(a_i) - \phi(x_0)) = \\ & = K(a_i, a_i) + K(x_0, x_0) - 2K(a_i, x_0) \geq 0 \quad \forall i = 1, ..., m \\ \hat{d}_l & = (\phi(b_l) - \phi(x_0))^T(\phi(b_l) - \phi(x_0)) = \\ & = K(b_l, b_l) + K(x_0, x_0) - 2K(b_l, x_0) \geq 0 \quad \forall l = 1, ..., k \end{array}$$

The problem (17) is a Linear Program of the same type as problem (6). As in section 3 its dual is stated in the form :

$$\begin{array}{rlll} \hat{f}_D = & \max\limits_{\hat{\alpha},\hat{\beta}} & \sum\limits_{i=1}^{m} \hat{c}_i\hat{\alpha}_i - \sum\limits_{l=1}^{k} \hat{d}_l\hat{\beta}_l & \\ & \text{s.t.} & \sum\limits_{i=1}^{m} \hat{\alpha}_i - \sum\limits_{l=1}^{k} \hat{\beta}_l = 1 & \\ & & 0 \leq \hat{\alpha}_i \leq C & \forall i = 1, ..., m \\ & & 0 \leq \hat{\beta}_l \leq C & \forall l = 1, ..., k. \end{array} \tag{18}$$

and can be solved by the algorithm (1).

Once the optimal solutions $(\hat{\alpha}^*, \hat{\beta}^*)$ and $(\hat{z}^*, \hat{\xi}^*, \hat{\mu}^*)$ for (18) and (17), respectively, have been calculated, recalling the substitution $\hat{R}^2 = \hat{z}$, the sphere $S(\phi(x_0), \sqrt{\hat{z}^*})$ can be used for classification purposes, in the sense that a new sample point $x \in \mathbb{R}^n$ will be classified as follows:

$$\begin{aligned}
&x \text{ is a point of the type } \mathcal{A} \text{ if} \\
&(\phi(x) - \phi(x_0))^T(\phi(x) - \phi(x_0)) = \\
&= K(x,x) + K(x_0,x_0) - 2K(x,x_0) < \hat{z}^*
\end{aligned}$$

$$\begin{aligned}
&x \text{ is a point of the type } \mathcal{B} \text{ if} \\
&(\phi(x) - \phi(x_0))^T(\phi(x) - \phi(x_0)) = \\
&= K(x,x) + K(x_0,x_0) - 2K(x,x_0) > \hat{z}^*.
\end{aligned}$$

We observe that if $x_0 \in \mathbb{R}^n$ is the barycenter of the set $\mathcal{A}$ (or of the set $\mathcal{A} \cup \mathcal{B}$), then $\phi(x_0)$ is not necessarily the barycenter of the set $\hat{\mathcal{A}}$ (or of the set $\hat{\mathcal{A}} \cup \hat{\mathcal{B}}$).

Let $Q = \{x_1, ..., x_q\}$ be a finite point set in the input space and $\hat{Q} = \{\phi(x_1), ..., \phi(x_q)\}$ the transformed set in the feature space. The barycenter of the sets $\hat{Q}$ is the vector

$$\phi_{\hat{Q}} = \frac{1}{q}\sum_{i=1}^{q} \phi(x_i).$$

As for all points in the feature space an explicit vector representation of this point is not available. However, despite of this apparent inaccessibility of the point $\phi_{\hat{Q}}$, we can compute its norm, and the distance of the image of any point x in the input space from it, by using only evaluations of the kernel on the inputs:

$$\phi_{\hat{Q}}^T \phi_{\hat{Q}} = \frac{1}{q^2}\sum_{i,j=1}^{q} K(x_i, x_j)$$

$$(\phi(x) - \phi_{\hat{Q}})^T(\phi(x) - \phi_{\hat{Q}}) = K(x,x) + \frac{1}{q^2}\sum_{i,j=1}^{q} K(x_i,x_j) - \frac{2}{q}\sum_{i=1}^{q} K(x,x_i).$$

Now we remark that if x_0 is not given and we look for a sphere in the feature space centered in the barycenter of the set $\hat{\mathcal{A}}$ (or of the set $\hat{\mathcal{A}} \cup \hat{\mathcal{B}}$), again we can use the algorithm (1) for solving the problem (18) with

$$\begin{aligned}\hat{c}_i &= (\phi(a_i) - \phi_{\hat{A}})^T(\phi(a_i) - \phi_{\hat{A}}) = \\ &= K(a_i, a_i) + \frac{1}{m^2}\sum_{j,s=1}^{m} K(a_j, a_s) - \frac{2}{m}\sum_{j=1}^{m} K(a_i, a_j) \geq 0 \quad \forall i = 1, ..., m\end{aligned}$$

$$\begin{aligned}\hat{d}_l &= (\phi(b_l) - \phi_{\hat{A}})^T(\phi(b_l) - \phi_{\hat{A}}) = \\ &= K(b_l, b_l) + \frac{1}{m^2}\sum_{j,s=1}^{m} K(a_j, a_s) - \frac{2}{m}\sum_{j=1}^{m} K(b_l, a_j) \geq 0 \quad \forall l = 1, ..., k\end{aligned}$$

Once the optimal solutions $(\hat{\alpha}^*, \hat{\beta}^*)$ and $(\hat{z}^*, \hat{\xi}^*, \hat{\mu}^*)$ for (18) and (17), respectively, have been calculated, the sphere $S(\phi_{\hat{A}}, \sqrt{\hat{z}^*})$ can be used for classification purposes, in the sense that any new sample point $x \in \mathbb{R}^n$ will be classified as follows:

x is a point of the type $\mathcal{A}$ if

$$\begin{aligned}&(\phi(x) - \phi_{\hat{A}})^T(\phi(x) - \phi_{\hat{A}}) = \\ &= K(x, x) + \frac{1}{m^2}\sum_{i,j=1}^{m} K(a_i, a_j) - \frac{2}{m}\sum_{i=1}^{m} K(x, a_i) < \hat{z}^*\end{aligned}$$

x is a point of the type $\mathcal{B}$ if

$$\begin{aligned}&(\phi(x) - \phi_{\hat{A}})^T(\phi(x) - \phi_{\hat{A}}) = \\ &= K(x, x) + \frac{1}{m^2}\sum_{i,j=1}^{m} K(a_i, a_j) - \frac{2}{m}\sum_{i=1}^{m} K(x, a_i) > \hat{z}^*.\end{aligned}$$

6 Numerical experiments

We have implemented the algorithm described in section 4 using Matlab 5.3 running on a Pentium IV 2.2 GHz Notebook. We have run it on several test problems available in the literature.

We have considered the following datasets:

- Six publicly available datasets from the UCI Machine Learning Repository [8], in particular, the Wisconsin Breast Cancer Prognosis (WBCP-old, WBCP-new), the Cleveland Heart Disease (Heart), Ionosphere (Ionosphere), Mushroom (Mushroom), Tic-Tac-Toe Endgame (Tic-Tac-Toe).
- The Galaxy Dim dataset (Galaxy Dim) used in galaxy discrimination with neural networks from [9].

In our implementation we have used the following kernel functions:

- *linear*: $K(x, y) = x^T y$;
- *polynomial*: $K(x, y) = (x^T y + 1)^{p_1}$;
- *radial basis function (RBF)*: $K(x, y) = \exp(-\|x - y\|^2)/2p_1^2$;
- *exponential radial basis function (ERBF)*: $K(x, y) = \exp(-\|x-y\|)/2p_1^2$;
- *sigmoidal*: $K(x, y) = \tanh(p_1 x^T y + p_2)$

with parameters p_1 and p_2.

We have run our algorithm for several values of the kernel parameters and of the positive weighting constant. The point x_0 has been selected as the barycenter of the set $\mathcal{A}$ and, whenever nonlinear kernel functions are adopted, the point $\phi(x_0)$ has been selected as the barycenter of the set $\hat{\mathcal{A}}$.

Furthermore, all features of the experiments, except those of the Cleveland Heart Disease dataset have been normalized to the range [-1; +1]. The results obtained by our method are reported in Table (1). For all datasets we report as benchmark, under the denomination "Other methods", the best result drawn from [5, 7], where extensive comparisons of several methods are presented. Only in the case of the dataset WBCP - old, an earlier version of the dataset WBCP no longer available at the UCI Machine Learning Repository, we compare our results with those obtained by our implementation of the RLP method [3] as reported in [1].

We have adopted the tenfold cross-validation protocol, which consists in splitting the dataset of interest into ten equally sized pieces. Nine of them are in turn used as training set and the remaining one as testing set. By correctness we intend the total percentage of well classified points (of both $\mathcal{A}$ and $\mathcal{B}$) when the algorithm stops.

For Mushroom a subset of the entire dataset is used. In particular, the final dataset, that we have considered, contains 22 features with 200 points in set $\mathcal{A}$ and 300 points in set $\mathcal{B}$.

Table 1: **Comparison of Training and Testing Correctness on standard datasets**

Dataset (m, k, n)	Method	Average Training Set Correctness	Average Testing Set Correctness
WBCP - old (46, 148, 32)	RLP [3]	70.57	57.76
	Spherical Sep. (Linear k.)	66.32	66.47
	Spherical Sep. (RBF k. - $p_1 = 1$)	74.58	73.18
WBCP - new (41, 69, 32)	Other methods [5, 7]	70.80	68.50
	Spherical Sep. (Linear k.)	66.26	67.20
	Spherical Sep. (RBF k. - $p_1 = 1$)	65.25	69.09
Hearth (83, 214, 13)	Other methods [5, 7]	87.70	86.50
	Spherical Sep. (Linear k.)	75.08	74.50
	Spherical Sep. (Polynomial k. - $p_1 = 8$)	83.39	82.12
	Spherical Sep. (Sigmoidal k. - $p_1 = 10$, $p_2 = 0.1$)	86.61	86.47
Ionosphere (126, 225, 34)	Other methods [5, 7]	97.00	95.80
	Spherical Sep. (Linear k.)	71.45	71.00
	Spherical Sep. (RBF k. - $p_1 = 0.69$)	92.85	88.05
Mushroom (3916, 4208, 22)	Other methods [5, 7]	90.91	88.20
	Spherical Sep. (Linear k.)	75.55	70.75
	Spherical Sep. (Polynomial k. - $p_1 = 30$)	88.18	85.00
	Spherical Sep. (ERBF k. - $p_1 = 0.2$)	89.60	87.60
Galaxy Dim (2082, 2110, 14)	Other methods [5, 7]	95.00	95.00
	Spherical Sep. (Linear k.)	86.16	84.16
	Spherical Sep. (Polynomial k. - $p_1 = 3$)	87.19	85.64
	Spherical Sep. (RBF k. - $p_1 = 0.1$)	88.75	87.19

References

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

[2] A. Astorino and M. Gaudioso, Ellipsoidal separation for classification problems, Optimization Methods and Software (to appear).

[3] K. P. Bennett and O. L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, Optimization Methods and Software, 1, 23-34, 1992.

[4] N. Cristianini and J. Shawe-Taylor, An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.

[5] G. Fung and O. L. Mangasarian, Proximal Support Vector Machine Classifiers, PDF Version Data Mining Institute Technical Report 01-02, February 2001 - Proceedings KDD-2001, San Francisco August 26-29, 2001 - Association for Computing Machinery, New York, 77-86, 2001.

[6] O.L. Mangasarian, Linear and nonlinear separation of patterns by linear programming, Operations Research, 13, 444-452, 1965.

[7] O. L. Mangasarian and David R. Musicant, Lagrangian Support Vector Machines, Journal of Machine Learning Research 1, 161-177, 2001.

[8] P. M. Murphy and D. W. Aha, UCI repository of machine learning databases, www.ics.uci.edu/∼mlearn/MLRepository.html, 1992.

[9] S. Odewahn, E. Stockwell, R. Pennington, R. Humphreys and W. Zumach, Automated star/galaxy discrimination with neural networks, Astronomical Journal, 103, 1, 318-331, 1992.

[10] J.B. Rosen, Pattern separation by convex programming, Journal of Mathematical Analysis and Applications, 10, 123-134, 1965.

[11] D. M. J. Tax and R. P. W. Duin, Data domain description using support vectors, ESANN'1999 proceedings Bruges (Belgium), 21 - 23 April 1999, D - Facto public, ISBN 2-600049-9-x, 251-256.

[12] V. Vapnik, The Nature of the statistical learning theory, Springer Verlag, New York, 1995.

[13] B. Schölkopf, C.J.C. Burges and A. J. Smola eds, Advances in Kernel Methods:Support Vector Learning, The MIT Press, Cambridge, Mass., 1999.