Comput Manag Sci (2009) 6:357–372
DOI 10.1007/s10287-007-0051-2

ORIGINAL PAPER

A fixed-center spherical separation algorithm with kernel transformations for classification problems

A. Astorino · M. Gaudioso

Published online: 14 March 2007

Abstract We consider a special case of the optimal separation, via a sphere, of two discrete point sets in a finite dimensional Euclidean space. In fact we assume that the center of the sphere is fixed. In this case the problem reduces to the minimization of a convex and nonsmooth function of just one variable, which can be solved by means of an "ad hoc" method 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 support vector machine (SVM) approach. Despite of its simplicity the method has provided interesting results on several standard test problems drawn from the binary classification literature.

Keywords Classification · Separability · Kernel methods · Support vector machine

Mathematics Subject Classification (2000): 90C90

This research has been partially supported by the Italian "Ministero dell'Istruzione, dell'Università e della Ricerca Scientifica", under PRIN project *Numerical Methods for Global Optimization and for some classes of Nonsmooth Optimization Problems* (2005017083.002).

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

M. Gaudioso(✉)
Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italy
e-mail: gaudioso@deis.unical.it

1 Introduction

In many supervised machine learning problems the objective is to assign elements to a finite set of classes or categories. For a given set of sample points coming from two classes, we want to construct a function for discriminating between the classes. The goal is to select a function that will efficiently and correctly classify future points. Classification techniques can be used for data mining or pattern recognition, where many applications require a categorization. A few examples of application are text categorization, object recognition in machine vision, cancer diagnosis and many others.

The classical binary classification problem is to discriminate between two finite sets of points in the n-dimensional space, by a separating surface. The problem consists in finding a separating surface minimizing an appropriate measure of the classification error. Several mathematical programming-based approaches for binary classification have been historically proposed (Rosen 1965, Mangasarian 1965, Bennett and Mangasarian 1992). Among the more recent ones we recall the support vector machine (SVM) technique (Cristianini and Shawe-Taylor 2000; Vapnik 1995), where a classifier is constructed by generating a hyperplane far away from the points of the two sets. By adopting kernel transformations within the SVM approach, we can obtain general nonlinear separation surfaces. In this case the basic idea is to map the data into a higher dimensional space (the feature space) and to separate the two transformed sets by means of one hyperplane, that corresponds to a nonlinear surface in the original input space.

In this paper we deal with discrimination of two datasets by means of a sphere, once the center of the sphere is given. This is a very simplified case of spherical separation (Tax and Duin 1999). For some alternative approaches see also Astorino and Gaudioso 2002, 2005.

We suppose that two nonempty and disjoint finite sets of sample 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*. As proposed by Tax and Duin (1999) the 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 $\mathcal{B}$). In general, $(n+1)$ parameters need to be selected: the center of the sphere (a point in $\mathbb{R}^n$) and the radius of the sphere (a scalar in $\mathbb{R}$).

In case the center of the sphere is given, the above problem reduces to the minimization of a function of just one variable (the radius), which is nonsmooth but convex. We develop for such particular case an "ad hoc" algorithm which finds the optimal solution in $O(p \log p)$ time, where $p = \max\{m, k\}$. The proposed algorithm consists basically of two phases: the "sorting" and the "cutting" ones. In first phase the sample points are sorted according to their distance from the center, while in second phase an optimal "cut" is found. Correctness of the algorithm is proved via reformulation of the univariate minimization problem as a structured Linear Program.

The simplification introduced is definitely drastic. Nevertheless, provided a judicious choice of the center is made, we obtain, at a very low computational cost, reasonably good separation results, thanks also to the kernel transformation that can be embedded into our approach in a rather straightforward way. Thus we believe that our approach can be fruitfully adopted at least as a “first-aid” attempt to deal with very large datasets.

The paper is organized as follows. In Sect. 2 we state the problem of finding a minimal volume separating sphere when the center of the sphere is given and describe our algorithm. In Sect. 3 we introduce kernel transformations. The results of some numerical experiments are described in Sect. 4, while some conclusions are drawn in Sect. 5. The proof of the correctness of the algorithm is in the Appendix.

Throughout the paper we adopt the following notations. We denote by $\|.\|$ the Euclidean norm in $\mathbb{R}^n$ and by $a^{\mathrm{T}}b$ 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 Minimal volume separating sphere

A possible 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 $\mathcal{B}$. 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.

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

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

The sets $\mathcal{A}$ and $\mathcal{B}$ are defined to be spherically separated by $S(x_0, R)$ if

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

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

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

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

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

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

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

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

We observe that spherical separation seems a particularly promising approach if compared with linear separation. In fact, linear separation can be considered a special case of spherical separation with infinite distance of the center of the sphere from the points of the dataset.

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

$$\min_{x_0,R} R^2 + C\sum_{i=1}^{m} \max\{0, (a_i - x_0)^{\mathrm{T}}(a_i - x_0) - R^2\} + \\ + C\sum_{l=1}^{k} \max\{0, R^2 - (b_l - x_0)^{\mathrm{T}}(b_l - x_0)\}, \tag{1}$$

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

We observe that the above problem requires minimization of a function which is nonsmooth and nonconvex and is, apart from the smooth quadratic term R^2, the sum of several functions of the max type.

It is possible to get rid of the nonsmoothness by transforming the unconstrained problem above into the following constrained optimization problem, where the additional variables ξ_i's and μ_l's have been introduced.

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

We obtain a drastic simplification of problem (1) if we do not consider any longer the center x_0 of the sphere as a decision variable. In fact we assume that it is known and fixed. Such simplification is based on the idea that, at least as a tentative approximation, any centroid for the set $\mathcal{A}$ is worth considering as a possible center of the sphere. In Sect. 5 we propose two different choices for x_0.

Once the center x_0 of the sphere is assumed known, by introducing the change of variable

$$z = R^2, \quad z \geq 0 \tag{3}$$

and by defining:

$$\begin{array}{ll} c_i \overset{\triangle}{=} (a_i - x_0)^{\mathrm{T}}(a_i - x_0) \geq 0 & \forall\, i = 1,\ldots,m \\ d_l \overset{\triangle}{=} (b_l - x_0)^{\mathrm{T}}(b_l - x_0) \geq 0 & \forall\, l = 1,\ldots,k \end{array} \tag{4}$$

the problem (1) becomes:

$$\min_{z\geq 0} z + C\left(\sum_{i=1}^{m}\max\{0, c_i - z\} + \sum_{l=1}^{k}\max\{0, z - d_l\}\right), \tag{5}$$

which is a convex, piecewise affine minimization problem in the scalar (non-negative) variable z.

Problem (5) can be approached by means of standard univariate minimization techniques. Nevertheless it is possible to devise a quite efficient algorithm which provides an exact solution in $O(p\log p)$ time, where $p = \max\{m, k\}$.

To state our algorithm and to prove its termination, it is useful to restate problem (5) in the form of a linear program as follows:

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

Of course z^*, the optimal value of the variable z, at any optimal solution for problem (6), provides an optimal solution to problem (5) too.

The dual of the above problem (6) is the following:

$$\begin{aligned} f_D = \max_{\alpha,\beta} \quad & \sum_{i=1}^{m} c_i\alpha_i - \sum_{l=1}^{k} d_l\beta_l & \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i - \sum_{l=1}^{k}\beta_l \leq 1 & \\ & 0 \leq \alpha_i \leq C & \forall\, i = 1,\dots,m \\ & 0 \leq \beta_l \leq C & \forall\, l = 1,\dots,k. \end{aligned} \tag{7}$$

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

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

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

$$\begin{cases} z\left(\sum_{i=1}^{m}\alpha_i - \sum_{l=1}^{k}\beta_l - 1\right) = 0 & \\ \xi_i\,(C-\alpha_i) = 0 & \forall\, i = l,\dots,m \\ \mu_l\,(C-\beta_l) = 0 & \forall\, l = l,\dots,k \end{cases} \tag{8}$$

$$\begin{cases} \alpha_i\,(z - c_i + \xi_i) = 0 & \forall\, i = l,\dots,m \\ \beta_l\,(-z + d_l + \mu_l) = 0 & \forall\, l = l,\dots,k. \end{cases}$$

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 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 $\sum_{i=1}^{m}\alpha_i - \sum_{l=1}^{k}\beta_l \le 1$ 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

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

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

$$\begin{array}{ll} \xi_i^* = c_i & \forall\, i = 1,\dots,m \\ \mu_l^* = 0 & \forall\, l = 1,\dots,k \end{array} \quad \text{and } f_P = C\sum_{i=1}^{m} c_i.$$

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

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

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

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

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

$$\begin{array}{ll}
\bar{\xi}_i = c_i - \bar{z} & \forall\, i = 1, \ldots, m \\
\bar{\mu}_l = 0 & \forall\, l = 1, \ldots, k.
\end{array}$$

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 $C > 1/m$. 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

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

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

2.1 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:

$$c_1 \geq c_2 \geq \cdots \geq c_m > 0 \quad \text{and } 0 < d_1 \leq d_2 \leq \cdots \leq d_k. \tag{12}$$

Now we describe an algorithm that finds the optimal solution to the dual problem (11) for $C > 1/m$. The basic idea is that, under the above numbering of the variables, there exists an optimal solution such that if any variable, say $\alpha_{\hat{\imath}}$ $(\beta_{\hat{l}})$, is positive at the optimum, then all the variables $\alpha_i,\ i < \hat{\imath}$ $(\beta_l,\ l > \hat{l})$ are equal to C.

For sake of notational simplicity we state the algorithm for the case $C \leq 1$. The algorithm for the case $C > 1$ is completely analogous and is described in the Appendix 1.

Algorithm 1 (Case $C \leq 1$)

```
Initialization
   Set
   – $r \overset{\triangle}{=} \left\lfloor \frac{1}{C} \right\rfloor$ (Remark: $m \geq r+1$)
   – $\bar{p} \overset{\triangle}{=} \min\{k, m-r\} \geq 1$
   – $\alpha_i = 0 \quad \forall\, i = 1, \ldots, m$
   – $\beta_l = 0 \quad \forall\, l = 1, \ldots, k$
Step 1. Set $\alpha_i = C \quad \forall\, i = 1, \ldots, r$
      If $(c_{r+1} \leq d_1)$ Set $\alpha_{r+1} = 1 - Cr$ and STOP [exit (a): the basic variable is $\alpha_{r+1}$].
        (Remark: If $C = 1$, then $\alpha_{r+1} = 0$ and the solution is a degenerate basic feasible solution).
      Endif
      If $(c_{r+i} > d_i \quad \forall\, i = 1, \ldots, \bar{p})$
            If $(\bar{p} > 1)$ Set $\alpha_{r+i} = C \quad \forall\, i = 1, \ldots, (\bar{p}-1)$
            Endif
         Set
         – $\alpha_{r+\bar{p}} = 1 - Cr$
         – $\beta_l = C \quad \forall\, l = 1, \ldots, (\bar{p}-1)$
         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
         – $\alpha_{r+i} = C \quad \forall\, i = 1, \ldots, (p^*-1)$
         – $\alpha_{r+p^*} = 1 - Cr$
```

```
    – $\beta_l = C \quad \forall\, l = 1,\ldots,(p^*-1)$
    and STOP [exit (c): the basic variable is $\alpha_{r+p^*}$].
Else Set $\alpha_{r+i} = C \quad \forall\, i = 1,\ldots,(p^*-1)$
        If $(p^* > 2)$ Set $\beta_l = C \quad \forall\, l = 1,\ldots,(p^*-2)$
        Endif
    Set $\beta_{p^*-1} = C(r+1)-1$ and STOP [exit (d): the basic variable is $\beta_{p^*-1}$].
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 1 *The algorithm (1) finds an optimal solution to problem (11).*

Proof See Appendix 2. □

Once the optimal solutions $(\bar{z}, \bar{\xi}, \bar{\mu})$ and $(\bar{\alpha}, \bar{\beta})$ for (6) and (11) respectively have been calculated, the optimal solution of problem (5) is also available. Recalling the substitution (3), $R^2 = z$, 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:

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

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

3 Using the kernels

Kernel transformation of the type adopted in SVM (see Vapnik 1995 and Schölkopf et al. 1999) 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

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

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

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

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 $Q = \{x_1, \dots, x_q\}$ be a finite point set in the input space and

$$\hat{Q} = \{\phi(x_1), \dots, \phi(x_q)\}$$

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:

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

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

Now we proceed looking for sphere in the feature space centered in the barycenter of the set $\hat{\mathcal{A}}$, with radius $\hat{R} \in \mathbb{R}$, with the objective of minimizing both the volume and the classification error.

We obtain the following problem

$$\begin{array}{lll}
\hat{f}_P = \min\limits_{\hat{z}, \hat{\xi}, \hat{\mu}} & \hat{z} + C \left(\sum\limits_{i=1}^{m} \hat{\xi}_i + \sum\limits_{l=1}^{k} \hat{\mu}_l \right) & \\
\text{s.t.} & \hat{z} - \hat{c}_i + \hat{\xi}_i \geq 0 & \forall\, i = 1, \dots, m \\
& \hat{d}_l - \hat{z} + \hat{\mu}_l \geq 0 & \forall\, l = 1, \dots, k \\
& \hat{z} \geq 0 & \\
& \hat{\xi}_i \geq 0 & \forall\, i = 1, \dots, m \\
& \hat{\mu}_l \geq 0 & \forall\, l = 1, \dots, k
\end{array} \qquad (13)$$

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

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

The problem (13) is a Linear Program of the same type as problem (6). As in Sect. 2 its dual is stated in the form:

$$\begin{array}{lll}
\hat{f}_D = \max\limits_{\hat{\alpha},\hat{\beta}} & \sum\limits_{i=1}^{m} \hat{c}_i\hat{\alpha}_i - \sum\limits_{l=1}^{k} \hat{d}_l\hat{\beta}_l & \\
\text{s.t.} & \sum\limits_{i=1}^{m} \hat{\alpha}_i - \sum\limits_{l=1}^{k} \hat{\beta}_l = 1 & \\
& 0 \leq \hat{\alpha}_i \leq C & \forall\, i = 1, \ldots, m \\
& 0 \leq \hat{\beta}_l \leq C & \forall\, l = 1, \ldots, k
\end{array} \tag{14}$$

and can be solved by the algorithm (1).

Once the optimal solutions $(\hat{\alpha}^*, \hat{\beta}^*)$ and $(\hat{z}^*, \hat{\xi}^*, \hat{\mu}^*)$ for (14) and (13), respectively, have been calculated, the sphere $S(\phi_{\hat{\mathcal{A}}}, \sqrt{\hat{z}^*})$ 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

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

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

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

4 Numerical experiments

We have implemented the algorithm described in Sect. 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:

- Four publicly available datasets from the UCI Machine Learning Repository Murphy and Aha (1992), in particular, the Wisconsin Breast Cancer Prognosis (WBCP), the Cleveland Heart Disease (Heart), Ionosphere (Ionosphere), Mushroom (Mushroom).
- The Galaxy Dim dataset (Galaxy Dim) used in galaxy discrimination with neural networks from Odewahn et al. (1992).

In our implementation we have used the following kernel functions:

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

with parameter p_1.

We have run our algorithm for several values of the kernel parameters and of the positive weighting constant C. The center x_0 has been defined either as the barycenter of the set $\mathcal{A}$:

$$x_0^{(1)} = \frac{1}{m}\sum_{i=1}^{m} a_i,$$

or as a point "far" from both the sets $\mathcal{A}$ and $\mathcal{B}$:

$$x_0^{(2)} = \frac{1}{m}\sum_{i=1}^{m} a_i + M\left(\frac{1}{m}\sum_{i=1}^{m} a_i - \frac{1}{k}\sum_{l=1}^{k} b_l\right),$$

for some sufficiently large positive constant M. Whenever nonlinear kernel functions have been adopted, the point $\phi(x_0)$ has been selected as $\phi_{\hat{\mathcal{A}}}$, the barycenter of the set $\hat{\mathcal{A}}$.

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.

In Table 1 we have reported for each dataset the results obtained in terms of averages on the tenfold cross-validation with the choice of the linear kernel. To provide a comparison opportunity we have reported in the same Table 1 also some results drawn from the literature, in particular those reported in Fung and Mangasarian (2001) related to the use of the SVM-light (Linear Kernel)

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 (41, 69, 32)	SVM-light (Linear K.)	62.7	62.7
	Spherical Sep. (Linear K. - $C = 10, x_0^{(1)}$)	66.70	67.20
	Spherical Sep.	68.61	67.70
	(Linear K. - $C = 10, x_0^{(2)}, M = 10^4$)		
Heart (83, 214, 13)	SVM-light (Linear K.)	87.7	86.5
	Spherical Sep.	75.08	74.50
	(Linear K. - $C = 10, x_0^{(1)}$)		
	Spherical Sep.		
	(Linear K. - $C = 10, x_0^{(2)}, M = 10^4$)	87.58	86.80
Ionosphere	SVM-light (Linear K.)	91.4	88.0
(126, 225, 34)	Spherical Sep.	71.54	71.00
	(Linear K. - $C = 10, x_0^{(1)}$)		
	Spherical Sep.	83.32	82.41
	(Linear K. - $C = 0.03, x_0^{(2)}, M = 10^4$)		
Mushroom	SVM-light (Linear K.)	81.5	81.5
(3916, 4208, 22)	Spherical Sep.	75.55	70.75
	(Linear K. - $C = 10, x_0^{(1)}$)		
	Spherical Sep.	77.11	75.37
	(Linear K. - $C = 0.01, x_0^{(2)}, M = 10^{14}$)		
Galaxy Dim	SVM-light (Linear K.)	94.2	94.1
(2082, 2110, 14)	Spherical Sep.	86.16	84.16
	(Linear K. - $C = 10, x_0^{(1)}$)		
	Spherical Sep.	89.60	88.62
	(Linear K. - $C = 0.03, x_0^{(2)}, M = 10^{14}$)		

Table 2 Training and testing correctness with nonlinear kernel functions

Dataset	Method	Average training set	Average testing set
(m, k, n)		correctness	correctness
Ionosphere	Spherical Sep.	92.85	88.05
(126, 225, 34)	(RBF k. - $C = 10, \phi_{\hat{\mathcal{A}}}, p_1 = 0.7$)		
Mushroom	Spherical Sep.	89.60	87.60
(3916, 4208, 22)	(ERBF k. - $C = 10, \phi_{\hat{\mathcal{A}}}, p_1 = 0.2$)		

approach (Joachims 1999). We remark that for Galaxy Dim, a relatively large dataset, the entire dataset has been used.

In Table 2 just the improving results, obtained with a different choice of kernel function, are reported.

The different parameters have been set after a tuning procedure aimed at finding the value which ensures on the average the best performance.

5 Conclusions

The results provided by our implementation show that the use of nonlinear kernel transformations in connection with the spherical separation approach improves on the classification correctness with respect to the spherical separation approach with linear kernel, i.e. spherical separation in the sample space. On the other hand our algorithm has provided, at a very low computational cost, results which appear comparable with those obtained by a well established method. This fact would suggest to consider our approach as one of the election tools to deal with very large datasets.

Appendix 1

We restate algorithm (1) of Sect. 2 for the case $C > 1$.

```
Initialization
  Set
  – $\bar{p} \overset{\triangle}{=} \min\{k,m\} \geq 1$
  – $\alpha_i = 0 \quad \forall\, i = 1,\dots,m$; $\beta_l = 0 \quad \forall\, l = 1,\dots,k$
Step 1.
      If $(c_1 \leq d_1)$ Set $\alpha_1 = 1$ and STOP [exit (a): the basic variable is $\alpha_1$].
      Endif
      If $(c_i > d_i \quad \forall\, i = 1,\dots,\bar{p})$
            If $(\bar{p} > 1)$ Set $\alpha_i = C \quad \forall\, i = 1,\dots,(\bar{p}-1)$
            Endif
         Set
         – $\alpha_{\bar{p}} = 1$
         – $\beta_l = C \quad \forall\, l = 1,\dots,(\bar{p}-1)$
         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
         – $\alpha_i = C \quad \forall\, i = 1,\dots,(p^*-1)$
         – $\alpha_{p^*} = 1$
         – $\beta_l = C \quad \forall\, l = 1,\dots,(p^*-1)$
         and STOP [exit (c): the basic variable is $\alpha_{p^*}$].
      Else Set $\alpha_i = C \quad \forall\, i = 1,\dots,p^*$
            If $(p^* > 2)$ Set $\beta_l = C \quad \forall\, l = 1,\dots,(p^*-2)$
            Endif
         Set $\beta_{p^*-1} = C - 1$ and STOP [exit (d): the basic variable is $\beta_{p^*-1}$].
      Endif
```

Appendix 2

Proof of Theorem 1 We assume that $\bar{\alpha} \geq 0$ and $\bar{\beta} \geq 0$ are those obtained on exit from the algorithm (1). 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

$$\sum_{i=1}^{m} \bar{\alpha}_i - \sum_{l=1}^{k} \bar{\beta}_l = 1 \tag{15}$$

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

$$\bar{\xi}_i = \begin{cases} 0 & \text{if } \bar{\alpha}_i = 0 \\ c_i - \bar{z} & \text{if } \bar{\alpha}_i = C \end{cases} \quad \text{for } i = 1, \ldots, m; i \neq h \tag{16}$$

$$\bar{\mu}_l = \begin{cases} 0 & \text{if } \bar{\beta}_l = 0 \\ \bar{z} - d_l & \text{if } \bar{\beta}_l = C \end{cases} \quad \text{for } l = 1, \ldots, k. \tag{17}$$

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

$$\bar{\xi}_i = \begin{cases} 0 & \text{if } \bar{\alpha}_i = 0 \\ c_i - \bar{z} & \text{if } \bar{\alpha}_i = C \end{cases} \quad \text{for } i = 1, \ldots, m \tag{18}$$

$$\bar{\mu}_l = \begin{cases} 0 & \text{if } \bar{\beta}_l = 0 \\ \bar{z} - d_l & \text{if } \bar{\beta}_l = C \end{cases} \quad \text{for } l = 1, \ldots, k; l \neq s. \tag{19}$$

It easy to verify that the complementary slackness conditions (8) are satisfied as consequence of (15) and of the variable setting (16), (17), (18) and (19).

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 $\bar{z}-c_i+\bar{\xi}_i \geq 0 \ \forall\ i=1,\ldots,m$ and $d_l-\bar{z}+\bar{\mu}_l \geq 0 \ \ \forall\ l=1,\ldots,k$ 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. □

References

Astorino A, Gaudioso M (2002) Polyhedral separability through successive LP. J OptimTheory Appl 112(2):265–293

Astorino A, Gaudioso M (2005) Ellipsoidal separation for classification problems. Optim Methods Softw 20(2–3):261–270

Bennett KP, Mangasarian OL (1992) Robust linear programming discrimination of two linearly inseparable sets. Optim Methods Softw 1:23–34

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

Fung G, Mangasarian OL (2001) 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, pp 77–86

Joachims T (1999) Making large-scale support vector machine learning practical. In: Schölkopf B, Burges CJC, Smola AJ (eds) Advances in kernel methods: support vector learning. MIT, Cambridge, pp 169–184

Mangasarian OL (1965) Linear and nonlinear separation of patterns by linear programming. Oper Res 13:444–452

Murphy PM, Aha DW (1992) UCI repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html

Odewahn S, Stockwell E, Pennington R, Humphreys R, Zumach W (1992) Automated star/galaxy discrimination with neural networks. Astron J 103(1):318–331

Rosen JB (1965) Pattern separation by convex programming. J Math Anal Appl 10:123–134

Schölkopf B, Burges CJC, Smola AJ (eds) (1999) Advances in kernel methods: support vector learning. MIT, Cambridge

Tax DMJ, Duin RPW (1999) Data domain description using support vectors. In: ESANN'1999 proceedings Bruges (Belgium), 21–23 April 1999, D-Facto public, ISBN 2-600049-9-x, pp 251–256

Vapnik V (1995) The nature of the statistical learning theory. Springer, New York