Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 173 (2006) 746–761
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Two-group classification via a biobjective margin maximization model ☆

Emilio Carrizosa *, Belen Martin-Barragan

Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Avda. Reina Mercedes, s/n, 41012 Sevilla, Spain

Received 15 October 2004; accepted 15 June 2005
Available online 18 November 2005

Abstract

In this paper we propose a biobjective model for two-group classification via margin maximization, in which the margins in both classes are simultaneously maximized. The set of Pareto-optimal solutions is described, yielding a set of parallel hyperplanes, one of which is just the solution of the classical SVM approach.

In order to take into account different misclassification costs or a priori probabilities, the ROC curve can be used to select one out of such hyperplanes by expressing the adequate tradeoff for sensitivity and specificity. Our result gives a theoretical motivation for using the ROC approach in case misclassification costs in the two groups are not necessarily equal.

Keywords: Multiple objective programming; Support vector machines; Biobjective; ROC curve; Classification; Data mining

1. Introduction

In the last decade, support vector machine (SVM), e.g., [1–5], has shown to be a powerful tool for the two-group classification problem. This method attempts to build a hyperplane with maximal margin that linearly separates the two groups, in the sense that it correctly classifies the whole set I of objects in the database, and the distance to the closest point is maximized. When linear separation is not possible, perturbations in the model are allowed. Moreover, the use of SVM is theoretically justified by the dependence on the margin of certain bounds on the probability of misclassifying a forthcoming object, e.g., [6–8].

☆ This research was partially supported by projects BFM2002-11282-E and BFM2002-04525-C02-02 of Ministerio de Ciencia y Tecnología (Spain) and FQM-329 of Plan Andaluz de Investigación (Andalucía, Spain).
* Corresponding author.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), belmart@us.es (B. Martin-Barragan).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.06.059

However, crude SVM cannot take into account different misclassification costs or known a priori probabilities. In this work, we formulate a new model in which margins between each class and the hyperplane are dealt independently. We study the simultaneous maximization of both margins, i.e., the distance to the closest point in each group. This yields a biobjective problem, whose Pareto-optimal solutions are sought. In other words, we seek the set of hyperplanes such that there is not any other hyperplane having greater margin for both classes, thus we expect their performance cannot be improved simultaneously with respect to both classes.

Our main result states that the set of all Pareto-optimal solutions is described as a set of parallel hyperplanes, which can be easily computed.

The paper is organized as follows: In Section 2, the problem is formally introduced, and notation is set. As in classical SVM approaches, we deal separately with the linearly separable case (Section 3) and the case in which the two classes are not linearly separable (Section 4). Some illustrative examples, as well as a visual procedure for choosing β based on the receiver operating characteristic (ROC) curves are given in Section 5, ending with some concluding remarks.

2. The problem

We have a set of objects Ω, each object u having two components $u = (x^u, y^u)$. The first component x^u is called the *predictor vector* and takes values in $\mathbb{R}^p$, whose components x_l, $l = 1, 2, \ldots, p$, are called *predictor variables*. The other component y^u, takes values in the set of classes $\mathscr{C} = \{-1, 1\}$ and is called the *class-membership* of object u. Object u is said to belong to class y^u.

In general, class-membership of objects in Ω is known only for a subset I, called the *training sample*: both predictor vector and class-membership are known for $u \in I$, whereas only x^u is known for $u \in \Omega \backslash I$.

Denote by $I_c = \{u \in I : y^u = c\}$, the set of objects of class c, for every $c \in \mathscr{C}$. We assume throughout this paper that each class is represented in the training sample, i.e., $I_c \neq \emptyset$, $\forall c \in \mathscr{C}$.

Our framework to classify objects is as follows. We are seeking a score function of the form:

$$f(x) = \sum_{k=1}^{p} \omega_k x_k + \beta = \omega^\top x + \beta, \tag{1}$$

where $\omega = (\omega_1, \omega_2, \ldots, \omega_p) \neq 0$. Such a function defines a hyperplane in the predictor space $\mathbb{R}^p$, given by $\{x \in \mathbb{R}^p : \omega^\top x + \beta = 0\}$. We identify with (ω, β) such a hyperplane.

An object $u \in \Omega$ will be allocated to class -1 if $f(x^u) < 0$, i.e., if x^u belongs to the halfspace $\{x \in \mathbb{R}^p : \omega^\top x + \beta < 0\}$. Analogously, u will be allocated to class 1 if $f(x^u) > 0$, i.e., if x^u belongs to the halfspace $\{x \in \mathbb{R}^p : \omega^\top x + \beta > 0\}$. In case of ties, i.e., when x^u belongs to the hyperplane, objects can be allocated randomly or by a prefixed order.

Our problem is to find a hyperplane (ω, β) correctly classifying all (or at least, most) objects $u \in I$, and enjoying good generalization properties, in the sense that one can expect the good behavior obtained in I to be generalized to Ω.

In this paper, as done in classical SVM [2–4,6,7], we address first the case in which a hyperplane correctly classifying all objects in the training sample exists, and consider later the case in which such hyperplane does not exist. Let us give a formal definition of linear separability.

Definition 1. A hyperplane (ω, β), is said to *separate* linearly I if

$$y^u(\omega^\top x^u + \beta) > 0, \quad \forall u \in I. \tag{2}$$

Moreover, I is said to be linearly separable if there exists a hyperplane (ω, β) linearly separating I.

3. The linearly separable case

In this section, we address the case in which I is linearly separable, i.e., (2) is satisfied by some (ω, β).

Although this condition may be rather restrictive, it is usually fulfilled in practice by mapping the data into a vector space of higher dimension, in such a way that the predictor vectors become linearly separable. See, e.g., [4,5].

The theoretical basis of SVM was developed by Vapnik et al. [3,6,7], where they gave bounds on the generalization ability of a linear classifier in terms of the margin, as defined, e.g., in [4].

Definition 2. Given the hyperplane (ω, β), the *margin of an object* u, is defined as the Euclidean distance between x^u and the hyperplane (ω, β), with positive sign if u is correctly classified, and negative sign otherwise, i.e.,

$$\rho^u(\omega, \beta) = \frac{y^u(\omega^\top x^u + \beta)}{\|\omega\|}, \tag{3}$$

where $\|\cdot\|$ stands for the Euclidean norm. The margin of a class $c \in \mathscr{C}$ in the training sample I is the minimum margin over all objects $u \in I_c$,

$$\rho^c(\omega, \beta) = \min_{u \in I_c} \rho^u(\omega, \beta), \tag{4}$$

and the margin of the training sample I, is the minimum over all the objects u in I

$$\rho^I(\omega, \beta) = \min_{u \in I} \rho^u(\omega, \beta). \tag{5}$$

Under the assumption that I is linearly separable, we can construct the so-called *hard margin hyperplane* [3,6,7], which is the hyperplane which linearly separates I and has maximal margin ρ^I. This is usually obtained by solving the problem

$$\begin{aligned} &\min \ \|\omega\|^2 \\ &\text{s.t.} \ \ y^u(\omega^\top x^u + \beta) \geqslant 1, \quad \forall u \in I, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}. \end{aligned} \tag{6}$$

It is shown in [6,7] that (6) always has just one optimal solution.

This approach does not take into account different misclassification costs or known a priori probabilities.

We propose a novel approach, in which instead of maximizing the margin on I, we simultaneously maximize the margin on both classes as defined in (4). This yields the following biobjective optimization problem with open feasible region:

$$\begin{aligned} &\max \ \{\rho^1(\omega, \beta), \rho^{-1}(\omega, \beta)\} \\ &\text{s.t.} \ \ y^u(\omega^\top x^u + \beta) > 0, \quad \forall u \in I, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \omega \neq 0. \end{aligned} \tag{7}$$

We seek the set of Pareto-optimal solutions to (7), i.e., the set of feasible solutions $(\bar{\omega}, \bar{\beta})$ such that no (ω, β) exists such that

$$\begin{aligned} &\rho^1(\omega, \beta) \geqslant \rho^1(\bar{\omega}, \bar{\beta}), \\ &\rho^{-1}(\omega, \beta) \geqslant \rho^{-1}(\bar{\omega}, \bar{\beta}), \end{aligned} \tag{8}$$

with at least one inequality strict. Note that, for $u \in I$,

$$\rho^u(\mu\omega, \mu\beta) = \rho^u(\omega, \beta), \quad \forall \mu > 0, \ \forall \omega \in \mathbb{R}^p, \ \omega \neq 0, \ \forall \beta \in \mathbb{R}. \tag{9}$$

Hence, for all $\mu > 0$, $\omega \in \mathbb{R}^p$, $\omega \neq 0$, $\beta \in \mathbb{R}$, one has that

$$\begin{aligned} &\rho^1(\mu\omega, \mu\beta) = \rho^1(\omega, \beta), \\ &\rho^{-1}(\mu\omega, \mu\beta) = \rho^{-1}(\omega, \beta). \end{aligned} \tag{10}$$

Hence, if (ω, β) is a Pareto-optimal solution to (7), then, for any $\mu > 0$, $(\mu\omega, \mu\beta)$ is also feasible for (7), and, by (10), it is also a Pareto-optimal solution to (7).

Our final aim is to construct classifiers with adequate tradeoff of misclassification costs in the two groups *in* Ω. In other words, we ideally would solve the biobjective problem

$$\begin{aligned} &\max \ \{\rho^1(\omega, \beta), \rho^{-1}(\omega, \beta)\} \\ &\text{s.t.} \quad y^u(\omega^\top x^u + \beta) > 0, \quad \forall u \in \Omega, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \omega \neq 0, \end{aligned} \tag{11}$$

by describing the set of Pareto-optimal solutions.

Since the class y^u of $u \in \Omega$ is known only for the objects $u \in I$, we consider (7) as a surrogate of (11), and thus the set of Pareto-optimal solutions of (7) is seen as an approximation to the set of Pareto-optimal solutions of (11).

First let us recall that $(\bar{\omega}, \bar{\beta})$ is a weakly efficient solution of Problem (7) if no feasible (ω, β) exists that is strictly better than $(\bar{\omega}, \bar{\beta})$ for both objectives, i.e.,

$$\begin{aligned} &\rho^1(\omega, \beta) > \rho^1(\bar{\omega}, \bar{\beta}), \\ &\rho^{-1}(\omega, \beta) > \rho^{-1}(\bar{\omega}, \bar{\beta}). \end{aligned} \tag{12}$$

We refer the reader to, e.g., [9] for further details on these concepts of vector optimization.

Since all feasible solutions (ω, β) satisfy that $\rho^1(\omega, \beta) > 0$ and $\rho^{-1}(\omega, \beta) > 0$ one can generate all weakly efficient solutions by solving max–min type scalarizations [9].

For the sake of completeness we state the following technical result:

Lemma 3. *The set of weakly efficient solutions of Problem (7) is obtained as the set of optimal solutions of*

$$\begin{aligned} &\max\min \ \{\rho^1(\omega, \beta), \theta\rho^{-1}(\omega, \beta)\} \\ &\text{s.t.} \qquad y^u(\omega^\top x^u + \beta) > 0, \quad \forall u \in I, \\ &\qquad\quad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \omega \neq 0, \end{aligned} \tag{13}$$

when $\theta \in (0, +\infty)$, *in the sense that*

(1) *any optimal solution of Problem (13) is weakly efficient for Problem (7),*
(2) *for every weakly efficient solution* $(\omega_\theta, \beta_\theta)$ *of Problem (7) there exists* $\theta \in (0, +\infty)$, *such that* $(\omega_\theta, \beta_\theta)$ *is optimal for Problem (13).*

For $\theta \in (0, +\infty)$, define

$$A_\theta = \frac{2\theta}{\theta + 1},$$

$$y^u_\theta = \begin{cases} 1 & \text{if } u \in I_1, \\ -\theta & \text{if } u \in I_{-1}, \end{cases}$$

and consider the convex quadratic problem

$$\begin{aligned} &\min \ \|\omega\|^2 \\ &\text{s.t.} \quad y_\theta^u(\omega^\top x^u + \beta) \geqslant A_\theta, \quad \forall u \in I, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}. \end{aligned} \tag{P_θ}$$

Observe that Problem (6) is a particular case, since it corresponds to the case $\theta = 1$.

Lemma 4. *One has*:

(1) *For each* $\theta \in (0, +\infty)$, *Problem* (P_θ) *has a unique optimal solution* $(\omega_\theta, \beta_\theta)$, *satisfying* $\omega_\theta \neq 0$.
(2) *Given* $(\omega, \beta) \in \mathbb{R}^p \times \mathbb{R}$, *the following statements are equivalent*:
   (a) *There exists* $\theta \in (0, +\infty)$ *such that* (ω, β) *is optimal for Problem* (P_θ).
   (b) $\omega = \omega_1$ *and* $|\beta - \beta_1| < 1$, *where* (ω_1, β_1) *is the optimal solution for Problem* (P_θ) *for* $\theta = 1$.

Proof. For θ given, since the function $\omega \mapsto \|\omega\|^2$ is strictly convex, there exists a unique ω_θ such that any optimal solution (ω, β) for Problem (P_θ) has $\omega = \omega_\theta$. We show now that the set of optimal solutions of Problem (P_θ) is a singleton.

For Problem (P_θ), KKT conditions at (ω_θ, β), which are necessary and sufficient for optimality, are given by

$$\begin{aligned} &\lambda_\theta^u \geqslant 0, \quad \forall u \in I, \\ &2\omega_\theta - \sum_{u \in I} \lambda\theta^u y_\theta^u x^u = 0, \\ &\sum_{u \in I} \lambda_\theta^u y_\theta^u = 0, \\ &\lambda_\theta^u [y_\theta^u(\omega_\theta^\top x^u + \beta) - A_\theta] = 0, \quad \forall u \in I. \end{aligned} \tag{14}$$

First of all, note that $\lambda_\theta \neq 0$. Indeed, if $\lambda_\theta^u = 0$ for all $u \in I$, one would have $\omega_\theta = 0$, which simultaneously implies, since (ω_θ, β) is feasible, that $\beta \geqslant A_\theta > 0$ and $\beta \leqslant \frac{A_\theta}{-\theta} < 0$. This is a contradiction, and hence $\lambda_\theta \neq 0$.

Hence, for any (ω_θ, β), optimal for (P_θ) there exists $u \in I$ such that

$$y_\theta^u\left(\omega_\theta^\top x^u + \beta\right) - A_\theta = 0, \tag{15}$$

i.e.,

$$\beta = \frac{A_\theta}{y_\theta^u} - \omega_\theta^\top x^u. \tag{16}$$

This means that the set of optimal solutions of Problem (P_θ) is finite. On the other hand, by convexity, for any two different optimal solutions of Problem (P_θ), all the solutions in the segment between them are optimal. This contradicts the finiteness of the set of optimal solutions of Problem (P_θ), yielding the conclusion that such a set has an unique solution, $(\omega_\theta, \beta_\theta)$, with β_θ of the form (16) for some $u \in I$.

In order to prove the second part of the Lemma we show that for $\theta \in (0, +\infty)$ the unique optimal solution $(\omega_\theta, \beta_\theta)$ of Problem (P_θ) is given by

$$\begin{aligned} &\omega_\theta = \omega_1, \\ &\beta_\theta = \beta_1 + \frac{\theta - 1}{\theta + 1}. \end{aligned}$$

To show this, let λ_1 be the multipliers in (14) for $\theta = 1$ and define the multipliers λ_θ as

$$\lambda_\theta^u = \lambda_1^u, \quad \forall u \in I_1,$$
$$\lambda_\theta^u = \frac{1}{\theta}\lambda_1^u, \quad \forall u \in I_{-1}.$$

It is easy to see that $(\omega_\theta, \beta_\theta, \lambda_\theta)$ satisfies (14). Indeed,

$$\sum_{u\in I} \lambda_\theta^u y_\theta^u x^u = \sum_{u\in I_1} \lambda_1^u x^u - \sum_{u\in I_{-1}} \frac{1}{\theta}\lambda_1^u \theta x^u = \sum_{u\in I} \lambda_1^u y_1^u x^u = 2\omega_1,$$

and

$$\sum_{u\in I} \lambda_\theta^u y_\theta^u = \sum_{u\in I_1} \lambda_1^u - \sum_{u\in I_{-1}} \frac{1}{\theta}\lambda_1^u \theta = \sum_{u\in I} \lambda_1^u y_1^u = 0.$$

Since, λ_θ^u is equal to zero iff $\lambda_1^u = 0$, and in such a case, they trivially satisfy the last set of equations of (14), we just have to prove that for all $u \in I$ with $\lambda_1^u > 0$, it holds

$$y_\theta^u\left(\omega_1^\top x^u + \beta_\theta\right) - A_\theta = 0. \tag{17}$$

Let $u \in I_1$ such that $\lambda_1^u \neq 0$. By (14), one has $\omega_1^\top x^u + \beta_1 = 1$. After substituting it into (17), it yields

$$y_\theta^u\left(\omega_1^\top x^u + \beta_\theta\right) - A_\theta = \omega_1^\top x^u + \beta_1 + \frac{\theta-1}{\theta+1} - \frac{2\theta}{\theta+1} = 1 - 1 = 0. \tag{18}$$

Analogously, let $u \in I_{-1}$ such that $\lambda_1^u \neq 0$. Then, $\omega_1^\top x^u + \beta_1 = -1$, and by substituting it into (17) it yields

$$y_\theta^u(\omega_1^\top x^u + \beta_\theta) - A_\theta = -\theta\left(\omega_1^\top x^u + \beta_1 + \frac{\theta-1}{\theta+1}\right) - \frac{2\theta}{\theta+1} = 0. \tag{19}$$

Hence, we conclude that (ω_1, β_θ) with $\beta_\theta = \beta_1 + \frac{\theta-1}{\theta+1}$, is the unique optimal solution of Problem (P_θ). It means that the set of all optimal solutions of Problem (P_θ) for all $\theta \in (0, +\infty)$, is given by

$$\{(\omega_1, \beta_\theta) : \theta \in (0, +\infty)\} = \left\{(\omega_1, \beta) : \beta = \beta_1 + \frac{\theta-1}{\theta+1} \text{ for some } \theta \in (0, +\infty)\right\}$$
$$= \{(\omega_1, \beta) : |\beta - \beta_1| < 1\}. \qquad \square$$

Theorem 5. *The set W of weakly efficient solutions of the biobjective Problem (7) is given by*

$$W = \{(\mu\omega_1, \mu\beta) : |\beta - \beta_1| < 1, \mu > 0\},$$

where (ω_1, β_1) is the optimal solution of Problem (P_1).

Proof. Let $(\bar\omega, \bar\beta) \in \mathbb{R}^p \times \mathbb{R}$. By Lemma 3, $(\bar\omega, \bar\beta)$ is weakly efficient for Problem (7) if and only if there exists $\theta \in (0, +\infty)$ such that $(\bar\omega, \bar\beta)$ is an optimal solution of Problem (13). This is equivalent to $(\bar\omega, \bar\beta)$ being optimal for

$$\begin{array}{ll} \min & \dfrac{\|\omega\|}{\min\left\{\min_{u\in I_1} y^u(\omega^\top x^u + \beta), \theta \min_{u\in I_{-1}} y^u(\omega^\top x^u + \beta)\right\}} \\ \text{s.t.} & y^u(\omega^\top x^u + \beta) > 0, \quad \forall u \in I, \\ & \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \omega \neq 0. \end{array} \tag{20}$$

Observe that (ω, β) is optimal for (20) if and only if $(\mu\omega, \mu\beta)$ is optimal for (20) for any $\mu > 0$. Hence, by normalizing the denominator in the objective of (20) we have that $(\bar{\omega}, \bar{\beta})$ is optimal for (20) if and only if there exist $\theta \in (0, +\infty)$ and $\mu > 0$ such that $(\mu\bar{\omega}, \mu\bar{\beta})$ is optimal for the following problem:

$$\begin{aligned} &\min \ \|\omega\| \\ &\text{s.t.} \quad \min\left\{\min_{u\in I_1}(\omega^\top x^u + \beta), \theta \min_{u\in I_{-1}}(-\omega^\top x^u - \beta)\right\} = A_\theta, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \omega \neq 0, \end{aligned} \tag{21}$$

where $A_\theta = \frac{2\theta}{\theta+1}$. Such a problem is equivalent to the following one

$$\begin{aligned} &\min \ \|\omega\|^2 \\ &\text{s.t.} \quad \min\left\{\min_{u\in I_1}(\omega^\top x^u + \beta), \min_{u\in I_{-1}} \theta(-\omega^\top x^u - \beta)\right\} \geqslant A_\theta, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \end{aligned} \tag{22}$$

which can be rephrased as

$$\begin{aligned} &\min \ \|\omega\|^2 \\ &\text{s.t.} \quad y_1^u(\omega^\top x^u + \beta) \geqslant A_\theta, \quad \forall u \in I_1, \\ &\qquad \theta y_1^u(\omega^\top x^u + \beta) \geqslant A_\theta, \quad \forall u \in I_{-1}, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \omega \neq 0. \end{aligned} \tag{23}$$

Since

$$\begin{aligned} &y_1^u = y_\theta^u, \quad \forall u \in I_1 \\ &\theta y_1^u = y_\theta^u, \quad \forall u \in I_{-1}. \end{aligned}$$

Problem (23) is actually Problem (P_θ). Hence, $(\bar{\omega}, \bar{\beta})$ is weakly efficient iff there exists $\mu > 0$ such that $(\mu\bar{\omega}, \mu\bar{\beta})$ solves (P_θ) for some $\theta \in (0, +\infty)$. By Lemma 4, this is equivalent to $(\bar{\omega}, \bar{\beta})$ having the form $(\mu\omega_1, \mu\beta)$ with $|\beta - \beta_1| < 1$. □

Corollary 6. *The set of Pareto-optimal solutions of the biobjective Problem (7) is given by W,*

$$W = \{(\mu\omega_1, \mu\beta) : |\beta - \beta_1| < 1, \mu > 0\}.$$

Proof. Any Pareto-optimal solution is, by definition, weakly efficient. Let us show the converse. Let $(\bar{\omega}, \bar{\beta})$ be weakly efficient. By Lemma 4 and Theorem 5, there exist $\theta \in (0, +\infty)$ and $\mu > 0$ such that $(\mu\bar{\omega}, \mu\bar{\beta})$ solves (P_θ).

Suppose $(\bar{\omega}, \bar{\beta})$ is not Pareto-optimal. Then $(\mu\bar{\omega}, \mu\bar{\beta})$ would not be Pareto-optimal either. Hence there would exist (ω', β') such that

$$\begin{aligned} &\rho^1(\omega', \beta') \geqslant \rho^1(\mu\bar{\omega}, \mu\bar{\beta}) = \rho^1(\bar{\omega}, \bar{\beta}), \\ &\rho^{-1}(\omega', \beta') \geqslant \rho^{-1}(\mu\bar{\omega}, \mu\bar{\beta}) = \rho^{-1}(\bar{\omega}, \bar{\beta}), \end{aligned} \tag{24}$$

with at least one of those inequality strict.

Without loss of generality we can suppose that $\|\omega'\| = \|\mu\bar{\omega}\|$. Then (24) is equivalent to

$$\begin{aligned} &\min_{u\in I_1} y^u(\omega'^\top x^u + \beta') \geqslant \min_{u\in I_1} y^u(\mu\bar{\omega}^\top x^u + \mu\bar{\beta}) \geqslant A_\theta, \\ &\min_{u\in I_{-1}} y^u(\omega'^\top x^u + \beta') \geqslant \min_{u\in I_{-1}} y^u(\mu\bar{\omega}^\top x^u + \mu\bar{\beta}) \geqslant A_\theta. \end{aligned} \tag{25}$$

Hence, (ω', β') would be feasible for Problem (P_θ). Since its objective value at (ω', β') is $\|\omega'\|^2 = \|\mu\bar{\omega}\|^2$, we would have that (ω', β') is also optimal for Problem (P_θ). By Lemma 4, (P_θ) has a unique optimal solution. Thus $\omega' = \mu\bar{\omega}$, contradicting that at least one of the inequalities in (24) is strict. □

4. The nonseparable case

When the set I is not linearly separable, no hyperplane exists classifying correctly all data points, and thus problem (6) is infeasible. One can try to find a hyperplane minimizing the number of misclassified points. However, this problem is known to be NP-hard, and very difficult to solve in practice [10].

For these cases, the margin maximization approach can be extended to the so-called soft margin approach, e.g., [3–5], which consists of allowing some objects in I to be misclassified, by perturbing (6) in order to make it feasible. In particular, one can replace (6) by its soft counterpart

$$\begin{aligned} \min\ & \|\omega\|^2 + C\sum_{u\in I}(\xi^u)^2, \\ \text{s.t.}\ & y^u(\omega^\top x^u + \beta) + \xi^u \geqslant 1, \quad \forall u \in I, \\ & \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \xi \in \mathbb{R}^{|I|}, \end{aligned} \tag{26}$$

where C is a constant which is usually chosen by crossvalidation techniques, see, e.g., [11–13] and is used in order to tradeoff the perturbations ξ^u and the classification scores $\omega^\top x^u + \beta$.

More generally, we can follow [14], where a more general approach is proposed, in which the perturbations are weighed by different parameters C_1 and C_{-1}, yielding the problem

$$\begin{aligned} \min\ & \|\omega\|^2 + C_1\sum_{u\in I_1}(\xi^u)^2 + C_{-1}\sum_{u\in I_{-1}}(\xi^u)^2, \\ \text{s.t.}\ & y^u(\omega^\top x^u + \beta) + \xi^u \geqslant 1, \quad \forall u \in I, \\ & \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \xi \in \mathbb{R}^{|I|}. \end{aligned} \tag{27}$$

The parameters C_1 and C_{-1} allow the incorporation of different a priori probabilities or misclassification costs in an approximate way [5]. The class c having smaller a priori probability (or classification cost) should have the large C_c value. For instance in [5] $C_c = \frac{1}{n_c}$, where n_c denotes the number of objects in I_c, for $c \in \{1, -1\}$ is suggested. With this, a priori probabilities, as well as different misclassification costs for each class, can be taken into account to weight the perturbations, but not the margin itself, which is the main aim of this paper.

As a generalization of Definition 2, now, the margin of an object $u \in I$, is defined as

$$\rho^u(\omega, \beta, \xi) = \frac{y^u(\omega^\top x^u + \beta) + \xi^u}{\|(\omega, \xi)\|^*}, \tag{28}$$

where $\|\cdot\|^*$ stands for the weighted Euclidean norm given by

$$\|(\omega, \xi)\|^* = \sqrt{\|\omega\|^2 + C_1\sum_{u\in I_1}(\xi^u)^2 + C_{-1}\sum_{u\in I_{-1}}(\xi^u)^2}.$$

The margin of a class $c \in \{+1, -1\}$ in a training sample I, and the margin of a training sample are defined as in Definition 2 for the linearly separable case.

We will study now the Pareto-optimal solutions for the problem of simultaneous maximization of the margin in both classes, under the constraint that all the objects in the training sample are correctly classified in the feature space

$$\begin{aligned} &\max \ \{\rho^1(\omega,\beta,\xi), \rho^{-1}(\omega,\beta,\xi)\} \\ &\text{s.t.} \quad y^u(\omega^\top x^u + \beta) + \xi^u > 0, \quad \forall u \in I, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \xi \in \mathbb{R}^{|I|}. \end{aligned} \tag{29}$$

This is a nonlinear nonconvex biobjective problem whose feasible region is not closed.

Analogously to the hard-margin approach one obtains the following result:

Theorem 7. *The set of Pareto solutions of the biobjective Problem (29) is given by* W^*

$$W^* = \{(\mu\omega_1, \mu\beta, \mu\xi_1) : |\beta - \beta_1| < 1, \mu > 0\}$$

for $(\omega_1, \beta_1, \xi_1)$ *optimal solution of Problem (27).*

Proof. Since all feasible solutions (ω, β, ξ) satisfy that $\rho^1(\omega,\beta,\xi) > 0$ and $\rho^{-1}(\omega,\beta,\xi) > 0$ one can generate, as in Lemma 3, all weakly efficient solutions by solving max–min type scalarizations [9] of the form

$$\begin{aligned} &\max\min \ \{\rho^1(\omega,\beta,\xi), \theta\rho^{-1}(\omega,\beta,\xi)\} \\ &\text{s.t.} \qquad y^u(\omega^\top x^u + \beta) + \xi^u > 0, \quad \forall u \in I, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \xi \in \mathbb{R}^{|I|}, \end{aligned} \tag{30}$$

for θ varying in $(0, +\infty)$. For θ given, this problem is homogeneous. Hence, as we did in the linearly separable case, (ω, β, ξ) is an optimal solution of (30) iff there exists $\mu > 0$ such that $(\mu\omega, \mu\beta, \mu\xi)$ is optimal for the following problem:

$$\begin{aligned} &\min \ \|\omega\|^2 + C_1 \sum_{u\in I_1} (\xi^u)^2 + C_{-1} \sum_{u\in I_{-1}} (\xi^u)^2 \\ &\text{s.t.} \quad (\omega^\top x^u + \beta) + \xi^u \geqslant A_\theta, \quad \forall u \in I_1, \\ &\qquad (-\theta)(\omega^\top x^u + \beta) + \theta\xi^u \geqslant A_\theta, \quad \forall u \in I_{-1}, \\ &\qquad \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}, \quad \xi \in \mathbb{R}^{|I|}, \end{aligned} \tag{31}$$

where $A_\theta = \frac{2\theta}{\theta+1}$.

Problem (31) is convex quadratic with linear constraints and KKT conditions are necessary and sufficient. Such conditions at the unique optimal solution $(\omega_\theta, \beta_\theta, \xi_\theta)$ can be expressed as

$$\begin{aligned} &\lambda^u_\theta \geqslant 0, \quad \forall u \in I, \\ &2\omega_\theta - \sum_{u\in I_1} \lambda^u_\theta x^u + \sum_{u\in I_{-1}} \theta\lambda^u_\theta x^u = 0, \\ &2C_1\xi^u_\theta - \lambda^u_\theta = 0, \quad \forall u \in I_1, \\ &2C_{-1}\xi^u_\theta - \theta\lambda^u_\theta = 0, \quad \forall u \in I_{-1}, \\ &\sum_{u\in I} \lambda^u_\theta y^u_\theta = 0, \\ &\lambda^u_\theta\left[(\omega^\top_\theta x^u + \beta_\theta) + \xi^u_\theta - A_\theta\right] = 0, \quad \forall u \in I_1, \\ &\lambda^u_\theta\left[(-\theta)(\omega^\top_\theta x^u + \beta_\theta) + \theta\xi^u_\theta - A_\theta\right] = 0, \quad \forall u \in I_{-1}. \end{aligned} \tag{32}$$

Table 1
Parameters of the databases

Database	Filename	$\|\Omega\|$	p
bupa	bupa.data	345	5
ionosphere	ionosphere.data	351	34
pima	pima-indians-diabetes.data	768	8
sonar	sonar.all-data	208	60
wdbc	wdbc.data	569	30

Let $(\omega_1, \beta_1, \xi_1)$ be the optimal solution of Problem (27), which coincides with Problem (31) for $\theta = 1$. Then, there exist $(\omega_1, \beta_1, \xi_1)$ and λ_1 satisfying (32) for $\theta = 1$. For each $\theta \in (0, +\infty)$, let $(\omega_\theta, \beta_\theta, \lambda_\theta)$ be given by

$$\begin{aligned}
&\omega_\theta = \omega_1,\\
&\xi_\theta^u = \xi^u, \quad \forall u \in I,\\
&\beta_\theta = \beta_1 + \frac{\theta - 1}{\theta + 1},\\
&\lambda_\theta^u = \lambda_1^u, \quad \forall u \in I_1,\\
&\lambda_\theta^u = \frac{1}{\theta}\lambda_1^u, \quad \forall u \in I_{-1}.
\end{aligned} \tag{33}$$

It is easy to see, that such choice of $(\omega_\theta, \beta_\theta, \xi_\theta)$ and λ_θ satisfy (32). Hence, we conclude that $(\omega_1, \beta_\theta, \xi_1)$ with $\beta_\theta = \beta_1 + \frac{\theta-1}{\theta+1}$, is the unique optimal solution of Problem (31). By homogeneity, all optimal solutions of Problem (31) for all $\theta \in (0, +\infty)$, are given by W^* and then the set of weakly efficient solutions of Problem (29) coincide with W^*. The proof of the equality between W^* and the set of Pareto-optimal solutions of Problem (29) is identical to the proof of Corollary 6, and will not be repeated here. □

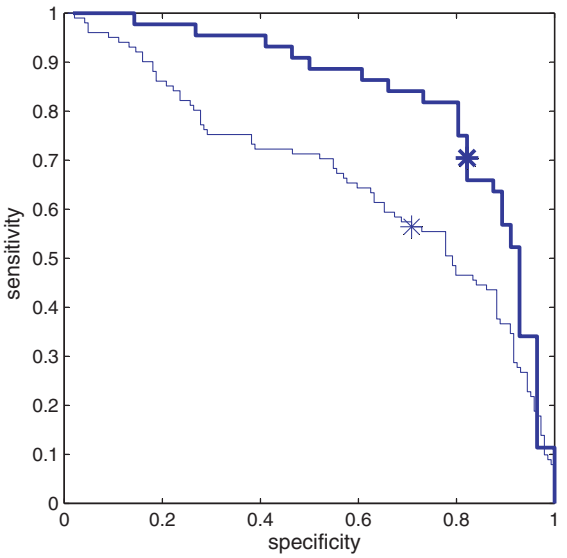

Fig. 1. ROC curve. Database: bupa, $C = 0.03125$.

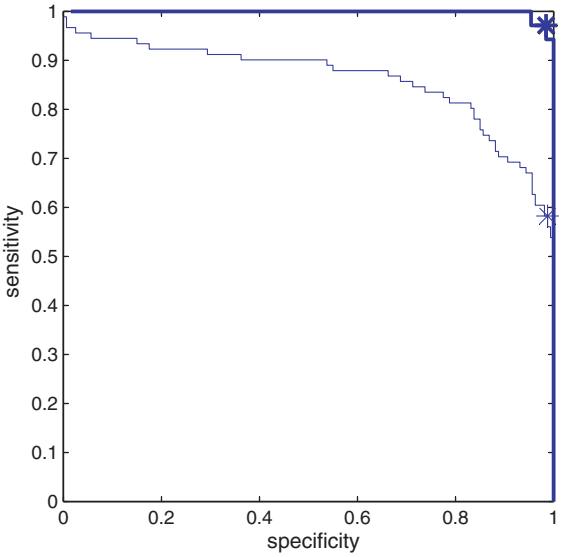

Fig. 2. ROC curve. Database: `ionosphere`, $C = 2.0$.

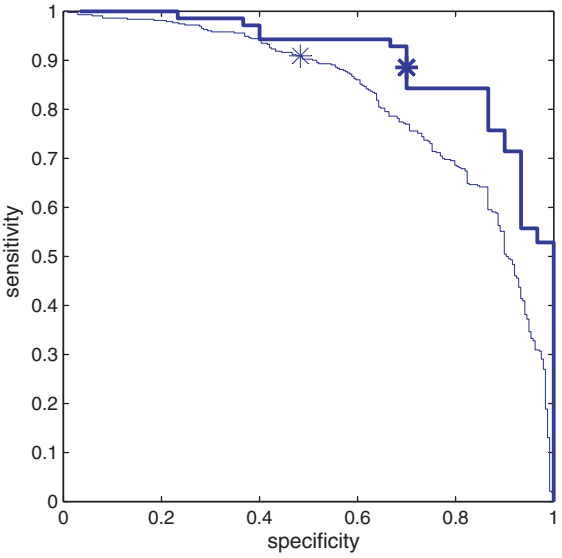

Fig. 3. ROC curve. Database: `pima`, $C = 0.03125$.

5. Illustrative examples

As has been proven in the preceding sections, the best classification rules, in the Pareto sense, are obtained by varying the value β in the optimal solution of the classical SVM. The choice of such a value β is up to the decision maker, who should take into account the tradeoff between the misclassification costs and a priori probabilities in both classes.

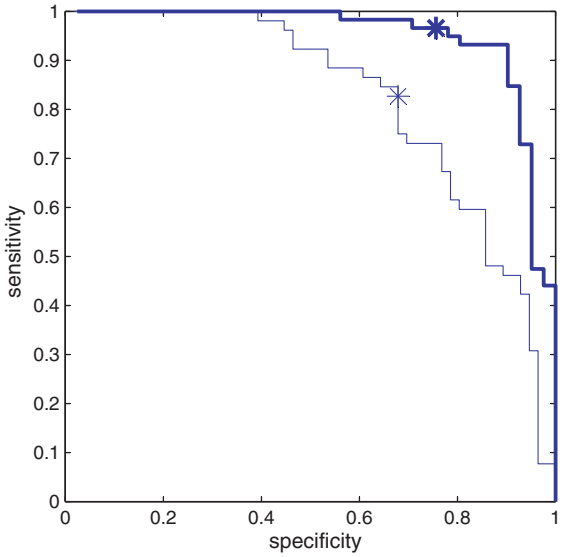

Fig. 4. ROC curve. Database: `sonar`, $C = 0.5$.

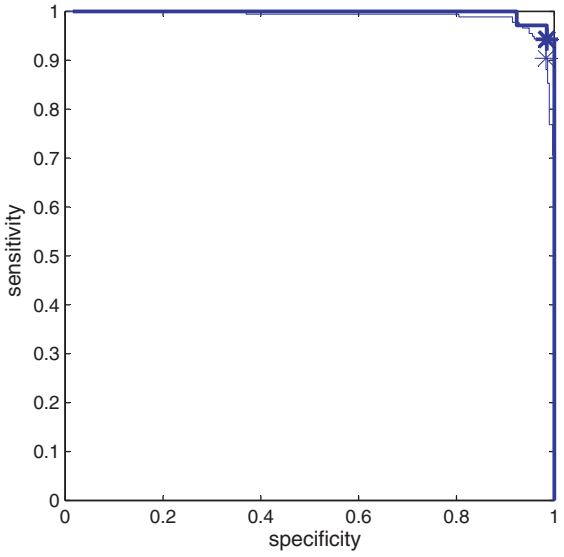

Fig. 5. ROC curve. Database: `wdbc`, $C = 2.0$.

In order to choose a value for the parameter β, some authors (see, e.g., [15]) have suggested the use of the ROC curve. The ROC curve shows the *sensitivity*, i.e., the proportion of correctly classified objects of the positive class, against the *specificity*, proportion of correctly classified objects of the negative class, for different values of the parameter β. The ROC curves can help the decision maker in the choice of β, due to the fact that the only free parameter is the scalar β, as shown by the characterization given in Corollary 6 and Theorem 7.

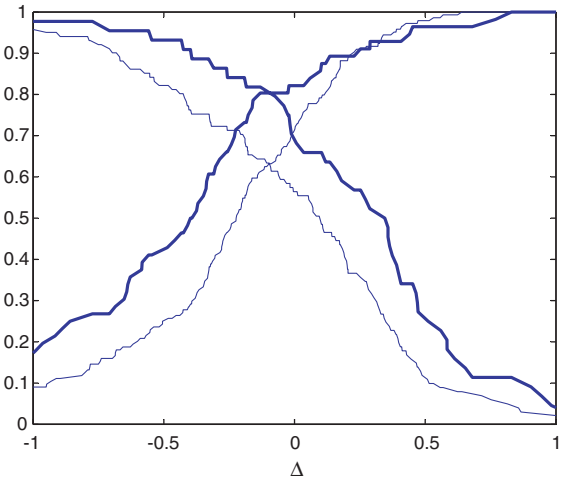

Fig. 6. Specificity and sensitivity of $(\omega_1, \beta_1 - \Delta)$, for a threshold Δ and (ω_1, β_1) optimal solution of (27). Database: `bupa`, $C = 0.03125$.

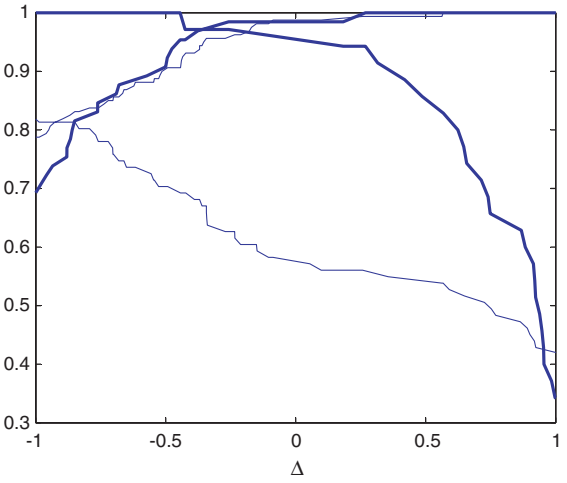

Fig. 7. Specificity and sensitivity of $(\omega_1, \beta_1 - \Delta)$, for a threshold Δ and (ω_1, β_1) optimal solution of (27). Database: `ionosphere`, $C = 2.0$.

In order to show how to guide the choice of β in real-life settings, we have performed various experiments using databases publicly available at the UCI Machine Learning Repository [16]. We have used those databases having two classes and no missing data whose predictor variables are all continuous, as detailed in the summary table [16], namely, the BUPA Liver-disorders Database, called here `bupa`; the Ionosphere Database, called here `ionosphere`; the Pima Indians Diabetes Database, called here `pima`; the Sonar Database, called here `sonar`; and the New Diagnostic Database contained in the Wisconsin Breast Cancer Databases, called here `wdbc`.

For each database, the filename (as called in the database repository [16]), the total number of objects $|\Omega|$ and the number of variables (all quantitative) p is given in Table 1.

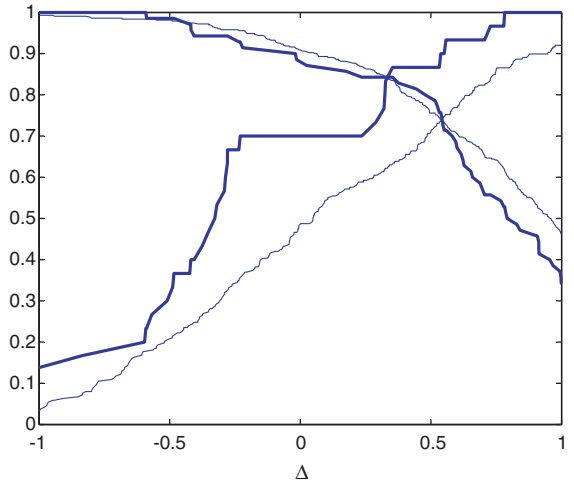

Fig. 8. Specificity and sensitivity of $(\omega_1, \beta_1 - \Delta)$, for a threshold Δ and (ω_1, β_1) optimal solution of (27). Database: `pima`, $C = 0.03125$.

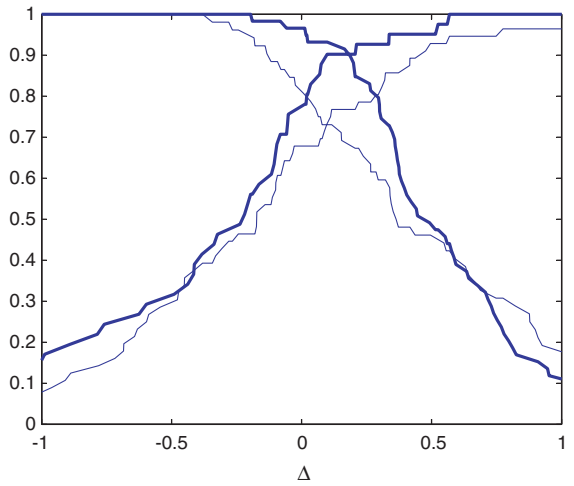

Fig. 9. Specificity and sensitivity of $(\omega_1, \beta_1 - \Delta)$, for a threshold Δ and (ω_1, β_1) optimal solution of (27). Database: `sonar`, $C = 0.5$.

From each database, a random sample of 100 objects is drawn and used as training sample I and the remaining is used as testing sample in order to validate the model.

All the numerical results have been performed by using the SVM toolbox for Matlab [17]. Data were not preprocessed and a linear kernel was used in all the experiments. The parameters C_1 and C_{-1} where set to be equal, and their value chosen by crossvalidation, as implemented in the popular SVM library LIBSVM [18].

With this information at hand, we can draw the ROC curve *in the training sample*, i.e., the plot, when β varies, of the proportions of misclassified objects in both classes in the available set of data. This is not the ROC curve for the whole population, which is unknown in real applications. We then use the former as a surrogate of the latter. In Figs. 1–5 we give the ROC curves for the training sample (thick lines) and testing sample (thin lines). The SVM solution is marked with a star. However, it is not evident to see from ROC

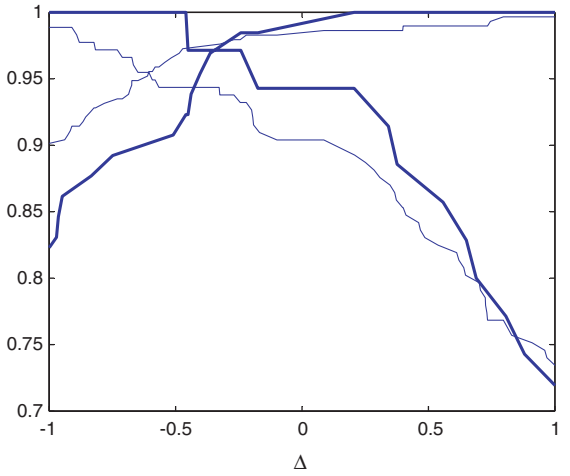

Fig. 10. Specificity and sensitivity of $(\omega_1, \beta_1 - \Delta)$, for a threshold Δ and (ω_1, β_1) optimal solution of (27). Database: `wdbc`, $C = 2.0$.

curves the effect of the β in the tradeoff between sensitivity and specificity, since the value β yielding each pair is not plotted.

In Figs. 6–10, specificity and sensitivity are shown for both training and testing sample (training in thick) in such a way that the decision maker can choose a value for the parameter β. Sensitivity and specificity values for the classical SVM correspond to the case $\Delta = 0$. The higher Δ, the higher $\rho^{-1}(\omega, \beta, \xi)$, at the expense of decreasing $\rho^1(\omega, \beta, \xi)$. This, as empirically illustrated in the graphics, translates into saying that the higher the value chosen for Δ, the higher the specificity and the lower the sensitivity.

6. Conclusion

In this paper, the concept of margin in a training sample I has been generalized to the margin in a class, in order to deal separately with them via a biobjective program. Then, it has been shown that the set of hyperplanes which are Pareto-optimal in the simultaneous optimization of the margin in both classes, is given by a set of parallel hyperplanes, one of which is just the optimal margin hyperplane as defined by the usual SVM approaches [3].

This paper proposes a simple way for taking into account different misclassification costs, or known a priori probabilities of the classes. Our main result gives a theoretical foundation for the commonly used ROC approach for tuning the parameter β.

References

[1] C. Bennet, C. Campbell, Support vector machines: Hype or hallelujah? SIGKDD Explorations 2 (2) (2000) 1–13.
[2] C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 18 (8) (1999) 675–685.
[3] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
[4] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000.
[5] R. Herbrich, Learning Kernel Classifiers: Theory and Algorithms, MIT Press, Cambridge, MA, 2002.
[6] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[7] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[8] V. Vapnik, An overview of statistical learning theory, IEEE Transactions on Neural Networks 10 (1999) 988–999.
[9] M. Ehrgott, M. Wiecek, in: Multiobjective Programming, Springer Science, New York, 2000, pp. 667–722 (Chapter 17).
[10] L. Devroye, L. Györfi, G. Lugosi, Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[11] B. Efron, R. Tibshirani, Improvements on cross-validation: The .632+ bootstrap method, Journal of the American Statistical Association 92 (438) (1997) 548–560.
[12] G. Wahba, Y. Lin, H. Zhang, Generalized approximate cross validation for support vector machines, in: Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, 2000, pp. 297–311.
[13] S. Weiss, C. Kulikowski, Computer Systems That Learn, Morgan Kaufmann, Los Altos, CA, 1999.
[14] K. Veropoulos, N. Cristianini, C. Campbell, Controlling the sensitivity of support vector machines, in: Proceeding of SVM workshop at IJCAI99, 1999.
[15] M. Kupinski, M. Anastasio, Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characteristic curves, IEEE Transactions On Medical Imaging 18 (8) (1999) 675–685.
[16] C. Blake, C. Merz, UCI Repository of Machine Learning Databases, Downloadable from website http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Irvine, Dept. of Information and Computer Sciences, 1998.
[17] A. Schwaighofer, SVM toolbox for Matlab, Downloadable from website http://www.cis.tugraz.at/igi/aschwaig/software.html, 2002.
[18] C. Chang, C. Lin, LIBSVM: A library for support vector machines, Downloadable from website http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.