*International Journal of Applied Mathematical Research*, *2 (2) (2013) 175-187*
*©Science Publishing Corporation*
*www.sciencepubco.com/index.php/IJAMR*

# A minmax Chebyshev approach to optimal binary classification

**Roberto Ragona**

*ENEA, Dept. of Advanced Technologies for Energy and Industry, Via Anguillarese, 301 - 00123 Rome (Italy)*
*Corresponding Author E-mail: roberto.ragona@enea.it*

## Abstract

Linear programming (LP) techniques for optimal binary classification have inspired research studies in recent years; they pose an alternative to the quadratic programming (QP) approach, which is usually credited with having greater complexity. In this paper, we describe an LP approach that is based on the minmax Chebyshev criterion, for which we demonstrate that it can determine an optimal solution with competitive properties. The approach is then extended so that two of the most attractive properties of the traditional QP approach (the direct formulation of the optimal classifier in higher dimensions and the sparseness of its coefficients) are preserved. The proposed method demonstrates its capabilities to successfully address situations that have separable and inseparable classes.

***Keywords***: *Minmax Chebyshev criterion, support vector machines, pattern classification, binary classifier, linear programming*.

## 1 Introduction

Classic Support Vector Machine (SVM) techniques have led to successful solutions to classification problems following a non-parametric approach oriented by the criterion of margin maximisation; in fact the SVM techniques do not assume knowledge of the forms of the underlying probability distributions. SVMs are now well developed and have been presented in a series of foundational papers and books, i.e., [1], [2], [3], [4].
The related applications are formulated as constrained quadratic optimisations, which, under conditions of semi-definite positiveness of the kernel matrix [3, page 38], lead to concave quadratic programming, and can assure a global maximum (minimum) of the objective function, although there might be cases in which the solution is not unique.
In this paper, we present a different approach, which is based on a minmax criterion for which the results are capable of assuring an optimal solution based on linear programming techniques.
Other authors report optimal methods of supervised classification when using linear programming techniques, i.e., [5], [6], [7, page 230], but from another point of view and with different results. Often, an optimisation for the $L_1$ or $L_\infty$ weight vector norm is pursued.
In section 2, we will discuss the basis of our minmax procedure with regard to linear and non-linear classifiers. In section 3, we will show how to gain sparseness and formulate the classifier directly in higher dimensions. In subsequent sections, relevant properties and computational details will be presented. Finally, in section 7, computational comparisons will be described.

## 2 Geometric motivations for linearly separable classes

We begin with considerations that were suggested by the solid geometry in $R^3$. If we regard a typical two separable class situation (represented in terms of the classes **A** and **B**, which are constituted respectively of $n_{\mathbf{A}}$ and $n_{\mathbf{B}}$ points P(x,y) on the x-y plane embedded in $R^3$, see fig.1), then the margin maximisation problem for a linear classifier

$$D(x,y) = w_1 * x + w_2 * y + b$$

is expressed analytically as the definition of the optimal coefficients $w_1^*$, $w_2^*$, $b^*$ such that the objective function

$$\|\mathbf{w}\|^2 = (w_1^2 + w_2^2)$$

becomes minimum, subject to the constraints

$$c_i \ast D(x_i,y_i) = c_i\,(w_1 \ast x_i + w_2 \ast y_i + b) \geq 1, \qquad (1)$$

where

$c_i = +1$, if $P(x_i,y_i) \in \mathbf{A}$,
$c_i = -1$, if $P(x_i,y_i) \in \mathbf{B}$,
$i = 1, 2,\ldots, p$; $p$ = sample size = $(n_\mathbf{A} + n_\mathbf{B})$.

Recall that, to produce discriminating results, the optimal classifier $D^*(x,y)$ is to be such that:

$\text{sign}(D^*(P(x_i,y_i) \in \mathbf{A})) > 0$,
$\text{sign}(D^*(P(x_i,y_i) \in \mathbf{B})) < 0$.

In this paper, the arrangement of the sample set $\{\mathbf{A} \cup \mathbf{B}\}$ follows this order: first is class **A**, then class **B**. The subscript i, which identifies points $P(x_i, y_i) \in \{\mathbf{A} \cup \mathbf{B}\}$, spans the integer interval $[1, (n_\mathbf{A} + n_\mathbf{B}) = p]$.
The related optimal classifier will be denoted equivalently by QP or QP_SVM, because its definition depends on the solution of a quadratic programming problem [1].
Moving from similar considerations, we define a linear *minmax* classifier for two separable classes **A** and **B** in $R^2$ as the plane $D(x,y) = w_1 \ast x + w_2 \ast y + b$ defined in $R^3$ in such a way that

$$m = \max |D(x_i,y_i)| = \max |w_1 \ast x_i + w_2 \ast y_i + b| = \max\{c_i\,(w_1 \ast x_i + w_2 \ast y_i + b)\},\ (1 \leq i \leq p)$$

be minimised (Chebyshev criterion), subject to the constraints in (1). Therefore, the minimisation of the maximum absolute value of $D(x,y) = w_1 \ast x + w_2 \ast y + b$ on the sample set $\{\mathbf{A} \cup \mathbf{B}\}$ is pursued and is constrained to render $|D(x_i,y_i)| \geq 1$, $i = 1, 2, \ldots, p$. Analytically, the problem can be expressed as the following LP optimisation:

$$\left.\begin{array}{l} \min m,\ \text{subject to (s.t.)} \\ m - c_i \ast D(x_i,y_i) = m - c_i\,(w_1 \ast x_i + w_2 \ast y_i + b) \geq 0, \\ \quad c_i \ast D(x_i,y_i) = c_i\,(w_1 \ast x_i + w_2 \ast y_i + b) \geq 1, \\ i = 1, 2,\ldots, p \end{array}\right\} \qquad (2)$$

Thus, we must solve an LP problem with 2p constraints.
The optimal solution to (2), in terms of the weights $w_i{}^*$ and of the bias term $b^*$, will be denoted as LP_MM. The results are usually different from the optimal QP_SVM solution defined on the same classes.

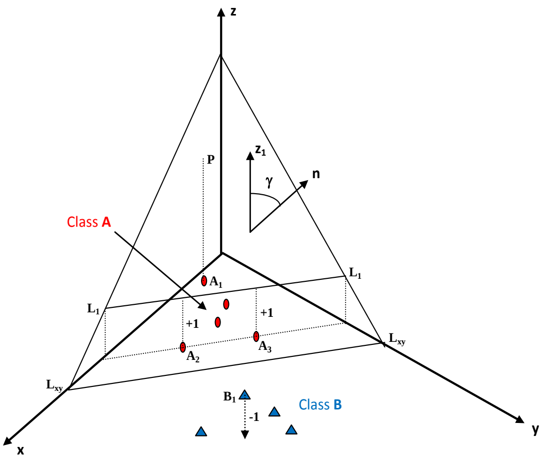

Fig. 1: Linear classification in $R^2$

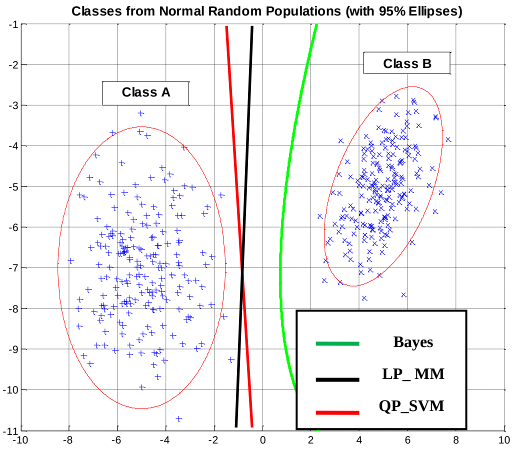

Fig. 2: Results on a Gaussian two-class sample

Fig. 2 presents a comparison between the trace $D^*(x,y) = 0$ of a linear LP_MM classifier (the black line) for a random Gaussian sample of two separable classes **A** and **B** generated in $R^2$ (mean values and covariance matrices assigned arbitrarily) with the corresponding trace of a linear QP_SVM classifier (the red line) of the same sample. The minimum-error-rate Bayes classifier trace [8, chapter 2, page 9] (the green line, with prior probabilities fixed to 0.5) and the 95% probability elliptic contours that enclose **A** and **B** are represented. The Bayes classifier is constituted in the case of Gaussian samples of hyperquadrics [8, chapter 2, page 25], and is usually credited with superior capabilities of classification, because the underlying class probability distributions are involved in its definition.

The generalisation to a linear minmax classifier on the points of the space $R^n$, $n > 2$, is straightforward: $D(\mathbf{x})$ is expressed as a sum of the type $\sum_1^n w_i * x_i + b = \mathbf{w}^T \bullet \mathbf{x}$, where $\mathbf{x} = [x_1, x_2, ..., x_n]^T \in R^n$, and problem (2) can be consequently reformulated. The operator $(\bullet)^T$ denotes a matrix/vector transposition.

To generalise to *non*-linear minmax classifiers on points of the space $R^n$, we resort, as in the QP_SVM case [3, page 25], to a decision function $D(\mathbf{x})$, which is defined through N non-linear functions $\varphi_i(\mathbf{x})$

$$D(\mathbf{x}) = \sum_{i=1}^{N} w_i \varphi_i(\mathbf{x}) + b = [w_1\ w_2 \ldots w_N] \bullet \boldsymbol{\varphi}(\mathbf{x}) + b = \mathbf{w}^T \bullet \boldsymbol{\varphi}(\mathbf{x}) + b \tag{3}$$
$$\mathbf{x} \in R^n,\ \varphi_i(\mathbf{x}): R^n \to R,\ \boldsymbol{\varphi}(\mathbf{x}) = [\varphi_1(\mathbf{x})\ \varphi_2(\mathbf{x}) \ldots \varphi_N(\mathbf{x})]^T:\ R^n \to R^N.$$

$D(\mathbf{x})$ implies in this case linearity in the adjustable parameters $w_i$ but not in $\mathbf{x}$. By a simple substitution of the relation (3) in problem (2), we obtain the following definition of the non-linear minmax problem in the case of two separable classes.

**Definition 2.1:** *The non-linear minmax problem for the binary classification in* $R^n$ *is formulated as:*

$$\left.\begin{array}{l} \min m,\ \text{s.t.} \\ m - c_i D(\mathbf{x}_i) = m - c_i\,[\,w_1\,\varphi_1(\mathbf{x}_i) + w_2\,\varphi_2(\mathbf{x}_i) + \ldots + w_N\,\varphi_N(\mathbf{x}_i) + b] \geq 0 \\ \quad c_i D(\mathbf{x}_i) = c_i\,[\,w_1\,\varphi_1(\mathbf{x}_i) + w_2\,\varphi_2(\mathbf{x}_i) + \ldots + w_N\,\varphi_N(\mathbf{x}_i) + b] \geq 1, \\ i = 1, 2, \ldots, p,\ \mathbf{x}_i \in \{\mathbf{A} \cup \mathbf{B}\}. \end{array}\right\} \tag{4}$$

Formulation (4) does not show a situation of sparseness in the set $\{w_i\}$; in general, all of the N terms can be present in $D(\mathbf{x})$. To gain a result of sparseness, together with the possibility of operating directly in higher dimensions, we must develop further considerations.

## 3 A close classifier and its properties

Let $M_{LP\text{-}(4)}$ be the sum of the squared optimal coefficients of the solution $D^*(\mathbf{x})$ to the problem (4) relative to a two separable class sample, assuming that a unique solution exists:

$$M_{LP\text{-}(4)} = \sum_1^N (w_j^*)^2$$

Let us consider now a *modified minmax problem*, to produce a different optimal $D_m^*(\mathbf{x})$; this new problem is composed of the original problem (4) augmented by an additional non-linear equality constraint:

$$\left.\begin{array}{l} \min m,\ \text{s.t.} \\ m - c_i D_m(\mathbf{x}_i) = m - c_i\,[\,w_1\,\varphi_1(\mathbf{x}_i) + w_2\,\varphi_2(\mathbf{x}_i) + \ldots + w_N\,\varphi_N(\mathbf{x}_i) + b] \geq 0 \\ \quad c_i D_m(\mathbf{x}_i) = c_i\,[\,w_1\,\varphi_1(\mathbf{x}_i) + w_2\,\varphi_2(\mathbf{x}_i) + \ldots + w_N\,\varphi_N(\mathbf{x}_i) + b] \geq 1 \\ \sum_1^N w_j^2 = M_{LP\text{-}(4)} + \varepsilon \quad (\varepsilon \text{ arbitrary and positive}) \\ i = 1, 2, \ldots., p,\ \mathbf{x}_i \in \{\mathbf{A} \cup \mathbf{B}\}. \end{array}\right\} \tag{5}$$

In other words, we impose additionally a constraint that the squared sum of the optimal coefficients is larger than $M_{LP\text{-}(4)}$ by an arbitrary $\varepsilon > 0$.

We will demonstrate that the solution to (5) possesses properties of sparseness and the possibility to operate directly in higher dimensions regardless of the value of $\varepsilon > 0$.

For $\varepsilon > 0$, the problem (5) is non-linear in $w_i$; next, we analyse the properties of its solution.

To proceed, we consider the Lagrange function L [9, page 315]

$$L = m - \sum_1^p \nu_i [\, m - c_i D_m(\mathbf{x}_i)] - \sum_1^p \eta_i [\, c_i D_m(\mathbf{x}_i) - 1] - \beta\, [\sum_1^N w_i^2 - M_{LP\text{-}(4)} - \varepsilon\,],$$

where $\nu_i \geq 0$, $\eta_i \geq 0$ and $\beta$ are Lagrange multipliers.
Necessary conditions of the optimum, in addition to the constraints, will be [9, page 315]:

$$\partial L / \partial w_i = 0,$$

and an analytic evaluation produces the result

$$w_i^* = [c_1(\nu_1 - \eta_1)\, \varphi_i(\mathbf{x}_1) + \ldots + c_p(\nu_p - \eta_p)\, \varphi_i(\mathbf{x}_p)] \,/\, (2\beta)$$

The replacement in (3) with all of the optimal $w_i^*$ yields the optimal *modified* minmax classifier $D_m^*(\mathbf{x})$:

$$\begin{aligned} D_m^*(\mathbf{x}) &= \sum_{i=1}^{N} w_i^* \varphi_i(\mathbf{x}) + b^* = \\ &= 1/(2\beta)\, \{[c_1(\nu_1-\eta_1)\varphi_1(\mathbf{x}_1) + \ldots + c_p(\nu_p-\eta_p)\varphi_1(\mathbf{x}_p)]\ \varphi_1(\mathbf{x}) + \\ &\quad + \ldots + [c_1(\nu_1-\eta_1)\varphi_N(\mathbf{x}_1) + \ldots + c_p(\nu_p-\eta_p)\varphi_N(\mathbf{x}_p)]\ \varphi_N(\mathbf{x})\} + b = \\ &\qquad \text{(rearranging the order of summation)} \\ &= 1/(2\beta)\, \{c_1(\nu_1-\eta_1)\, [\varphi_1(\mathbf{x}_1)\varphi_1(\mathbf{x}) + \ldots + \varphi_N(\mathbf{x}_1)\varphi_N(\mathbf{x})] + \ldots \\ &\quad + c_p(\nu_p-\eta_p)[\varphi_1(\mathbf{x}_p)\varphi_1(\mathbf{x}) + \ldots + \varphi_N(\mathbf{x}_p)\varphi_N(\mathbf{x})]\} + b^* = \\ &= \sum_1^p \theta_i^* K(\mathbf{x}_i,\mathbf{x}) + b^*, \end{aligned} \tag{6}$$

having denoted

$$\begin{aligned} \theta_i^* &= c_i(\nu_i - \eta_i)/(2\beta) \\ K(\mathbf{x}_i,\mathbf{x}) &= \varphi_1(\mathbf{x}_i)\varphi_1(\mathbf{x}) + \varphi_2(\mathbf{x}_i)\varphi_2(\mathbf{x}) + \ldots + \varphi_N(\mathbf{x}_i)\varphi_N(\mathbf{x}) = \\ &= [\varphi_1(\mathbf{x}_i)\ \ \varphi_2(\mathbf{x}_i)\ \ldots\ldots\ \varphi_N(\mathbf{x}_i)] \bullet [\varphi_1(\mathbf{x})\ \ \varphi_2(\mathbf{x})\ \ldots\ \varphi_N(\mathbf{x})\,]^T = \boldsymbol{\varphi}^T(\mathbf{x}_i) \bullet \boldsymbol{\varphi}(\mathbf{x}). \end{aligned}$$

Based on related theory, the Lagrange multipliers $\eta_i$, $\nu_i$ and $\beta$ can assume the following values:

- $\eta_i > 0$ and $\nu_i = 0$, when $D_m^*(\mathbf{x}_i) = \pm 1$, i.e., when the constraint $c_i D_m^*(\mathbf{x}_i) = 1$ is active
- $\eta_i = 0$ and $\nu_i > 0$, when $D_m(\mathbf{x}_i) = \pm m^*$, i.e., when the constraint $c_i D_m^*(\mathbf{x}_i) = m^*$ is active
- $\eta_i = 0$ and $\nu_i = 0$, in all of the other cases;
- $\beta \neq 0$, because the additional constraint in (5) is active regardless of $\varepsilon > 0$.

The function $K(\mathbf{x}_i,\mathbf{x})$ is representable as a dot product between N-dimensional vectors and, similar to the QP_SVM case, allows us to treat higher dimensional problems defined in the space $R^N$. In fact the function $K(\mathbf{x}_i,\mathbf{x})$ can be in favourable situations defined in closed form, also if $N \to \infty$. Additionally it can specifically be chosen to be a positive semidefinite kernel [2] when the Mercer's condition is fulfilled, although in the LP context the condition of positive semidefiniteness is not required (later, in section 7, this situation will be illustrated with an example).
The set of coefficients $\theta_i$ is sparse because usually only a reduced number of terms from (6) survives in the sum. The surviving terms are those pertaining to the special situations $D_m^*(\mathbf{x}_i) = \pm 1$ or $\pm m^*$. Therefore, the solution to (5) possesses properties of sparseness and can be formulated through dot products or kernel functions. The problem resides in the non-linearity of (5), which creates difficulties when a solution is to be found. We now present the key assumption of our approach.

**Assumption 3.1:** *There is evidence that for $\varepsilon \to 0$ the problems (4) and (5) tend to coincide. We assume from now on that the LP solution to problem (4) represents a tight approximation to the solution to the problem (5), when we consider a value that is infinitesimal and positive for $\varepsilon$.*

Therefore, this assumption implies that the following tight approximation holds, because the problems (4) and (5) are coincident except for a constraint which reports an infinitesimal difference $\varepsilon$ :

$$D^*(\mathbf{x})_{\text{problem (4)}} \cong D_m^*(\mathbf{x})_{\text{problem (5)}} = \sum_1^p \theta_i^* \; \mathrm{K}(\mathbf{x}_i,\mathbf{x}) + b^*.$$

More specifically, this approach permits us to solve, up to a tight approximation, the problem in (5) by means of the problem in (4), which can be reformulated in the light of (6) by means of p coefficients $\theta_i$ and terms relative to the dot products $\mathrm{K}(\mathbf{x}_i,\mathbf{x}_k)$:

$$\left.\begin{array}{l} \min m,\ \text{s.t.} \\ m - c_i D(\mathbf{x}_i) = m - c_i\,[\theta_1 \mathrm{K}(\mathbf{x}_1,\mathbf{x}_i) + \theta_2 \mathrm{K}(\mathbf{x}_2,\mathbf{x}_i) + \ldots + \theta_p \mathrm{K}(\mathbf{x}_p,\mathbf{x}_i) + b] \geq 0 \\ c_i D(\mathbf{x}_i) = c_i\,[\theta_1 \mathrm{K}(\mathbf{x}_1,\mathbf{x}_i) + \theta_2 \mathrm{K}(\mathbf{x}_2,\mathbf{x}_i) + \ldots + \theta_p \mathrm{K}(\mathbf{x}_p,\mathbf{x}_i) + b] \geq 1 \\ i = 1, 2, .., p \end{array}\right\} \quad \begin{array}{r} \\ (7.a) \\ (7.b) \\ \\ \end{array}$$

We would like to point out the fact that assumption 3.1 allows us to define $D(\mathbf{x})$ either in N-dimensional space $\boldsymbol{\varphi}(\bullet)$, or (up to a tight approximation) in the space of the dot products $\mathrm{K}(\bullet,\bullet)$, maintaining the sparseness intrinsically present in the set $\{\theta_i\}$; proposition 4.2 will give a measure of the sparseness.

What is known in the QP_SVM context as a "kernel trick" [3, page 25 & page 317] continues to be valid also in the LP_MM context in a more relaxed formulation, without any prescription on the dot product $\mathrm{K}(\bullet,\bullet)$.

In matrix/vector form, the minmax LP problem with its constraints (7.a) – (7.b) can be expressed as reported in (8), where K(i,j) denotes an abbreviation for $\mathrm{K}(\mathbf{x}_i,\mathbf{x}_j)$.

$$\left.\begin{array}{l} \min\,[1\;0\;0\;\ldots\ldots..0]\;\;[m\;\theta_1\;\theta_2\;\ldots\ldots\ldots\theta_p\;b]^{T} = \;\min \mathbf{s}^{T}\bullet\mathbf{p},\ \text{s.t.} \\ \left|\begin{array}{lllll} 1 & -c_1K(1,1) & -c_1K(2,1) \ldots\ldots\ldots\ldots & -c_1K(p,1) & -c_1 \\ 1 & -c_2K(1,2) & -c_2K(2,2) \ldots\ldots\ldots\ldots & -c_2K(p,2) & -c_2 \\ \ldots\ldots & \ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots & \\ 1 & -c_pK(1,p) & -c_pK(2,p) \ldots\ldots\ldots\ldots & -c_pK(p,p) & -c_p \\ 0 & c_1K(1,1) & c_1K(2,1) \ldots\ldots\ldots\ldots & c_1K(p,1) & c_1 \\ 0 & c_2K(1,2) & c_2K(2,2) \ldots\ldots\ldots\ldots & c_2K(p,2) & c_2 \\ \ldots\ldots & \ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots & \ldots \\ 0 & c_pK(1,p) & c_pK(2,p) \ldots\ldots\ldots\ldots & c_pK(p,p) & c_p \end{array}\right| \bullet \left|\begin{array}{c} m \\ \theta_1 \\ \theta_2 \\ . \\ \\ \\ \theta_p \\ b \end{array}\right| \geq \left|\begin{array}{c} 0 \\ 0 \\ \\ 0 \\ 1 \\ 1 \\ \\ 1 \end{array}\right| \end{array}\right\} \quad (8)$$

The LP problem has 2p constraints and (p+2) variables.

Synthetically, we can write problem (8) in the following form:

$$\begin{array}{l} \min \mathbf{s}^{T}\bullet\mathbf{p},\ \text{s.t.} \\ \mathbf{H}\bullet\mathbf{p} \geq \mathbf{h}; \end{array}$$

where **H** is the primal constraint matrix, and **h** is the right-hand vector of the constraint system.

# 4 The dual solution to separable classes (hard minmax optimisation) and the support vectors

It is well known that duality is a fruitful concept in linear programming.

Recalling the correspondence rules between primal and dual LPs [10, page 131], the primal problem (8) is transformed into its dual (9), which assumes (p+2) *equality* constraints and 2p variables.

$$\left.\begin{array}{l} \qquad\quad |\text{ -- p terms -- }|\text{ -- p terms -- }| \\ \max\,[0\;0\;\ldots\ldots0\;1\;1\;\ldots\ldots..\;1]\;\;[v_1\;v_2\;\ldots\ldots\ldots\ldots\ldots\;v_{2p}]^{T} = \max \mathbf{h}^{T}\bullet\mathbf{v},\ \text{s.t.} \\ \left|\begin{array}{cccc} 1 & 1 & . & 0 \\ -c_1K(1,1) & -c_2K(1,2) & . & c_pK(1,p) \\ -c_1K(2,1) & -c_2K(2,2) & . & c_pK(2,p) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ -c_1K(p,1) & -c_2K(p,2) & . & c_pK(p,p) \\ -c_1 & -c_2 & . & c_p \end{array}\right| \bullet \left|\begin{array}{c} v_1 \\ v_2 \\ . \\ \\ \\ \\ . \\ v_{2p} \end{array}\right| = \left|\begin{array}{c} 1 \\ 0 \\ 0 \\ . \\ \\ \\ \\ 0 \end{array}\right| \end{array}\right\} \quad (9)$$

$v_i \geq 0$, i = 1, 2, ...., 2p

Synthetically, we can represent problem (9) in the following form:

$$\max \mathbf{h}^T \bullet \mathbf{v}, \text{ s.t.}$$
$$\mathbf{H}^T \bullet \mathbf{v} = \mathbf{s}$$
$$\mathbf{v} \geq \mathbf{0}.$$

$\mathbf{H}^T$ is the transpose of **H**, and **s** is the right-hand vector of the constraint system; the matrix $\mathbf{H}^T$ has a rectangular structure, with resulting dimensions of [(p+2) *x* 2p].

Before introducing, in our context, the concept of support vectors, we present an introductory proposition; moreover, from now on, the rank of the matrix **H** (or equivalently of $\mathbf{H}^T$) will be denoted by $r_\mathbf{H}$.

**Proposition 4.1:** *The characteristic set $S_{id}$*

*Let us suppose that the primal/dual pair (8) - (9) possesses optimal solutions* $\mathbf{p}^*$ *and* $\mathbf{v}^*$, *respectively. The set of indexes* $\{i: v_i^* > 0\}$ *of all of the strictly positive components of the dual solution* $\mathbf{v}^*$ *furnishes a set (called the characteristic set)* $S_{id} = \{i_1, i_2, \ldots\}$, *which in the primal (8) identifies the constraints assuming the equality sign (primal binding or active constraints)*

This proposition relies completely on the Complementary Slackness (C.S.) theorem [9, page 96], prescribing for any primal/dual LP pair expressed in *symmetric form* that iff $\mathbf{p}^*$ and $\mathbf{v}^*$ are the respective optimal solutions, then:

1) the dual variable $v_i^* > 0 \Rightarrow$ the i-th constraint in the primal assumes the sign “=”;
2) the dual variable $v_i^* = 0 \Leftarrow$ the i-th constraint in the primal assumes the sign “>”;
3) the primal variable $p_i^* > 0 \Rightarrow$ the i-th constraint in the dual assumes the sign “=”;
4) the primal variable $p_i^* = 0 \Leftarrow$ the i-th constraint in the dual assumes the sign “<”.

The symmetric form [9, page 94] prescribes that all of the primal variables be restricted in sign (greater than or equal to zero), but it is a simple matter to express (8). Thus, for example, $\theta_i$ can be represented as $\theta_i = \theta_i^+ - \theta_1^-$, where both $\theta_i^+$ and $\theta_1^-$ are greater than or equal to zero. The variables are duplicated, but all of the restrictions in sign are respected. Moreover, because the primal constraint system assumes the sign "$\geq$", the dual constraint system assumes the opposite sign "$\leq$". In conclusion, in symmetric form, the dual problem takes a different form, but it clearly remains equivalent to a more synthetic form (9).

The first relation demonstrates fully proposition 4.1.

The importance of the characteristic set $S_{id}$ resides in the fact that it identifies, in our context, the set of *support vectors*, which we are now ready to present formally by means of proposition 4.2.

**Proposition 4.2:** *Support vectors in the minmax context*

*If the primal/dual pair (8)-(9) possesses unique, non-degenerate and finite solutions, then there exist* $r_\mathbf{H}$ *support vectors, which are intended as special points* $\mathbf{x}_i \in \{\mathbf{A} \cup \mathbf{B}\}$, *where either* $D^*(\mathbf{x}_i) = \pm 1$ *(class-boundary points) or* $D^*(\mathbf{x}_i) = \pm m^*$ *(minmax points);* $m^*$ *is the optimal value relative to the primal LP problem. Moreover, the value* $r_\mathbf{H}$ *measures the sparseness of the set* $\{\theta_i\}$ *of the primal coefficients.*

The demonstration of the first part results from the properties of the solution to the dual LP problem (9). In fact, the supposed unique dual optimal solution **v*** is found among the basic feasible solutions (extreme points of the feasible polyhedron), and, if not *degenerate* (that is, if exactly $r_\mathbf{H}$ values $v_i^* > 0$ exist), it furnishes a set of $r_\mathbf{H}$ indexes (the characteristic set $S_{id}$) that identify all of the components of $\mathbf{v}^*$ that are greater than zero:

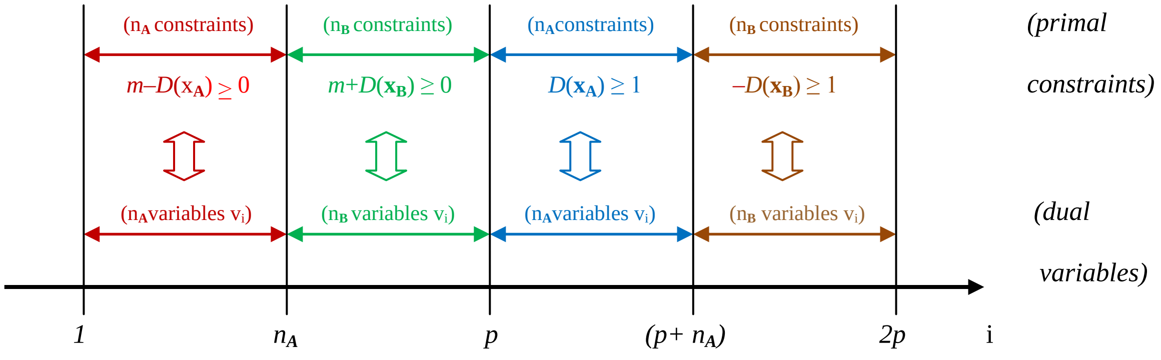

Figure 3: The correspondence (8) - (9)

$S_{id} = \{i_1, i_2, \ldots, i_{rH}\textbf{:}\ v_{ij}^* > 0\},\ 1 \le i_j \le 2p.$

It is worthwhile, at this point, to clarify the correspondence among the primal constraints (7.a) - (7.b) and the dual variables $v_i$. Presented in fig. 3 is the arrangement implied in (8) and (9).
Recall that $(n_\mathbf{A} + n_\mathbf{B}) = p$, that $c_i = +1$ if $\mathbf{x}_i \in \mathbf{A}$, and that $c_i = -1$ if $\mathbf{x}_i \in \mathbf{B}$.
Moreover, recall that a generic constraint (7.a) becomes

$m - D(\mathbf{x}_i) \geq 0$ if $\mathbf{x}_i \in \mathbf{A}$, and
$m + D(\mathbf{x}_i) \geq 0$ if $\mathbf{x}_i \in \mathbf{B}$;

analogous consequences arise for the constraints (7.b).
It follows that an index $i \in S_{id}$ can be associated to a value $v_i^* > 0$ and, when $1 \le i \le n_\mathbf{A}$, implies by virtue of proposition 4.1 a primal active constraint of the type (7.a) operating on a point of the class **A** (see fig. 3):

$$m^* - D^*(\mathbf{x}_i) = 0 \qquad \Rightarrow \qquad D^*(\mathbf{x}_i) = +\,m^*$$

Given an index $i \in S_{id}$; the condition $(n_\mathbf{A}+1) \le i \le (n_\mathbf{A}+n_\mathbf{B}) = p$ implies that there is a primal active constraint of the type (7.a) operating on a point of the class **B** (see fig. 3):

$$m^* + D^*(\mathbf{x}_i) = 0 \qquad \Rightarrow \qquad D^*(\mathbf{x}_i) = -\,m^*$$

Given an index $i \in S_{id}$; the condition $(p+1) \le i \le (p+ n_\mathbf{A})$ implies that there is a primal active constraint of the type (7.b) operating on a point of class **A** (see fig. 3):

$$D^*(\mathbf{x}_{i\text{-}p}) = 1 \qquad \Rightarrow \qquad D^*(\mathbf{x}_{i\text{-}p}) = +\,1$$

The subscript (i-p) originates from the fact that the sample $\{\mathbf{A} \cup \mathbf{B}\}$ is constituted of p points, whereas the primal LP problem accounts for 2p constraints.
Given an index $i \in S_{id}$; the condition $(p+n_\mathbf{A}+1) \le i \le (p+ n_\mathbf{A}+ n_\mathbf{B}) = 2p$ implies that there is a primal active constraint of the type (7.b) operating on a point of class **B** (see fig. 3):

$$-\,D^*(\mathbf{x}_{i\text{-}p}) = 1 \qquad \Rightarrow \qquad D^*(\mathbf{x}_{i\text{-}p}) = -\,1$$

In conclusion, $S_{id}$ identifies $r_\mathbf{H}$ special points $\mathbf{x}_i$ that we call *support vectors*, where the optimal classifier $D^*(\mathbf{x})$ attains its extreme values, and the range (–1, +1) is forbidden by definition.
Proposition 4.2 leaves out of consideration the possible existence of degeneracy and multiplicity in the LP scenario, which can create special situations.
Unlike the QP context and because the minmax procedure acts on the range of $D(\mathbf{x})$, the set of the support vectors includes also the peripheral ones *(minmax support vectors* $\mathbf{x}_i$*)*, which define the two bounds of the optimised range: $D^*(\mathbf{x}_i) = \pm\, m^*$. The classification capabilities are mainly associated with the subset of the *class-boundary support vectors* $\mathbf{x}_k$, where $D^*(\mathbf{x}_k) = \pm\, 1$; in our experience, we typically found an almost equal subdivision between the two types of support vectors.
With regard to the sparseness of $\{\theta_i\}$ and its connection with the rank of $\mathbf{H}^T$, $r_\mathbf{H}$ represents the number of linearly independent dual constraints, and the possibility of redundancy cannot be ruled out. However, eliminating a redundant constraint (row) in the dual means cancelling a corresponding variable $\theta_i$ (column) in the primal; then, $r_\mathbf{H}$ is equal to the number of primal coefficients $\theta_i$ that are minimally necessary in $D^*(\mathbf{x})$.
As a first example to show the feasibility of our method and some of its characteristic results, fig. 4 presents the solution to a problem in $R^5$ that pertains to a training sample of the *thyroid* dataset (the 8-th sample out of 100 available, with 140 units) from a benchmark repository [11], which has two separable classes, **A** (41 points) and **B** (99 points); other samples of 140 units from the same collection are instead formed by inseparable classes. To take a guess at its distribution, a nonlinear projection from the original space $R^5$ into $R^2$ by means of the Sammon method is reported in the left panel of fig. 4. The values of the optimal ***RBF*** classifier $D^*(\mathbf{x})$ are presented in the right panel; the 64 red circled values pertain to the support vector set and are positioned at $\pm\, 1$ and at $\pm\, m^* = \pm\, 2.2333$.
The optimal primal (simplex) solution to (8) results in the following:

$$\mathbf{p}^* = [m^* \ \ \theta_1^*\theta_2^*\ldots\theta_{140}^* \ \ b^*]^T \ = [2.2333 \ \ \ldots\ 63 \text{ values} \neq 0 \ldots 78 \text{ values} = 0]^T,$$

which reflects a certain degree of sparseness in the solution (63 optimal coefficients instead of 140).

In the right panel of fig. 4, the sequence of points on the abscissa follows an arrangement by class in the above-defined order: first is class **A**, then class **B**. All of the red circled points on the lines $\pm 1$ pertain to the class-boundary support vectors, and those on the lines $\pm m^* = \pm 2.2333$ pertain to the minmax support vectors. The dataset is correctly classified, because we obtain that:

$$1 \leq D^*(\mathbf{x}_{i \in \mathbf{A}}) \leq 2.2333$$
$$-2.2333 \leq D^*(\mathbf{x}_{i \in \mathbf{B}}) \leq -1.$$

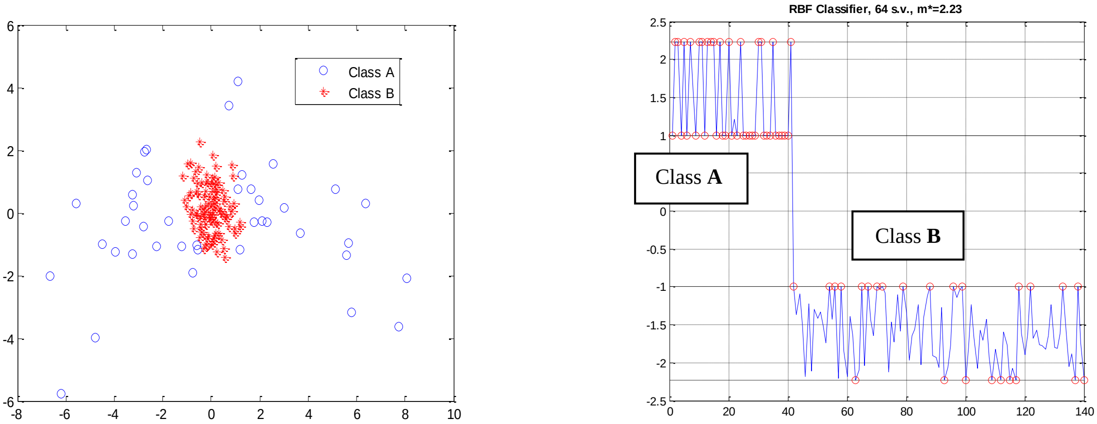

Fig. 4: Separable classes projected from $R^5$ into $R^2$ (left) with their optimal ***RBF*** classifier (right)

The results presented in fig. 5 were instead developed around a random Gaussian sample of two separable classes in $R^2$, class **A** (associated with the left yellow squares) and class **B** (associated with the right yellow circles), which are presented in the same way in the left and in the right panel. The regions covered by the red and the green points (from a sample of 2000 random points $\mathbf{x}_i \in R^2$ located inside a square with sides of length 20 at the origin O) pertain to the classification results of the optimal LP_MM classifier (the left panel) and of the optimal QP_SVM classifier (the right panel) trained on the two classes, both with an ***RBF*** kernel and proper values of $\gamma_{\text{RBF}}$, applied to the random sample. The regions are depicted in red (points assigned to class **A**) and green (points assigned to class **B**), as a result of the classification. Additionally, the trace of the minimum-error-rate Bayes classifier [8, chapter 2, page 9] is presented as the blue curve (with its prior probabilities fixed at 0.5).

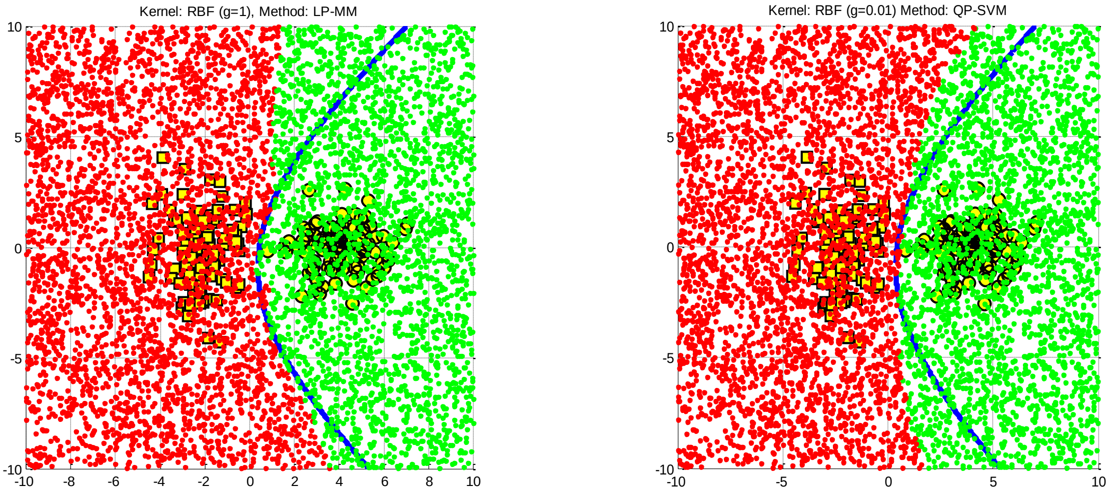

Fig. 5: The classification regions for the LP_MM (left) and the QP_SVM (right) classifier (the blue curve: the minimum-error-rate Bayes classifier)

# 5 The solution to inseparable classes (soft minmax optimisation)

In much the same way as in traditional QP_SVM, in our context, the case of inseparable classes can be treated by introducing nonnegative slack variables $\xi_i$.

Let us examine the following modified constraint of the type (7.b) and its implications, for example, pertaining to point $\mathbf{x}_i \in \mathbf{A}$ (where $c_i = +1$); analogous considerations hold for when point $\mathbf{x}_i \in \mathbf{B}$:

$$c_i D(\mathbf{x}_i \in \mathbf{A}) + \xi \geq 1 \Rightarrow D(\mathbf{x}_i \in \mathbf{A}) \geq 1 - \xi \quad (\xi \geq 0). \tag{10}$$

If the constraint (10) becomes active when the LP problem is solved, then the "greater than or equal to sign $\geq$" becomes an "equal sign =", and several results can arise as a consequence of the *equality* $D^*(\mathbf{x}_i \in \mathbf{A}) = 1 - \xi$ :

a) if $\xi = 0 \Rightarrow D^*(\mathbf{x}_i \in \mathbf{A}) = 1$: $\mathbf{x}_i$ is a (normal) class-boundary support vector;

b) if $0 < \xi < 1 \Rightarrow 0 < D^*(\mathbf{x}_i \in \mathbf{A}) < 1$: $\mathbf{x}_i$ is a *special* class-boundary support vector that assigns to $D^*$ a value in the forbidden region (-1, 1). The point $\mathbf{x}_i$ continues to be correctly classified, because sign($D^*(\mathbf{x}_i \in \mathbf{A})$) > 0. Recall that the sign of $D^*(\mathbf{x})$ decides the association of $\mathbf{x}$ with class **A** (if positive) or class **B** (if negative).

c) if $\xi = 1 \Rightarrow D^*(\mathbf{x}_i \in \mathbf{A}) = 0$: the class attribution is not possible.

d) if $\xi > 1 \Rightarrow D^*(\mathbf{x}_i \in \mathbf{A}) < 0$: $\mathbf{x}_i$ is a *special* class-boundary support vector that assigns to $D^*$ an incorrect value. In fact, sign($D^*(\mathbf{x}_i \in \mathbf{A})$) < 0 and $\mathbf{x}_i$ is associated with class **B**.

In other words, the term $\xi$ allows the constraint (10), when active, to be relaxed by choice; thus, if we add this degree of freedom in the problem and we impose in the LP problem its contemporary minimisation toward $0^+$, then the least possible amount of misclassification is assured, and a gain in classification can be expected in the situation of inseparable classes.

Thus, a formulation of the minmax problem for inseparable classes becomes the following, which considers slack nonnegative variables $\xi_i$ only at class-boundary constraints:

$$\left.\begin{array}{l} \min\,(m + \mathrm{C}\sum_{1}^{p}\xi_i), \quad \text{s.t.} \\ m - c_i D(\mathbf{x}_i) \geq 0 \\ c_i D(\mathbf{x}_i) + \xi_i \geq 1 \\ \xi_i \geq 0, \quad \mathrm{C} > 0, \quad i = 1, 2, \ldots, p \end{array}\right\} \tag{11}$$

As already stated, the contemporary minimisation of the term $\sum_{1}^{p}\xi_i\,(\xi_i \geq 0)$ should assure only the survival of the components $\xi_i$ that are necessary to recover, at the boundaries, situations of inseparability between the classes A and B. C is the usual parameter of trade-off between the two terms.

Accounting for the correspondence rules between primal and dual LPs [10, pag.131], in vector/matrix form, the dual of (11) is expressed exactly as in (9), with a series of additional constraints. In other words, we obtain:

$$\left.\begin{array}{l} \max\,[0\ 0 \ldots\ldots 0\ 1\ 1 \ldots\ldots\ 1]\ [v_1\ v_2 \ldots\ldots\ldots\ldots v_{2p}]^T, \text{ s.t.} \\ \mathbf{H}^T\ [v_1 v_2 \ldots\ldots\ldots\ldots v_{2p}]^T = [1\ 0\ 0 \ldots\ldots\ldots\ldots 0]^T \\ 0 \leq v_{p+i} \leq \mathrm{C}, \\ v_i \geq 0 \\ i = 1, 2, \ldots\ldots, p \end{array}\right\} \tag{12}$$

We observe only a difference in the upper bound C for all of the terms $v_{p+i}$, which constitute the upper half of the dual vector $\mathbf{v}$; $v_i$ is associated with the lower half of $\mathbf{v}$ instead, for i = 1, 2,..., p.

Fig. 6 presents an example in $R^2$ that pertains to a random sample of 40 points that are equally distributed between the partially overlapped classes **A** and **B** (the left panel). The optimal ***RBF*** classifier derived from the solution of (12) (the right panel, C = 8) shows one green circled point belonging to class B producing misclassification, because of assuming an incorrect sign; the remainder is correctly classified. The 31 black circled points of the right panel pertain to the set of support vectors (at $D^*(\mathbf{x}) = \pm 1$ and $D^*(\mathbf{x}) = m^* = \pm 2.72$).

# 6 Analysing the solution to inseparable classes

The dual structure (12) implies that there could be, in its solution $\mathbf{v}^*$, a certain number of components that assume the limit value C. To explore the possible implications, we also have, in this case, recourse to the C.S. theorem presented in section 4. As before, it is applied to the primal/dual pair when expressed in symmetric form after the simple operations of transformation of unrestricted sign variables into restricted sign variables and accounting for the fact that the variables $\xi_i$ now enter the primal vector of variables; see (11).

Likewise, the conditions $0 \leq v_{p+i} \leq C$, i = 1, 2,…, p, enter the dual constraint system; see (12).
Let us examine in detail four interesting situations, which involve primal and dual variables:

1) $\xi_i^* > 0 \quad \Rightarrow v_{p+i}^* = C$ (from property 3) of the C.S. theorem).

2) $v_{p+i}^* > 0 \;\Rightarrow\; c_i D^*(\mathbf{x}_i) + \xi_i^* = 1$ (from property 1) of the C.S. theorem): a slack variable is present.

3) $\xi_i^* = 0 \quad \Leftarrow v_{p+i}^* < C$ (from property 4) of the C.S. theorem).

4) $v_i^* > 0 \quad \Rightarrow m^* - c_i D^*(\mathbf{x}_i) = 0$ (from property 1) of the C.S.theorem) $\Rightarrow D^*(\mathbf{x}_i) = \pm\, m^*$(minmax support vector)

Combining the above results in 1) and 2), we obtain that

$$\xi_i^* > 0 \Rightarrow v_{p+i}^* = C > 0 \Rightarrow c_i D^*(\mathbf{x}_i) + \xi_i^* = 1.$$

Consequences to the classification can arise, depending on the value of $\xi_i^*$, which was discussed at the beginning of section 5 (points b), c), d) of the list given there).
Combining the above results in 2) and 3), we obtain that

$$0 < v_{p+i}^* < C \Rightarrow c_i D^*(\mathbf{x}_i) = 1 \Rightarrow D^*(\mathbf{x}_i) = \pm 1 \text{ (class-boundary support vector).}$$

In other words, an optimal dual value $0 < v_{p+i}^* < C$ identifies a normal class-boundary support vector, whereas an optimal primal value $\xi_i^* > 0$ gives rise to different possible situations (see points b), c), d) of the list given in section 5).

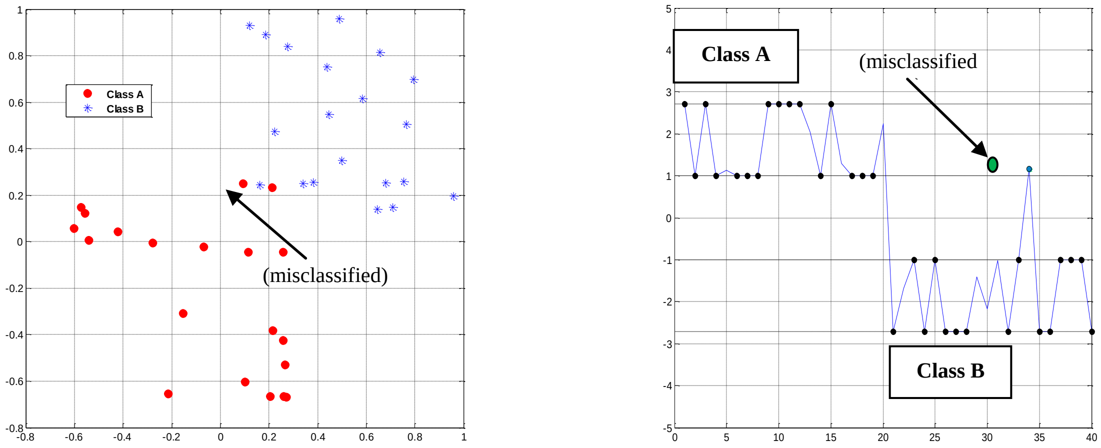

Fig. 6: An example of partially overlapped classes in $R^2$ (left) with their ***RBF*** classifier (right)

# 7 Computational comparisons

Alternative developments of SVM were proposed as solutions to maximising margins, which were implemented using, for example, the $L_1$ or $L_\infty$ weight vector norms, which led to LP formulations whose solutions can be searched by off-the-shelf LP solvers (i.e., CMPLX$^{TM}$). However, for large-scale data, such an approach can be very expensive or impractical, whereas for QP_SVM (based on the $L_2$ norm), several efficient implementations are now available [12], [13], [14]. Some algorithmic developments for large-size LP problems were proposed in recent years, inspired by chunking methods that claim significant progress [7]. In the actual absence of analogous developments for our LP_MM method, which we consider to be, in principle, possible, we will center our next discussion on the accuracy and methods of improving it, limiting the comparisons to samples with a small or medium size.

A first experimental comparison among classifiers was conducted using 2 datasets from a benchmark repository [11]: the *banana* (an artificial dataset) and the *thyroid* dataset, which are both available in a series of 100 partitions into training and test sets. On each training set, we built our LP_MM classifier and obtained the related *test* error by its application to the corresponding *test* sample; then, all of them were averaged, and this result compared with the average test error of an optimal QP_SVM classifier, which is described in [11] and pertains to the same datasets; the final results are reported in table 1.

To define the free parameters (C, $\gamma_{RBF}$) of our method, we conducted a tuning procedure on 5 training sets; as a result, the contents of table 1 pertain to the remaining 95. It is shown that almost the same amount of generalised error is attributable to both classifiers (the standard deviations are also quoted); the results show only a slight advantage in favour of our LP_MM method.

Another comparison was performed that considered the datasets referenced in [13]. There, the authors showed the best accuracy levels of the QP_SVM classifiers with regard to two real-world examples, which were obtained by the LIBSVM software after data scaling and an optimal search of the free parameters C and $\gamma_{RBF}$; table 2 summarises these results.

Table 3 reports our results, providing evidence in any case to high levels of accuracy; some words of explanation are appropriate. The rank $r_{\mathbf{H}}$ of the system matrix $\mathbf{H}^T$ from analysis of the first dataset (Astroparticle) decreased to 1158; as a result, the LP_MM procedure was given an equivalent number of constraints, which were randomly chosen within the available sample and assured the rank $r_{\mathbf{H}}$.

Table 1: Average generalisation errors

| Average Generalisation Error (%) | QP_SVM | LP_MM |
|---|---|---|
| Banana<br>$K(\mathbf{x},\mathbf{y}) = \exp(-\gamma \|\mathbf{x}-\mathbf{y}\|^2)$ | 11.5 ± 0.6 (C=316.2, $\gamma_{RBF}$ =1) | 11.3 ± 0.6 (C=1000, $\gamma_{RBF}$ =0.004) |
| Thyroid<br>$K(\mathbf{x},\mathbf{y}) = \exp(-\gamma \|\mathbf{x}-\mathbf{y}\|^2)$ | 4.8 ± 2.2 (C=10, $\gamma_{RBF}$ =3) | 4.35 ± 2.3 (C=10, $\gamma_{RBF}$ =3) |

This process of constraint reduction was next pursued and applied on a wider basis to produce classifiers that were built on reduced samples: one random subsample of 800 and 10 different random subsamples of 400 constraints each (with overlaps) were extracted from the training set.

Table 2: Accuracy results for two real-world examples [13]

| Applications | #training data | #testing data | # features | # classes | Testing Accuracy by LIBSVM |
|---|---|---|---|---|---|
| Astroparticle | 3,089(*) | 4,000 | 4 | 2 | 96.9% |
| Vehicle | 1,243 | 41 | 21 | 2 | 87.8% |

*(*) the training sample contains duplicate copies of data; the effective size goes down to 3,025*

The upper half of table 3 reports the training/testing accuracy of the related LP_MM classifiers; the accuracy losses of the subsample-based classifiers were very limited, which justifies the idea that a proper sample size reduction, even if related to a suboptimal solution, can retain an acceptable level of accuracy.

This measure can be seen, for example, for matrices $\mathbf{K} = \{K(i,j)\}$ with RBF kernel $K(\bullet,\bullet)$ (see (8)), as the *approximation* of a Gram matrix by a low-rank matrix [15].

The second dataset (Vehicle) yielded analogous results (see table 3, lower half) with regard to the subsample-based classifiers, therefore encouraging this practice.

On the other hand, often in our experience with practical medium and large size applications that involve ***RBF*** kernels, the system matrix $\mathbf{H}^T$ suffers from a rank deficiency; as a result, the proposed measure of reduction agrees with respect to this occurrence.

Also in our analyses, a data scaling and an optimal definition of C and $\gamma_{RBF}$ was preliminarily performed, and a Pentium dual-core CPU with 2.80 GHz and 4 GB was employed.

The GLPK (GNU Linear Programming Kit) solver was utilised for the LP solutions, which is freely available at http://ftp.gnu.org/gnu/glpk/.

Last, an interesting result is presented in table 4. This result refers to a classifier with a *sigmoid* kernel that is implemented on the *thyroid* dataset of table 1 and that produced, in every instance, a kernel matrix **K** = {K(i,j)} with at least one *negative* eigenvalue.
In fact, it is well known that the sigmoid kernel matrix is conditionally positive definite [16] in its parameters.
The LP procedure was unaffected by this situation, as expected. Thus, this procedure results in a solution, even though the solution has a degraded average performance compared with the levels given in table 1.

Table 3: Accuracy results for the LP_MM method

| Applications | C, $\gamma_{RBF}$, # of constraints | m* | Computing time (sec.) | Training accuracy | Testing accuracy |
|---|---|---|---|---|---|
| Astroparticle (size reduced to the rank$r_{\mathbf{H}}$ = 1158) | 2, 2, 1158 | 2.88 | 8611 | 97.52% | 95.30% |
| Astroparticle (subsample of 800 units) | 2, 2, 800 | 4.27 | 5455 | 97.49% | 95.45% |
| Astroparticle (averaged over 10 subsamples of 400) | 2, 2, 400 | 15.69 | 683 | 96.85% | 94.95% |
| Vehicle (full size) | 1, 1, 1243 | 1 | 4125 | 100% | 100% |
| Vehicle (subsample of 600 units) | 1, 1, 600 | 15.59 | 3027 | 99.92% | 100% |
| Vehicle (averaged over 10 subsamples of 400) | 1, 1, 400 | 104.52 | 59 | 98.39% | 97.56% |

## 8 Conclusions

A method for supervised binary classification based on the minmax Chebyshev criterion has been presented. This method is shown to be competitive with the maximum margin method. The method implies the solution of a linear programming problem and, in general, requires simpler computational procedures. In our case, a customisation to address large size applications has not yet been developed. The method, deriving from an approximation to a non-linear problem (section 3), has demonstrated its coherence and validity on a series of experimental situations and comparisons.
A number of questions are still open, and deeper investigations are required; we mention, among these, the treatment of situations with rank deficiency, which is often encountered in practical applications and is motivated by the fact that usually many eigenvalues of the matrix **H** approach zero (in practical situations, the concept of *numerical rank* is to be adopted). This question is important because its exploitation could allow us a reduction of the constraints and computing time, which is also significant. However, other investigations are necessary.

Table 4: The average generalisation error for a sigmoid kernel with negative eigenvalues

| Average Generalisation Error (%) | QP_SVM | LP_MM |
|---|---|---|
| Thyroid<br>K($\mathbf{x}$,$\mathbf{y}$)=tanh[$\alpha(\mathbf{x}_*\mathbf{y}^T)$/dim($\mathbf{x}$)+$\beta$] | ---<br>Not Positive Definiteness | 13.28 ± 3.85 (C=1, $\alpha$=0.01, $\beta$ =0.1) |

The proposed method offers the following properties: being insensitive to possible negative eigenvalues of the matrix **K,** maintaining at the same time two important points of convenience of the QP solution, i.e., the sparseness of its solution and the achievement of the well-appreciated kernel trick.
A Matlab code is freely available upon request.

## References

[1] B.E. Boser, I.M.Guyon, V.Vapnik, A training algorithm for optimal margin classifiers, Proceedings *Fifth* ACM *Workshop on Computational Learning Theory*, Pittsburgh, 1992

[2] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000

[3] S. Abe, Support Vector Machines for Pattern Classification, Springer-Verlag London, 2005

[4] C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 2, 121–167 (1998)

[5] W. Zhou, L. Zhang, L. Jiao, Linear programming support vector machines, Pattern Recognition, 35(12):2927-36, 2002

[6] V. Kecman, T. Arthanari, I. Hadzic, LP and QP based learning from empirical data, Proceedings of International Joint Conference on Neural Networks (IJCNN), Como, Italy, 2000

[7] M.C. Ferris, O.L. Mangasarian, S.J. Wright, Linear Programming with MATLAB, MPS-SIAM Series on Optimization, 2007

[8] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley Interscience, 2000; freely available at: http://neuron.tuke.sk/hudecm/helps/NN/DUDA_HART_STORK/
[9] D.G. Luenberger, Linear and Nonlinear Programming, Addison-Wesley Publishing Company, 2nd ed., 1984
[10] G.B. Dantzig, M.N. Thapa, Linear Programming, vol.1: Introduction, Springer series in operations research, 1997
[11] G. Ratsch, T. Onoda, K-R. Muller, Soft Margins for AdaBoost, NeuroCOLT2 Technical Report Series, NC-TR-1998-021, 1998
[12] Chih-Chung Chang, Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011
[13] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, A practical guide to Support Vector Classification, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2010
[14] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges and A. Smola (Eds.), MIT Press, 1999.
[15] P. Drineas, M.W. Mahoney, Approximating a Gram Matrix for Improved Kernel-Based Learning, COLT 2005, P. Auer and R. Meir (Eds.), pp. 323-337, Springer-Verlag 2005
[16] Hsuan-Tien Lin, Chih-Jen Lin, A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods, Department of Computer Science and Information Engineering, National Taiwan University; available on-line: www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf