A Geometric Approach to Support Vector Machine (SVM) Classification

Michael E. Mavroforakis, and Sergios Theodoridis, *Senior Member, IEEE*

***Abstract*—The geometric framework for the SVM classification problem provides an intuitive ground for the understanding and the application of geometric optimization algorithms, leading to practical solutions of real world classification problems. In this work, the notion of "reduced convex hull" is employed and supported by a set of new theoretical results. These results allow existing geometric algorithms to be directly and practically applied to solve not only separable, but also non-separable classification problems both accurately and efficiently. As a practical application of the new theoretical results, a known geometric algorithm has been employed and transformed accordingly to solve non-separable problems successfully.**

***Index Terms*—Support vector machines, reduced convex hulls, classification, pattern recognition, kernel methods.**

I. Introduction

Support Vector Machine (SVM) formulation of pattern recognition (binary) problems brings along a bunch of advantages over other approaches, e.g., [1], [2], some of which are: 1) the assurance that once a solution has been reached, it is the unique (global) solution, 2) good generalization properties of the solution, 3) sound theoretical foundation based on learning theory (*Structural Risk Minimization* (SRM)) and Optimization theory, 4) common ground / formulation for the class separable and the class non-separable problems (through the introduction of appropriate penalty factors of arbitrary degree in the optimization cost function) as well as for linear and non-linear problems (through the so called "kernel trick") and, last but not least, 5) clear geometric intuition on the classification task. Due to the above nice properties, SVM have been successfully used to a number of applications, e.g., see [3] – [9].

The contribution of this work consists of the following: 1) It provides the theoretical background for the solution of the non-separable (both linear and non-linear) classification problems with linear (1st degree) penalty factors, by means of the reduction of the size of the convex hulls of the training patterns. This task, although is, in principle, of combinatorial complexity in nature, it is transformed to one of linear complexity, by a series of theoretical results deduced and presented in this work. 2) It exploits the intrinsic geometric intuition to the full extent, i.e., not only theoretically but also practically (leading to an algorithmic solution), in the context of classification through the SVM approach. 3) It provides an easy way to relate each class with a different penalty factor, i.e., to relate each class with a different risk (weight). 4) It applies a fast, simple and easily conceivable algorithm to solve the SVM task. Finally, 5) it opens the road for applying other geometric algorithms, finding the closest pair of points between convex sets in Hilbert spaces, for the non-separable SVM problem.

Although some authors have presented the theoretical background of the geometric properties of SVMs, exposed thoroughly in [10], the main stream of solving methods comes from the algebraic field (mainly decomposition). One of the best representative algebraic algorithms with respect to speed and ease of implementation, also presenting very good scalability properties, is the Sequential Minimal Optimization (SMO) [11]. The geometric properties of learning [12] and specifically of SVMs in the feature space, have been pointed out early enough, through the dual representation (i.e., the convexity of each class and finding the respective support hyperplanes that exhibit the maximal margin) for the separable case [13] and also for the non-separable case through the notion of the *Reduced Convex Hull* (RCH) [14]. However, the geometric algorithms presented until now ([15], [16]) are suitable only for solving directly the separable case. These geometric algorithms, in order to be useful, have been extended to solve indirectly the non-separable case through the technique proposed in [17], which transforms the non-separable problem to a separable one. However, this transformation (artificially extending the dimension of the input space by the number of training patterns) is equivalent to a quadratic penalty factor. Moreover, besides the increase of complexity due to the artificial expansion of the dimension of the input space, it has been reported that the generalization properties of the resulting SVMs can be poor [15].

The content of the rest of the paper has been structured as follows: In Section II, some preliminary material on SVM classification has been presented. In Section III, the notion of the reduced convex hull is defined and a direct and intuitive

Manuscript received November 11, 2004. (Write the date on which you submitted your paper for review.)

M. Mavroforakis is with the Informatics and Telecommunications Dept. of the University of Athens, TYPA buildings, Univ. Campus, 15771, Athens, Greece (phone: +30-210-9648663, +30-210-9479417; e-mail: mmavrof@di.uoa.gr).

S. Theodoridis is with the Informatics and Telecommunications Dept. of the University of Athens, TYPA buildings, Univ. Campus, 15771, Athens, Greece (e-mail: stheodor@di.uoa.gr).

connection to the non-separable SVM classification problem is presented. In the sequel, the main contribution of this work is displayed, i.e., a complete mathematical framework is devised to support the RCH and therefore make it directly applicable to practically solve the non-separable SVM classification problem. Without this framework, the application of a geometric algorithm in order to solve the non-separable case through RCH is practically impossible, since it is a problem of combinatorial complexity. In Section IV, a geometric algorithm is rewritten in the context of this framework, therefore showing the practical benefits of the theoretical results derived herewith to support the RCH notion. Finally, in Section V, the results of the application of this algorithm to solve certain classification tasks are presented.

II. PRELIMINARY

The complex and challenging task of (binary) classification or (binary) pattern recognition in supervised learning can be described as follows [18]: given a set $\mathcal{X}$ of training objects (patterns) – each belonging to one of two classes – and their corresponding class identifiers, assign the correct class to a newly (not a member of $\mathcal{X}$) presented object; ($\mathcal{X}$ does not need any kind of structure except of being a non-empty set). For the task of learning, a measure of similarity between the objects of $\mathcal{X}$ is necessary, so that patterns of the same class are mapped "closer" to each other, as opposed to patterns belonging to different classes. A reasonable measure of similarity has the form $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$, $(x_1,x_2)\to k(x_1,x_2)$, where k is (usually) a real (symmetric) function, called a kernel. An obvious candidate is the inner product $(x_1\,|\,x_2)$[1], in case that $\mathcal{X}$ is an inner-product space (e.g., $\mathbb{R}^d$), since it leads directly to a measure of lengths through the norm derived from the inner product $\|x\|=\sqrt{(x\,|\,x)}$ and also to a measure of angles and hence to a measure of distances. When the set $\mathcal{X}$ is not an inner product space, it may be possible to map its elements x to an inner product space, $\mathcal{H}$, through a (nonlinear) function $\Phi:\mathcal{X}\to\mathcal{H}$ such that $\mathbf{x}\equiv\Phi(x)$, $\mathbf{x}\in\mathcal{H}$, $\forall x\in\mathcal{X}$. Under certain loose conditions (imposed by Mercer's theorem [19]), it is possible to relate the kernel function with the inner product of the feature space $\mathcal{H}$, i.e., $k(x_1,x_2)=(\Phi(x_1)\,|\,\Phi(x_2))$ for all $x_1,x_2\in\mathcal{X}$. Then, $\mathcal{H}$ is known as a Reproducing Kernel Hilbert Space (RKHS). RKHS is a very useful tool, because any Cauchy sequence converges to a limit in the space, which means that it is possible to approximate a solution (e.g., a point with maximum similarity) as accurately as needed.

[1] The notation $(x\,|\,y)$ will be used interchangeably with $\langle x,y\rangle$, for spaces which coincide with their dual.

A. SVM classification

Simply stated, a SVM finds the best separating (*maximal margin*) hyperplane between the two classes of training samples in the feature space, as it is shown in Fig. 1.

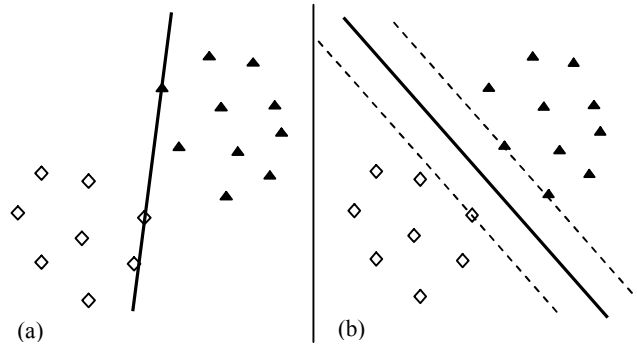

Fig. 1. A separating hyperplane exhibiting zero margin (a), compared to the maximal margin separating hyperplane (b) for the same classes of training samples presented in feature space.

A *linear discriminant function* has the form of the linear functional $f(x)=\langle w,x\rangle+c$, which corresponds to a hyperplane [20], dividing the feature space. If, for a given pattern mapped in the feature space to x, the value of $f(x)$ is a positive number, then the pattern belongs to the class labeled by the numeric value $+1$; otherwise to the class with value -1. Denoting as y_i the numeric value of the class label of pattern x_i and m the maximum (functional) margin, the problem of classification is equivalent to finding the functional f (satisfying $y_i(\langle w,x_i\rangle+c)\geq m$) that maximizes m.

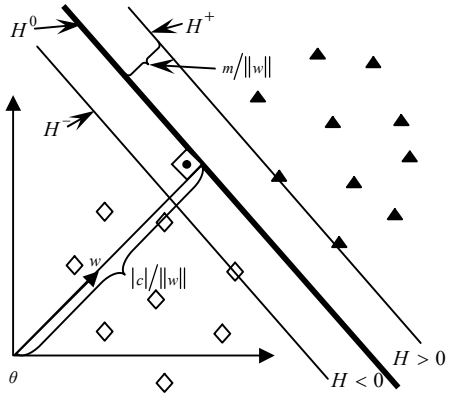

Fig. 2. Geometric interpretation of the maximal margin classification problem. Setting $H\equiv(\langle w,x\rangle/\|w\|)+(c/\|w\|)$ the hyperplanes $H^0:H=0$, $H^-:H=-(m/\|w\|)$ and $H^+:H=m/\|w\|$ are shown.

In geometric terms, expressing the involved quantities in "lengths" of w (i.e., $\|w\|$), the problem is restated as follows: find the hyperplane $H(w,c):\langle w,x\rangle+c=0$, maximizing the (geometric) margin $m/\|w\|$ and satisfying

$y_i\left(\frac{\langle w,x_i\rangle}{\|w\|}+\frac{c}{\|w\|}\right)\geq\frac{m}{\|w\|}$ for all the training patterns.

The *geometric margin* $m/\|w\|$ represents the minimum distance of the training patterns of both classes from the separating hyperplane defined by (w,c). The resulting hyperplane is called the *maximal margin hyperplane*. If the quantity $m/\|w\|$ is positive, then the problem is a linearly separable one. This situation is shown in Fig. 2.

It is clear that $H(w,c)=H(sw,sc)$, $s>0$ (because of the linearity of inner product) and since $\|sw\|=|s|\|w\|$, a scaling of the parameters w, c and m does not change the geometry. Therefore, assuming $m=1$ (*canonical hyperplane*), the classification problem takes the equivalent form: Find the hyperplane

$$H(w,c):\langle w,x\rangle+c=0 \tag{1}$$

maximizing the total (interclass) margin $2/\|w\|$, or equivalently minimizing the quantity

$$(1/2)\|w\|^2 \tag{2}$$

and satisfying

$$y_i\left(\langle w,x_i\rangle+c\right)\geq 1. \tag{3}$$

This is a quadratic optimization problem (if the Euclidean norm is adopted) with linear inequality constraints and the standard algebraic approach is to solve the equivalent problem of minimizing the Lagrangian

$$\mathcal{L}_P\equiv\frac{1}{2}\langle w,w\rangle-\sum_i a_i\left(y_i\left(\langle w,x_i\rangle+c\right)-1\right) \tag{4}$$

subject to the constraints $a_i\geq 0$. The corresponding dual optimization problem is to maximize

$$\mathcal{L}_D\equiv\sum_i a_i-\frac{1}{2}\sum_i\sum_j y_i y_j a_i a_j\langle x_i,x_j\rangle \tag{5}$$

subject to the constraints

$$a_i\geq 0 \tag{6}$$

and

$$\sum_i y_i a_i=0. \tag{7}$$

Denote, for convenience, by I^- and I^+ the sets of indices i, such that $y_i=-1$ and $y_i=+1$ respectively and by I the set of all indices, i.e., $I=I^-\cup I^+$.

The *Karush-Kuhn-Tucker* (KKT) *optimality conditions* provide the necessary and sufficient conditions that the unique solution has been found to the last optimization problem, i.e., (besides the initial constraints):

$$w=\sum_i y_i a_i x_i=\sum_{i\in I^+} a_i x_i-\sum_{i\in I^-} a_i x_i \tag{8}$$

$$\sum_i y_i a_i=0\Leftrightarrow\sum_{i\in I^+} a_i=\sum_{i\in I^-} a_i \tag{9}$$

and the KKT complementarity condition

$$a_i\left(y_i\left(\langle w,x_i\rangle+c\right)-1\right)=0 \tag{10}$$

which means that, for the *inactive* constraints there is $a_i=0$ and for the *active* ones (when $y_i\left(\langle w,x_i\rangle+c\right)-1=0$ is satisfied) there is $a_i\geq 0$. The points with $a_i>0$ lie on the canonical hyperplane and are called *support vectors*. The interpretation of the KKT conditions (especially (8) and (9) with the extra reasonable non-restrictive assumption that $\sum_{i\in I^+} a_i=\sum_{i\in I^-} a_i=1$) is very intuitive [1] and leads to the conclusion that the solution of the linearly separable classification problem is equivalent to finding the points of the two convex hulls [21] (each generated by the training patterns of each class) which are closest to each other and the maximum margin hyperplane a) bisects and b) is normal to the line segment joining these two closest points, as seen in Fig. 3. The formal proof of this is presented in [13].

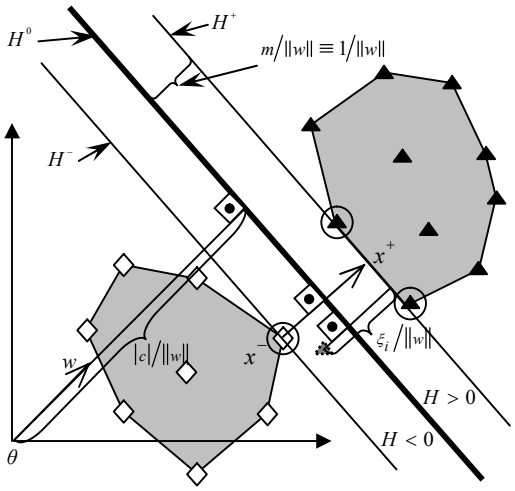

Fig. 3. Geometric interpretation of the maximal margin classification problem. Closest points are denoted by circles.

To address the (most common in real world applications) case of linearly non-separable classification problem, for which any effort to find a separating hyperplane is hopeless, the only way for someone to reach a solution is to relax the data constraints. This is accomplished through the addition of *margin slack variables* ξ_i, which allow a controlled violation of the constraints [22]. Therefore, the constraints in (3) become:

$$y_i\left(\langle w,x_i\rangle+c\right)\geq 1-\xi_i \tag{11}$$

where $\xi_i\geq 0$. It is clear that if $\xi_i>1$, then the point x_i is misclassified by the hyperplane $H(w,c)$. The quantity $\xi_i/\|w\|$ has a clear geometric meaning: it is the distance of the point x_i (in lengths of w) from the supporting hyperplane of its corresponding class; since ξ_i is positive, x_i lies in the opposite direction of the supporting hyperplane of its class, i.e., the corresponding supporting hyperplane separates x_i from its own class. A natural way to incorporate the cost for the errors in classification is to augment the cost function (2) by the term $C\sum_i\xi_i$ (although terms of the form $C\sum_i\xi_i^d$

have also been proposed), where C is a free parameter (known also as *regularization parameter* or *penalty factor*) indicating the penalty imposed to the 'outliers', i.e. higher value of C corresponds to higher penalty for the 'outliers' [23]. Therefore, the cost function (2) for the non-separable case becomes:

$$\frac{1}{2}\|w\|^2 + C\sum_i \xi_i \,. \tag{12}$$

Consequently, the Langrangian of the primal problem is

$$\begin{aligned}\mathcal{L}_P \equiv & \frac{1}{2}\langle w, w\rangle + C\sum_i \xi_i \\ & - \sum_i a_i \left(y_i \left(\langle w, x_i \rangle + c \right) - 1 + \xi_i \right) - \sum_i \nu_i \xi_i\end{aligned} \tag{13}$$

subject to the constraints $a_i \geq 0$ and $\nu_i \geq 0$ (introduced to ensure positivity of ξ_i). The corresponding *dual* optimization problem has again the form of (5) i.e. to maximize

$$\mathcal{L}_D \equiv \sum_i a_i - \frac{1}{2}\sum_i \sum_j y_i y_j a_i a_j \langle x_i, x_j \rangle \tag{14}$$

but now subject to the constraints

$$0 \leq a_i \leq C \tag{15}$$

and

$$\sum_i y_i a_i = 0 \,. \tag{16}$$

It is interesting that neither the slack variables, ξ_i, nor their associated Lagrange multipliers, ν_i, are present in the Wolfe dual formulation of the problem (a result of choosing $d = 1$ as the exponent of the penalty terms) and that the only difference from the separable case is the impose of the upper bound C to the Lagrange multipliers a_i.

However, the clear geometric intuition of the separable case has been lost; it is regained through the work presented in [14], [13] and [10], where the notion of the reduced convex hull, introduced and supported with new theoretical results in the next section, plays an important role.

III. Reduced Convex Hulls (RCH)

The set of all convex combinations of points of some set X, with the additional constraint that each coefficient a_i is upper-bounded by a non-negative number $\mu < 1$, is called the *reduced convex hull* of X and denoted by $\mathrm{R}(X, \mu)$:

$$\mathrm{R}(X,\mu) = \left\{ w : w = \sum_{i=1}^{k} a_i x_i,\ x_i \in X,\ \sum_{i=1}^{k} a_i = 1,\ 0 \leq a_i \leq \mu \right\}$$

Therefore, for the non-separable classification task, the initially overlapping convex hulls, with a suitable selection of the bound μ, can be reduced so that to become separable. Once separable, the theory and tools developed for the separable case can be readily applied. The algebraic proof is found in [14] and [13]; a totally geometric formulation of SVM leading to this conclusion is found in [10].

The effect of the value of bound μ to the size of the RCH is shown in Fig. 4.

In the sequel, we will prove some theorems and propositions that shed further intuition and usefulness to the RCH notion and at the same time form the basis for the development of the novel algorithm which is proposed in this paper.

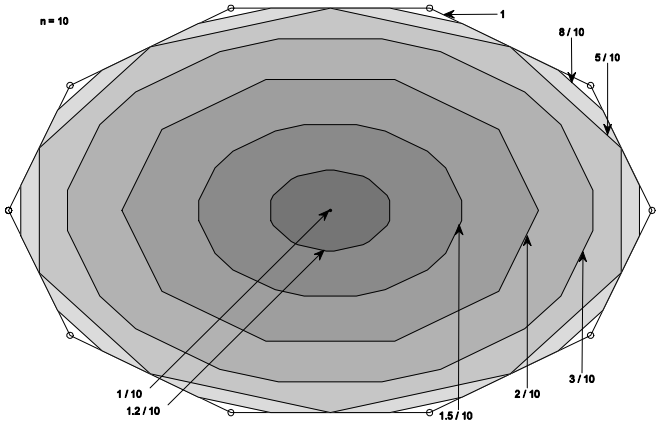

Fig. 4. Evolution of a convex hull with respect to μ. (The corresponding μ of each RCH are the values indicated by the arrows.) The initial convex hull ($\mu = 1$), generated by 10 points ($n = 10$), is successively reduced, setting μ to 8/10, 5/10, 3/10, 2/10, 1.5/10, 1.2/10 and finally 1/10, which corresponds to the centroid. Each smaller (reduced) convex hull is shaded with a darker color.

Proposition 1 : If all the coefficients a_i of all the convex combinations forming the RCH $\mathrm{R}(X, \mu)$ of a set X with k elements, are less than $1/k$ (i.e., $\mu < 1/k$), then $\mathrm{R}(X, \mu)$ will be empty.

Proof: $a_i < 1/k \Rightarrow \sum_{i=1}^{k} a_i < \sum_{i=1}^{k} (1/k) = k(1/k) = 1 \Rightarrow \sum_{i=1}^{k} a_i < 1$. Since $\sum_{i=1}^{k} a_i = 1$ is needed to be true, it is clear that $\mathrm{R}(X, \mu) = \{\varnothing\}$.□

Proposition 2: If for every i, there is $a_i = 1/k$ in a RCH $\mathrm{R}(X, \mu)$ of a set X with k different points as elements, then $\mathrm{R}(X, \mu)$ degenerates to a set of one single point, the *centroid* point (or *barycenter*) of X.

Proof: From the definition of the RCH, it is:

$\mathrm{R}(X,\mu) = \left\{ w : w = \sum_{i=1}^{k} a_i x_i \right\} = \left\{ w : w = (1/k) \sum_{i=1}^{k} x_i \right\} =$

$\left\{ w : w = (1/k) z \right\}$, where $z \equiv \sum_{i=1}^{k} x_i$ is a single point.□

Remark: It is clear that in a RCH $\mathrm{R}(X, \mu)$, a choice of $\mu > 1$ is equivalent with $\mu = 1$ as the upper bound for all a_i, because it must be $\sum_{i=1}^{k} a_i = 1$ and therefore $a_i \leq 1$. As a consequence of this and the above proposition, it is deduced that the RCH $\mathrm{R}(X, \mu)$ of a set X will be either empty (if

$\mu < 1/k$), or grows from the centroid ($\mu = 1/k$) to the convex hull ($\mu \geq 1$) of X .

For the application of the above to real life algorithms, it is absolutely necessary to have a clue about the extreme points of the RCH. In the case of the convex hull, generated by a set of points, only a subset of these points constitute the set of extreme points, which, in turn, is the *minimal representation* of the convex hull. *Therefore, only a subset of the original points is needed to be examined and not every point of the convex hull* [24]. In contrast, as it will soon be seen, for the case of RCH, its extreme points are the result of combinations of the extreme points of the original convex hull, which, however, *do not belong* to the RCH, as it was deduced above.

In the sequel, it will be shown that not any combination of the extreme points of the original convex hull leads to extreme points of the RCH, but only a small subset of them. *This is the seed for the development of the novel efficient algorithm to be presented later in this paper.*

Lemma 1: For any point $w \in \mathrm{R}(X,\mu)$, if there exists a reduced convex combination $w = \sum_{i=1}^{k} a_i x_i$, with $x_i \in X$, $\sum_{i=1}^{k} a_i = 1$, $0 \leq a_i \leq \mu$ and at least one coefficient a_r , $1 \leq r \leq k$, not belonging in the set $S = \{0,\ 1-\lfloor 1/\mu \rfloor \mu,\ \mu\}$, where $\lfloor 1/\mu \rfloor$ is the integer part of the ratio $1/\mu$, then there exists at least another coefficient a_s , $1 \leq s \leq k$, $r \neq s$, not belonging in the set S , i.e., there cannot be a reduced convex combination with just one coefficient not belonging in S .

Proof: The lengthy proof of this Lemma, is found in Appendix.

Theorem 1: The extreme points of a RCH $\mathrm{R}(X,\mu) = \left\{ w : w = \sum_{i=1}^{k} a_i x_i,\ x_i \in X,\ \sum_{i=1}^{k} a_i = 1,\ 0 \leq a_i \leq \mu \right\}$ have coefficients a_i belonging to the set $S = \{0,\ 1-\lfloor 1/\mu \rfloor \mu,\ \mu\}$.

Proof: In the case that $\mu = 1$ the theorem is obviously true since $\mathrm{R}(X,1)$ is the convex hull of X i.e. $\mathrm{R}(X,1) = \mathrm{conv}X$ and, therefore, all the extreme points belong to the set X . Hence, if x_i is an extreme point, its j -th coefficient $a_{i,j}$ is $a_{i,j} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$.

For $0 < \mu < 1$ the theorem will be proved by contradiction: Assuming that a point $w \in \mathrm{R}(X,\mu)$ is an extreme point, with some coefficients not belonging in S , a couple of other points $w_1, w_2 \in \mathrm{R}(X,\mu)$ is needed to be found and then to be proved that w belongs to the line segment $[w_1, w_2]$. As two points are needed, two coefficients have to be found not belonging in S . However, this is the conclusion of Lemma 1, which ensures that if there exists a coefficient of a reduced convex combination not belonging in S there exists a second one not belonging in S too.

Therefore, let an extreme point $w \in \mathrm{R}(X,\mu)$, where $w = \sum_{i=1}^{k} a_i x_i$, that have at least two coefficients a_r and a_s , $r \neq s$, such that $0 < a_r, a_s < \mu$ and $a_r, a_s \notin \{0, 1-\lfloor 1/\mu \rfloor \mu, \mu\}$. Let also $\delta > 0$ such that $\delta < \min\{a_r, a_s\}$ and $\delta < \mu - \max\{a_r, a_s\}$, i.e., it is $0 \leq a_r \pm \delta, a_r \pm \delta \leq \mu$. Consequently, the points w_1, w_2 are constructed as follows: $w_1 = \sum_{i \neq r,s} a_i x_i + (a_r + \delta) x_r + (a_s - \delta) x_s$ and $w_2 = \sum_{i \neq r,s} a_i x_i + (a_r - \delta) x_r + (a_s + \delta) x_s$. For the middle point of the line segment $[w_1, w_2]$, it is $1/2\, w_1 + 1/2\, w_2 = \sum_{i \neq r,s} a_i x_i + 1/2(a_r + \delta + a_r - \delta) x_r + 1/2(a_s - \delta + a_s + \delta) x_s =$ $1/2\left(\sum_{i \neq r,s} a_i x_i + (a_r + \delta) x_r + (a_s - \delta) x_s \right)$ $+ 1/2\left(\sum_{i \neq r,s} a_i x_i + (a_r - \delta) x_r + (a_s + \delta) x_s \right) =$ $\sum_{i=1}^{k} a_i x_i = w$ which is a contradiction to the assumption that w is an extreme point. This proves the theorem. □

Proposition 3: Each of the *extreme points* of a RCH $\mathrm{R}(X,\mu) = \left\{ w : w = \sum_{i=1}^{k} a_i x_i,\ x_i \in X,\ \sum_{i=1}^{k} a_i = 1,\ 0 \leq a_i \leq \mu \right\}$ is a reduced convex combination of $m = \lceil 1/\mu \rceil$ (distinct) points of the original set X , where $\lceil 1/\mu \rceil$ is the smallest integer for which it is $\lceil 1/\mu \rceil \geq 1/\mu$. Furthermore, if $1/\mu = \lceil 1/\mu \rceil$ then all $a_i = \mu$; otherwise, $a_i = \mu$ for $i = 1, \ldots, m-1$ and $a_m = 1 - \lfloor 1/\mu \rfloor \mu$.

Proof: Theorem 1 states that the only coefficients through which a point from the original set X contributes to an extreme point of the RCH $\mathrm{R}(X,\mu)$ are either μ or $1 - \lfloor 1/\mu \rfloor \mu$.

If $\lfloor 1/\mu \rfloor = 1/\mu$ then $1 - \lfloor 1/\mu \rfloor \mu = 0$, hence the only coefficient valid is μ and, since $\sum_i a_i = 1$ and $a_i = \mu$, it is $m\mu = 1 \Leftrightarrow\ m = 1/\mu = \lceil 1/\mu \rceil$.

If $(1/\mu) - \lfloor 1/\mu \rfloor \equiv \rho$ with $0 < \rho < 1$ then $1 - \lfloor 1/\mu \rfloor \mu = \mu\rho$ and therefore $1 - \lfloor 1/\mu \rfloor \mu < \mu$. Let w be an extreme point of $\mathrm{R}(X,\mu)$, q be the number of points contributing to w with coefficient μ and p the number of points with coefficient $1 - \lfloor 1/\mu \rfloor \mu$, i.e. $w = \mu \sum_{i=1}^{q} x_i + \left(1 - \lfloor 1/\mu \rfloor \mu\right) \sum_{j=1}^{p} x_j$ (17). Since $\sum_i a_i = 1$, there is $p\left(1 - \lfloor 1/\mu \rfloor \mu\right) + q\mu = 1 \Rightarrow$ $p(1/\mu) - p\lfloor 1/\mu \rfloor + q = 1/\mu$ (18). If $p = 1$ then (18) becomes

$(1/\mu)-\lfloor 1/\mu \rfloor + q = 1/\mu \Rightarrow \quad q = \lfloor 1/\mu \rfloor$, hence $m = p+q = 1+\lfloor 1/\mu \rfloor = \lceil 1/\mu \rceil$ which is the desired result.

Therefore, the remaining case, is when $p>1$. Assuming that there exist at least two initial points x_u and x_v with coefficient $1-\lfloor 1/\mu \rfloor \mu$, the validity of the proposition will be proved by contradiction. Since it is true $0<1-\lfloor 1/\mu \rfloor \mu<\mu$ for this case, there exists a real positive number $\delta>0$ s.t. $\delta < \min\left(1-\lfloor 1/\mu \rfloor \mu, \mu-\left(1-\lfloor 1/\mu \rfloor \mu\right)\right)$. Let $a_r = 1-\lfloor 1/\mu \rfloor \mu - \delta$ and $a_s = 1-\lfloor 1/\mu \rfloor \mu + \delta$; using them, let $w_1 = \mu\sum_{i=1}^{q} x_i + \left(1-\lfloor 1/\mu \rfloor \mu\right)\sum_{j=1}^{p-2} x_j + a_r x_u + a_s x_v$ and $w_2 = \mu\sum_{i=1}^{q} x_i + \left(1-\lfloor 1/\mu \rfloor \mu\right)\sum_{j=1}^{p-2} x_j + a_s x_u + a_r x_v$.

Obviously, since $0<a_r, a_s<\mu$, the points w_1 and w_2 belong in the RCH $\mathrm{R}(X,\mu)$. Taking into consideration that $a_r + a_s = 1-\lfloor 1/\mu \rfloor \mu - \delta + 1-\lfloor 1/\mu \rfloor \mu + \delta = 2\left(1-\lfloor 1/\mu \rfloor \mu\right)$, the middle point of the line segment $[w_1, w_2]$, is $(1/2)w_1 + (1/2)w_2 =$ $\mu\sum_{i=1}^{q} x_i + \left(1-\lfloor 1/\mu \rfloor \mu\right)\sum_{j=1}^{p-2} x_j + \left((a_r+a_2)/2\right)(x_u+x_v) =$ $\mu\sum_{i=1}^{q} x_i + \left(1-\lfloor 1/\mu \rfloor \mu\right)\sum_{j=1}^{p} x_j = w$. Therefore w cannot be an extreme point of the RCH $\mathrm{R}(X,\mu)$, which contradicts with the assumption that $p>1$ and this concludes the proof. □

Remark: For the coefficients $\lambda \equiv 1-\lfloor 1/\mu \rfloor \mu$ and μ, it holds $0 \le \lambda < \mu$. This is a byproduct of the proof of the above Proposition 3.

Remark: The separation hyperplane depends on the pair of closest points of the convex hulls of the patterns of each class, and each such point is a convex combination of some extreme points of the RCHs. As, according to the above Theorem, each extreme point of the RCHs depends on $\lceil 1/\mu \rceil$ original points (training patterns), it follows directly that the number of support vectors (points with non-zero Lagrange multipliers) is at least $\lceil 1/\mu \rceil$, i.e, *the lower bound of the number of initial points contributing to the discrimination function is* $\lceil 1/\mu \rceil$ (Fig. 5).

Remark: Although the above Theorem 1, along with Proposition 3, restrict considerably the candidates to be extreme points of the RCH, since they should be reduced convex combinations of $\lceil 1/\mu \rceil$ original points and also *with specific coefficients* (belonging to the set S), the problem is still of combinatorial nature, because each extreme point is a combination of $\lceil 1/\mu \rceil$ out of k initial points for each class. This is shown in Fig. 5. Theorem 1 provides the *necessary* but not the *sufficient* condition for a point to be extreme in a RCH. The set of points satisfying the condition is larger than the set of extreme points; these are the "candidate to be extreme points", shown in Fig. 5. Therefore, the solution of the problem of finding the closest pair of points of the two reduced convex hulls essentially entails the following three stages:

1. Identifying all the extreme points of each of the RCHs, which are actually subsets of the candidates to be extreme points pointed out by Theorem 1.
2. Finding the subsets of the extreme points that contribute to the closest points, one for each set.
3. Determining the specific convex combination of each subset of the extreme points for each set, which gives each of the two closest points.

However, in the algorithm proposed herewith, it is not the extreme points themselves that are needed, but their inner products (projections onto a specific direction). This case can be significantly simplified, through the next theorem.

Lemma 2: Let $S = \{s_i \mid s_i \in \mathbb{R}, i = 1, \cdots, n\}$, $\lambda \ge 0$, $\mu > 0$ and $\lambda \ne \mu$, with $k\mu + \lambda = 1$. The minimum weighted sum on S (for k elements of S if $\lambda = 0$, or $k+1$ elements of S if $\lambda > 0$) is the expression $\lambda s_{i_1} + \mu\sum_{j=2}^{k+1} s_{i_j}$ if $0<\mu<\lambda$ or $\mu\sum_{j=1}^{k} s_{i_j} + \lambda s_{i_{k+1}}$ if $0<\lambda<\mu$ or $\mu\sum_{j=1}^{k} s_{i_j}$ if $\lambda = 0$, where $s_{i_p} \le s_{i_q}$ if $p<q$.

Proof: The proof of this Lemma is found in the Appendix.

Theorem 2: The minimum projection of the extreme points of a RCH

$$\mathrm{R}(X,\mu) = \left\{ w : w = \sum_{i=1}^{k} a_i x_i,\ x_i \in X,\ \sum_{i=1}^{k} a_i = 1,\ 0 \le a_i \le \mu \right\}$$

in the direction p (setting $\lambda = 1-\lfloor 1/\mu \rfloor \mu$ and $m = \lfloor 1/\mu \rfloor$) is:

- $\mu\sum_{j=1}^{m} s_{i_j}$ if $0<\mu$ and $\lambda = 0$
- $\mu\sum_{j=1}^{m} s_{i_j} + \lambda s_{i_{m+1}}$ if $0<\lambda<\mu$

where $s_{i_j} = (p \mid x_j)/\|p\|$ and s_i is an ordering, such that $s_{i_p} \le s_{i_q}$ if $p<q$.

Proof: The extreme points of $\mathrm{R}(C,\mu)$ are of the form $z_i = \sum_{j=1}^{k} \alpha_j x_j$, where $k = |X|$, $x_j \in X$, $\sum_{j=1}^{k} \alpha_j = 1$, $\alpha_j \in \{0,\ 1-\lfloor 1/\mu \rfloor \mu,\ \mu\}$. Therefore, taking into account that if $\lambda > 0$, it is always $\lambda < \mu$, as it follows from the Corollary of Proposition 3, the projection of an extreme point has the form:

$$(p \mid z_i)/\|p\| = \left(p \mid \sum_{j=1}^{k} \alpha_j x_j\right) = \sum_{j=1}^{k} \alpha_j (p \mid x_j) = \sum_{j=1}^{k} \alpha_j s_{t_j}$$

and, according to the above Lemma 2, proves the theorem. □

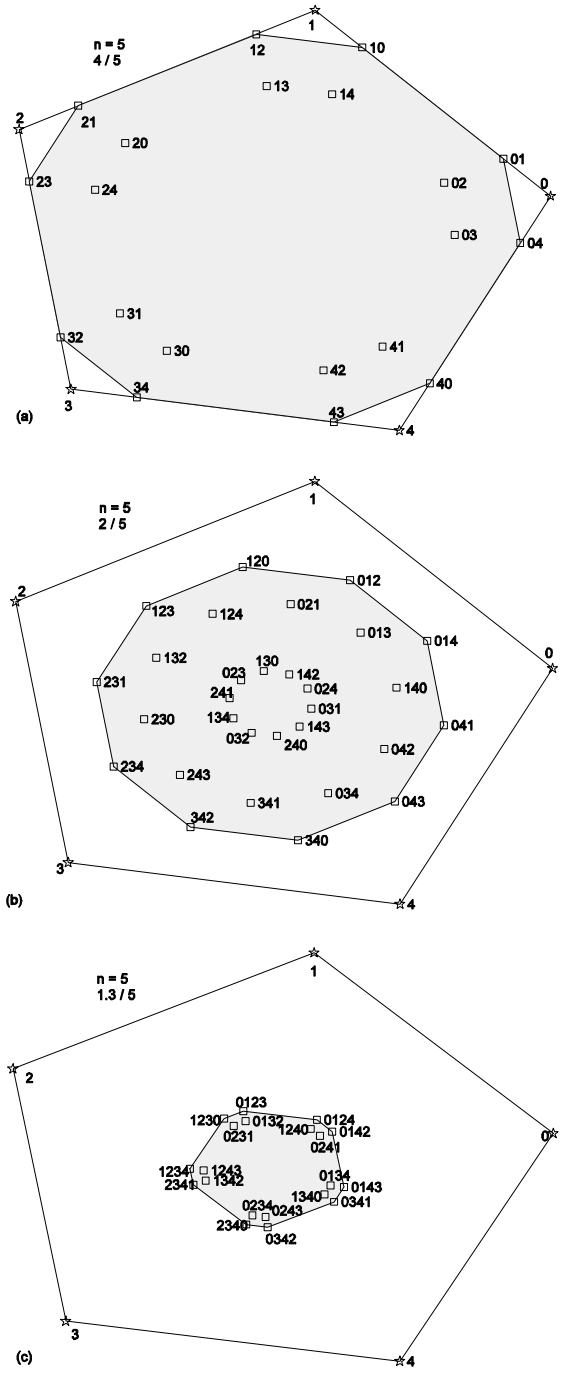

Fig. 5. Three RCHs ((a) $\mathrm{R}(\mathrm{P}5,4/5)$ [2], (b) $\mathrm{R}(\mathrm{P}5,2/5)$ and (c) $\mathrm{R}(\mathrm{P}5,1.3/5)$) are shown, generated by 5 points (stars), to present the points that are candidates to be extreme, marked by small squares. Each candidate to be extreme point in the RCH is labeled so as to present the original points from which it has been constructed, i.e., point (01) results from points (0) and (1); the last label is the one with the smallest coefficient.

Remark: In other words, the above Theorem states that the calculation of the minimum projection of a RCH onto a specific direction does not need the direct formation of all the possible extreme points of RCH, but only the calculation of the projections of the n original points and then the summation of the first least $\lceil 1/\mu \rceil$ of them, each multiplied with the corresponding coefficient imposed by Theorem 2. This is illustrated in Fig. 6.

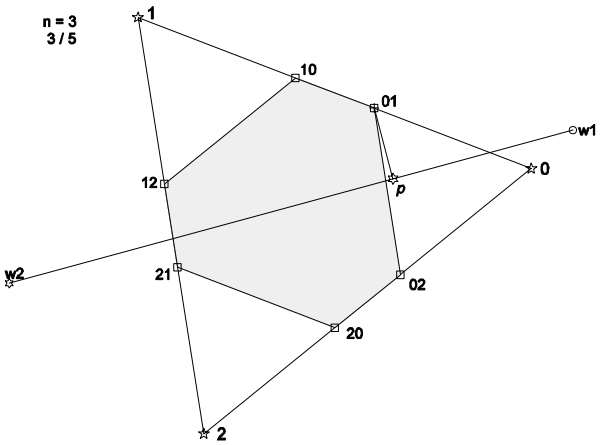

Fig. 6. The minimum projection p of the RCH $\mathrm{R}(\mathrm{P}3,3/5)$, generated by 3 points and having $\mu = 3/5$, onto the direction $w_2 - w_1$ belongs to the point (01), which is calculated, according to Theorem 2, as the ordered weighted sum of the projection of only $\lceil 5/3 \rceil = 2$ points ((0) and (1)) of the 3 initial points. The magnitude of the projection, in lengths of $\|w_2 - w_1\|$ is $(3/5)(x_0 \mid w_2 - w_1) + (2/5)(x_1 \mid w_2 - w_1)$.

Summarizing, the computation of the minimum projection of a RCH onto a given direction, entails the following steps:

1. Compute the projections of all the points of the original set.
2. Sort the projections in ascending order.
3. Select the first (smaller) $\lceil 1/\mu \rceil$ projections.
4. Compute the weighted average of these projections, with weights suggested in Theorem 2.

Proposition 4: A linearly non-separable SVM problem can be transformed to a linearly separable one through the use of RCHs (by a suitable selection of the reduction factor μ for each class) if and only if the centroids of the classes do not coincide.

Proof: It is a direct consequence of Proposition 2, found in [14]. □

IV. Geometric Algorithm for SVM Separable and Non-Separable Tasks.

As it has already been pointed out, an iterative, geometric algorithm for solving the linearly separable SVM problem has been presented recently in [16]. This algorithm, initially proposed by Kozinec for finding a separating hyperplane and

[2] Pn stands for a (convex) Polygon of n vertices.

improved by Schlesinger for finding an ε-optimal separating hyperplane, can be described by the following three steps (found and explained in [16], reproduced here for completeness):

1. *Initialization*: Set the vector w_1 to any vector $x \in X_1$ and w_2 to any vector $x \in X_2$.
2. *Stopping condition*: Find the vector x_t closest to the hyperplane as $x_t = \arg\min_{i \in I_1 \cup I_2} m(x_i)$ where
$$m(x_i) = \begin{cases} \dfrac{\langle x_i - w_2, w_1 - w_2 \rangle}{\|w_1 - w_2\|}, & \text{for } i \in I_1 \\ \dfrac{\langle x_i - w_1, w_2 - w_1 \rangle}{\|w_1 - w_2\|}, & \text{for } i \in I_2 \end{cases}. \quad (19)$$
If the ε-optimality condition $\|w_1 - w_2\| - m(x_t) < \varepsilon$ holds, then the vector $w = w_1 - w_2$ and $b = 1/2\left(\|w_1\|^2 - \|w_2\|^2\right)$ defines the ε-solution; otherwise go to step 3.
3. *Adaptation*: If $x_t \in X_1$ set $w_2^{new} = w_2$ and compute $w_1^{new} = (1-q)w_1 + qx_t$, where $q = \min\left(1, \dfrac{\langle w_1 - w_2, w_1 - x_t \rangle}{\|w_1 - x_t\|^2}\right)$; otherwise, set $w_1^{new} = w_1$ and compute $w_2^{new} = (1-q)w_2 + qx_t$, where $q = \min\left(1, \dfrac{\langle w_2 - w_1, w_2 - x_t \rangle}{\|w_2 - x_t\|^2}\right)$. Continue with step 2.

The above algorithm, which is shown in work schematically in Fig. 7, is easily adapted to be expressed through the kernel function of the input space patterns, since the vectors of the feature space are present in the calculations only through norms and inner products. Besides, a caching scheme can be applied with only $O(|I_1| + |I_2|)$ storage requirements.

The adaptation of the above algorithm is easy, with the mathematical toolbox for RCHs presented above and after making the following observations:

1. w_1 and w_2 should be initialized in such a way that it is certain they belong to the RCHs of X_1 and X_2 respectively. An easy solution is to use the centroid of each class as such. The algorithm secures that w_1 and w_2 evolve in such a way that they are always in their respective RCHs and converge to the nearest points.
2. Instead of the initial points (i.e. $x \in X_1 \cup X_2$), all the candidates to be extreme points of the RCH have to be examined. However, actually what matters is not the absolute position of each extreme point but their projection onto $w_1 - w_2$ or to $w_2 - w_1$, if the points to be examined belong to the RCHs of X_1 and X_2 respectively.
3. The minimum projection belongs to the point which is formed according to Theorem 2.

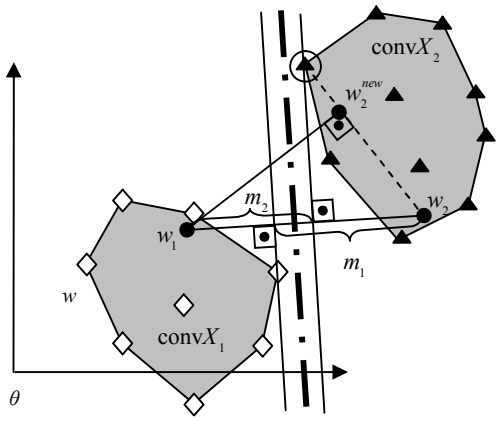

Fig. 7. The quantities involved in S-K algorithm, are shown here for simplicity for (not reduced) convex hulls: w_1 is the best (until current step) approximation to the closest point of convX_1 to convX_2; m_2 is the distance of w_1 from the closest projection of points of X_2 onto $(w_1 - w_2)$ in lengths of $(w_1 - w_2)$. The new w belongs to the set with the least m (e.g., in this case in convX_2) and it is the closest point of the line segment with one end the old w and the other end the point presenting the closest projection (m_2), which in the figure is circled; this new w is shown in the figure as w_2^{new}.

According to the above, and for the clarity of the adapted algorithm to be presented, it will be helpful that some definitions and calculations of the quantities involved to be provided beforehand.

At each step, the points w_1 and w_2, representing the closest points (up to that step) for each class respectively, are known through the coefficients a_i, i.e., $w_1 = \sum_{i \in I_1} a_i x_i$ and $w_2 = \sum_{i \in I_2} a_i x_i$. However, the calculations do not involve w_1 and w_2 directly, *but only through inner products*, which is also true for all points. This is expected, since the goal is to compare distances and calculate projections and not to examine absolute positions. This is the point where the "kernel trick" comes into the scene, allowing the transformation of the linear to a non-linear classifier.

The aim at each step is to find the point z_r, belonging to any of the RCHs of both classes, which minimizes the margin $m(z_r)$, defined (as in (19)) as:
$$m(z_r) = \begin{cases} \dfrac{\langle z_{1r} - w_2, w_1 - w_2 \rangle}{\|w_1 - w_2\|}, & z_{1r} \in \mathrm{R}(X_1, \mu_1) \\ \dfrac{\langle z_{2r} - w_1, w_2 - w_1 \rangle}{\|w_1 - w_2\|}, & z_{2r} \in \mathrm{R}(X_2, \mu_2) \end{cases}. \quad (20)$$

The quantity $m(z_r)$ is actually the distance, in lengths of $\|w_1 - w_2\|$, of one of the closest points (w_1 or w_2) from the closest projection of the RCH of the other class, onto the line defined by the points w_1 and w_2. This geometric interpretation

is clearly shown in Fig. 7. The intermediate calculations, required for (20), are given in the Appendix.

According to the above, the algorithm becomes:

1. *Initialization*:
   a. Set $\lambda_1 \equiv 1-\lfloor 1/\mu_1 \rfloor \mu_1$, $m_1 \equiv \lfloor 1/\mu_1 \rfloor$, $\lambda_2 \equiv 1-\lfloor 1/\mu_2 \rfloor \mu_2$, $m_2 \equiv \lfloor 1/\mu_2 \rfloor$ and secure that $\mu_1 \geq 1/|I_1|$ and $\mu_2 \geq 1/|I_2|$.
   b. Set the vectors w_1 and w_2 to be the centroids of the corresponding convex hulls, i.e., set $a_i = 1/|I_1|$, $i \in I_1$ and $a_i = 1/|I_2|$, $i \in I_2$.
2. *Stopping condition*: Find the vector $z_r = \begin{cases} z_{1r} = \sum_{i \in I_1} b_i x_i, & b_i \in \{0, \lambda_1, \mu_1\}, \ \sum_{i \in I_1} b_i = 1 \\ z_{2r} = \sum_{i \in I_2} b_i x_i, & b_i \in \{0, \lambda_2, \mu_2\}, \ \sum_{i \in I_2} b_i = 1 \end{cases}$ (actually the coefficients b_i) s.t. $z_r = \underset{z_{1r} \in R(X_1,\mu_1), z_{2r} \in R(X_2,\mu_2)}{\arg\min} \left(m(z_{1r}), m(z_{2r}) \right)$ where
$$m(z_r) = \begin{cases} \dfrac{\langle z_{1r} - w_2, w_1 - w_2 \rangle}{\|w_1 - w_2\|}, & z_{1r} \in \mathrm{R}(X_1, \mu_1) \\ \dfrac{\langle z_{2r} - w_1, w_2 - w_1 \rangle}{\|w_1 - w_2\|}, & z_{2r} \in \mathrm{R}(X_2, \mu_2) \end{cases}, \quad (21)$$
using (53) and (54).
   If the ε-optimality condition $\|w_1 - w_2\| - m(z_r) < \varepsilon$ (calculated after (44), (53) and (54)) holds, then the vector $w = w_1 - w_2$ and $c = 1/2\left(\|w_1\|^2 - \|w_2\|^2\right)$ defines the ε-solution; otherwise go to step 3.
3. *Adaptation*: If $z_r = z_{1r} \in R(X_1, \mu_1)$, set $w_2^{new} = w_2$ and compute $w_1^{new} = q_1 z_{1r} + (1-q_1) w_1$, where $q_1 = \min\left(1, \dfrac{\langle w_1 - w_2, w_1 - z_{1r} \rangle}{\|w_1 - z_{1r}\|^2}\right)$ and $\dfrac{\langle w_1 - w_2, w_1 - z_{1r} \rangle}{\|w_1 - z_{1r}\|^2} = \dfrac{A - \langle w_1, z_{1r} \rangle - C + \langle w_2, z_{1r} \rangle}{A + \langle z_{1r}, z_{1r} \rangle - 2\langle w_1, z_{1r} \rangle}$ (using (57)-(59)); hence $a_i^{new} = q_1 b_i + (1-q_1) a_i$, $i \in I_1$; otherwise, set $w_1^{new} = w_1$ and compute $w_2^{new} = q z_{2r} + (1-q) w_2$, where $q_2 = \min\left(1, \dfrac{\langle w_2 - w_1, w_2 - z_{2r} \rangle}{\|w_2 - z_{2r}\|^2}\right)$ and $\dfrac{\langle w_2 - w_1, w_2 - z_{2r} \rangle}{\|w_2 - z_{2r}\|^2} = \dfrac{B - \langle w_2, z_{2r} \rangle - C + \langle w_1, z_{2r} \rangle}{B + \langle z_{2r}, z_{2r} \rangle - 2\langle w_2, z_{2r} \rangle}$ (using (60)-(62)); hence $a_i^{new} = q_2 b_i + (1-q_2) a_i$, $i \in I_2$. Continue with step 2.

This algorithm (RCH-SK) has almost the same complexity as the Schlesinger–Kozinec (SK) one (the extra cost is the sort involved in each step to find the least $\lceil 1/\mu_1 \rceil$ and $\lceil 1/\mu_2 \rceil$ inner products, plus the cost to evaluate the inner product $\langle z_r, z_r \rangle$); the same caching scheme can be used, with only $O(|I_1| + |I_2|)$ storage requirements.

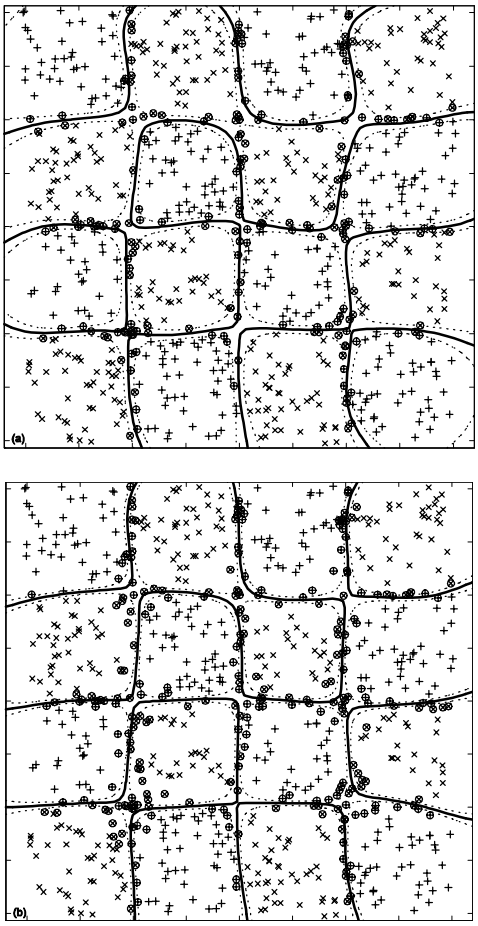

Fig. 8. Classification results for the checkerboard dataset for (a) SMO and (b) RCH-SK algorithms. Circled points are support vectors.

V. Results

In the sequel, some representative results of RCH-SK algorithm are included, concerning two known non-separable datasets, since the separable cases work in exactly the same way as the SK algorithm, proposed in [16]. Two datasets were chosen. One is an artificial dataset of a 2-dimensional checkerboard with 800 training points in 4×4 cells, similar to the dataset found in [25]. The reason that a 2-dimensional example was chosen is to make possible the graphical representation of the results. The second dataset is the Pima Indians Diabetes dataset, with 768 8-dimensional training patterns [26]. Each dataset was trained to achieve comparable success rates for both algorithms, the one presented here (RCH-SK) and the SMO algorithm presented in [11], using the

same model (kernel parameters). The results of both algorithms (total run time and number of kernel evaluations) were compared and summarized in Table I. An Intel Pentium M PC has been used for the tests.

1. *Checkerboard*: A set of 800 (Class A: 400, Class B: 400) randomly generated points on a 2-dimensional checkerboard of 4×4 cells was used. Each sample attribute ranged from -4 to 4 and the margin was -0.1 (the negative value indicating the overlapping between classes, i.e., the overlapping of the cells). A RBF kernel was used with $\sigma = 0.9$ and the success rate was estimated using 40-fold cross validation (40 randomly generated partitions of 20 samples each, the same for both algorithms). The classification results of both methods are shown in Fig. 8.
2. *Diabetes*: The 8-dimensional 768 samples dataset was used to train both classifiers. The model (RBF kernel with $\sigma = 20$), as well as the error rate estimation procedure (cross validation on 100 realizations of the samples) that was used for both algorithms, is found in [26]. Both classifiers (SMO and RCH-SK) closely approximated the success rate 76.47% (±1.73), reported in [26].

Method	Dataset	Time (sec)	Kernel evaluations	Success Rate (%)
SMO	Checkerboard	2276	84496594	88.17
RCH-SK		**618**	**28916632**	**90.13**
SMO	Diabetes	811	1945488	**76.66**
RCH-SK		**180**	**1902735**	76.20

Table I : Comparative results for the SMO algorithm [11] with the algorithm presented in this work (RCH-SK).

As it is apparent from Table I, substantial reductions with respect to run-time and kernel evaluations can be achieved using the new geometric algorithm (RCH-SK) proposed here. These results indicate that exploiting the theorems and propositions presented in this paper can lead to geometric algorithms that can be considered as viable alternatives to already known decomposition schemes.

VI. Conclusion

The SVM approach to machine learning is known to have both theoretical and practical advantages. Among these, are the sound mathematical foundation of SVM (supporting their generalization bounds and their guaranteed convergence to the global optimum unique solution), their overcoming of the "curse of dimensionality" (through the "kernel trick") and the intuition they display. The geometric intuition is intrinsic to the structure of SVM and has found application in solving both the separable and non-separable problem. The iterative geometric algorithm of Schlesinger and Kozinec, modified here to work for the non-separable task employing RCHs, resulted in a very promising method of solving SVM. The algorithm presented here does not use any heuristics and provides a clear understanding of the convergence process and the role of the parameters used. Furthermore, the penalty factor μ (which has clear meaning corresponding to the reduction factor of each convex hull) can be set different for each class, reflecting the importance of each class.

Appendix

● *Proof* of **Lemma 1**: In case that $\mu = 1$ the lemma is obviously true since $S = \{0,1\}$.

The other case will be proved by contradiction. So, let $0 < \mu < 1$ and $w \in \mathrm{R}(X, \mu)$ be a point of this RCH. Furthermore, suppose that w is a reduced convex combination with M the number of coefficients for which $a_i = \mu$, Λ the number of coefficients for which $a_i = 1 - \lfloor 1/\mu \rfloor \mu$ and r the position of the only coefficient of w such that $a_r \notin S$ (with

$$0 < a_r < \mu \tag{22}$$

by assumption).

Clearly, $\mathrm{M} + \Lambda + 1 \le k \Rightarrow \mathrm{M} + \Lambda < k$. Since $0 < \mu < 1$, it is $1/\mu > 1 \Rightarrow \lfloor 1/\mu \rfloor \ge 1$. Besides, it is

$$\lfloor 1/\mu \rfloor \le 1/\mu \le \lfloor 1/\mu \rfloor + 1. \tag{23}$$

From the first inequality of (23) it is $\lfloor 1/\mu \rfloor \mu \le 1 \Leftrightarrow 1 - \lfloor 1/\mu \rfloor \mu \ge 0$ and from the second inequality of (23) it is $1 - \lfloor 1/\mu \rfloor \mu < \mu$. These inequalities combined become:

$$0 \le 1 - \lfloor 1/\mu \rfloor \mu < \mu \tag{24}$$

According to the above and since $\sum_{i=1}^{k} a_i = 1$, it is:

$$\mathrm{M}\mu + \Lambda\left(1 - \lfloor 1/\mu \rfloor \mu\right) + a_r = 1. \tag{25}$$

Two distinct cases need to be examined: 1) $\lfloor 1/\mu \rfloor = 1/\mu$ and 2) $\lfloor 1/\mu \rfloor < 1/\mu$.

1) Let $\lfloor 1/\mu \rfloor = 1/\mu$ (26). Then $\lfloor 1/\mu \rfloor \mu = 1$ and $1 - \lfloor 1/\mu \rfloor \mu = 0$ (27). Substituting the above to (25), it becomes $\mathrm{M}\mu + a_r = 1$ and therefore $\mathrm{M}\mu = 1 - a_r$ (28). a) If $\mathrm{M} = \lfloor 1/\mu \rfloor$ then $\mathrm{M}\mu = \lfloor 1/\mu \rfloor \mu$, which, substituted into (28) and using (27), gives $a_r = 0$, which contradicts to the assumption that $0 < a_r$. b) If $\mathrm{M} > \lfloor 1/\mu \rfloor$ then $\mathrm{M}\mu > \lfloor 1/\mu \rfloor \mu$ and from (26) it gives $\mathrm{M}\mu > 1$ which is a contradiction. c) If $\mathrm{M} < \lfloor 1/\mu \rfloor$ then $(\mathrm{M}+1)\mu \le \lfloor 1/\mu \rfloor \mu$, or $(\mathrm{M}+1)\mu \le 1$ (29). But since $\mu > a_r \Rightarrow \mu - a_r > 0$, which through (28) gives $\mu - 1 + \mathrm{M}\mu > 0$ or $(\mathrm{M}+1)\mu > 1$, a contradiction to (29).
2) Let $\lfloor 1/\mu \rfloor < 1/\mu$ (30). Then $1 - \lfloor 1/\mu \rfloor \mu > 0$ (31) and $\lfloor 1/\mu \rfloor \mu < 1$ (32). The cases when $\Lambda = 1$ and $\Lambda > 1$ will be considered separately. a) Let $\Lambda = 1$ (33), which,

substituted to (25) gives $\mathrm{M}\mu+1-\lfloor 1/\mu\rfloor\mu+a_r=1 \Rightarrow a_r=\lfloor 1/\mu\rfloor\mu-\mathrm{M}\mu$ (34). i) Let $\mathrm{M}=\lfloor 1/\mu\rfloor$; consequently (34) gives by substitution $a_r=\lfloor 1/\mu\rfloor\mu-\lfloor 1/\mu\rfloor\mu=0$ which is a contradiction. ii) Let $\mathrm{M}>\lfloor 1/\mu\rfloor \Rightarrow \mathrm{M}\mu>\lfloor 1/\mu\rfloor\mu$; substituting this value in (34) gives $a_r<0$ which is a contradiction. iii) Let $\mathrm{M}<\lfloor 1/\mu\rfloor \Rightarrow (\mathrm{M}+1)\mu\le\lfloor 1/\mu\rfloor\mu$ and, using (34), gives $a_r\ge(\mathrm{M}+1)\mu-\mathrm{M}\mu \Rightarrow a_r\ge\mu$ which is a contradiction. b) Let $\Lambda>1$ (35). i) If $\mathrm{M}=\lfloor 1/\mu\rfloor$ then, setting $x\equiv 1-\lfloor 1/\mu\rfloor\mu$ and observing that $x>0$ from (31), (25) becomes $x(1-\Lambda)=a_r$ which is a contradiction, since the LHS is negative whereas the RHS is positive. ii) Similarly, if $\mathrm{M}>\lfloor 1/\mu\rfloor$ then $\mathrm{M}\mu>\lfloor 1/\mu\rfloor\mu$ (36). Setting $x\equiv 1-\lfloor 1/\mu\rfloor\mu$ and observing from (31) that $x>0$, (25) through (36) becomes $x(1-\Lambda)>a_r$ which is a contradiction, since the LHS is negative whereas the RHS is positive. iii) If $\mathrm{M}<\lfloor 1/\mu\rfloor$ then there exists a positive integer $\mathrm{K}\ge 1$ such that $\mathrm{M}+\mathrm{K}=\lfloor 1/\mu\rfloor \Rightarrow \mathrm{M}\mu+\mathrm{K}\mu=\lfloor 1/\mu\rfloor\mu$ (37). This relation, through (25), becomes $1-\Lambda(1-\lfloor 1/\mu\rfloor\mu)-a_r+\mathrm{K}\mu=\lfloor 1/\mu\rfloor\mu \Rightarrow a_r=\mathrm{K}\mu-(\Lambda-1)(1-\lfloor 1/\mu\rfloor\mu)$ (38). Substituting (38) into (25) gives $\mathrm{M}\mu+\Lambda(1-\lfloor 1/\mu\rfloor\mu)+\mathrm{K}\mu-(\Lambda-1)(1-\lfloor 1/\mu\rfloor\mu)=1 \Rightarrow (\mathrm{M}+\mathrm{K})\mu+1-\lfloor 1/\mu\rfloor\mu=1$ (39). This last relation states that, in this case, there is an alternative configuration to construct w (other than (25)), which does not contain the coefficient a_r but only coefficients belonging to the set S. This contradicts to the initial assumption that there exists an extreme point w in a RCH that is a reduced convex combination of points of X with all except one (a_r) coefficients belonging in S, since a_r is not necessary to construct w.

Therefore, the lemma has been proved. □

● *Proof* of **Lemma 2**: Let $0<\mu<\lambda$, $w_1=\lambda s_{i_1}+\mu\sum_{m=2}^{k+1}s_{i_m}$, where $s_{i_p}\le s_{i_q}$ if $p<q$ and $w_2=\lambda s_{j_1}+\mu\sum_{m=2}^{k+1}s_{j_m}$, where no ordering is imposed on the s_{j_m}. It is certain that $s_{i_1}\le s_{j_1}\Rightarrow\lambda s_{i_1}\le\lambda s_{j_1}$. $\sum_{m=2}^{k+1}s_{j_m}$ is minimum if the k additives are the k minimum elements of S. In such a case $i_1\in\{j_2,\cdots,j_{k+1}\}$ and, in the best case, $j_1\in\{i_2,\cdots,i_{k+1}\}$, particularly $j_1=i_{k+1}$ (since $j_m=i_l$ where $m\in\{1,\cdots,k\}$ and $l\in\{2,\cdots,k+1\}$).

Then $w_1-w_2=\lambda s_{i_1}+\mu\sum_{m=2}^{k+1}s_{i_m}-\lambda s_{j_1}-\mu\sum_{m=2}^{k+1}s_{j_m}$
$=\lambda s_{i_1}+\mu s_{i_{k+1}}-\lambda s_{i_{k+1}}-\mu s_{i_1}=\lambda(s_{i_1}-s_{i_{k+1}})-\mu(s_{i_1}-s_{i_{k+1}})$
$=(\lambda-\mu)(s_{i_1}-s_{i_{k+1}})\le 0$.

Each of the remaining cases (i.e., $\lambda=0$, or $0<\lambda<\mu$) is proved similarly as above. □

● *Calculation of the intermediate quantities involved in the algorithm*:

For the calculation of $\|w_1-w_2\|$, it is:

$$\|w_1-w_2\|=\sqrt{\langle w_1,w_1\rangle+\langle w_2,w_2\rangle-2\langle w_1,w_2\rangle}. \tag{40}$$

Setting:

$$\langle w_1,w_1\rangle=\left\langle\sum_{i\in I_1}a_ix_i,\sum_{j\in I_1}a_jx_j\right\rangle=\sum_{i\in I_1}\sum_{j\in I_1}a_ia_j\langle x_i,x_j\rangle\equiv A \tag{41}$$

$$\langle w_2,w_2\rangle=\left\langle\sum_{i\in I_2}a_ix_i,\sum_{j\in I_2}a_jx_j\right\rangle=\sum_{i\in I_2}\sum_{j\in I_2}a_ia_j\langle x_i,x_j\rangle\equiv B \tag{42}$$

and

$$\langle w_1,w_2\rangle=\left\langle\sum_{i\in I_1}a_ix_i,\sum_{j\in I_2}a_jx_j\right\rangle=\sum_{i\in I_1}\sum_{j\in I_2}a_ia_j\langle x_i,x_j\rangle\equiv C \tag{43}$$

it is:

$$\|w_1-w_2\|=\sqrt{A+B-2C}. \tag{44}$$

According to the above Proposition 3, any extreme point of the RCHs has the form:

$$z_r=\begin{cases}z_{1r}=\sum_{i\in I_1}b_ix_i,\\ \quad b_i\in\{0,\lambda_1,\mu_1\},\quad \sum_{i\in I_1}b_i=1\\ z_{2r}=\sum_{i\in I_2}b_ix_i,\\ \quad b_i\in\{0,\lambda_2,\mu_2\},\quad \sum_{i\in I_2}b_i=1\end{cases}. \tag{45}$$

The projection of z_r onto the direction w_1-w_2 is needed. According to Theorem 2, the minimum of this projection is formed as the weighted sum of the projections of the original points onto the direction w_1-w_2.

Specifically, the projection of z_{1r} onto w_1-w_2, where $z_{1r}\in\mathrm{R}(X_1,\mu_1)$ and $z_{1r}=\sum_{i\in I_1}b_ix_i$, $b_i\in\{0,\lambda_1,\mu_1\}$, $\sum_{i\in I_1}b_i=1$, is $p(z_{1r})=\langle z_{1r},w_1-w_2\rangle/\|w_1-w_2\|$, and by (44):

$$p(z_{1r})=\langle z_{1r},w_1-w_2\rangle/\left(\sqrt{A+B-2C}\right). \tag{46}$$

Since the quantity $\sqrt{A+B-2C}$ is constant at each step, for the calculation of $\min(p(z_{1r}))$, the ordered inner products of x_i, $x_i\in I_1$, with w_1-w_2 must be formed. From them, the

smallest $\lceil 1/\mu_1 \rceil$ numbers, each multiplied by the corresponding coefficient (as of Theorem 2), must be summed. Therefore, using $\lambda_1 \equiv 1-\lfloor 1/\mu_1 \rfloor \mu_1$, $m_1 \equiv \lfloor 1/\mu_1 \rfloor$, $\lambda_2 \equiv 1-\lfloor 1/\mu_2 \rfloor \mu_2$, $m_2 \equiv \lfloor 1/\mu_2 \rfloor$ it is:

$$\begin{aligned}\langle x_r, w_1 \rangle &= \left\langle x_r, \sum_{i\in I_1} a_i x_i \right\rangle \\ &= \sum_{i\in I_1} a_i \langle x_i, x_r \rangle \equiv D_r\end{aligned} \tag{47}$$

$$\begin{aligned}\langle x_r, w_2 \rangle &= \left\langle x_r, \sum_{i\in I_2} a_i x_i \right\rangle \\ &= \sum_{i\in I_2} a_i \langle x_i, x_r \rangle \equiv E_r\end{aligned} . \tag{48}$$

In the sequel, the numbers D_r must be ordered, for each set of indices, separately:

$$\tilde{D}_r^1 \equiv \left\{ D_{r_1}^1, \cdots, D_{r_{m_1}}^1 \right\}, \quad r \in I_1, \quad i<j \Rightarrow D_{r_i}^1 \le D_{r_j}^1 \tag{49}$$

$$\tilde{D}_r^2 \equiv \left\{ D_{r_1}^2, \cdots, D_{r_{m_2}}^2 \right\}, \quad r \in I_2, \quad i<j \Rightarrow D_{r_i}^2 \le D_{r_j}^2 \tag{50}$$

and

$$\tilde{E}_r^1 \equiv \left\{ E_{r_1}^1, \cdots, E_{r_{m_1}}^1 \right\}, \quad r \in I_1, \quad i<j \Rightarrow E_{r_i}^1 \le E_{r_j}^1 \tag{51}$$

$$\tilde{E}_r^2 \equiv \left\{ E_{r_1}^2, \cdots, E_{r_{m_2}}^2 \right\}, \quad r \in I_2, \quad i<j \Rightarrow E_{r_i}^2 \le E_{r_j}^2 . \tag{52}$$

With the above definitions ((47) - (52)) and applying Theorem 2, it is

$$\begin{aligned}&\min\left(\langle z_{1r}, w_1 - w_2 \rangle\right) \\ &= \mu_1 \left(\sum_{j=1}^{m_1} \left(\tilde{D}_{r_j}^1 - \tilde{E}_{r_j}^1 \right) \right) + \lambda_1 \left(\tilde{D}_{r_{m_1+1}}^1 - \tilde{E}_{r_{m_1+1}}^1 \right)\end{aligned}$$ and

consequently (using definitions (20), (40)-(43) and (47)-(52)),

$$\min\left(m\left(z_{1r}\right)\right) = \frac{\mu_1 \sum_{j=1}^{m_1} \tilde{D}_{r_j}^1 + \lambda_1 \tilde{D}_{r_{m_1+1}}^1 - \mu_1 \sum_{j=1}^{m_1} \tilde{E}_{r_j}^1 - \lambda_1 \tilde{E}_{r_{m_1+1}}^1 - C + B}{\sqrt{A+B-2C}} , \tag{53}$$

and, respectively:

$$\min\left(m\left(z_{2r}\right)\right) = \frac{\mu_2 \sum_{j=1}^{m_2} \tilde{E}_{r_j}^2 + \lambda_2 \tilde{E}_{r_{m_2+1}}^2 - \mu_2 \sum_{j=1}^{m_2} \tilde{D}_{r_j}^2 - \lambda_2 \tilde{D}_{r_{m_2+1}}^2 - C + A}{\sqrt{A+B-2C}} . \tag{54}$$

Finally, for the adaptation phase, the scalar quantities

$$\frac{\langle w_1 - w_2, w_1 - z_{1r} \rangle}{\|w_1 - z_{1r}\|^2} = \frac{A - \langle w_1, z_{1r} \rangle - C + \langle w_2, z_{1r} \rangle}{A + \langle z_{1r}, z_{1r} \rangle - 2\langle w_1, z_{1r} \rangle} \tag{55}$$

and

$$\frac{\langle w_2 - w_1, w_2 - z_{2r} \rangle}{\|w_2 - z_{2r}\|^2} = \frac{B - \langle w_2, z_{2r} \rangle - C + \langle w_1, z_{2r} \rangle}{B + \langle z_{2r}, z_{2r} \rangle - 2\langle w_2, z_{2r} \rangle} \tag{56}$$

are needed in the calculation of q. Therefore, the inner products $\langle w_1, z_{1r} \rangle$, $\langle w_2, z_{1r} \rangle$, $\langle z_{1r}, z_{1r} \rangle$, $\langle w_1, z_{2r} \rangle$, $\langle w_2, z_{2r} \rangle$ and $\langle z_{2r}, z_{2r} \rangle$ need to be calculated. The result is:

$$\langle w_1, z_{1r} \rangle = \mu_1 \sum_{j=1}^{m_1} \tilde{D}_{r_j}^1 + \lambda_1 \tilde{D}_{r_{m_1+1}}^1 \tag{57}$$

$$\langle w_2, z_{1r} \rangle = \mu_1 \sum_{j=1}^{m_1} \tilde{E}_{r_j}^1 + \lambda_1 \tilde{E}_{r_{m_1+1}}^1 \tag{58}$$

$$\begin{aligned}\langle z_{1r}, z_{1r} \rangle &= (\mu_1)^2 \sum_{j=1}^{m_1} \sum_{i=1}^{m_1} \langle x_{r_j}, x_{r_i} \rangle \\ &+ (\lambda_1)^2 \langle x_{r_{m_1+1}}, x_{r_{m_1+1}} \rangle + 2\lambda_1 \mu_1 \sum_{j=1}^{m_1} \langle x_{r_{m_1+1}}, x_{r_j} \rangle\end{aligned} \tag{59}$$

$$\langle w_1, z_{2r} \rangle = \mu_2 \sum_{j=1}^{m_2} \tilde{D}_{r_j}^2 + \lambda_2 \tilde{D}_{r_{m_2+1}}^2 \tag{60}$$

$$\langle w_2, z_{2r} \rangle = \mu_2 \sum_{j=1}^{m_2} \tilde{E}_{r_j}^2 + \lambda_2 \tilde{E}_{r_{m_2+1}}^2 \tag{61}$$

$$\begin{aligned}\langle z_{2r}, z_{2r} \rangle &= (\mu_2)^2 \sum_{j=1}^{m_2} \sum_{i=1}^{m_2} \langle x_{r_j}, x_{r_i} \rangle \\ &+ (\lambda_2)^2 \langle x_{r_{m_2+1}}, x_{r_{m_2+1}} \rangle + 2\lambda_2 \mu_2 \sum_{j=1}^{m_2} \langle x_{r_{m_2+1}}, x_{r_j} \rangle\end{aligned} \tag{62}$$

REFERENCES

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

[2] S. Theodoridis, K. Koutroumbas, *Pattern Recognition, 2nd edition*, Academic Press, 2003.

[3] C.Cortes, V. N. Vapnik "Support vector networks", Machine Learning, vol. 20, 1995, pp. 273–297.

[4] I. El-Naqa, Y. Yang, M. Wernik, N. Galatsanos, R. Nishikawa "A Support Vector Machine Approach for Detection of Microcalsifications", *IEEE Transactions on Medical Imaging*, vol. 21, No 12, 2002, pp. 1552–1563.

[5] T. Joachims "Text categorization with support vector machines: learning with many relevant features", *Proceedings of the European Conference on Machine Learning* (ECML), Chemnitz, Germany, 1998.

[6] E. Osuna, R. Freund, F. Girosi "Training support vector machines: An application to face detection", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR'97, Puerto Rico, 1997, pp. 130–136.

[7] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T. S. Furey, M. Ares Jr., D. Haussler "Knowledge-based analysis of microarray gene expression data by using support vector machines", *Proc. Nat. Acad. Sci. 97*, 2000, pp. 262–267.

[8] A. Navia-Vasquez, F. Perez-Cuz, A. Artes-Rodriguez "Weighted least squares training of support vector classifiers leading to compact and adaptive schemes", *IEEE Transactions on Neural Networks*, vol. 12(5), 2001, pp. 1047–1059.

[9] D. J. Sebald, J. A. Buklew "Support vector machine techniques for nonlinear equalization", *IEEE Transactions on Signal Processing*, vol. 48 (11), 2000, pp. 3217–3227.

[10] D. Zhou, B. Xiao, H. Zhou, R. Dai "Global Geometry of SVM Classifiers", Technical Report in AI Lab, Institute of Automation, Chinese Academy of Sciences. Submitted to NIPS 2002.

[11] J. Platt "Fast training of support vector machines using sequential minimal optimization" in B. Schölkopf, C. Burges and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pp. 185–208. MIT Press, 1999.

[12] K. P. Bennett, E. J. Bredensteiner "Geometry in Learning" in C. Gorini, E. Hart, W. Meyer and T. Phillips, editors, *Geometry at Work*, Mathematical Association of America, 1998.

[13] K. P. Bennett, E. J. Bredensteiner "Duality and Geometry in SVM classifiers" in Pat Langley, editor, *Proc. 17th International Conference on Machine Learning*, Morgan Kaufmann, 2000, pp. 57–64.

[14] D. J. Crisp, C. J. C. Burges "A geometric interpretation of ν-SVM classifiers" NIPS 12, 2000, pp. 244–250.

[15] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy "A fast iterative nearest point algorithm for support vector machine classifier design", Technical Report No. TR-ISL-99-03, Department of CSA, IISc, Bangalore, India, 1999.

[16] V. Franc, V. Hlaváč "An iterative algorithm learning the maximal margin classifier", *Pattern Recognition* 36, 2003, pp. 1985–1996.

[17] T. T. Friess, R. Harisson "Support vector neural networks: the kernel adatron with bias and soft margin" Technical Report ACSE-TR-752, University of Sheffield, Department of ACSE, 1998.
[18] B. Schölkopf, A. Smola, *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press, 2002.
[19] V. N. Vapnik, *Statistical Learning Theory,* John Wiley & Sons, Inc., 1998.
[20] David G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, Inc., 1969.
[21] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.
[22] S. G. Nash, A. Sofer, *Linear and Nonlinear Programming*, McGraw-Hill, 1994.
[23] C. J. C. Burges "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery* 2, pp. 121–167, 1998.
[24] J.-B. Hiriart-Urruty, C. Lemaréchal, *Convex Analysis and Minimization Algorithms I*, Springer-Verlag, 1991.
[25] L. Kaufman, "Solving the quadratic programming problem arising in Support Vector Classification", in B. Schölkopf, C. Burges and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pp. 147 – 167, MIT Press, 1999.
[26] G. Rätsch, T. Onoda, K. – R. Müller, "Soft Margins for AdaBoost", Machine Learning, vol. 42 (3), Kluwer Academic Publishers, 2000, p.p. 287 – 320.