Geometric Approach to Statistical Learning Theory through Support Vector Machines (SVM) with Application to Medical Diagnosis

Michael E. Mavroforakis*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
mmavrof@di.uoa.gr

Abstract. This dissertation deals with problems of Pattern Recognition in the framework of Machine Learning (ML) and, specifically, Statistical Learning Theory (SLT), using Support Vector Machines (SVMs). The focus of this work is on the geometric interpretation of SVMs, which is accomplished through the notion of Reduced Convex Hulls (RCHs), and its impact on the derivation of new, efficient algorithms for the solution of the general, i.e., linear, nonlinear, separable and non-separable, SVM optimization task. The contributions of this work are i) the extension of the mathematical framework of RCHs (which allow the restriction of the expression of the extreme points of the RCHs and provide an analytic form of their projection onto a specific direction), ii) the development of novel geometric algorithms for SVMs (based on Schlesinger-Kozinec and Gilbert nearest point algorithms), which were tested using public benchmark datasets and outperformed the existing algebraic SVM algorithms and, finally, iii) the derivation and assessment of a set of qualitative and quantitative mammographic textural and morphological features (using methods of statistical and fractal analysis) and the application of the SVM algorithms (as well as other machine learning paradigms) to the field of Medical Image Analysis and Diagnosis (Mammography) with very encouraging practical results.

Keywords: Classifier, Support vector machine, Geometric algorithm, Reproducing kernel Hilbert space, Reduced convex hull, Mammography, Image processing, Fractal analysis

1 Introduction

The contribution[1] of this dissertation is twofold: i) the extension of the geometric framework of the Support Vector Machine (SVM) paradigm, which is a fundamental derivative of the Statistical Learning Theory (SLT) and is used to

* Dissertation Advisor: Sergios Theodoridis, Professor

[1] Published parts of this work have been awarded with the following international distinctions:

accomplish a wide range of Machine Learning tasks, and the development of efficient and theoretically sound algorithms to practically solve the general SVM problem and ii) the derivation and assessment of a set of qualitative and quantitative mammographic textural and morphological features (using methods of statistical and fractal analysis) and the application of the SVM algorithms (as well as other machine learning paradigms) to the field of Medical Image Analysis and Diagnosis (Mammography).

Geometry provides a very intuitive background for the understanding and the solution of many problems in the fields of Pattern Recognition and Machine Learning. The SVM paradigm in pattern recognition presents a lot of advantages over other approaches (e.g., [4,21]), some of which are: 1) the uniqueness of the solution (as it is guaranteed to be the global minimum of the corresponding optimization problem), 2) good generalization properties of the solution, 3) rigid theoretical foundation based on SLT and optimization theory, 4) common 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 of the classification problem. Due to these very attractive properties, SVM have been successfully used in a number of applications. Although some authors have presented the theoretical background of the geometric properties of SVM, exposed thoroughly in [23], the main stream of solving methods comes from the algebraic field (mainly decomposition). One of the most popular algebraic algorithms, combining speed and ease of implementation with very good scalability properties, is the Sequential Minimal Optimization (SMO) [19]. The geometric properties of learning [1] and specifically of SVM 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 [2] and also for the non-separable case through the notion of the Reduced Convex Hull (RCH) [3]. Actually, the geometric algorithms presented until the work of this thesis ([11,5]) are suitable only for solving directly the separable case and indirectly the non-separable case through the technique proposed in [6]. However, the latter incorporates not linear, but quadratic penalty factors and it has been reported to lead to poor results in practical cases [11].

The main contribution of this work is the development of a complete mathematical framework 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-

---

– **Outstanding Paper Award of the IEEE Transactions on Neural Networks for the year 2008** (IEEE Computational Intelligence Society (CIS) Awards Committee).

– **1st prize of the student paper competition of European Signal Processing Conference (EUSIPCO) 2005**.

separable case through RCH is practically impossible, since it leads to a problem of combinatorial complexity. Subsequently, two known and well studied geometric algorithms, namely Schlesinger-Kozinec's and Gilbert's algorithms, have been rewritten in the context of this framework, therefore showing the practical benefits of the theoretical results derived to support the RCH notion.

2 Geometric Support Vector Machines

2.1 Support Vector Machines

A SVM finds the best separating (*maximal margin*) hyperplane between two classes of training samples in the feature space, which is in line with optimizing bounds concerning the generalization error [22,20]. The playground for SVM is the *feature space* $\mathcal{H}$, which is a *Reproducing Kernel Hilbert Space* (RKHS), where the mapped patterns reside ($\Phi : \mathcal{X} \to \mathcal{H}$). It is not necessary to know the mapping Φ itself analytically, but only its kernel, i.e., the value of the inner products of the mappings of all the samples ($K(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$ for all $x_1, x_2 \in \mathcal{X}$) [20]. Through the "kernel trick", it is possible to transform a nonlinear classification problem to a linear one, but in a higher (maybe infinite) dimensional space $\mathcal{H}$[2]. Once the patterns are mapped in the feature space, provided that the problem for the given model (kernel) is separable, the target of the classification task is to find the maximal margin hyperplane. This classification task, expressed in its dual form, is equivalent with finding the closest points between the convex hulls generated by the (mapped) patterns of each class in the feature space [2], i.e., it is a Nearest Point Problem (NPP). Finally, in case the classification task deals with non-separable datasets, i.e., the convex hulls of the (mapped) patterns in the feature space are overlapping, the problem is still solvable, provided that the corresponding hulls are reduced, so that to become non-overlapping [3,17]. This is illustrated in Figure 1. Therefore, the need to resort to the notion of the reduced convex hulls becomes apparent. However, in order to work in this RCH geometric framework, one has to extend the available palette of tools by a set of new RCH-related mathematical properties.

2.2 Reduced Convex Hulls (RCHs)

Definition 1. *(*Reduced Convex Hull*): The set of all convex combinations of points in some set C (of cardinality $|C|$), with the additional constraint that each coefficient α_i is upper-bounded by a non-negative number $\mu < 1$ is called the* reduced convex hull (RCH*) of C and it is denoted as $R(C, \mu)$:*

$$R(C,\mu) \doteq \left\{ w | w = \sum_{i=1}^{|C|} \alpha_i x_i,\ x_i \in C,\ \sum_{i=1}^{|C|} \alpha_i = 1,\ 0 \leq \alpha_i \leq \mu \right\}. \qquad (1)$$

[2] In the rest of this work, for keeping the notation clearer and simpler, the quantities x will be used instead of $\Phi(x)$, since in the final results, the patterns enter only through inner products and not individually, therefore making the use of kernels readily applicable.

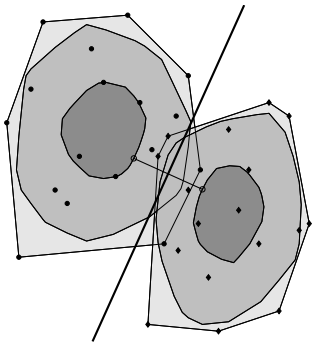


Fig. 1. The initial convex hulls (light gray), generated by the two training datasets (of disks and diamonds respectively) are overlapping; still overlapping are the RCHs with $\mu = 0.4$ (darker gray); however, the RCHs with $\mu = 0.1$ (darkest gray) are disjoint and, hence, separable. The nearest points of the RCHs, found by the Nearest Point Algorithms (NPAs) presented here, are shown as circles and the separating hyperplane as the bold line.

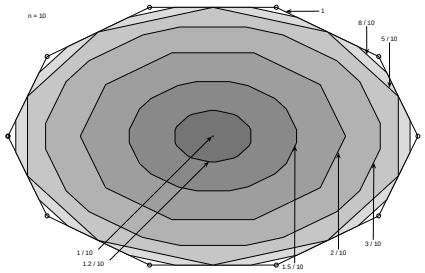


Fig. 2. Evolution of a reduced convex hull with respect to μ

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. The corresponding μ of each RCH are the values indicated by the arrows.

Every combination of the above form, i.e., of points belonging in $R(C, \mu)$, is called a reduced convex combination.

The effect of the upper-bound parameter μ to the size of RCH is very intuitive and is presented in Figure 2.

In this way, 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 [3] and [2] and the geometric one in [23]. The bound μ, for a given set of original points, plays the role of a reduction factor, since it controls the size of the generated RCH; the effect of the value of bound μ to the size of the RCH is shown in Figures 1 and 2.

Although, at a first glance, this is a nice result, that paves the way to a geometric solution, i.e., finding the nearest points between the RCH, it turns out not to be such a straightforward task: The search for nearest points between the two (one for each class) convex hulls depends directly on their extreme points [10], which, for the separable case are some of the points in the original dataset. However, in the non-separable case, each extreme point of the RCH turns out to be a reduced convex combination of the original points. This fact makes the direct application of a geometric NPA impractical, since an intermediate step of combinatorial complexity has been introduced.

In the sequel, we present a mathematical framework of theorems and propositions (which has been development as a result of this thesis) that shed further intuition and usefulness to the RCH notion and at the same time form the basis for the development of the novel geometric SVM algorithms which we have developed.

Proposition 1. *If all the coefficients α_i of all the reduced convex combinations, forming the RCH $R(\boldsymbol{X}, \mu)$ of a set $\boldsymbol{X}$ with k elements, are less than $1/k$ (i.e., $\mu < 1/k$), then $R(\boldsymbol{X}, \mu) = \emptyset$. [18]*

Proposition 2. *If for every i , it is $\alpha_i = 1/k$ in a RCH $R(\boldsymbol{X}, \mu)$ of a set $\boldsymbol{X}$ with k different points as elements, then $R(\boldsymbol{X}, \mu)$ degenerates to a set of one single point, the* centroid *point (or* barycenter*) of $\boldsymbol{X}$. [18]*

Remark 1. It is clear that in a RCH $R(C, \mu)$, a choice of $\mu > 1$ is equivalent with $\mu = 1$, as the upper bound for all α_i, because, from the Definition 1, it must be $\sum_{i=1}^{k} \alpha_i = 1$ and, therefore, $\alpha_i \leq 1$. As a consequence of this and the above Proposition 2, it is deduced that the RCH $R(C, \mu)$ of a set C will be either empty (if $\mu < 1/k$), or grows from the centroid ($\mu = 1/k$), to the convex hull ($\mu \geq 1$) of C.

Proposition 3. *The set $-R(C, \mu)$ is still a RCH; actually, it is $R(-C, \mu)$. [16]*

Proposition 4. *Scaling is a RCH-preserving property, i.e., for any $s \in \mathbb{R} \setminus \{0\}$, it is $sR(C, \mu) = R(sC, \mu)$. [16]*

Proposition 5. *The Minkowski sum (or difference) of two RCH $R(C_1, \mu_1) - R(C_2, \mu_2)$ is a convex set. [16]*

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 a convex hull generated by a set of points, as stated before, the set of extreme points consists of a subset of the set of points, which, it turns out to be the minimal representation of the convex hull. *Therefore, as it clear (e.g., [10]), only a subset of the original points is needed to be examined and not every point of the convex hull.* In contrast, for the case of RCH, its extreme points are the result of (reduced convex) 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 efficient algorithms presented in this dissertation.

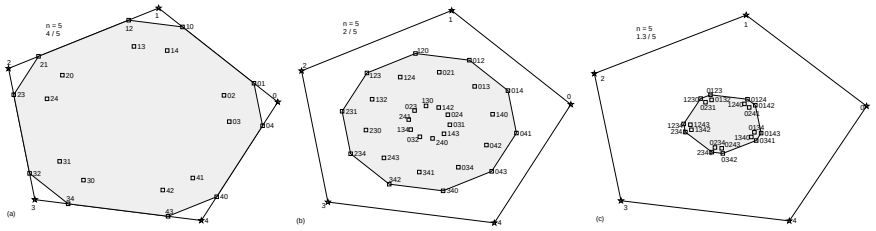
Lemma 1. *For any point $\mathbf{w} \in R(\boldsymbol{X}, \mu)$, if there exists a reduced convex combination $\sum_{i=1}^{k} \alpha_i \mathbf{x}_i$, with $\mathbf{x}_i \in \boldsymbol{X}$, $k = |\boldsymbol{X}|$, $\sum_{i=1}^{k} \alpha_i = 1$, $0 \leq \alpha_i \leq \mu$ and at least one coefficient α_r, $1 \leq r \leq k$, not belonging in the set $S = \{0,\ 1 - \lfloor 1/\mu \rfloor \mu,\ \mu\}$, then there exists at least another coefficient α_s, $1 \leq s \leq k$, $s \neq r$, also not belonging in the set S, i.e., there cannot be a reduced convex combination with just one coefficient not belonging in S. [18]*

Theorem 1. *The extreme points of a RCH* $R(\boldsymbol{X}, \mu)$ *of a set* $\boldsymbol{X}$*, with* $\mathbf{x}_i \in \boldsymbol{X}$ *and* $k = |\boldsymbol{X}|$*, have coefficients* α_i *belonging to the set* $S = \{0,\ 1 - \lfloor 1/\mu \rfloor \mu,\ \mu\}$*. [18]*

Proposition 6. *Each of the* extreme points *of a RCH* $R(\boldsymbol{X}, \mu)$ *of a set* $\boldsymbol{X}$*, with* $\mathbf{x}_i \in \boldsymbol{X}$ *and* $k = |\boldsymbol{X}|$*, is a reduced convex combination of* $m = \lceil 1/\mu \rceil$ *(distinct) points of the original set* $\boldsymbol{X}$*. Furthermore, if* $\lceil 1/\mu \rceil = 1/\mu$ *then all* $\alpha_i = \mu$*; otherwise,* $\alpha_i = \mu$ *for* $i = 1, 2, \cdots, m-1$ *and* $\alpha_m = 1 - \lfloor 1/\mu \rfloor \mu$*.[18]*

Remark 2. For the coefficients $\lambda \doteq 1 - \lfloor 1/\mu \rfloor \mu$ and μ, it holds $0 \leq \lambda < \mu$. This is a byproduct of the proof of the above Proposition 6 [18].

Remark 3. 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 1, 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* $\lceil 1/\mu \rceil$ (Figure 3)[3].



(a) $R(P5, 4/5)$ (b) $R(P5, 2/5)$ (c) $R(P5, 1.3/5)$

Fig. 3. Extreme points of an evolving (shrinking) RCH
Three RCHs ((a) $R(P5, 4/5)$, (b) $R(P5, 2/5)$ and (c) $R(P5, 1.3/5)$) are shown, generated by 5 points (stars), to present the points that are candidates to be extreme (small squares). Each point in the RCH is labeled so, as to present the original points from which it has been constructed; the last label is the one with the lowest coefficient.

Remark 4. Although the above Theorem 1, along with Proposition 6, restricts 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 a combinatorial nature, since each extreme point is a combination of $\lceil 1/\mu \rceil$ out of k initial points for each class. This is shown in Figure 3. 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

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

"candidate to be extreme points", shown in Figure 3. 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 algorithms 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* $S = \{s_i | s_i \in \mathbb{R},\ i = 1, 2, \cdots, n\}$, $\lambda \geq 0$, $\mu > 0$ *and* $\lambda \neq \mu$, *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* $s_{i_p} \leq S_{i_q}$ *if* $p < q$. *[18]*

Theorem 2. *The minimum projection of the extreme points of a RCH* $R(\boldsymbol{X}, \mu)$ *of a set* $\boldsymbol{X}$, *with* $\mathbf{x}_i \in \boldsymbol{X}$ *and* $k = |\boldsymbol{X}|$, *in the direction* $\mathbf{p}$ *(setting* $\lambda \doteq 1 - \lfloor 1/\mu \rfloor \mu$ *and* $m \doteq \lfloor 1/\mu \rfloor$*) 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 $s_{i_j} \doteq \frac{\langle \mathbf{p}, \mathbf{x}_j \rangle}{\|\mathbf{p}\|}$ *and* s_i *is an ordering, such that* $s_{i_p} \leq S_{i_q}$ *if* $p < q$. *[18]*

Remark 5. In other words, the above Theorem 2 states that the calculation of the minimum projection of a RCH onto a specific direction does not depend on the knowledge of all the possible extreme points of RCH, but only on the projections of the k original points and then a subsequent 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 Figure 4.

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 sum of these projections, with weights suggested in Theorem 2.

Proposition 7. *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. [3]*

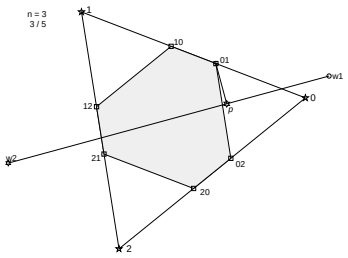


Fig. 4. Minimum projection of a RCH onto a given direction

The minimum projection $\boldsymbol{p}$ of the RCH $R\,(P3.{}^{3}/_{5})$, generated by 3 points and having $\mu = {}^{3}/_{5}$, onto the direction $\mathbf{w}_2 - \mathbf{w}_1$ 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 $\|\mathbf{w}_2 - \mathbf{w}_1\|$ is $({}^{3}/_{5})\,\langle \mathbf{x}_0, \mathbf{w}_2 - \mathbf{w}_1 \rangle + ({}^{2}/_{5})\,\langle \mathbf{x}_1, \mathbf{w}_2 - \mathbf{w}_1 \rangle$.

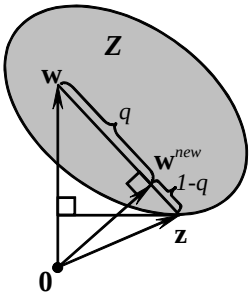


Fig. 5. Elements involved in Gilbert's algorithm

2.3 Geometric SVM Algorithms

The framework that we have developed and presented in the previous Subsection 2.2 around the notion of RCH, paves the way to adapt existing geometric algorithms, solving NPPs or the equivalent Minimum Norm Problems (MNPs), for the solution of the general (nonlinear, non-separable) SVM optimization task. This approach presents many clear advantages, compared to the algebraic approach used until now, as, the geometric algorithms are very intuitive in the way they work, have been extensively and rigorously examined regarding the convergence to the solution and do not rely on obscure and some times inefficient heuristics.

In the course of this dissertation, we have chosen to apply our theoretical results concerning the RCHs to adapt two of the well-known geometric NPP algorithms, namely a) Schlesinger-Kozinec's and b) Gilbert's algorithms, which will be presented in the sequel.

2.4 Schlesinger - Kozinec's Algorithm

An iterative, geometric algorithm for solving the linearly separable SVM problem has been presented recently in [5]. This algorithm, initially proposed by Kozinec for computing a separating hyperplane and improved by Schlesinger for searching for an ϵ-optimal separating hyperplane, can be described by the following three steps[4] (given and explained in [5]).

[4] assuming that the training classes consist of the sets $\boldsymbol{X}_+$ and $\boldsymbol{X}_-$ of I_+ and I_- elements respectively

Step 1 *Initialization*: Set the vector $\mathbf{w}_-$ to any vector $\mathbf{x} \in \boldsymbol{X}_-$ and $\mathbf{w}_+$ to any vector $\mathbf{x} \in \boldsymbol{X}_+$.

Step 2 *Stopping condition*: Find the vector $\mathbf{x}_t$ closest to the hyperplane as $\mathbf{x}_t = \arg\min_{i \in I_- \bigcup I_+} m(\mathbf{x}_i)$ where

$$m(\mathbf{x}_i) = \begin{cases} \frac{\langle \mathbf{x}_i - \mathbf{w}_+, \mathbf{w}_- - \mathbf{w}_+ \rangle}{\|\mathbf{w}_- - \mathbf{w}_+\|}, & \text{for } i \in I_- \\ \frac{\langle \mathbf{x}_i - \mathbf{w}_-, \mathbf{w}_+ - \mathbf{w}_- \rangle}{\|\mathbf{w}_- - \mathbf{w}_+\|}, & \text{for } i \in I_+ \end{cases}. \tag{2}$$

If the ϵ-optimality condition $\|\mathbf{w}_- - \mathbf{w}_+\| - m(\mathbf{x}_t) < \epsilon$ holds, then the vector $\mathbf{w} = \mathbf{w}_- - \mathbf{w}_+$ and $b = 1/2\left(\|\mathbf{w}_-\|^2 - \|\mathbf{w}_+\|^2\right)$ defines the ϵ-solution; otherwise, go to step 3.

Step 3 *Adaptation*: If $\mathbf{x}_t \in \boldsymbol{X}_-$ set $\mathbf{w}_+^{new} = \mathbf{w}_+$ and compute $\mathbf{w}_-^{new} = (1-q)\,\mathbf{w}_- + q\mathbf{x}_t$, where $q = \min\left\{1, \frac{\langle \mathbf{w}_- - \mathbf{w}_+, \mathbf{w}_- - \mathbf{x}_i \rangle}{\|\mathbf{w}_- - \mathbf{x}_t\|^2}\right\}$; otherwise, set $\mathbf{w}_-^{new} = \mathbf{w}_-$ and compute $\mathbf{w}_+^{new} = (1-q)\,\mathbf{w}_+ + q\mathbf{x}_t$, where $q = \min\left\{1, \frac{\langle \mathbf{w}_+ - \mathbf{w}_-, \mathbf{w}_+ - \mathbf{x}_i \rangle}{\|\mathbf{w}_+ - \mathbf{x}_t\|^2}\right\}$. Continue with step 2.

The modified Schlesinger-Kozinec's algorithm to solve the general SVM optimization problem is described in [18,17].

2.5 Gilbert's Algorithm

Another well known (well studied and applied) geometric algorithm, is the MNP algorithm proposed originally by Gilbert [9]. Although Gilbert's algorithm is a MNP algorithm, while the SVM optimization task corresponds to a NPP, the two formulations are equivalent, as it has already been proved, e.g., in [11]. Hence, the general (non-separable) SVM optimization task can be formulated as a MNP as follows: Find $\mathbf{z}^*$ such that $\mathbf{z}^* = \arg\min_{\mathbf{z} \in \boldsymbol{Z}}(\|\mathbf{z}\|)$, where $\boldsymbol{Z} = \{\mathbf{z} | \mathbf{z} = \mathbf{x}_+ - \mathbf{x}_-,\ \mathbf{x}_- \in R(\boldsymbol{X}_-, \mu),\ \mathbf{x}_+ \in R(\boldsymbol{X}_+, \mu)\}$.

Obviously, the restriction that the RCHs do not overlap means equivalently that $\|\mathbf{z}\| > 0$, $\forall \mathbf{z} \in \boldsymbol{Z}$, i.e., the null vector does not belong to $\boldsymbol{Z}$.

A brief description of the standard Gilbert's algorithm, (provided that $\boldsymbol{Z}$ is a convex set, of which we need to find the minimum norm member $\mathbf{z}^*$), is given below:

Step 1 Choose $\mathbf{w} \in \boldsymbol{Z}$.

Step 2 Find the point $\mathbf{z} \in \boldsymbol{Z}$ with the minimum projection onto the direction of $\mathbf{w}$. If $\|\mathbf{w}\| \cong \|\mathbf{z}\|$ then $\mathbf{z}^* \leftarrow \mathbf{w}$; stop.

Step 3 Find the point $\mathbf{w}^{new}$ of the line segment $[\mathbf{w}, \mathbf{z}]$, with minimum norm (closest to the origin). Set $\mathbf{w} \leftarrow \mathbf{w}^{new}$ and go to Step 2.

The idea behind the algorithm is very simple and intuitive and the elements involved in the above steps of the algorithm are illustrated in Figure 5. The modified Gilbert's algorithm to solve the general SVM optimization problem is described in [15,16].

Results The results of the new geometric algorithm presented in [15,18,17,16], compared to the most popular and fast algebraic ones, are very impressive, differing even to order(s) of magnitude: Their advantage with respect to the number of kernel evaluations, *for the same level of accuracy*, compared to the most popular algebraic techniques, is readily noticeable. The enhanced performance is justified by the fact that, although the algebraic algorithms (especially SMO with improvements described in [11]) make a clever utilization of the cache, where kernel values are stored, they cannot avoid repetitive searches in order to find the best couple of points (working set selection) to be used in the next iteration of the optimization process. Furthermore, the enhanced performance of the new geometric algorithms against its algebraic competitors can be explained from the fact that they are straightforward optimization algorithms, with a clear optimization target at each iteration step, always aiming at the global minimum and at the same time being independent of obscure and sometimes inefficient heuristics. Besides, they are well studied concerning convergence, a property that is hardly proved for the algebraic algorithms.

3 Application to Medical Image Analysis – Mammography

The field of Medical Image Analysis and Diagnosis (and particularly of Mammography, studied in this work) is very crucial for social reasons[5] and very demanding from the computational point of view. In this thesis, a set of qualitative [13] and quantitative mammographic textural [12,14] and morphological [8,7] features have been assessed (using methods of statistical and fractal analysis); besides, several machine learning paradigms, e.g., Artificial Neural Networks (ANNs) and SVMs, have been used to discriminate benign from malignant mammographic masses. SVMs outperformed the other classifiers.

4 Conclusions

The work accomplished in the context of this Ph.D. dissertation, was mainly twofold: First, the exploration of the SLT field through its most well-known derivative, i.e., the SVM learning, and the investigation of the effect of the geometric interpretation of the SVM framework for the derivation of more effective algorithms; and, second, to apply SVM classification algorithms to the field of Medical Analysis, compare the results with other state-of-the-art classification tools, e.g., Artificial Neural Networks and derive new image statistical features that can be helpful in the Computer Aided Detection and Diagnosis of masses on mammographic images.

The first objective has been met through the creation of a mathematical toolbox around the notion of RCHs and the derivation of several theoretical

[5] Breast cancer remains a major cause of death among female population and mammography the main tool of its early diagnosis.

corollaries that made possible the incorporation (through adaptation) of geometric (nearest point) algorithms for the solution of the general, i.e., non-linear, non-separable SVM classification task, without the penalty of the combinatorial complexity that such approaches suffered until now. As a practical result of this novel theoretical framework, the transformation and adaptation of two well-known geometric NPAs, namely Gilbert's and Schlesinger-Kozinec's algorithms, has been accomplished. Both converted algorithms (to work with RCHs and, hence, being appropriate for the SVM classification task) have been implemented (in Matlab) and compared to other state-of-the-art algebraic implementations, using several publicly available benchmark datasets. The results that have been obtained from these comparisons were highly encouraging, as the geometric algorithms impressively outperformed their algebraic counterparts in terms of speed and (sometimes also) of accuracy.

The second objective was accomplished through the implementation of a robust image feature set, which was compared to the qualitative feature set that experienced radiologists use, in order to diagnose mammographic images. This feature set include textural and morphological features, that describe both the textural content of mammographic masses and its deviation from the textural content of normal tissue, as well as the morphology of the mass boundary, that is informative of the benignancy or malignancy of the particular mass. The information content of the datasets, that were produced from a mammographic image database which was created for this purpose in the context of this work, has been assessed through fractal analysis and comparison with the qualitative information available to the experienced physicians when proceeding to a mammographic diagnosis.

Besides, several classification schemes and architectures have been used, including SVMs (that have been implemented during the first objective of this study), ANNs and k-NN; as it was expected, SVMs presented the best overall performance regarding the success rate of the classification results.

References

1. Kristin P. Bennett and Erin J. Bredensteiner. Geometry in learning, September 27 1997.
2. Kristin P. Bennett and Erin J. Bredensteiner. Duality and geometry in SVM classifiers. In Pat Langley, editor, *ICML*, pages 57–64. Morgan Kaufmann, 2000.
3. David J. Crisp and Christopher J. C. Burges. A geometric interpretation of ν-SVM classifiers. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *NIPS*, pages 244–250. The MIT Press, 1999.
4. N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, March 01 2000.
5. Vojtěch Franc and Václav Hlaváč. An iterative algorithm learning the maximal margin classifier. *Pattern recognition*, 36(9):1985–1996, September 2003.
6. T.-T. Frieß and R. Harisson. Support vector neural networks: the kernel adatron with bias and soft margin. Technical Report ACSE-TR-752, Department of ACSE, University of Sheffield, Sheffield, UK, 1998.

7. H. Georgiou, M. Mavroforakis, N. Dimitropoulos, D. Cavouras, and S. Theodoridis. Multi-scaled morphological features for the characterization of mammographic masses using statistical classification schemes. *Artificial Intelligence In Medicine*, 41(1):39–55, 2007.
8. H. V. Georgiou, M. E. Mavroforakis, D. Cavouras, N. Dimitropoulos, and S. Theodoridis. Multi-resolution morphological analysis and classification of mammographic masses using shape, spectral and wavelet features. *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on*, 1, 2002.
9. E. Gilbert. Minimizing the quadratic form on a convex set. *SIAM J. Control*, 4:61–79, 1966.
10. Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I.* Springer-Verlag, 1996.
11. S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. A fast iterative nearest point algorithm for support vector machine classifier design. Technical Report TR-ISL-99-03, Dept of CSA, IISc, Department of CSA, IISc, Bangalore, India, 1999.
12. M. E. Mavroforakis, H. V. Georgiou, D. Cavouras, N. Dimitropoulos, and S. Theodoridis. Mammographic mass classification using textural features and descriptive diagnostic data. *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on*, 1, 2002.
13. Michael Mavroforakis, Harris Georgiou, Nikos Dimitropoulos, Dionisis Cavouras, and Sergios Theodoridis. Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines. *European Journal of Radiology*, 54(1):80–89, Apr 2005.
14. Michael E. Mavroforakis, Harris V. Georgiou, Nikos Dimitropoulos, Dionisis Cavouras, and Sergios Theodoridis. Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers. *Artificial Intelligence in Medicine*, 37(2):145–162, 2006.
15. Michael E. Mavroforakis, Margaritis Sdralis, and Sergios Theodoridis. A novel SVM geometric algorithm based on reduced convex hulls. In *ICPR*, pages 564–568. IEEE Computer Society, 2006.
16. Michael E. Mavroforakis, Margaritis Sdralis, and Sergios Theodoridis. A geometric nearest point algorithm for the efficient solution of the SVM classification task. *IEEE Transactions on Neural Networks*, 18(5):1545–1549, 2007.
17. Michael E. Mavroforakis and Sergios Theodoridis. Support Vector Machine (SVM) classification through Geometry. In *Proceedings of EUSIPCO 2005*, Antalya, Turkey, 2005.
18. Michael E Mavroforakis and Sergios Theodoridis. A geometric approach to support vector machine (svm) classification. *IEEE Trans Neural Netw*, 17(3):671–682, May 2006.
19. J. Platt. *Advances in Kernel Methods - Support Vector Learning*, chapter Fast training of Support Vector Machines using Sequential Minimal Optimization, pages 185–208. MIT Press, 1999.
20. B. Schölkopf and A.J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, USA, 2002.
21. Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition.* Academic Press, 3rd edition, 2006.
22. V. N. Vapnik. *Statistical Learning Theory.* John_Wiley, September 1998.
23. Dengyong Zhou, Baihua Xiao, Huibin Zhou, and Ruwei Dai. Global geometry of SVM classifiers, July 11 2002.