SUPPORT VECTOR MACHINE (SVM) CLASSIFICATION THROUGH GEOMETRY

Michael E. Mavroforakis, and Sergios Theodoridis

Informatics and Telecommunications Dept., University of Athens
TYPA buildings, Univ. Campus, 15771, Athens, Greece
phone: + (30) 210-9648663, email: mmavrof@di.uoa.gr
web: www.di.uoa.gr/~idsp

ABSTRACT

Support Vector Machines is a very attractive and useful tool for classification and regression; however, since they rely on subtle and complex algebraic notions of optimization theory, lose their elegance and simplicity when implementation is concerned. It has been shown that the SVM solution, for the case of separate classes, corresponds to the minimum distance between the respective convex hulls. For the non-separable case, this is true for the Reduced Convex Hulls (RCH). In this paper a new geometric algorithm is presented, applied and compared with other non-geometric algorithms for the non-separable case.

1. INTRODUCTION

Geometry provides a very intuitive background for the understanding and the solution of many problems in the fields of Pattern Recognition and Machine Learning, which, in turn, play a decisive role in Signal and Image Processing.

Support Vector Machine (SVM) paradigm in pattern recognition presents a lot of advantages over other approaches (e.g., [4], [10]), 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) reduced number of tuning parameters and, last but not least, 4) clear geometric intuition on the classification procedure.

The contribution of this work consists of the following: 1) It exploits the intrinsic geometric intuition to the full extend, i.e., not only theoretically but also practically (leading to a novel algorithmic solution), in the context of classification through the SVM approach, 2) it provides, for the first time, the theoretical background for a geometric 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, 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 develops, for the first time, an *efficient* algorithm for the computation of the minimum distance between the RCHs and 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.

2. SUPPORT VECTOR MACHINES

Simply stated, a SVM finds the best separating (*maximal margin*) hyperplane between the two classes of training samples in the feature space, which leads to maximal generalization. The patterns in the original, low dimensional space $\mathcal{X}$, are mapped ($\Phi:\mathcal{X}\rightarrow\mathcal{H}$) in a high-dimensional *feature space* $\mathcal{H}$, which is a *Reproducing Kernel Hilbert Space* (RKHS). It is not necessary to know the map 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 pairs of samples $x_1,x_2\in\mathcal{X}$) [9]. Through this “kernel trick”, it is possible to transform a nonlinear classification problem to a linear one, but in a higher (maybe infinite) dimensional space.

Although some authors have presented the theoretical background of the geometric properties of SVMs, exposed thoroughly in [11], 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) [8]. The geometric properties of learning [1] 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 [2] and also for the non-separable case through the notion of the *Reduced Convex Hull* (RCH) [3]. Actually, the geometric algorithms presented until now ([7], [5]) *are suitable only for solving directly the separable case* and indirectly the non-separable case through the trick proposed in [6]. However, the latter (artificially extending the dimension of the input space by the number of training patterns) is equivalent to a quadratic penalty factor and, 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 [7].

In this work, we support the notion of the RCH with the sufficient mathematical background, so that to overcome the combinatorial complexity problems inherent in RCH constructs and, therefore, making it suitable for solving efficiently the SVM problem, employing geometric arguments.

3. REDUCED CONVEX HULLS (RCH)

The set of all convex combinations of points in some set C, 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 C and denoted $\mathrm{R}(C, \mu)$:

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

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 [11].

The effect of the value of bound μ to the size of the RCH is presented in Fig. 1.

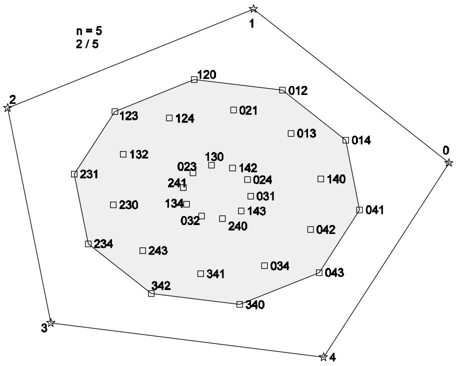

Fig. 1. The RCHs $\mathrm{R}(\mathrm{P5}, 2/5)$ is shown – generated by 5 points (stars) and having $\mu = 2/5$ – to present the points that are candidates to be extreme (marked by small squares). Each point in the RCH is labeled in order to present the original points from which it has been constructed; the last label is the one with the lowest coefficient.

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, proposed in this paper.

The main rationale of our methodology consists of the following steps: 1) Any convex hull is defined by its extreme points, 2) We prove that the extreme points of a RCH are computed by a specific finite linear combination of the points of the originally convex hull, 3) We prove that the minimum projection of a RCH onto a certain direction is a *specific* linear combination of the projections of its extreme points. The respective proofs will only be sketched here due to limited space.

Theorem 1: The extreme points of a RCH $\mathrm{R}(C,\mu)=\left\{w: w=\sum\nolimits_{i=1}^{k} a_i x_i,\ x_i \in X,\ \sum\nolimits_{i=1}^{k} a_i=1,\ 0 \le a_i \le \mu\right\}$ have coefficients a_i belonging to the set $S=\left\{0,\ \ 1-\lfloor 1/\mu \rfloor \mu,\ \ \mu\right\}$, where $\lfloor 1/\mu \rfloor$ is the integral part of the ratio $1/\mu$.

Proof: The proof is rather lengthy, so suitable sketch of it is presented here. In the case that $\mu = 1$ the theorem is obviously true. For $0 < \mu < 1$ the theorem will be proved by contradiction: Assuming that a point $w \in \mathrm{R}(C,\mu)$ is an extreme point, with some coefficients not belonging in S, a couple of other points $w_1, w_2 \in \mathrm{R}(C,\mu)$ is needed to be found and then to be proved that w belongs to the line segment $[w_1, w_2]$. But since two other points are needed, at least two coefficients have to be found not belonging in S. Therefore, the first aim is to prove, by contradiction, that any point $w \in \mathrm{R}(C,\mu)$ cannot have only one coefficient not belonging in S. Afterwards, using these coefficients, it is easy to construct a couple of points $w_1, w_2(C,\mu)$, such that w is the middle point of the line segment joining them. □

Proposition 1: Each of the *extreme points* of a RCH $\mathrm{R}(C,\mu)=\left\{w: w=\sum\nolimits_{i=1}^{k} a_i x_i,\ x_i \in X,\ \sum\nolimits_{i=1}^{k} a_i=1,\ 0 \le a_i \le \mu\right\}$ 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 $\lceil 1/\mu \rceil \ge 1/\mu$. Furthermore, if $1/\mu = \lceil 1/\mu \rceil$ then all $a_i = \mu$; otherwise, $a_i = \mu$ for $i = 1, \ldots, m-1$ and $a_m = 1 - \lfloor 1/\mu \rfloor \mu$.

Sketch of the 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 $\mathrm{R}(C,\mu)$ are either μ or $1 - \lfloor 1/\mu \rfloor \mu$. Furthermore, the fact that only one coefficient with value $1 - \lfloor 1/\mu \rfloor \mu > 0$ can be present, is proved by contradiction. □

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

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

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

- $\mu \sum\nolimits_{j=1}^{m} s_{i_j}$ if $0 < \mu$ and $\lambda = 0$
- $\mu \sum\nolimits_{j=1}^{m} s_{i_j} + \lambda s_{i_{m+1}}$ 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 $p < q$. □

The effect of Theorem 2 is illustrated in Fig. 2.

Proposition 2: 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: This is a direct consequence of Proposition 2, found in [3]. □

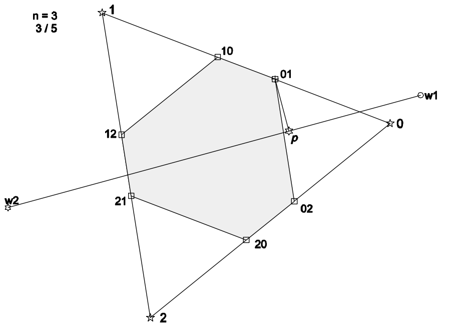

Fig. 2. The minimum projection p of the RCH $\mathrm{R}(P3,3/5)$, generated by 3 points and having $\mu = 3/5$, onto the direction $w_2 - w_1$ belongs to the point (01), which is calculated, according to Theorem 2, as the ordered weighted sum of the projection of only $\lceil 5/3 \rceil = 2$ points ((0) and (1)) of the 3 initial points. The magnitude of the projection, in lengths of $\|w_2 - w_1\|$ is $(3/5)(x_0 \mid w_2 - w_1) + (2/5)(x_1 \mid w_2 - w_1)$.

4. GEOMETRIC ALGORITHM FOR THE SVM

An iterative, geometric algorithm for solving the linearly separable SVM problem has been presented recently in [5]. This algorithm is adapted here, with the mathematical toolbox for RCHs presented above, to solve the non-separable SVM problem and can be described by the following three steps:

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
$$z_r = \begin{cases} z_{1r} = \sum_{i \in I_1} b_i x_i, & b_i \in \{0, \lambda_1, \mu_1\}, \quad \sum_{i \in I_1} b_i = 1 \\ z_{2r} = \sum_{i \in I_2} b_i x_i, & b_i \in \{0, \lambda_2, \mu_2\}, \quad \sum_{i \in I_2} b_i = 1 \end{cases}$$
(actually the coefficients b_i) such that $z_r = \underset{z_{1r} \in R(X_1,\mu_1), z_{2r} \in R(X_2,\mu_2)}{\arg\min} \left(m(z_{1r}), m(z_{2r}) \right)$ where
$$m(z_r) = \begin{cases} \dfrac{\langle z_{1r} - w_2, w_1 - w_2 \rangle}{\|w_1 - w_2\|}, & z_{1r} \in \mathrm{R}(X_1, \mu_1) \\ \dfrac{\langle z_{2r} - w_1, w_2 - w_1 \rangle}{\|w_1 - w_2\|}, & z_{2r} \in \mathrm{R}(X_2, \mu_2) \end{cases}.$$
The quantity $m(z_r)$ actually represents the distance 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.

   If the ε-optimality condition $\|w_1 - w_2\| - m(z_r) < \varepsilon$ 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 $q_1 = \min\left(1, \dfrac{\langle w_1 - w_2, w_1 - z_{1r} \rangle}{\|w_1 - z_{1r}\|^2}\right)$, which means $a_i^{new} = q_1 b_i + (1 - q_1) a_i$, $i \in I_1$; otherwise, set $w_1^{new} = w_1$ and compute $w_2^{new} = q z_{2r} + (1 - q) w_2$, where $q_2 = \min\left(1, \dfrac{\langle w_2 - w_1, w_2 - z_{2r} \rangle}{\|w_2 - z_{2r}\|^2}\right)$, which means $a_i^{new} = q_2 b_i + (1 - q_2) a_i$, $i \in I_2$. Continue with step 2.

The quantities to be calculated involve minimum projections of the RCHs onto $w = w_1 - w_2$ or inner products of the RCH points presenting such minimum projections. Therefore, the calculations are done efficiently after the application of the mathematical background presented above.

This algorithm has almost the same complexity as the Schlesinger – Kozinec one (the extra cost is the sort 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 $\langle z_r, z_r \rangle$) and the same caching scheme can be used, with only $O(|I_1| + |I_2|)$ storage requirements.

5. RESULTS

Some representative results are included, concerning only non-separable cases, since the separable cases work in exactly the same way as the initial algorithm. The results were compared with the SMO algorithm for the training total run time and the number of kernel evaluations and summarized in Table I. The test cases were run in an Intel Pentium 4 PC.

- *Linear non-separable case*: A 2-dimensional sample space of 390 (Class A) and 395 (Class B) randomly generated samples was used. Each sample attribute ranged from -0.5 to 0.5 and the margin was -0.1 (negative margin indicates the overlapping between classes).

- *Non-linear non-separable cas*A 2-dimensional sample space of 390 (Class A) and 395 (Class B) randomly generated samples were used, based on the checkers' board pattern. Each sample attribute ranged from -4 to 4 and the margin was -0.2. The kernel used was RBF with $\sigma = 1.8$.

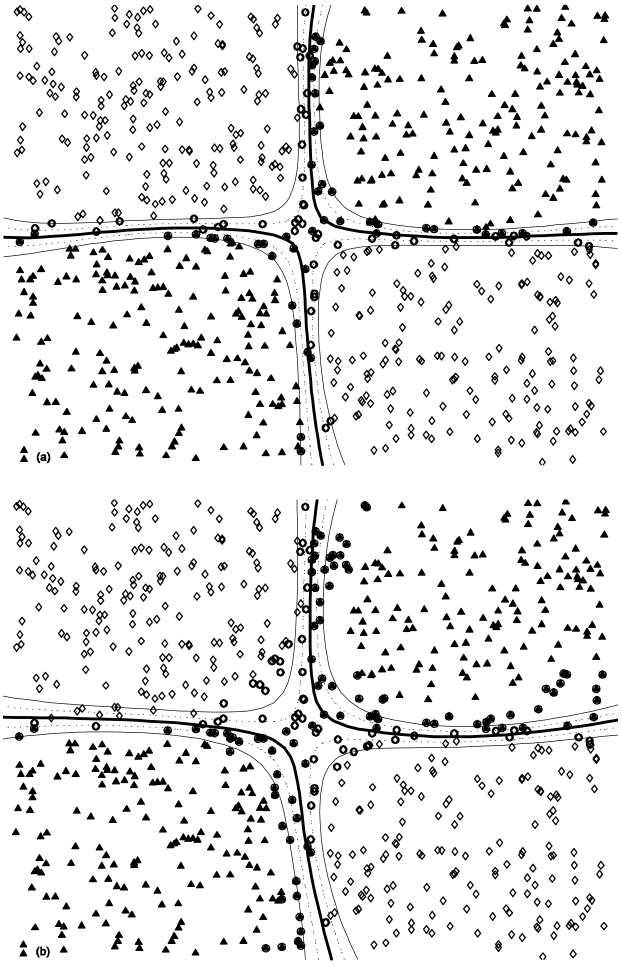

Fig. 1. Classification results for the non-linear non-separable case for SMO (a) and RCH-SK (b) algorithms.

The resulting separating surfaces, (shown in Fig. 1 only for the non-linear case), were very close for both methods ((a) SMO and (b) RCH-SK proposed here). The bold solid line represents the separating hypersurface (value 0), the thin solid lines correspond to values -0.5 and 0.5 and the dashed thin lines to the values -1.0 and 1.0 respectively. The circled patterns correspond to support vectors. The computational time requirements (along with the parameters used for each method) are shown in Table I, from which the advantages of our new method are apparent.

Method	**Kernel**	**Time (sec)**	**Kernel evalutions**	**Parameters**
SMO	Linear	885.6	19459278	C=50, tol=0.001
RCH-SK	Linear	**156.7**	**5453801**	$\mu_1 = \mu_2 = 0.006$, $\varepsilon = 0.0001$
SMO	RBF	641.5	10264378	C=10, tol=0.001
RCH-SK	RBF	**179.5**	**3699596**	$\mu_1 = \mu_2 = 0.03$, $\varepsilon = 0.07$

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

6. CONCLUSION

A new geometric algorithm for implementing a SVM classifier has been presented. The algorithm computes the minimum distance between the RCHs of the two classes. It is the first time in the literature that such an algorithm is presented for the non-separable classification task. Also, the paper presented the proofs of new results concerning RCHs and projections on a direction. These theorems were necessary for the development of the new algorithm.

REFERENCES

[1] 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.

[2] 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.

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

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

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

[6] 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.

[7] 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.

[8] 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.

[9] B. Schölkopf, A. Smola, *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*, The MIT Press, 2002.

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

[11] 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*.