Analysis of Generic Perceptron-Like Large Margin Classifiers

Petroula Tsampouka and John Shawe-Taylor

School of Electronics and Computer Science,
University of Southampton, UK

Abstract. We analyse perceptron-like algorithms with margin considering both the standard classification condition and a modified one which demands a specific value of the margin in the augmented space. The new algorithms are shown to converge in a finite number of steps and used to approximately locate the optimal weight vector in the augmented space. As the data are embedded in the augmented space at a larger distance from the origin the maximum margin in that space approaches the maximum geometric one in the original space. Thus, our procedures exploiting the new algorithms can be regarded as approximate maximal margin classifiers.

1 Introduction

Rosenblatt's perceptron algorithm [6] is the simplest on-line learning algorithm for binary linear classification [3]. A variant of the perceptron also exists which unlike the original algorithm aims at a solution hyperplane with respect to which the data possess a non-zero margin. The problem, however, of finding the optimal hyperplane has been successfully addressed only with the advent of the Adatron algorithm [1] and later by the Support Vector Machines (SVMs) [7, 2].

Our purpose here is to address the problem of maximal margin classification using the less time consuming, compared to SVMs, perceptron-like algorithms. We work in a space augmented by one additional dimension [3] in which we embed the data by placing them at a distance ρ in the extra dimension and replace the perceptron classification condition with a new one insisting on a specific value of the margin in this augmented space. We show that the algorithms with the modified condition converge in a finite number of steps and use them to approximately locate the solution with maximum margin in the augmented space. As $\rho \to \infty$ the maximum margin in the augmented space approaches the maximum geometric one in the original space. Thus, our algorithmic procedures can be considered as approximate maximal margin classifiers.

Whilst proving convergence of the new algorithms we found it useful to introduce the notion of stepwise convergence, the property of the algorithms that approach the optimal solution vector at each step. Through a formulation involving stepwise convergence we provide a unified approach in establishing convergence for a large class of algorithms with additive perceptron-like update rules.

J. Gama et al. (Eds.): ECML 2005, LNAI 3720, pp. 750–758, 2005.

Section 2 contains our theoretical analysis. In Sect. 3 we describe algorithmic implementations aiming at an approximate determination of the maximum margin. Finally, Sect. 4 contains our conclusions.

2 Theoretical Analysis

In what follows we make the assumption that we are given a training set which, even if not initially linearly separable can, by an appropriate feature mapping into a space of a higher dimension, be classified into two categories by a linear classifier. This higher dimensional space in which the patterns are linearly separable will be the considered space. By adding one additional dimension and placing all patterns in the same position $\rho_0 = \rho > 0$ in that dimension we construct an embedding of our data into the so-called augmented space. The advantage of this embedding is that the linear hypothesis in the augmented space becomes homogeneous.

We concentrate on algorithms that update the augmented weight vector $\boldsymbol{a}_t$ by adding a suitable positive amount in the direction of the misclassified (according to an appropriate condition) training pattern $\boldsymbol{y}_k$. The general form of such an update rule is

$$\boldsymbol{a}_{t+1} = \boldsymbol{a}_t + \eta f_t \boldsymbol{y}_k \ , \tag{1}$$

where η is the (constant) learning rate and f_t a function of the current step (time) t which we require to be positive and bounded, i.e.

$$0 < f_{\min} \leq f_t \leq f_{\max} \ . \tag{2}$$

For the special case of the perceptron algorithm $f_t = 1$. Each time the predefined misclassification condition is satisfied by a training pattern the algorithm proceeds to the update of the weight vector. Throughout our discussion a reflection with respect to the origin in the augmented space of the negatively labelled patterns is assumed in order to allow for a common classification condition for both categories of patterns [3]. Also, we use the notation $R = \max_k \|\boldsymbol{y}_k\|$ and $r = \min_k \|\boldsymbol{y}_k\|$.

The relation characterising optimally correct classification of the training patterns by a weight vector $\boldsymbol{u}$ of unit norm in the augmented space is

$$\boldsymbol{u} \cdot \boldsymbol{y}_k \geq \gamma_{\mathrm{d}} \quad \forall k \ . \tag{3}$$

The quantity γ_{d}, which we call the optimal directional margin, is defined as

$$\gamma_{\mathrm{d}} = \max_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \min_k \{\boldsymbol{u} \cdot \boldsymbol{y}_k\} \tag{4}$$

and is obviously bounded from above by r. The optimal directional margin determines the maximum distance from the origin in the augmented space of the hyperplane normal to $\boldsymbol{u}$ placing all training patterns on the positive side. In the determination of this hyperplane only the direction of $\boldsymbol{u}$ is exploited with no

reference to its projection onto the original space. As a consequence the above maximum margin in the augmented space is not necessarily realised with the same weight vector that gives rise to the optimal geometric margin in the original space. Notice, however, that the existence of a directional margin means that there exists a geometric margin at least as large as the directional one.

First, in Sect. 2.1, we examine algorithms in which the misclassification condition takes the form

$$\boldsymbol{a}_t \cdot \boldsymbol{y}_k \leq b \ , \tag{5}$$

where b is a positive parameter. A slight transformation of (5) to

$$\boldsymbol{u}_t \cdot \boldsymbol{y}_k \leq \frac{b}{\|\boldsymbol{a}_t\|} \ , \tag{6}$$

where $\boldsymbol{u}_t$ is the weight vector $\boldsymbol{a}_t$ normalised to unity, reveals that the minimum directional margin required by the standard margin condition is lowered as the length of the weight vector grows.

Subsequently, in Sect. 2.2, we examine algorithms with a misclassification condition of the form

$$\boldsymbol{u}_t \cdot \boldsymbol{y}_k \leq \beta \ , \tag{7}$$

where β is a positive parameter. Notice that the above condition amounts to requiring a minimum directional margin which is not lowered with the number of steps. Therefore, successful termination of the algorithm leads to a solution with a guaranteed geometric margin at least as large as the directional margin β found. This is an important difference from the misclassification condition of (5) which, as (6) illustrates, cannot by itself guarantee a minimum margin. Obviously, convergence of the algorithm is not possible unless

$$\beta < \gamma_{\rm d} \ . \tag{8}$$

The condition (7) involving only the direction of the weight vector motivates new positive and bounded functions f_t like the function $f_t = (\beta_u - \boldsymbol{u}_t \cdot \boldsymbol{y}_k)/\|\boldsymbol{y}_k\|$ with $\beta_u > \beta$. We consider two cases depending on whether the length of the augmented weight vector is free to grow or is kept constant throughout the algorithm. In the last category of algorithms a fixed-length weight vector is achieved by a renormalisation of the newly produced weight vector to the target margin value β each time an update according to (1) takes place.

A very desirable property of an algorithm is certainly progressive convergence at each step meaning that at each update $\boldsymbol{u}_t$ moves closer to the optimal direction $\boldsymbol{u}$. Let us assume that

$$\boldsymbol{u}_t \cdot \boldsymbol{u} > 0 \ . \tag{9}$$

Because of (9) the criterion for stepwise angle convergence, namely

$$\Delta \equiv \boldsymbol{u}_{t+1} \cdot \boldsymbol{u} - \boldsymbol{u}_t \cdot \boldsymbol{u} > 0 \ , \tag{10}$$

can be equivalently expressed as a demand for positivity of D

$$D \equiv (\boldsymbol{u}_{t+1} \cdot \boldsymbol{u})^2 - (\boldsymbol{u}_t \cdot \boldsymbol{u})^2 = 2\frac{\eta f_t}{\|\boldsymbol{a}_t\|}(\boldsymbol{u}_t \cdot \boldsymbol{u})\left\|\boldsymbol{u}_t + \frac{\eta f_t}{\|\boldsymbol{a}_t\|}\boldsymbol{y}_k\right\|^{-2} A \ , \tag{11}$$

where use has been made of the update rule (1) and A is defined by

$$A \equiv \boldsymbol{y}_k \cdot \boldsymbol{u} - (\boldsymbol{u}_t \cdot \boldsymbol{u})(\boldsymbol{y}_k \cdot \boldsymbol{u}_t) - \frac{1}{2}\frac{\eta f_t}{\|\boldsymbol{a}_t\|}\left(\|\boldsymbol{y}_k\|^2 (\boldsymbol{u}_t \cdot \boldsymbol{u}) - \frac{(\boldsymbol{y}_k \cdot \boldsymbol{u})^2}{(\boldsymbol{u}_t \cdot \boldsymbol{u})}\right) \ . \quad (12)$$

Positivity of A leads to positivity of D on account of (2) and (9) and consequently to stepwise convergence. Actually, convergence occurs in a finite number of steps provided that after some time $\|\boldsymbol{a}_t\|$ and A become bounded from below by a positive constant and $\|\boldsymbol{a}_t\|$ increases at most linearly with t. Following this rather unified approach one can show that sooner or later the algorithms under consideration enter the stage of stepwise convergence and terminate successfully in a finite number of steps. Better time bounds are, however, obtainable by alternative methods.

Finally, Sect. 2.3 contains our derivations which place an upper bound on the optimal geometric margin of a training set in terms of the optimal directional one, thereby leading to an estimate of the optimal geometric margin.

2.1 Algorithms with the Standard Margin Condition

We first analyse the algorithms with the general update rule (1) by calculating an upper bound on the number of updates until a solution is found, thereby extending Novikoff's theorem [5, 4]. From the difference $\boldsymbol{a}_{t+1} \cdot \boldsymbol{u} - \boldsymbol{a}_t \cdot \boldsymbol{u}$ we obtain a relation whose repeated application, assuming $\boldsymbol{a}_0 = \boldsymbol{0}$, implies

$$\|\boldsymbol{a}_t\| \geq \boldsymbol{a}_t \cdot \boldsymbol{u} \geq \eta f_{\min} \gamma_{\mathrm{d}} t \ . \quad (13)$$

Also the difference $\|\boldsymbol{a}_{t+1}\|^2 - \|\boldsymbol{a}_t\|^2$ gives a relation whose repeated application leads to

$$\|\boldsymbol{a}_t\| \leq \sqrt{(\eta^2 f_{\max}^2 R^2 + 2\eta f_{\max} b)t} \ . \quad (14)$$

Combining (13) and (14) we get Novikoff's time bound

$$t \leq t_{\mathrm{N}} \equiv \frac{f_{\max}^2}{f_{\min}^2}\frac{R^2}{\gamma_{\mathrm{d}}^2}\left(1 + \frac{2}{\eta f_{\max}}\frac{b}{R^2}\right) \ . \quad (15)$$

We next turn to a discussion of stepwise convergence. From (13) it is clear that for $t > 0$ (9) holds. Also, $\boldsymbol{y}_k \cdot \boldsymbol{u}$ appearing in A is definitely positive due to (3) whereas $\|\boldsymbol{a}_t\|$ increases with time because of (13), thereby making the term of A linear in η negligible. Moreover, (6) shows that the term $(\boldsymbol{u}_t \cdot \boldsymbol{u})(\boldsymbol{y}_k \cdot \boldsymbol{u}_t)$ is suppressed with time. Thus, for time t larger than a critical time t_{c} positivity of A and consequently of D is accomplished. By using (3), (5) and (13) we obtain

$$A \geq \gamma_{\mathrm{d}} - \frac{1}{2\eta f_{\min}\gamma_{\mathrm{d}} t}\left(2b + \eta f_{\max}(R^2 - \gamma_{\mathrm{d}}^2)\right) \ . \quad (16)$$

From the above inequality the time sufficient for stepwise convergence to begin is

$$t_{\mathrm{c}} \equiv \frac{1}{2}\frac{f_{\max}}{f_{\min}}\frac{R^2}{\gamma_{\mathrm{d}}^2}\left(1 + \frac{2}{\eta f_{\max}}\frac{b}{R^2} - \frac{\gamma_{\mathrm{d}}^2}{R^2}\right) < \frac{1}{2}\frac{f_{\min}}{f_{\max}} t_{\mathrm{N}} \ . \quad (17)$$

Therefore, unless the algorithm terminates much before Novikoff's time $t_{\rm N}$ is exhausted, it will definitely enter the phase of stepwise convergence. Actually, because of (13), (14) and (16) an alternative proof of convergence in a finite number of steps is obtained.

It would be interesting to estimate the margin that the algorithm is able to achieve [4]. For $t = t_{\rm N}$ (13) and (14) hold as equalities leading to the largest possible value of $\|\boldsymbol{a}_t\|$, namely $\|\boldsymbol{a}_{t_{\rm N}}\| = \eta f_{\min}\gamma_{\rm d} t_{\rm N}$, which provides a lower bound $\beta_{\min} = b/\|\boldsymbol{a}_{t_{\rm N}}\|$ on the directional margin $\beta = b/\|\boldsymbol{a}_t\|$ appearing in (6)

$$\beta_{\min} = \frac{f_{\min}}{f_{\max}} \frac{\gamma_{\rm d}}{(2 + f_{\max}(\eta R^2/b))} = \frac{1}{2}\frac{f_{\min}}{f_{\max}}\gamma_{\rm d}\left(1 - \frac{f_{\max}^2}{f_{\min}^2}\frac{R^2}{\gamma_{\rm d}^2} t_{\rm N}^{-1}\right) \ . \tag{18}$$

The above guaranteed value of the directional margin acquires a maximum of $\frac{1}{2}\frac{f_{\min}}{f_{\max}}\gamma_{\rm d} \leq \frac{1}{2}\gamma_{\rm d}$ for $b \gg \eta R^2$ or $t_{\rm N} \gg R^2/\gamma_{\rm d}^2$.

2.2 Algorithms with Fixed Directional Margin Condition

Algorithms with Free-Length Weight Vector. In the case that $\boldsymbol{a}_t$ is free to grow indefinitely and $\boldsymbol{a}_0 = \boldsymbol{0}$ (13) is again obtained and as a consequence for $t > 0$ (9) is once more recovered. Therefore, positivity of D is equivalent to stepwise convergence. By using (3) and (7) we get a lower bound on the η-independent part of A

$$\boldsymbol{y}_k \cdot \boldsymbol{u} - (\boldsymbol{u}_t \cdot \boldsymbol{u})(\boldsymbol{y}_k \cdot \boldsymbol{u}_t) \geq \gamma_{\rm d} - \beta \ , \tag{19}$$

which is definitely positive on account of (8). Furthermore, because of (13) the terms of A linear in η, which are not necessarily positive, become less important with time leading to positivity of A and consequently of D for t larger than a critical time $t_{\rm c}$. More formally, employing (3), (7) and (13) we can place a lower bound on A

$$A \geq \gamma_{\rm d} - \beta - \frac{1}{2}\frac{f_{\max}}{f_{\min}}\frac{1}{\gamma_{\rm d} t}(R^2 - \gamma_{\rm d}^2) \tag{20}$$

and demanding positivity estimate the time $t_{\rm c}$ sufficient for the onset of stepwise convergence

$$t_{\rm c} \equiv \frac{1}{2}\frac{f_{\max}}{f_{\min}}\frac{R^2}{\gamma_{\rm d}^2}\left(1 - \frac{\gamma_{\rm d}^2}{R^2}\right)\left(1 - \frac{\beta}{\gamma_{\rm d}}\right)^{-1} \ . \tag{21}$$

Notice the crucial dependence of $t_{\rm c}$ on $\gamma_{\rm d} - \beta$. Since we initially set the weight vector to zero, $\boldsymbol{a}_t$ is entirely generated by the first t updates and its norm satisfies the obvious bound

$$\|\boldsymbol{a}_t\| \leq \eta f_{\max} R t \ . \tag{22}$$

Then, stepwise convergence along with (13), (20) and (22) lead to convergence in a finite number of steps.

Following a Novikoff-like procedure and provided $f_{\min}\gamma_{\rm d} - f_{\max}\beta > 0$ (which always holds if $f_t = 1$) we can obtain for every positive integer N a relation

$$\frac{t - N}{C_N + \ln\sqrt{t-1}} \leq \left(\frac{f_{\max}}{f_{\min}}\frac{R}{\gamma_{\rm d}}\right)^2 \left(1 - \frac{f_{\max}}{f_{\min}}\frac{\beta}{\gamma_{\rm d}}\right)^{-1} \tag{23}$$

constraining the growth of t. Here

$$C_N = N\frac{f_{\min}}{f_{\max}}\frac{\gamma_{\rm d}}{R}\left(1-\frac{f_{\min}}{f_{\max}}\frac{\gamma_{\rm d}}{R}\right)-\frac{1}{2}\left(\ln N-\frac{1}{N}\right)\ . \tag{24}$$

If $[x]$ denotes the integer part of x the optimal value of N is given by

$$N_{\rm opt}=\left[\frac{1}{2}\frac{f_{\max}}{f_{\min}}\frac{R}{\gamma_{\rm d}}\left(1-\frac{\beta}{R}\right)^{-1}\right]+1\ . \tag{25}$$

Notice that both (21) and (23) are independent of η. This is an interesting property of all algorithms of this class with $\boldsymbol{a}_0=\boldsymbol{0}$ under the additional assumption that f_t depends on $\boldsymbol{a}_t$ only through $\boldsymbol{u}_t$. This may be understood by observing that a rescaling of η results in a rescaling of $\boldsymbol{a}_t$ by the same factor which does not affect either the hyperplane normal to $\boldsymbol{a}_t$ or the classification condition.

Algorithms with Fixed-Length Weight Vector. We demand that $\boldsymbol{u}_t\cdot\boldsymbol{u}>0$ for all t which requires an appropriate choice of the initial condition. Notice that in this particular class of algorithms $\boldsymbol{a}_t$ cannot be set initially to zero since $\|\boldsymbol{a}_t\|=\beta$. We propose that $\boldsymbol{u}_0$ be chosen in the direction of one of the $\boldsymbol{y}_k$'s. Then, due to the form of the update rule and the positivity of f_t, it is obvious that $\boldsymbol{a}_t$ is a linear combination with positive coefficients of the training patterns. Therefore, since according to (3) $\boldsymbol{y}_k$ satisfies $\boldsymbol{y}_k\cdot\boldsymbol{u}>0$ the same is true for $\boldsymbol{a}_t$ and consequently for $\boldsymbol{u}_t$. Positivity of $\boldsymbol{u}_t\cdot\boldsymbol{u}$ allows us to use positivity of D as a criterion for stepwise convergence. Taking a closer look at A reveals that according to (8) and (19) the η-independent term remains positive throughout the algorithm. For the term linear in η which has no definite sign we conclude that an appropriate choice of η can render it smaller than the η-independent one, thereby leading to stepwise convergence from the first step of the algorithm. More specifically, using (3), (7) and the fact that $\|\boldsymbol{a}_t\|=\beta$ we have

$$A\geq\gamma_{\rm d}-\beta-\frac{\eta f_{\max}}{2\beta}(R^2-\gamma_{\rm d}^2)\ . \tag{26}$$

Positivity of A and D is achieved for η smaller than the critical value

$$\eta_{\rm c}\equiv\frac{2}{f_{\max}}\frac{(\gamma_{\rm d}-\beta)\beta}{R^2}\left(1-\frac{\gamma_{\rm d}^2}{R^2}\right)^{-1}\ . \tag{27}$$

Taking into account (9) and (26) and given that $\|\boldsymbol{a}_t\|=\beta$ stepwise convergence from the first step implies convergence in a finite number of steps.

By placing a t-independent lower bound on Δ defined in (10) and repeatedly applying the resulting inequality it is possible to derive an upper bound on t. For the optimal value of the learning rate

$$\eta_{\rm opt}\simeq\frac{1}{f_{\max}}\frac{(\gamma_{\rm d}-\beta)\beta}{R^2}\left(1+\frac{2\beta}{R}\right)^{-1} \tag{28}$$

we have

$$t < 2\frac{f_{\max}}{f_{\min}}\frac{R^2}{(\gamma_{\rm d}-\beta)^2}\left(1+\frac{2\beta}{R}\right)\left(1-\frac{(\gamma_{\rm d}-\beta)R}{(R+2\beta)^2}\right)^{-1} \ . \tag{29}$$

This bound is rather analogous to the one of the perceptron without margin. The main differences are a factor of 2 and the replacement of $\gamma_{\rm d}^2$ by $(\gamma_{\rm d}-\beta)^2$.

2.3 Estimating the Optimal Geometric Margin

If we denote by $\boldsymbol{a}=[\boldsymbol{w}\ w_0]$ a weight vector in the augmented space that classifies the patterns correctly the geometric margin $\gamma(\boldsymbol{a})$ of the set is

$$\gamma(\boldsymbol{a})=\frac{\|\boldsymbol{a}\|}{\|\boldsymbol{w}\|}\gamma_{\rm d}(\boldsymbol{a})=\frac{1}{\|\boldsymbol{w}\|}\min_k\{\boldsymbol{a}\cdot\boldsymbol{y}_k\}=\frac{1}{\|\boldsymbol{w}\|}\min_k\{\boldsymbol{w}\cdot\boldsymbol{x}_k+w_0\rho_0\} \ , \tag{30}$$

where $\gamma_{\rm d}(\boldsymbol{a})$ is the corresponding directional margin and $\boldsymbol{y}_k=[\boldsymbol{x}_k\ \rho_0]$. Notice that $|w_0|\rho/\|\boldsymbol{w}\|$ (with $\rho=|\rho_0|$) is the distance from the origin of the hyperplane normal to $\boldsymbol{w}$ which cannot exceed $R_x=\max_k\|\boldsymbol{x}_k\|$. Hence, $|w_0|/\|\boldsymbol{w}\|\le R_x/\rho$. As a consequence, $\|\boldsymbol{w}\|\le\|\boldsymbol{a}\|=\sqrt{\|\boldsymbol{w}\|^2+w_0^2}\le\|\boldsymbol{w}\|\sqrt{1+R_x^2/\rho^2}=\|\boldsymbol{w}\|\,R/\rho$ given that $R^2=\rho^2+R_x^2$. Then, (30) leads to $\gamma_{\rm d}(\boldsymbol{a})\le\gamma(\boldsymbol{a})$ but also to

$$\gamma(\boldsymbol{a})\le\frac{R}{\rho}\gamma_{\rm d}(\boldsymbol{a}) \ . \tag{31}$$

In the case that the weight vector $\boldsymbol{a}$ is the optimal one $\boldsymbol{a}_{\rm opt}$ maximising the geometric margin and taking into account that $\gamma_{\rm d}=\max_{\boldsymbol{a}}\gamma_{\rm d}(\boldsymbol{a})\ge\gamma_{\rm d}(\boldsymbol{a}_{\rm opt})$ and $\gamma\equiv\gamma(\boldsymbol{a}_{\rm opt})=\max_{\boldsymbol{a}}\gamma(\boldsymbol{a})\ge\max_{\boldsymbol{a}}\gamma_{\rm d}(\boldsymbol{a})=\gamma_{\rm d}$ the inequality (31) leads to

$$1\le\frac{\gamma}{\gamma_{\rm d}}\le\frac{R}{\rho} \ . \tag{32}$$

In the limit $\rho\to\infty$, $R/\rho\to 1$ and from (32) $\gamma_{\rm d}\to\gamma$. Thus, with ρ increasing the optimal directional margin $\gamma_{\rm d}$ approaches the optimal geometric one γ.

3 Algorithmic Implementation

In this section we present algorithms seeking the optimal directional margin which, however, due to the analysis of Sect. 2.3 could be used to approximately obtain the optimal geometric margin.

A first implementation makes repeated use of the algorithms of Sect. 2.2. In each round of its application the algorithm looks for a fixed directional margin β according to the condition $\boldsymbol{u}_t\cdot\boldsymbol{y}_k>\beta$. Each round lasts until the condition is satisfied by all $\boldsymbol{y}_k$'s or until an upper bound on the number of checks is reached. The range of feasible β values and therefore the interval that the algorithm should search extends from 0 to r. The search can be performed efficiently by a Bolzano-like bisection method with an initial target margin $\beta=\frac{r}{2}$ and a step parameter

set initially to $\frac{r}{2}$. If the algorithm comes up with a solution without exhausting the upper number of checks the round is considered successful. The weight vector is stored as the best solution found so far and is exploited as the initial condition of the next trial, thereby speeding up the procedure substantially. One could also envisage using the final weight vector of an unsuccessful previous round as the initial weight vector of a subsequent one until the first successful trial is reached. At the end of each trial the step is divided by 2. A successful (unsuccessful) trial is followed by an increase (decrease) of the target margin β by the current step value. Therefore, for a sufficiently large upper number of checks, the procedure guarantees that the deviation of β from the maximum margin is halved in each round. Termination occurs when the step reaches a certain predefined value.

A second possibility is to first use the standard perceptron algorithm with margin of Sect. 2.1 in order to obtain a solution with a guaranteed fraction of the existing directional margin given by (18) and then attempt to incrementally boost the margin obtained by repeatedly employing the algorithms of Sect. 2.2. The initial condition of each round of boosting will be the final weight vector of the previous round and the step by which the target margin increases will be determined as a fraction of the margin found in the first stage. The algorithm ends with the first unsuccessful trial. An analogous boosting procedure could follow a first stage of successful employment of the Bolzano-like method.

The above procedures were tested on artificial as well as real-life data with encouraging preliminary results.

4 Conclusions

We examined perceptron-like algorithms with margin and developed a criterion for the stronger requirement of stepwise convergence which allowed us to adopt a unified approach in the analysis. We also proposed a new class of such algorithms in which the standard classification condition is replaced by a more stringent one insisting on a fixed value of the directional margin and proved that they converge in a finite number of steps. Two implementations made possible a fast search for the optimal directional margin. We finally showed that as the data are placed increasingly far in the augmented space the optimal directional margin approaches the optimal geometric one. This observation transforms our procedures into fast and simple approximate maximal margin classifiers.

References

1. Anlauf, J. K., Biehl, M.: The adatron: an adaptive perceptron algorithm. Europhysics Letters **10** (1989) 687–692
2. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines (2000) Cambridge, UK: Cambridge University Press
3. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classsification (2000) Wiley-Interscience, 2nd edition

4. Li, Y., Zaragoza, H., Herbrich, R., Shawe-Taylor, J., Kandola, J.: The perceptron algorithm with uneven margins. In ICML'02 379–386
5. Novikoff, A. B. J.: On convergence proofs on perceptrons. In Proceedings of the Symposium on the Mathematical Theory of Automata Volume 12 (1962) 615–622
6. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review **65**(6) (1958) 386–408
7. Vapnik, V. N.: The Nature of Statistical Learning Theory (1995) Springer Verlag