ELSEVIER

May 1995

Pattern Recognition Letters 16 (1995) 491–497

Pattern Recognition Letters

On the rate of convergence of perceptron learning

U. Bhattacharya, S.K. Parui *

Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta 700 035, India

Received 5 August 1993; revised 14 November 1994

Abstract

In neural networks, convergence in iterative learning is a common problem. For fast learning one should be able to control the rate of convergence. In the present paper, the single-layer perceptron model for two classes is considered where the rate of convergence is studied with several choices of the gain term in the updation rule. Experimental results on a number of two-class problems are reported.

Keywords: Pattern classification; Perceptron learning; Rate of convergence; Gain term; Rate of misclassification; Bayes classifier

1. Introduction

For perceptron learning (Rosenblatt, 1959) in a linearly separable case for a two-class problem, the following rule is used to update the components w_i of the weight vector W on the basis of the feature vector X:

$$w_i(t+1) = w_i(t) + \eta\, f(X(t), W(t)) \tag{1}$$

where η, called the gain term, is a small positive quantity, t is the number of iterations and f is a function of $X(t)$ and $W(t)$, the feature and weight vectors respectively at the tth iteration. $W(t)$ determines the perceptron classifier at time t.

Though, in a linearly separable case, $W(t)$ converges, it will not in general converge when the two classes are not linearly separable. In such cases, to achieve convergence, η is made dependent on t such that $\eta(t)$ decreases with increasing t under certain conditions. However, little is known about the rate of convergence of $W(t)$ in such situations (Fukunaga, 1990). In this paper, we compare, on an empirical basis, the rates of convergence for different choices of $\eta(t)$. For this, we simulate a number of two-class problems using Gaussian distributions. It is found that, for a faster rate of convergence of $W(t)$, $\eta(t)$ in many situations should depend on the problem at hand in general and on how well $W(t)$ discriminates between the two classes, in particular.

2. Background

Let us consider a two-class problem in the p-dimensional feature space where the classes are indicated by 0 and 1. Suppose $W = (w_1, w_2, \ldots, w_p, w_{p+1})^{\mathrm{T}}$ is a $(p+1)$-dimensional vector, where T indicates transpose. Let, for a feature vector $X = (x_1, x_2, \ldots, x_p)^{\mathrm{T}}$,

$$y = g(X, W) = \begin{cases} 0 & \text{if } \sum_{i=1}^{p+1} w_i x_i < 0, \\ 1 & \text{if } \sum_{i=1}^{p+1} w_i x_i \geqslant 0 \end{cases} \tag{2}$$

[*] Corresponding author.

SSDI 0167-8655(95)00124-3

where $x_{p+1} = -1$. Let $(x_1(t), x_2(t), \ldots, x_p(t), d(t))^{\mathrm{T}}$, $t = 1, 2, \ldots$ be a sequence of observation vectors where $x_i(t)$ indicates the value of the ith feature at time t $(i = 1, 2, \ldots, p)$ and $d(t) = 0$ or 1 depending on whether $X(t) = (x_1(t), x_2(t), \ldots, x_p(t))^{\mathrm{T}}$ comes from class 0 or class 1 respectively.

Perceptron learning means getting hold of a weight vector W for which the y values match the d values in as many observations as possible (Hush and Horne, 1993). The perceptron algorithm that iteratively learns this W using the above sequence of observation vectors updates the weight vector in the following way (Lippmann, 1987):

$$W(t+1) = W(t) + \eta[d(t) - y(t)]X(t) \tag{3}$$

where (3) is a particular form of (1) and $y(t) = g(X(t), W(t))$.

Now, $W(t)$ converges in linearly separable cases, but not in cases which are not linearly separable. However, $W(t)$ will converge in all cases if η decreases with increasing t satisfying

$$\lim_{t \to \infty} \eta(t) = 0, \tag{4}$$

$$\sum_{t=1}^{\infty} \eta(t) = \infty, \tag{5}$$

$$\sum_{t=1}^{\infty} \eta^2(t) < \infty. \tag{6}$$

The physical meaning of these conditions can be described in the following way. Condition (4) allows $W(t)$ to settle down in the limit. Condition (5) guarantees that there is enough corrective action to avoid stopping short of the optimum W. Condition (6) ensures that the variance of the accumulated noise is finite so that the effect of noise can be corrected (Fukunaga, 1990).

The conventional choice of $\eta(t) = 1/t$ satisfies each of the conditions (4), (5) and (6). But this type of predetermined decrease of $\eta(t)$ ignoring the problem at hand, often leads to very slow convergence of $W(t)$ (Duda and Hart, 1973). In this paper we study the nature of convergence of $W(t)$ for a number of choices of $\eta(t)$ satisfying (4), (5) and (6) in two-class problems with varying amounts of overlap between the classes. It has been found that $\eta(t) = 1/t$ performs poorly in terms of the rate of convergence of $W(t)$.

3. Convergence of $W(t)$

The following four choices of $\eta(t)$ are considered in this paper:

$$\eta_1(t) = (t)^{-1}, \tag{7}$$

$$\eta_2(t) = (t)^{-0.51}, \tag{8}$$

$$\eta_3(t) = (h(t))^{-1}, \tag{9}$$

$$\eta_4(t) = (h(t))^{-0.51} \tag{10}$$

where $h(t) = t \times p(t)$ and $p(t) =$ the probability of misclassification of the classifier obtained at time t. $h(t)$ tends to give a higher value if the overlap between the two classes is larger. In other words, the change in $W(t)$ is large when the class overlap is small. Thus, cases (9) and (10) depend on the classification problem at hand while cases (7) and (8) do not.

Here, we deal with synthetic data where the two classes follow Gaussian distributions with known parameters. For learning, a sequence of observation vectors is generated in the following way. At time t, first class 0 or 1 is randomly selected each with probability 0.5 and then a feature vector $X(t)$ is randomly selected from the selected class. $W(t)$ is updated with $X(t)$ using Eq. (3). For the study of convergence of $W(t)$, two quantities are computed – percentage of misclassification $a(t)$ and a measure of disparity $b(t)$ between $W(t)$ and the weight vector corresponding to the Bayes classifier for the two-class problem under consideration. For computing the rate of misclassification, a testing set S of size $2n$ is generated a priori where n feature vectors are randomly selected from each of the two classes. Suppose $n_1(t)$ is the number of feature vectors in S coming from class 0 and satisfying $\sum_{i=1}^{p+1} w_i(t)x_i \geqslant 0$ and $n_2(t)$ is the corresponding number of feature vectors coming from class 1 and satisfying $\sum_{i=1}^{p+1} w_i(t)x_i < 0$. $(n_1(t) + n_2(t))$ gives the number of misclassified feature vectors in S resulting from the classifier determined by $W(t)$. $a(t)$ is computed as $100(n_1 + n_2)/2n$. The quantity $a(t)/100$ can be taken as an estimate of $p(t)$. The values of $b(t)$ are studied to see if $W(t)$ asymptotically gives the same classifier as the Bayes classifier.

4. Simulation results

We will simulate here two-class problems using Gaussian distributions in two dimensions. Two cases are considered here depending on whether the Bayes classifier is linear or non-linear. In the first case, the two classes have the same covariance matrices while in the second case they have different covariance matrices.

4.1. Linear case

The means of the two classes are $\mu_1 = (20, 40)^T$ and $\mu_2 = (80, 60)^T$ and the common covariance matrix is $\sigma^2 I$. Here, $p = 2$. Five different values of σ, namely, 5, 10, 15, 20 and 25 are considered in the present case. Before learning starts, a testing set S_1 of 2000 samples (1000 randomly selected from each of class 0 and class 1) is formed. From the population parameters, the Bayes classifier is computed as (w_1^*, w_2^*, w_3^*) where a feature vector $(x_1, x_2)^T$ is classified in class 0 if $w_1^* x_1 + w_2^* x_2 - w_3^* < 0$ and in class 1 otherwise. In the present example, $w_1^* = 0.015$, $w_2^* = 0.005$, $w_3^* = 1$.

It can be seen that, for updating $W(t)$, it is not possible to estimate $p(t)$ by computing $a(t)$ for all t since it will drastically slow down the learning process. A less accurate estimate of $p(t)$ is computed in the following way. We only take note of real updations, that is, the number of times when $d(t)$ and $y(t)$ (defined in Section 2) are different. Let $q(t)$ denote the total number of updations up to time t. Then $p(t)$ is approximated as $q(t)/t$ which asymptotically gives $p(t)$.

Now, at selected numbers of iterations, the following quantities are computed.

(i) $a(t)$ = percentage of misclassification of the perceptron classifier obtained at the tth iteration with respect to the testing set S_1. This is computed as $100(n_1 + n_2)/2n$.

(ii) $b(t) = \{[w_1(t)/w_3(t) - w_1^*/w_3^*]^2 + [w_2(t)/w_3(t) - w_2^*/w_3^*]^2\}^{1/2}$.

The value of $b(t)$ shows how close the perceptron classifier at the tth iteration is to the Bayes classifier. The values of $a(t)$ and $b(t)$ are computed at iterations 1, 20, 50, 250, 500, 1000, 5000, 10 000, 50 000, 100 000, 200 000, 300 000, 400 000, 500 000, 600 000, 700 000, 800 000, 900 000, 1 000 000 which are marked as 1, 2, ..., 19 respectively in Figs. 1 and 2 where the graphs corresponding to η_i are indicated by symbols with legends i ($i = 1, 2, 3, 4$). Figs. 1(a)–(c) show the values of $a(t)$ for all four choices of $\eta(t)$ for $\sigma = 5$, 15 and 25 respectively. Similarly, Figs. 2(a)–(c) show the values of $b(t)$ in such cases. In all these cases, the starting vector $W(0)$ is always taken as (0.01, −0.03, 1.0) which in fact represents the straight line joining μ_1 and μ_2 and is perpendicular to the straight line giving the Bayes boundary. This is to let the iterative process start from a position with high rate of misclassification (in fact, 50% in this case). From Figs. 1 and 2 it can be said on the whole that in terms of the rate of convergence of both $a(t)$ and $b(t)$, η_2,η_3 and η_4 perform nearly equally well and perform better than η_1. After 1 000 000 iterations the percentages of misclassification achieved by the four perceptron classifiers and that by Bayes classifier are given in Table 1 for $\sigma = 5$, 10, 15, 20 and 25. It can be concluded that the performance of the best perceptron is nearly as good as that of the Bayes classifier.

Table 1
Misclassification (%)

(σ)	Perceptron				Bayes
	η_1	η_2	η_3	η_4	
5	0.35	0.35	0.35	0.6	0.00
10	7.2	0.6	4.05	0.4	0.25
15	15.6	1.75	2.1	1.9	1.75
20	5.95	6.15	6.6	6.35	6.0
25	10.95	11.65	11	11.35	11.0

4.2. Nonlinear case

The means of the two classes are $\mu_1 = (40, 0)^T$ and $\mu_2 = (100, 0)^T$ and the covariance matrices are $\Sigma_1 = \sigma_1^2 I$ and $\Sigma_2 = \sigma_2^2 I$. The feature space here is made 5-dimensional indicated by $(x_1, x_2, x_1^2, x_2^2, x_1 x_2)$ making what is quadratic in 2 dimensions, linear in 5-dimensional space. Thus, $p = 5$ here. Three different sets of values for the pair (σ_1, σ_2) are considered – (10, 15), (15, 20) and (20, 25). Before learning starts, a testing set S_2 of 4000 samples (2000 randomly selected from each of the two classes 0 and 1) is formed. In the present case also $p(t)$ is approximated as in the linear case above.

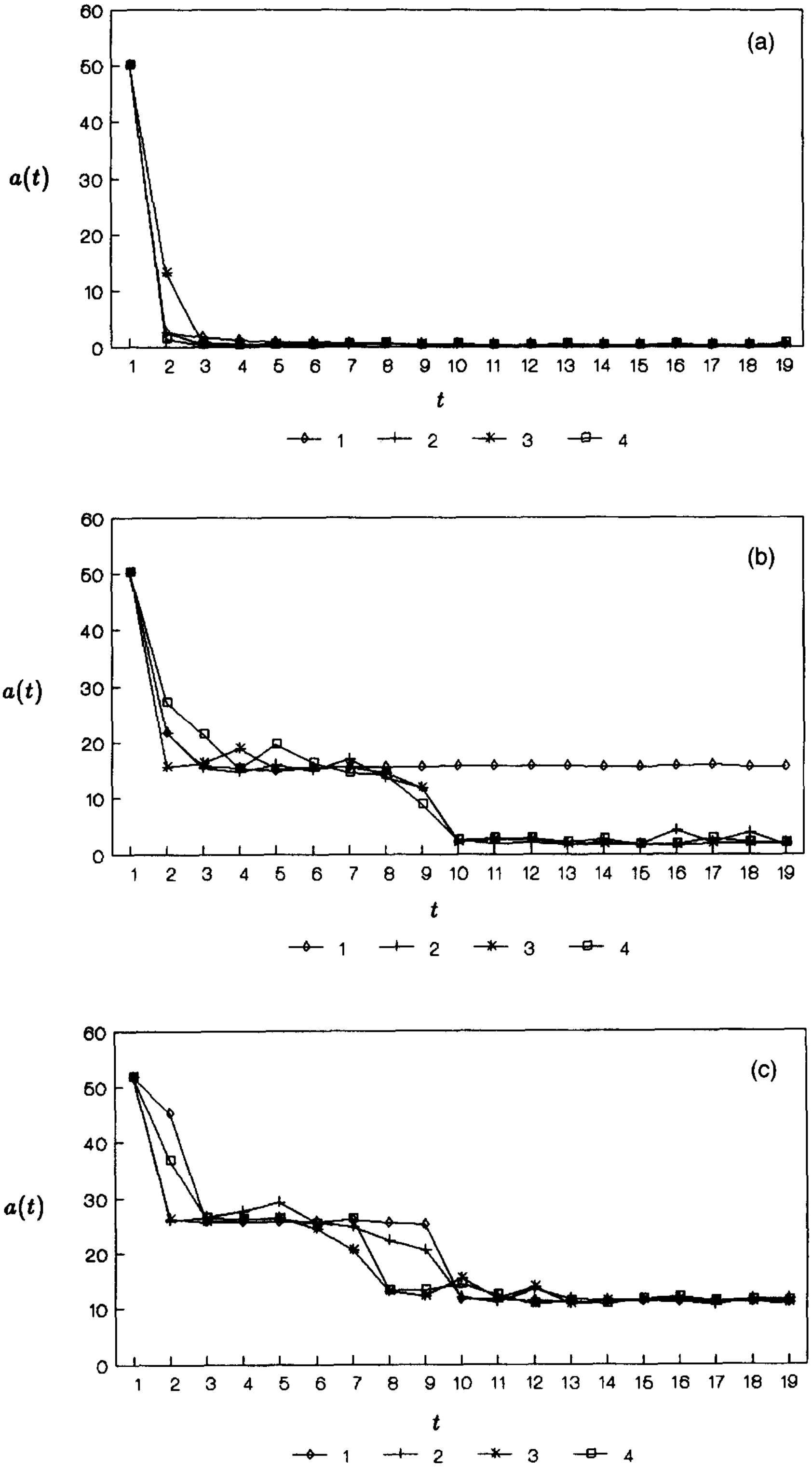

Fig. 1. t vs $a(t)$ graphs with (a) $\sigma = 5$, (b) $\sigma = 15$ and (c) $\sigma = 25$.

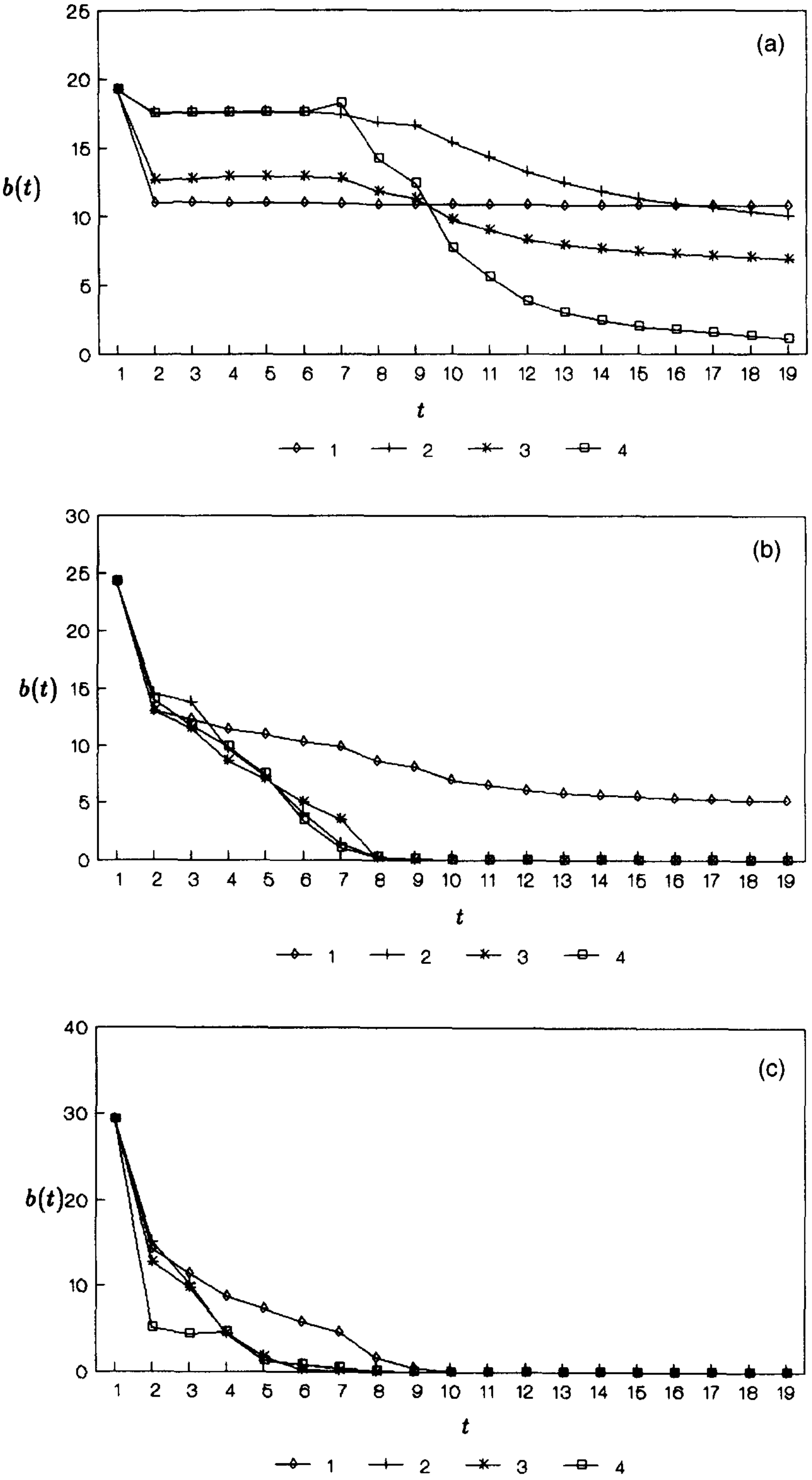

Fig. 2. t vs $b(t)$ graphs with (a) $\sigma = 5$, (b) $\sigma = 15$ and (c) $\sigma = 25$.

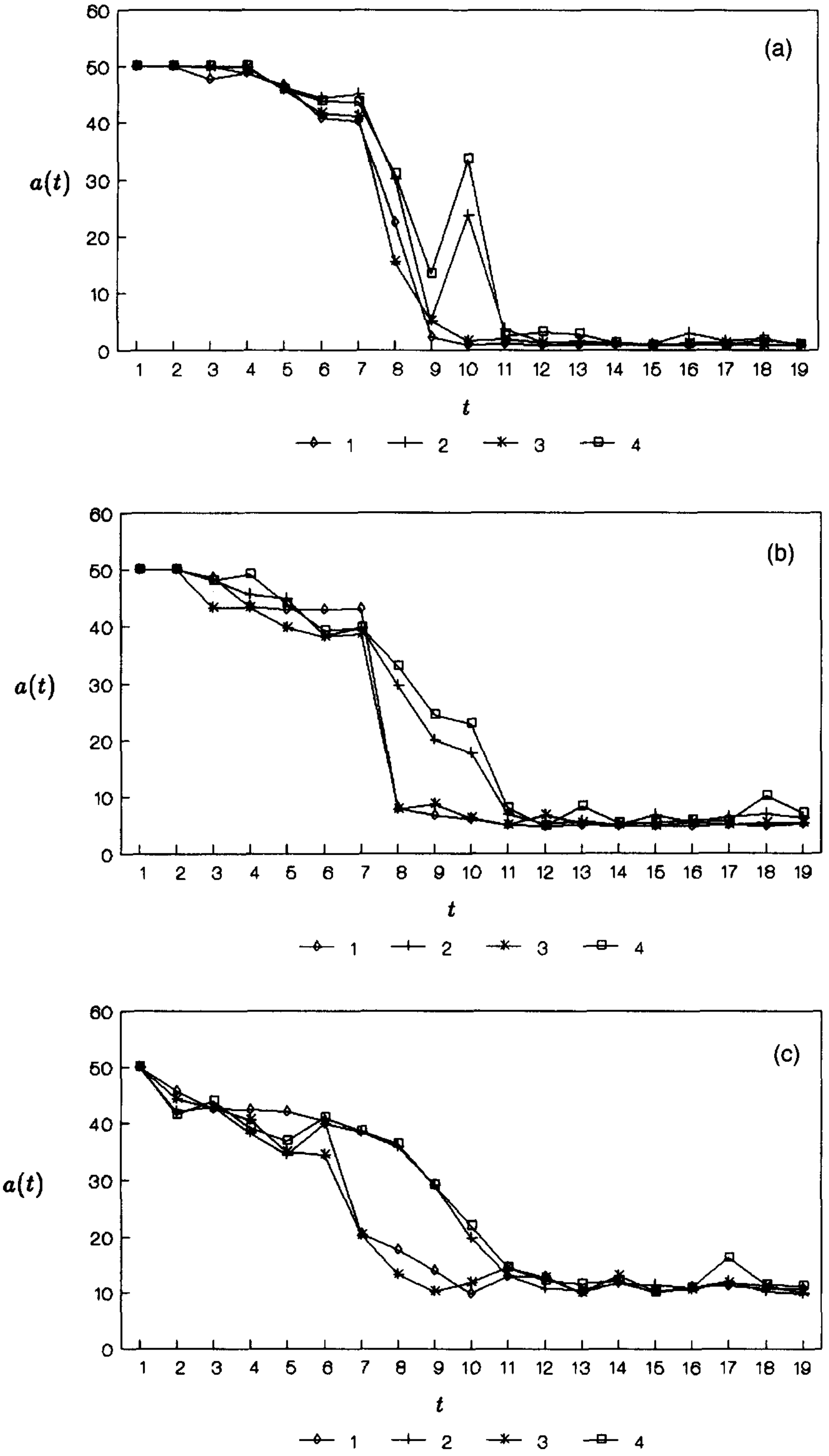

Fig. 3. t vs $a(t)$ graphs with (a) $\sigma_1 = 10$, $\sigma_2 = 15$, (b) $\sigma_1 = 15$, $\sigma_2 = 20$ and (c) $\sigma_1 = 20$, $\sigma_2 = 25$.

Table 2
Misclassification (%)

(σ_1, σ_2)	Perceptron				Bayes
	η_1	η_2	η_3	η_4	
(10, 15)	0.875	0.925	0.925	1.025	0.850
(15, 20)	5.15	6.225	5.225	7.05	4.400
(20, 25)	10.85	9.95	10.125	11.375	8.875

Here also at the same selected number of iterations the quantities $a(t)$ and $b(t)$ are computed with respect to the testing set S_2. It is seen from the values of $b(t)$ in each of the three cases that $W(t)$ does not converge to the Bayes classifier (quadratic). Figs. 3(a)–(c) show the values of $a(t)$ for η_1, η_2, η_3 and η_4 for the three different pairs of (σ_1, σ_2). In terms of the rate of convergence of $a(t)$, the four choices of $\eta(t)$ form two distinct groups – η_1, η_3 and η_2, η_4. The first group performs almost uniformly better than the second one. After 1 000 000 iterations the percentages of misclassification achieved by the four perceptron classifiers and that by the Bayes classifier are given in Table 2.

5. Conclusions

In both linear and non-linear cases above, η_3 performs better than (or at least equally well as) η_1, η_2 and η_4 in terms of rate of convergence of $W(t)$ measured on the basis of classification error. We have so far considered $\eta(t)$ which satisfies the conditions (4), (5) and (6). Now, there may be choices of $\eta(t)$ violating condition (5) or (6) where $W(t)$ can still converge. Our next plan is to consider such choices of $\eta(t)$ and the corresponding rate of convergence of $W(t)$. In the present paper the choice of g (Eq. (2)) has been taken as hard limiting non-linearity. Other choices of g (sigmoid non-linearity, for example) can be investigated.

Acknowledgement

The authors are thankful to Prof. B.B. Chaudhuri, the Head, Nodal Centre for Knowledge Based Computing, ISI for allowing us to use the VAX 2000 system of NCKBC.

References

Duda, R.O. and P.E. Hart (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, 2nd edition. Academic Press, New York.

Hush, D.R. and B. Horne (1993). Progress in supervised neural networks: what's new since Lippmann? *IEEE Signal Process. Mag.*, January.

Lippmann, R.P. (1987). An introduction to computing with neural nets. *IEEE Acoust. Speech Signal Process. Mag.* 4 (2), 4–22.

Rosenblatt, R. (1959). *Principles of Nuerodynamics*. Spartan Books, New York.