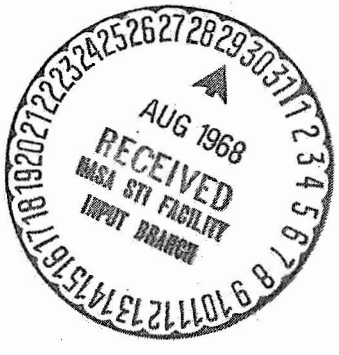
PURDUE UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

RECURSIVE ALGORITHMS FOR PATTERN CLASSIFICATION USING MISCLASSIFIED SAMPLES*

R. L. Kashyap

Technical Report No. TR-EE 68-14

June 1968



Lafayette, Indiana 47907

RECURSIVE ALGORITHMS FOR PATTERN CLASSIFICATION USING MISCLASSIFIED SAMPLES*

R. L. Kashyap

Technical Report No. TR-EE 68-14

June 1968

School of Electrical Engineering

Purdue University

Lafayette, Indiana 47907

*This work was partially supported by NSF under grant GK-1970 and by NASA under Grant NGR 15-005-021.

RECURSIVE ALGORITHMS FOR PATTERN CLASSIFICATION USING MISCLASSIFIED SAMPLES

R. L. Kashyap

ABSTRACT

We consider the samples $x(i)$ belonging to one of the two non-overlapping classes, ω_1 and ω_o, which possess a separating function $f(x)$. The observed membership of pattern $x(i)$ is represented by the variable $z(i)$ which can assume only one of two values, ± 1, or $z(i) = [\text{sgn } f(x(i))]\eta(i)$ where $\eta(i)$ is the measurement noise and $E(\eta)$ is known. Thus the membership of the training samples may be erroneous. Using only the available sample pairs $\{x(i), z(i)\}$, $i=1,2,\ldots$, we will obtain an optimal approximation to the separating function $f(x)$.

I. Introduction

There are a number of ways of posing the pattern classification problem. Here we assume that there are only two classes, ω_1 and ω_o, and that we are given a sequence of patterns whose classification is known to us. The given information regarding the classification may be erroneous. Based only on this information, we would like to develop decision functions to classify new patterns. Since the number of available samples may be large, it is convenient to insist that the method of computation of the decision function be recursive. After the appearance of each training pattern, we should be able to update the decision function in real time.

It is convenient to associate every pattern with a m-dimensional vector x known as the pattern vector. The components $x_1, x_2, \ldots, x_m$ of the pattern vector are referred to as the attributes or features of the pattern. The selection of the features associated with a set of patterns is a problem of considerable complexity. However, we assume the knowledge of some of the features associated with the patterns of the given problem. With every pattern we associate a scalar variable y, the so-called true class indicator, whose value indicates the class to which the pattern x belongs. The variable y assumes only one of two values, $\pm$ 1.

$$y = 1 \quad \text{if } x \in \omega_1$$

$$y = -1 \quad \text{if } x \in \omega_o$$

Without any loss of generality, we can regard x and y as random variables with an unknown probability density function $p(x,y)$. Based on the nature of the conditional probability function $p(y/x)$, we can divide the classification problems into two categories. The first category consists of examples in which the class indicates variable y is a <u>deterministic</u> function of x; i.e., the probability function $p(y/x)$ can assume only one

of two values, 0 or 1. In the second category, the variable y is a probabilistic function of x; i.e., $0 < p(y/x) < 1$. A typical example of the first category is the alpha-numeric character recognition problem in which every pattern has a unique class membership. A typical example of the second category is the detection of a stochastic signal in the presence of noise.

We shall be concerned here only with the problems of the first category, the so-called deterministic classification problems. In other words, the classes are not overlapping in these problems. Hence, in every problem there exists a separating function f(x) (not necessarily unique) so that

$$\begin{aligned} f(x) &> \epsilon > 0 \qquad x \in \omega_1 \\ &< -\epsilon \qquad x \in \omega_o \end{aligned} \tag{1.1}$$

Recalling the definition of the variable y

$$\begin{aligned} y = \operatorname{sgn} f(x) &= +1 \quad \text{iff } x \in \omega_1 \\ &= -1 \quad \text{iff } x \in \omega_o \end{aligned} \tag{1.2}$$

Most of the problems posed in the literature can be summarized as follows: given a sequence of patterns with known classification; i.e., the sample pairs $\{x(i), y(i)\}$, $i=1,2,\ldots$, find a decision function d(x) which is either a separating function satisfying (1.1) or a "best" approximation to a separating function f(x) satisfying (1.1). A number of recursive algorithms have been developed to solve this problem like the fixed increment or perceptron algorithm of Novikov[1], the relaxation and orthogonal projection algorithms of Agmon[2], the modified minimum mean square error algorithms of Ho and Kashyap[3], etc. It is important to note that these algorithms lead to meaningful results only if the classification of all the patterns x(1), x(2), etc, is known exactly. One does not know the behavior of the

algorithms if the information regarding the classification of even a few of the samples is in error. An interesting question is whether we can develop algorithms which lead to meaningful results even if there are a few errors in the classification of the given training samples. This paper is concerned only with this topic.

II. Statement of the Problem

Let z represent the observed (noisy) class indicator variable

$$z = +1 \quad \text{if } x \text{ is observed to be in class } \omega_1$$
$$= -1 \quad \text{if } x \text{ is observed to be in class } \omega_0$$

Recalling the definition of y, the true class indicator, we can represent z as

$$z = y\,\eta$$
$$= [\operatorname{sgn} f(x)]\,\eta \tag{2.1}$$

where f(x) is the separating function defined in (1.1) and η is the measurement noise which can take only one of two values, ± 1. The properties of the process η will be mentioned later.

The problem is to determine a "best" approximation $\hat{f}(x)$ to a separating function f(x) obeying (1.1) based only on the available sample pairs $\{x(i), z(i)\}$, i=1,2,..., where z(i) is the measurement associated with the i^{th} training pattern x(i). The function $\hat{f}(x)$ mentioned herein will be used as the decision function for classifying new samples in the following manner:

$$\hat{f}(x) > 0 \qquad x \text{ will be placed in class } \omega_1$$
$$< 0 \qquad x \text{ will be placed in class } \omega_0 \tag{2.2}$$

To determine $\hat{f}(x)$, we adopt the following procedure. We chose $\hat{f}(x)$ among the functions defined in (2.3).

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i \, \omega_i(x) = \alpha^T \varphi(x) \tag{2.3}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are the undetermined coefficients and $\omega_1(x), \ldots, \omega_n(x)$ are known linearly independent functions in the variables $x_1, x_2, \ldots, x_m$. The coefficients $\alpha_1, \ldots, \alpha_n$ have to be chosen to minimize the error caused by the use of $\hat{f}(x)$ as the decision function for classifying the new patterns using the rule (2.2). It is clear that all samples x for which sgn $f(x) \neq$ sgn $\hat{f}(x)$ will be classified erroneously by the decision function $\hat{f}(x)$. Consider the expectation of the misclassified patterns in the 'φ' space

$$E\,[(\text{sgn } f(x) - \text{sgn } \hat{f}(x))\, \varphi(x)] \tag{2.4}$$

A reasonable way of choosing α is to select it so that the expected value of the misclassified patterns given in (2.4) is zero.

$$E\,[(\text{sgn } f(x) - \text{sgn } (\alpha^T x))\, \varphi_i(x)] = 0 \quad , \quad i=1,\ldots,n \tag{2.5}$$

Let a solution of (2.5) be α^* and the corresponding optimal decision function is given in (2.6).

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i^* \, \varphi_i(x) \tag{2.6}$$

Out intention is to solve the equation (2.5) for α^* using only the available sample pairs $\{x(i), z(i)\}$, $i=1,2,\ldots$.

III. Algorithm

To find the solution α^* of equations (2.5) in a recursive manner, we could use Newton's method.

$$\alpha(i+1) = \alpha(i) + \rho(i)\, E[\{\text{sgn } f(x) - \text{sgn}(\alpha^T(i)\varphi(x(i))^T\}\varphi(x(i))/\alpha(i)] \tag{3.1}$$

where $\rho(i)$ is the scalar representing the step size. Directly, (3.1) is of very little use since there is no way of evaluating the expectation

term in it. Hence, we will replace the expectation term by an observable random vector whose expectation is identical with that in (3.1). To do this we need the following assumptions on the stochastic processes $\eta(i)$ and $x(i)$.

(A1) The patterns $x(1), x(2), x(3)$ ... and the noise variables $\eta(1), \eta(2), \ldots$ are all statistically independent of one another.

(A2) $E(\eta) \triangleq \hat{\eta} \neq 0$

Under these conditions we can replace the expectation term in (3.1) by the following random variable.

$$\left[\frac{1}{\hat{\eta}} \{ z(i) - \hat{\eta} \, \mathrm{sgn}(\alpha^T(i)\varphi(x(i))\} \varphi(x(i)) \right] \tag{3.2}$$

To demonstrate the equivalence of the two expectations, we note that

$$E\left[\frac{1}{\hat{\eta}} \{ z - \hat{\eta} \, \mathrm{sgn} \, (\alpha^T \varphi(x))\} \, \varphi(x)/\alpha \right]$$

$$= E\left[\{ \mathrm{sgn} \, f(x) - \mathrm{sgn} \, (\alpha^T \varphi(x))\} \, \varphi(x)/\alpha \right]$$

$$+ \frac{1}{\hat{\eta}} E\left[(\eta - \hat{\eta}) \, (\mathrm{sgn} \, f(x)) \, \varphi(x)/\alpha \right]$$

$$= E\left[\{ \mathrm{sgn} \, f(x) - \mathrm{sgn} \, (\alpha^T \varphi(x))\} \, \varphi(x)/\alpha \right]$$

by assumptions (A1) and (A2). Hence, algorithm (3.1) becomes

$$\alpha(i+1) = \alpha(i) + \rho(i) \frac{1}{\hat{\eta}} [z(i) - \hat{\eta} \, \mathrm{sgn} \, (\alpha^T(i)\varphi(i))] \, \varphi(i) \tag{3.3}$$

To "compensate" for the error involved in replacing the expected value by the random variable in (3.2), the gain sequence $\rho(i)$ should obey (3.4)

$$\rho(i) > 0 \quad , \quad \sum_i \rho(i) = \infty \quad , \quad \sum_i \rho^2(i) < \infty \tag{3.4}$$

In order to establish the convergence of the algorithm (3.3), we need an additional assumption (A3) in addition to the assumptions (A1) and (A2) already made.

$$\text{(A3)} \quad \inf_{x \in \omega_o \cup \omega_1} |\alpha^{*T}\varphi(x) - \alpha^T\varphi(x)| = g_1(\alpha, \alpha^*) > 0 \qquad \forall\, \alpha \neq \alpha^*$$

$$E\,\|\varphi(x)\| < \infty$$

$$E\,\|f(x)\| < \infty$$

We will briefly remark on the assumptions. Assumptions (A1) and (A3) do not require any additional explanation. Consider assumption (A2). When $\hat{\eta} = 0$, the probability of a training sample having correct classification is 0.5. In other words, we are not given any information at all. It is not surprising that the algorithm may not converge to the desired value in this case. Let us consider the cases $\hat{\eta} > 0$ and $\hat{\eta} < 0$ separately. When $\hat{\eta} > 0$, the percentage of training samples with correct classification is greater than that with wrong classification. It is easy to guess that the algorithm will work here. But what may be surprising, at first sight, is the convergence of the algorithm to the desired value when $\hat{\eta} < 0$. However, we notice that when $\hat{\eta} < 0$, we work with the complements $\bar{z}$ and $\bar{\eta}$ instead of working directly with z and η and $E(\bar{\eta}) = -E(\eta) > 0$. This is possible since we are dealing with only two classes.

The convergence properties of the algorithm (3.3) can be represented in the form of a proposition.

Proposition: Consider the algorithm (3.3) with the gain $\rho(i)$ obeying the condition (3.4). Under assumptions (A1)-(A3), $\alpha(i)$ will tend to α^* in the mean square sense; i.e.,

$$\lim_{i \to \infty} E\,|\mathrm{sgn}(\alpha^{*T}\varphi(x)) - \mathrm{sgn}(\alpha^T(i)\varphi(x))|^2 = 0 \tag{3.5}$$

where α^* is the solution of the equation

$$E\left[\{\mathrm{sgn}\, f(x) - \mathrm{sgn}\,(\alpha^T\varphi(x))\}\,\varphi_i(x)\right] = 0 \tag{3.6}$$

$$i = 1, 2, \ldots, n$$

In particular, if

$$f(x) = \sum_{i=1}^{n} \theta_i \varphi_i(x) \tag{3.7}$$

then $\alpha^* = \theta$

IV. Proof of the Proposition

Let $\tilde{\alpha}(i) = \alpha^* - \alpha(i)$

$\tilde{\eta}(i) = (\eta(i)-\hat{\eta})/\hat{\eta}$

$\tilde{f}(x) = [(\text{sgn } f(x) - \text{sgn } (\alpha^{*T}\varphi(x))]$

$\varphi(x(i)) \triangleq \varphi(i)$

We can rewrite (3.3) as

$$\tilde{\alpha}(i+1) = \tilde{\alpha}(i) - \rho(i)\left[\text{sgn}(\alpha^{*T}\varphi(i)) - \text{sgn}\,(\alpha^{T}(i)\varphi(i))\right]\varphi(i)$$

$$- \rho(i)\,\tilde{\eta}(i)\left[\text{sgn}(\alpha^{*T}\varphi(i))\right]\varphi(i) - \rho(i)\,\frac{\eta(i)}{\hat{\eta}}\,\tilde{f}(x(i))\varphi(i) \tag{3.7}$$

Scalar multiply (3.7) with itself and take conditional expectation keeping $\alpha(i)$ fixed.

$$E\left\{\|\tilde{\alpha}(i+1)\|^2/\alpha(i)\right\} = \|\alpha(i)\|^2 - 2\rho(i)\,E\left[\left\{\text{sgn}(\alpha^{*T}\varphi(i)) - \text{sgn}\,(\alpha^{T}(i)\varphi(i))\right\}\tilde{\alpha}^{T}(i)\varphi(i)/\alpha(i)\right] -$$

$$2\rho(i)\,E\left[\left\{\text{sgn}(\alpha^{*T}\varphi(i))\right\}\tilde{\alpha}^{T}(i)\varphi(i)\tilde{\eta}(i)/\alpha(i)\right] -$$

$$2\rho(i)\,E\left[\frac{\eta(i)\tilde{\alpha}^{T}(i)}{\hat{\eta}}\,\tilde{f}(x(i))\varphi(i)/\alpha(i)\right] +$$

$$\rho^2(i)\Big[E\Big\{(\text{sgn}(\alpha^{*T}\varphi(i)) - \text{sgn}(\alpha^{T}(i)\varphi(i)))^2\|\varphi(i)\|^2/\alpha(i)\Big\} +$$

$$E\Big\{(\tilde{\eta}(i))^2\|\varphi(i)\|^2/\alpha(i)\Big\} + E\Big\{\frac{\eta^2(i)}{(\hat{\eta})^2}(\tilde{f}(x(i)))^2\|\varphi(i)\|^2/\alpha(i)\Big\}\Big]$$

... (3.8)

Recall that $\alpha(i)$ is a function of the samples $x(i-1)$, $x(i-2)$, $x(i-3)$,... and $\eta(i-1)$, $\eta(i-2)$.... By assumption (A1) and (A2), the third term in (3.8) is zero. By definition of α^* in (3.6) and (A1), the fourth term in (3.8) is zero. By assumption (A3)

$$\text{fourth term in (3.8)} \leq \rho^2(i)k_1 \tag{3.9}$$

where $\infty > k_1 > 0$.

Thus (3.8) becomes

$$E\left(\|\tilde{\alpha}(i+1)\|^2/\alpha(i)\right) \leq \|\tilde{\alpha}(i)\|^2 - 2\rho(i)\, E\Big[\left\{\text{sgn}(\alpha^{*T}\varphi(i)) - \text{sgn}(\alpha^T(i)\varphi(i))\right\} \tilde{\alpha}^T(i)\varphi(i)/\alpha(i)\Big] + \rho^2(i)k_1 \tag{3.10}$$

Let us simplify the second term in (3.10). By inspection

$$\left\{\text{sgn}(\alpha^{*T}\varphi(x)) - \text{sgn}(\alpha^T\varphi(x))\right\}\left\{\alpha^{*T}\varphi(x) - \alpha^T\varphi(x)\right\}$$

$$= 0 \qquad \text{if sgn } (\alpha^{*T}\varphi(x)) = \text{sgn } (\alpha^T\varphi(x))$$

$$= 2|\alpha^{*T}\varphi(x) - \alpha^T\varphi(x)| \qquad \text{if sgn } (\alpha^{*T}\varphi(x)) \neq \text{sgn } (\alpha^T\varphi(x))$$

$$= |\text{sgn}(\alpha^{*T}\varphi(x)) - \text{sgn } (\alpha^T\varphi(x))|\ |\alpha^{*T}\varphi(x) - \alpha^T\varphi(x)| \tag{3.11}$$

Using (A3) and (3.11), we get

$$E\left[\left\{\text{sgn}(\alpha^{*T}\varphi(x)) - \text{sgn}(\alpha^T\varphi(x))\right\}\left\{\alpha^{*T}\varphi(x) - \alpha^T\varphi(x)\right\}/\alpha\right]$$

$$\geq g_1(\alpha,\alpha^*)\ h(\alpha,\alpha^*) \tag{3.12}$$

where

$$h(\alpha,\alpha^*) = \underset{x}{E}\left\{|\text{sgn}(\alpha^{*T}\varphi(x)) - \text{sgn}(\alpha^T\varphi(x))|\ /\alpha\right\} \tag{3.13}$$

Substituting (3.12) in (3.10), we get

$$E\left\{\|\tilde{\alpha}(i+1)\|^2\ /\alpha(i)\right\} \leq \|\tilde{\alpha}(i)\|^2 - 2\rho(i)h(\alpha(i),\alpha^*)g_1(\alpha(i),\alpha^*) + \rho^2(i)k_1$$

... (3.14)

By using the property of $\rho(i)$ in (3.4), we can rewrite (3.14) as

$$E\left(\|\tilde{\alpha}(i+1)\|^2/\alpha(i)\right) \leq \|\tilde{\alpha}(i)\|^2 - (2-\epsilon_1)\rho(i)h(\alpha(i),\alpha^*)g_1(\alpha(i),\alpha^*)$$

$$\forall\, i > i_1, \tag{3.15}$$

where $0 < \epsilon_1 < 2$.

Taking expectations over $\alpha(i)$ in (3.18), we get

$$E\left(\|\tilde{\alpha}(i+1)\|^2\right) \leq E\,\|\tilde{\alpha}(i)\|^2-(2-\epsilon_1)\rho(i)\;E\left[h(\alpha(i),\alpha^*)g_1(\alpha(i),\alpha^*)\right] \tag{3.16}$$

Using (3.16) repeatedly, we get

$$(2-\epsilon_1)\sum_{j=1}^{i}\rho(j)\;E(h(\alpha(j),\alpha^*)g_1(\alpha(j),\alpha^*)) \leq E\,\|\tilde{\alpha}(1)\|^2-E\,\|\tilde{\alpha}(i+1)\|^2$$

Thus

$$\lim_{i\to\infty}\sum_{j=1}^{i}\rho(j)\;E(h(\alpha(j),\alpha^*)g_1(\alpha(j),\alpha^*)) < \infty \tag{3.17}$$

Since $\sum_j \rho(j) = \infty$, (3.17) implies that

$$\lim_{j\to\infty} E(h(\alpha(j),\alpha^*)g_1(\alpha(j),\alpha^*)) = 0 \tag{3.18}$$

Since $h(\alpha(j),\alpha^*), g(\alpha(j),\alpha^*)$ are non-negative for all $\alpha(j)$, (3.18) is valid if <u>one</u> of the following equalities (3.19) <u>or</u> (3.20) is valid.

$$\lim_{j\to\infty} E(h(\alpha(j),\alpha^*)) = 0 \tag{3.19}$$

$$\lim_{j\to\infty} E(g_1(\alpha(j),\alpha^*) = 0 \tag{3.20}$$

Among these, (3.19) is the weaker condition. Hence,

$$\lim_{j\to\infty} E\,|\mathrm{sgn}(\alpha^T(j)\varphi(x)) - \mathrm{sgn}(\alpha^{*T}\varphi(x))| = 0$$

This completes the proof of the proposition.

V. Discussion

We have developed an algorithm for finding an optimal decision function for classification using training samples which may be erroneously classified. It is necessary to have a knowledge of the average value of the corrupting noise. The use of the gain sequence $\rho(i)$ as in (3.4) should also be mentioned. If $\rho(i)$ were a constant for all i, algorithm (3.3) would be nothing but the perceptron algorithm[1] and it would not converge to any meaningful result when we have erroneous training samples.

We would like to stress the fact that our algorithm does not need the knowledge of the structure of the separating function. It always converges to a surface which renders the expected values of the erroneously classified patterns zero. If, in addition, the separating function has a known linear representation as in (3.7), the algorithm (3.3) will lead to a separating function.

A comment on the proof of convergence may be in order. Even though the algorithm (3.3) with the gains as in (3.4) may appear to belong to the family of stochastic approximation algorithms, its proof of convergence does not directly follow from the proofs of stochastic approximation algorithms [Venter[4], Gladyshev[5], Albert and Gardner[6]] on account of the discontinuous nature of the correction term in (3.3).

REFERENCES

1. A. Novikoff, "On Convergence Proofs for Perceptrons", Proc. 1962, Symp. on Mathematical Theory of Automata, Polytechnic Institute of Brooklyn, pp. 615-622.

2. S. Agmon, "The Relaxation Method for Linear Inequalities", Canadian J. Math, Vol. 6, (1956), pp. 382-392.

3. Y. C. Ho and R. L. Kashyap, "A Class of Iterative Procedures for Linear Inequalities", J. SIAM on Control, Vol. 4, No. 1, 1966, pp. 112-115.

4. J. H. Venter, "On Dvoretzsky Stochastic Approximation Theorems", Ann. Math. Stat., Vol. 37, No. 4, 1966, pp. 1534-1544.

5. E. G. Gladyshev, "On Stochastic Approximation", Theory of Probability and Its Applications, Vol. 10 (1965), No. 2.

6. A. E. Albert and L. A. Gardner, Jr., Stochastic Approximation and Nonlinear Regression", M.I.T. Press, 1967.

7. E. M. Braverman, "On the Potential Function Method", Automation and Remote Control, Vol. 26 (1965), pp. 2205-2213.