INFORMATION SCIENCES 53, 191–201 (1991)

Effect of Wrong Samples on the Convergence of Learning Processes

AMITA PAL (PATHAK)

and

SANKAR K. PAL

Software Technology Branch, NASA, Johnson Space Center, Houston, Texas 77058

ABSTRACT

For the problem of parameter learning in pattern recognition, when there is a possibility of training samples being mislabeled, the authors have investigated the convergence of stochastic-approximation-based learning algorithms. In the cases considered, it is found that estimates converge to nontrue values in the presence of labeling errors. The general m-class, N-feature pattern recognition problem is considered.

I. INTRODUCTION

The learning of unknown parameters of classifiers is an indispensible part of pattern recognition problems. If a sufficiently large set of correctly labeled training samples is available, then "reasonably good" estimates of the parameters can generally be obtained. In many real-life situations, however, it is either difficult or expensive to obtain labels, so that mislabeling of training samples can become one of the spectres a pattern recognition scientist has to contend with. It is therefore useful to know how this problem can affect the learning procedure. A reasonable amount of work has been done for the two-class classification problem. The effects of random training errors on Fisher's discriminant function have been studied by Lachenbruch [1, 2], McLachlan [3], Michalek and Tripathi [4], O'Neill [5], and Krishnan [6]. They concluded that the effect is to underestimate distances, overestimate error rates, introduce bias into estimates of the discriminant function, make the maximum-likelihood estimates of the discriminant function converge to nontrue values, and affect the asymptotic relative efficiency (ARE) relative to a completely correctly classified sample of the same size.

655 Avenue of the Americas, New York, NY 10010 0020-0255/91/$03.50

In the context of recursive learning of parameters, the usefulness of stochastic approximation procedures cannot be over emphasized [7]. Briefly, a stochastic approximation procedure for recursively estimating by $\hat{\theta}_n$ a parameter θ by means of an unbiased statistic T, at the nth step, is

$$\hat{\theta}_{n+1} = \hat{\theta}_n - a_n(\hat{\theta}_n - \mathrm{T}_{n+1}),$$

where either $\hat{\theta}_1$ is a constant or $\hat{\theta}_1 = \mathrm{T}_1$, and $\{a_n\}$ is a suitably chosen sequence of positive numbers. For instance, as a recursive procedure for estimating the population mean μ of a variable x with the help of the sample mean $\bar{x}$, we can choose

$$\bar{\mathrm{x}}_{n+1} = \bar{\mathrm{x}}_n - \frac{1}{n}(\bar{\mathrm{x}}_n - \mathrm{X}_{n+1}),$$

X_{n+1} being the $(n+1)$th observation on X.

In this paper, the particular case in which errors occur in the labeling of training samples is studied for an m-class, N-feature pattern recognition problem. The effect of mislabeling is to cause wrong samples to be used in the recursive learning of the estimates, for any given class. A simple but realistic model [8] is adopted to describe this sort of situation. Under this model, the authors have investigated the convergence of recursive learning procedures of the type mentioned above. It is found that under certain conditions, these estimates do converge strongly (that is, with probability one), but to nontrue values—to be more specific, to convex linear combinations of true parameters of all m classeses. This is obtained using some results on multidimensional stochastic approximation [9].

II. STATEMENT OF THE PROBLEM

Let us consider a general m-class pattern recognition problem, C_i $(i = 1, \ldots, m)$ being the m classes, for which an N-dimensional feature vector

$$\underset{N\times 1}{\mathrm{X}} = [X_1, X_2, \ldots, X_N]', \qquad \mathrm{X} \in \mathbb{R}^N,$$

has been specified. Let us assume:

(A1) the distribution of X in each class is continuous;

(A2) the probability densities $p(\cdot|C_i)$ of X for the classes C_i, $i = 1, \ldots, m$, are of the same family and differ only in values of parameters.

Let $d(\cdot)$ be a decision function based on $\mathbf{X}$, i.e., let

$$d: \mathbb{R}^N \to \mathbb{R},$$

and let it depend only on the $p(\cdot\,|C_i)$, $i = 1, \ldots, m$. For each i, let

$$\underset{q \times 1}{\varphi_i} = [\varphi_{1i}, \varphi_{2i}, \ldots, \varphi_{qi}]'$$

be the vector of unknown parameters of $p(\cdot\,|C_i)$ and hence $d_i(\cdot)$.

Let us further assume:

(A3) an unbiased statistic exists for the parameter vector φ with respect to the probability density function p.

Let us suppose that for the purpose of learning we have been given a set of independent samples $\mathbf{X}_1^{(k)}, \mathbf{X}_2^{(k)}, \ldots, \mathbf{X}_{n_k}^{(k)}$, $k = 1, \ldots, m$, where the superscripts k denotes the labels given to the respective samples. For the learning itself, let us utilize a stochastic approximation algorithm LA1 defined below:

LA1. Let $\hat{\varphi}_t^{(k)}$ denote the estimate obtained at the tth step for the class C_k. Then

$$\hat{\varphi}_1^{(k)} = \mathbf{f}\big(\mathbf{X}_1^{(k)}\big), \tag{1a}$$

and for $t \geqslant 1$,

$$\hat{\varphi}_{t+1}^{(k)} = \hat{\varphi}_t^{(k)} - a_t\big[\hat{\varphi}_t^{(k)} - \mathbf{f}\big(\mathbf{X}_{t+1}^{(k)}\big)\big] \tag{1b}$$

for $k = 1, \ldots, m$, where $\{a_t\}$ is a sequence of positive real numbers such that $a_t \leqslant 1\ \forall t$, and $\mathbf{f}: \mathbb{R}^N \to \mathbb{R}^q$ is an unbiased statistic for φ.

III. A MODEL FOR LABELING ERRORS

The model to be used for this purpose was developed by Chittineni [8]. It can be specified as follows:

Let w and $\hat{w}$ denote respectively the true and the given labels. Clearly,

$$w, \hat{w} \in \{1, 2, \ldots, m\}.$$

Let $\pi_i = P[w = i]$ denote the *a priori* probability for the class C_i, $i = 1, \ldots, m$. Further, let $p_i(\mathbf{X}) = p(\mathbf{X}|w = i)$ be the class-conditional density of the feature

vector $\mathbf{X}$ for C_i. Also, let β_{ji} denote the probability that a sample from C_j has been given the label i, i.e.,

$$\beta_{ij} = P(\hat{w} = i | w = j), \qquad i, j = 1, \ldots, m. \tag{2}$$

Clearly, we must have

$$\sum_{i=1}^{m} \beta_{ij} = 1 \qquad \text{for every } j, \tag{3a}$$

i.e.,

$$\mathbf{B}'\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}, \tag{3b}$$

where

$$\underset{m \times 1}{\boldsymbol{\varepsilon}} = [1 \quad 1 \quad 1 \quad \cdots \quad 1]'$$

and

$$\mathbf{B} = \big((\beta_{ij})\big).$$

Under the above setup, from [8] we have

$$p(\mathbf{X}|\hat{w} = i) = \sum_{j=1}^{m} \alpha_{ij} p(\mathbf{X}|\hat{w} = i, w = j), \tag{4}$$

where

$$\alpha_{ij} = \frac{\pi_j \beta_{ij}}{\sum_{l=1}^{m} \pi_l \beta_{il}}. \tag{5}$$

If we are prepared to assume

(A4) $p(\mathbf{X}|w = j) = p(\mathbf{X}|\hat{w} = i, w = j)\ \forall i, j$

then Equation (4) becomes

$$p(\mathbf{X}|\hat{w} = i) = \sum_{j=1}^{m} \alpha_{ij} p(\mathbf{X}|w = j)$$

$$= \sum_{j=1}^{m} \alpha_{ij} p_j(\mathbf{X}). \tag{6}$$

IV. CONVERGENCE OF THE LEARNING ALGORITHM

The convergence of a recursive discrete algorithm $\hat{\theta}_n$ for estimating a parameter θ can be defined in various ways. For instance, we say that

DEFINITION. The sequence $\{\hat{\theta}_n\}$ converges to θ *with probability* 1 or *almost surely* (in symbols, $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta$) if

$$P\left[\lim_{n\to\infty} \|\hat{\theta}_n - \theta\| = 0\right] = 1.$$

For studying the behavior of the learning algorithms LA1 given in Section 2, use will be made of the following results, due to Schmetterer [9]:

LEMMA 1. *Let $\{a_n\}$ be a sequence of positive real numbers such that*

(C1) $\sum_{n=1}^{\infty} a_n^2 < \infty$.

Let $\mathbf{x}_n$ and $\mathbf{y}_n$ be k-dimensional random vectors which satisfy

(C2) $\mathbf{x}_{n+1} = \mathbf{x}_n - a_n \mathbf{y}_n$, $n \geq 1$.

Let M_n be a measurable mapping from $\mathbb{R}^k$ to $\mathbb{R}^k$ such that

(C3) $E(\mathbf{y}_n | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \mathbf{M}_n(\mathbf{x}_n)$ *a.e.*

Let a, b, c, be nonnegative real numbers, and let

(C4) $E(\|\mathbf{y}_n\|^2)\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \leq a + b\|\mathbf{x}_n\| + c\|\mathbf{x}_n\|^2$ *a.e.*

Also, for every $\mathbf{x} \in \mathbb{R}^k$ and $n \geq 1$, let

(C5) $\mathbf{x}'\mathbf{M}_n(\mathbf{x}) \geq 0$.

If $\mathbf{x}_1$ is chosen in such a way that

(C6) $E(\|\mathbf{x}_1\|^2)$ *exists,*

then the sequence $\{\mathbf{x}_n\}$ converges with probability 1, *i.e. almost surely, and the sequence $E\{\|\mathbf{x}_n\|^2\}$ converges also.*

LEMMA 2. *Suppose that conditions* (C1)–(C6) *hold. If, further, there exists for every $\eta > 0$ a $\delta > 0$ such that for $n \geq 1$*

(C7) $\inf_{\eta < \|\mathbf{x}\| \eta^{-1}} [\mathbf{x}'\mathbf{M}_n(\mathbf{x})] \geq \delta$,

then $\{\mathbf{x}_n\}$ converges almost surely to the k-dimensional null vector $\mathbf{0}$.

We shall prove the following:

PROPOSITION 1. *Consider the setup given in Sections* 2 *and* 3. *If, in addition to assumptions* (A1)–(A4), *we also have*

(A5) $\sum_{n=1}^{\infty} a_n^2 < \infty$,

(A6) $\rho_i = E(\|\mathbf{f}(\mathbf{X})\|^2 | w = i)$ *exists, with respect to each class-conditional density* $p_i(\mathbf{X})$,

then

$$\hat{\varphi}_i^{(k)} \xrightarrow{\text{a.s.}} \sum_{j=1}^{m} \alpha_{kj} \varphi_j,$$

where α_{ij}, $i, j = 1, \ldots, m$, *are as in Equation* (5).

Proof of Proposition. The validity of the proposition can be inferred directly from Lemmas 1 and 2, provided one can show that conditions (C1)–(C7) hold for

$$\psi_i^{(k)} = \varphi_i^{(k)} - \sum_{j=1}^{m} \alpha_{kj} \varphi_j.$$

We note that from Equations (1a) and (1b) we have, for $k = 1, \ldots, m$,

$$\psi_t^{(k)} = \begin{cases} \mathbf{g}_k\left(\mathbf{X}_1^{(k)}\right) & \text{for} \quad t = 1, \qquad (7a) \\ \psi_t^{(k)} - a_t\left(\psi_t^{(k)} - \mathbf{g}_k\left(\mathbf{X}_{t+1}^{(k)}\right)\right) & \text{otherwise}, \qquad (7b) \end{cases}$$

where

$$\mathbf{g}_k(\mathbf{X}) = \mathbf{f}(\mathbf{X}) - \sum_{j=1}^{m} \alpha_{kj} \varphi_j.$$

Condition (C1) is seen to be true because of (A5). Condition (C2) is equivalent to Equation (7b).

Condition (C3) also holds, with $M_t(\mathbf{x}) = \mathbf{x}$, since

$$E\left[\psi_t^{(k)} - \mathbf{g}_k\left(\mathbf{X}_{t+1}^{(k)}\right) \middle| \psi_1^{(k)}, \psi_2^{(k)}, \ldots, \psi_t^{k}\right]$$

$$= \psi_t^{(k)} - E\mathbf{g}_k\left(\mathbf{X}_{t+1}^{(k)}\right),$$

(since $\mathbf{X}_{t+1}^{(k)}$ is independent of $\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_t^{(k)}$ and hence of $\psi_1^{(k)}, \psi_2^{(k)}, \ldots, \psi_t^{(k)}$)

$$= \psi_t^{(k)} - E\left[\mathbf{g}_k(\mathbf{X}) \middle| \hat{w} = k\right]$$

$$= \psi_t^{(k)} - E\left(\mathbf{f}(\mathbf{X}) \middle| \hat{w} = k\right) - \sum_{j=1}^{m} \alpha_{kj} \varphi_j$$

$$= \psi_t^{(k)},$$

as, by Equation (6), $E(\mathbf{f}(\mathbf{X}) \ \hat{w} = k) = \sum_{j=1}^{m} \alpha_{kj} \varphi_j = \bar{\varphi}_k$, say. Similarly, we have

$$E\left(\left\|\psi_t^{(k)} - \mathbf{g}_k\left(\mathbf{X}_{t+1}^{(k)}\right)\right\|^2 \middle| \psi_1^{(k)}, \psi_2^{(k)}, \ldots, \psi_t^{(k)}\right)$$

$$= E\left(\|\psi_t^{(k)}\|^2 - 2\psi_t^{(k)\prime} \mathbf{g}_k\left(\mathbf{X}_{t+1}^{(k)}\right) + \left\|\mathbf{g}_k\left(\mathbf{X}_{t+1}^{(k)}\right)\right\|^2 \middle| \psi_1^{(k)}, \psi_2^{(k)}, \ldots, \psi_t^{(k)}\right)$$

$$= \|\psi_t^{(k)}\|^2 + E\left(\mathbf{g}_k(\mathbf{X}) \middle| \hat{w} = k\right) + E\left(\|\mathbf{g}_k(\mathbf{X})\|^2 \middle| \hat{w} = k\right)$$

(for the same reason as before)

$$= \|\psi_t^{(k)}\|^2 + E\left(\|\mathbf{f}(\mathbf{X}) - \bar{\varphi}\|^2 \middle| \hat{w} = k\right) \qquad \left[\text{since } E\left(\mathbf{g}_k(\mathbf{X}) \middle| \hat{w} = k\right) = \mathbf{0}\right]$$

$$= \|\psi_t^{(k)}\|^2 - \|\bar{\varphi}\|^2 + E\left(\|\mathbf{f}(\mathbf{X})\|^2 \middle| \hat{w} = k\right)$$

$$\leqslant \|\psi_t^{(k)}\|^2 + \|\bar{\varphi}_k\|^2 + \sum_{j=1}^{m} \alpha_{kj} \rho_j \qquad [\text{because of (A6)}]$$

$$\leqslant \|\psi_t^{(k)}\|^2 + \sum_{j=1}^{m} \left(\|\varphi_j\|^2 + \rho_j\right)^{j=1},$$

so that (C4) is seen to hold with

$$a = \sum_{j=1}^{m} \left(\|\varphi\|^2 + \rho_j\right), \qquad b = 0, \quad c = 1.$$

Condition (C5) is seen to be true because

$$\mathbf{x}'\mathbf{M}_t(\mathbf{x}) = \mathbf{x}'\mathbf{x} \geqslant 0.$$

The validity of (C6) follows from the fact that

$$\begin{aligned} E\|\psi_1^{(k)}\|^2 &= E\left\|\mathbf{g}\left(\mathbf{X}_1^{(k)}\right)\right\|^2 \\ &= E\left(\|\mathbf{g}(\mathbf{X})\|^2 \middle| \hat{w} = k\right) \\ &\leqslant \|\bar{\varphi}_k\|^2 + \sum_{j=1}^{m} \alpha_{kj}\rho_j < \infty. \end{aligned}$$

Finally, (C7) follows from the fact that

$$\inf_{\eta < \|\mathbf{x}\| < \eta^{-1}} \left[\mathbf{x}'\mathbf{M}_t(\mathbf{x})\right] = \inf_{\eta < \|\mathbf{x}\| < \eta^{-1}} \mathbf{x}'\mathbf{x} = \eta^2 > 0.$$

Hence the proposition.

Some implications of Proposition 1:

(1) If the matrix $\mathbf{B}$ is the identity matrix, i.e., if there is *no* mislabeling, then under our assumptions,

$$\hat{\varphi}_t^{(k)} \xrightarrow{\text{a.s.}} \varphi_k,$$

as expected.

(2) If $\mathbf{B} \neq \mathbf{I}_m$, then clearly, the estimates $\hat{\varphi}_t^{(k)}$ for the different classes converge to nontrue values

$$\bar{\varphi}_k = \sum_{j=1}^{m} \alpha_{kj}\varphi_j,$$

i.e., a convex linear combination of the parameter vectors of all the classes, as

$$\sum_{j=1}^{m} \alpha_{kj} = 1, \qquad j = 1, \ldots, m.$$

(3) Yet another implication can be stated formally as follows:

PROPOSITION 2. *Consider the setup specified in Sections 2 and 3. If assumptions* (A1)–(A6) *hold, then*

$$\sum_{j=1}^{m} \gamma_{kj} \hat{\varphi}_t^{(j)} \xrightarrow{\text{a.s.}} \varphi_k, \qquad k = 1, \ldots, m,$$

where $\Gamma_{m \times m} = ((\gamma_{jj}))$ *is a generalized inverse* [10] *of the matrix*

$$\underset{m \times m}{A} = \left((\alpha_{ij})\right)_{i=1,\ldots,m,\, j=1,\ldots,m}$$

satisfying

$$\Gamma A = I_m. \tag{8}$$

Proof. Let us write

$$A' = (\alpha_1 \mid \alpha_2 \mid \alpha_3 \mid \cdots \mid \alpha_m),$$

$$F = (\varphi_1 \mid \varphi_2 \mid \varphi_3 \mid \cdots \mid \varphi_m),$$

$$\hat{F}_t = \left(\hat{\varphi}_t^{(1)} \mid \hat{\varphi}_t^{(2)} \mid \hat{\varphi}_t^{(3)} \mid \cdots \mid \hat{\varphi}_m^{(t)}\right).$$

From Proposition 1, it is known that

$$\hat{\varphi}_t^{(k)} \xrightarrow{\text{a.s.}} F\alpha_k, \qquad k = 1, 2, \ldots, m,$$

i.e.,

$$\hat{F}_t \xrightarrow{\text{a.s.}} FA^T \qquad \text{columnwise}$$

(that is, every column of the matrix on the left-hand side converges a.s. to the corresponding column of the matrix on the right-hand side). By Lemma 3 below, this implies that

$$\hat{F}_t \xrightarrow{\text{a.s.}} FA^T \qquad \text{elementwise}$$

(i.e., every element of the matrix on the left-hand side converges a.s. to the corresponding element of the matrix on the right-hand side). By Lemma 4

below, this implies that

$$\hat{\mathbf{F}}_l \boldsymbol{\Gamma}^T \xrightarrow{\text{a.s.}} \mathbf{F} \qquad \text{elementwise},$$

and this, in turn, implies by lemma 3 that

$$\hat{\mathbf{F}}_l \boldsymbol{\Gamma}^T \xrightarrow{\text{a.s.}} \mathbf{F} \qquad \text{columnwise}.$$

Hence the proposition.

LEMMA 3. *Let* $\{\mathbf{x}_n\}$ *be a sequence defined over* $\mathbb{R}^N$, *and let* $\mathbf{a} \in \mathbb{R}^N$. *Then* $\mathbf{x}_n \xrightarrow{\text{a.s.}} \mathbf{a}$ *if and only if*

$$x_{ni} \xrightarrow{\text{a.s.}} a_i, \qquad i = 1, 2, \ldots, N,$$

where x_{ni} *and* a_i *are respectively the* ith *elements of* $\mathbf{x}_n$ *and* $\mathbf{a}$.

LEMMA 4. *Let*

$$x_{ij}^{(n)} \xrightarrow{\text{a.s.}} a_{ij}, \quad i = 1, 2, \ldots, p, \; j = 1, 2, \ldots, q, \qquad \text{as} \quad n \to \infty,$$

where $x_{ij}^{(n)}, a_{ij} \in \mathbb{R}$. *Let*

$$\mathbf{x}_n = \left(\left(x_{ij}^{(n)}\right)\right) \quad \text{and} \quad \mathbf{A} = \left((a_{ij})\right).$$

If

$$\underset{m \times l}{\mathbf{Z}_n} = \left(\left(Z_{ij}^{(n)}\right)\right) = \mathbf{P}\mathbf{X}_n\mathbf{Q} \qquad \text{for some } \underset{m \times p}{\mathbf{P}} \text{ and } \underset{q \times l}{\mathbf{Q}},$$

then

$$Z_{ij}^{(n)} \xrightarrow{\text{a.s.}} b_{ij},$$

where

$$\mathbf{B} = \left((b_{ij})\right) = \mathbf{PAQ}.$$

It may be mentioned in passing that one $\boldsymbol{\Gamma}$ satisfying Equation (8) is the Moore-Penrose inverse [10] of $\mathbf{A}$, viz., $\mathbf{A}^+$ defined as

$$\mathbf{A}^+ = \sum_{i=1}^{r} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i',$$

where λ_i is the ith nonzero eigenvalue of $\mathbf{A}$, and $\mathbf{u}_i$ the corresponding eigenvector, $i = 1, \ldots, r$.

V. CONCLUSION

For the general m-class, N-feature pattern recognition problem it is found that in the presence of labeling errors for training samples, the recursive estimates for class parameters φ_k, defined by means of Equation (1), do converge strongly under certain conditions. However, the values they converge to are not the true class-parameter values but certain convex linear combinations of true values for all m classes.

REFERENCES

1. P. A. Lachenbruch, Discriminant functions when the initial samples are misclassified, *Technometrics* 8:657–662 (1966).
2. _____, Discriminant functions when the initial samples are misclassified II: Nonrandom misclassification models, *Technometrics* 16:419–424 (1974).
3. G. J. McLachlan, Estimating the linear discriminant function from initial samples containing a small number of unclassified observations, *J. Amer. Statist. Assoc.* 72:403–406 (1977).
4. J. E. Michalek and R. C. Tripathi, The effect of errors in diagnosis and measurement on the estimation of probability of an event, *J. Amer. Statist. Assoc.* 75:713–721 (1980).
5. T. J. O'Neill, Normal discrimination with unclassified observations, *J. Amer. Statist. Assoc.* 73:821–825 (1978).
6. T. Krishnan, Efficiency of Normal Discrimination with Misclassified Initial Samples, Tech, Report ASC/85/3, Indian Statistical Inst., Calcutta, 1985.
7. M. B. Nevel'son and R. Z. Has'minskii, *Stochastic Approximation and Recursive Estimation*, Amer. Math. Soc. Providence, 1973.
8. C. B. Chittineni, Learning with imperfectly labelled samples, *Pattern Recognition* 12:281–291 (1980).
9. L. Schmetterer, Multidimensional stochastic approximation, in *Multivariate Analysis — II: Proceedings of the 2nd International Symposium on Multivariate Analysis, Dayton, Ohio* (P. R. Krishnaiah, Ed.), Academic, New York, 1968.
10. C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, Wiley, New York, 1971.

Received 17 December 1987; revised 31 May 1988