Lockheed Electronics Company, Inc.
A SUBSIDIARY OF LOCKHEED CORPORATION
1830 NASA Road 1, Houston, Texas 77058
Tel. 713-333-5411

JSC-14867
MAY 2 1979

CR 160183

Ref: 642-7612
Contract NAS 9-15800
Job Order 73-705

TECHNICAL MEMORANDUM

LEARNING WITH IMPERFECTLY LABELED PATTERNS

By

C. B. Chittineni

Approved By: J. C. Minter
T. C. Minter, Supervisor
Techniques Development Section

CONTENTS

FIGURES

1. INTRODUCTION*

In practical applications of pattern recognition, such as remote sensing, one of the difficult problems is obtaining the labels for training patterns. Labeling training patterns is costly, and very often the labels are imperfect.

In the recent literature, several authors investigated the problem of pattern recognition with imperfectly labeled patterns. Duda and Singleton [1] showed that, for orthogonal pattern vectors, the average weight vector of a threshold logical unit converges to a solution weight vector for the correctly labeled pattern set. Whitney and Dwyer [2] obtained error bounds in a two-class situation on the performance of a nearest neighbor rule with an imperfect teacher. Kashyap and Blaydon [3] proposed an iterative training procedure for a two-class case. Gimlin and Ferrell [4] studied the correction of labels using a nearest neighbor procedure. Shanmugam and Breipohl [5] proposed an error-correcting procedure for disjoint densities using Parzen estimators. Chittineni [6,7,8] investigated the applicability of probabilistic distance measures for feature selection with imperfectly labeled patterns.

This paper considers the problem of learning with imperfectly labeled patterns. In section 2, the author develops a model for the imperfectly labeled patterns. Section 3 presents an analysis of the Bayes classifier error with and without imperfections in the labels. In section 4, we obtain bounds on the performance of nearest neighbor classifiers for a multiclass case. The training of a classifier with and without imperfections is discussed in section 5, and schemes for the correction of imperfections in the labels are developed in section 6. Section 7 presents expressions for success probability as a function of time for the one-dimensional classifier, and section 8 treats feature selection criteria with imperfect labels.

---

*This document was prepared originally in January 1979 for submission to the Institute of Electrical and Electronics Engineers (IEEE) Journal. Thus, its format conforms to the IEEE requirements and is not consistent with standard Lockheed Electronics Company, Inc., document specifications.

2. A MODEL FOR LABELING IMPERFECTIONS

Let ω and $\hat{\omega}$ be the perfect and imperfect labels, respectively, each of which takes values of 1, 2, $\cdots$, M; where M is the number of classes. Let $p(\omega = i)$ and $p(X|\omega = i)$ be the a priori probabilities and class conditional densities of the patterns in classes $\omega = i$.

Assuming that the imperfections in the labels are described by the probabilities

$$\beta_{ji} = P(\hat{\omega} = i|\omega = j) \quad ; \quad i,j = 1,2,\cdots,M \tag{2-1}$$

where i and j indicate class, we have the constraint

$$\sum_{i=1}^{M} \beta_{ji} = 1 \tag{2-2}$$

Assume further that

$$p(X|\hat{\omega}_i,\omega_j) = p(X|\omega_j) \tag{2-3}$$

In order to find the relationship between $p(X|\omega = i)$ and $p(X|\hat{\omega} = i)$, consider

$$p(X|\hat{\omega} = i) = \frac{1}{P(\hat{\omega} = i)} \sum_{j=1}^{M} p(X,\hat{\omega} = i,\omega = j)$$

$$= \frac{1}{P(\hat{\omega} = i)} \sum_{j=1}^{M} p(X|\hat{\omega} = i,\omega = j)P(\hat{\omega} = i|\omega = j)P(\omega = j)$$

$$= \frac{1}{P(\hat{\omega} = i)} \sum_{j=1}^{M} \beta_{ji}P(\omega = j)p(X|\omega = j) \tag{2-4}$$

Cross-multiplying and dividing equation (2-4) by p(X) establishes the relationship between a posteriori probabilities

$$p(\hat{\omega} = i|X) = \sum_{j=1}^{M} \beta_{ji}p(\omega = j|X) \tag{2-5}$$

Similarly, the a priori probabilities are related as

$$P(\hat{\omega} = i) = \sum_{j=1}^{M} \beta_{ji} P(\omega = j) \tag{2-6}$$

Inverting equation (2-4) yields the following result for a two-class case.

$$\left.\begin{aligned} P(\omega = 1)p(X|\omega = 1) &= \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}} [\beta_{22} P(\hat{\omega} = 1)p(X|\hat{\omega} = 1) \\ &\quad - \beta_{21} P(\hat{\omega} = 2)p(X|\hat{\omega} = 2)] \\ P(\omega = 2)p(X|\omega = 2) &= \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}} [\beta_{11} P(\hat{\omega} = 2)p(X|\hat{\omega} = 2) \\ &\quad - \beta_{12} P(\hat{\omega} = 1)p(X|\hat{\omega} = 1)] \end{aligned}\right\} \tag{2-7}$$

Similarly, for the a priori and a posteriori probabilities,

$$P(\omega = i) = \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}} [\beta_{jj} P(\hat{\omega} = i) - \beta_{ji} P(\hat{\omega} = j)] \quad ; \quad \begin{matrix} i,j = 1,2 \\ i \neq j \end{matrix} \tag{2-8}$$

$$p(\omega = i|X) = \frac{1}{\beta_{11}\beta_{12} - \beta_{12}\beta_{21}} [\beta_{jj} p(\hat{\omega} = i|X) - \beta_{ji} p(\hat{\omega} = j|X)] \quad ; \quad \begin{matrix} i,j = 1,2 \\ i \neq j \end{matrix} \tag{2-9}$$

For a symmetric case, when

$$\beta_{11} = \beta_{22} = \beta \text{ and } \beta_{12} = \beta_{21} = 1 - \beta \tag{2-10}$$

then

$$\beta_{11}\beta_{22} - \beta_{12}\beta_{21} = (2\beta - 1) \tag{2-11}$$

From equations (2-7), (2-10), and (2-11),

$$[P(\omega = 1)p(X|\omega = 1) - P(\omega = 2)p(X|\omega = 2)] = \frac{1}{2\beta - 1} [P(\hat{\omega} = 1)p(X|\hat{\omega} = 1) - P(\hat{\omega} = 2)p(X|\hat{\omega} = 2)] \tag{2-12}$$

The densities for symmetric and nonsymmetric labeling errors are illustrated in figures 2-1 and 2-2.

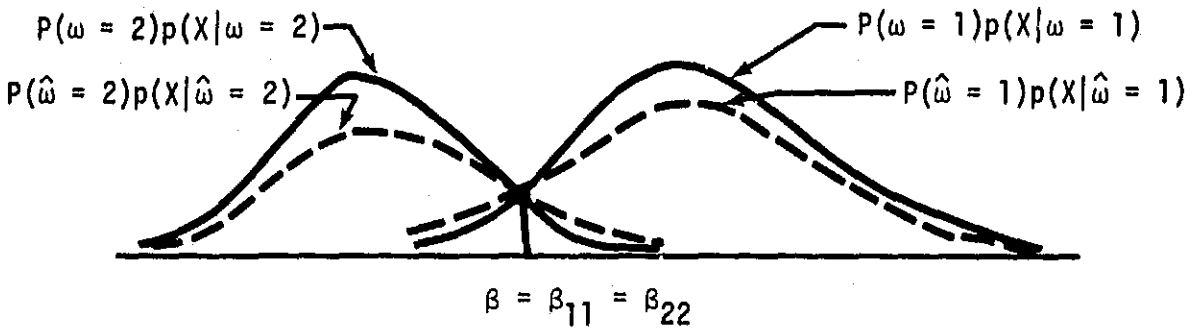

Figure 2-1.— Illustration of densities for symmetric labeling errors.

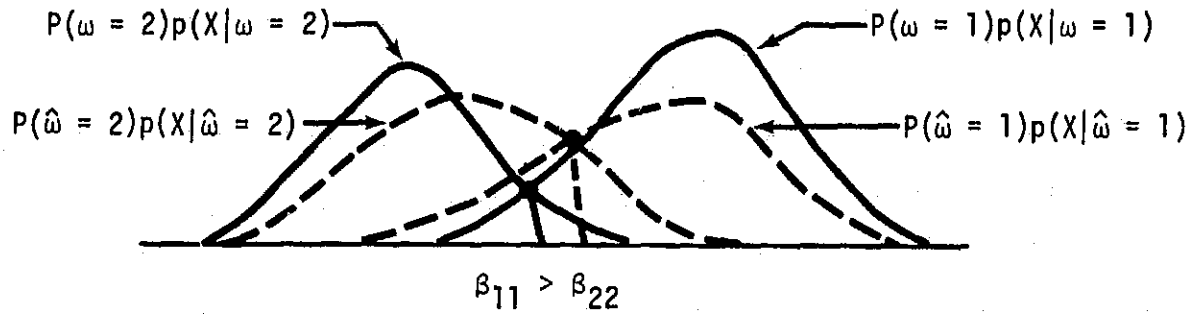

Figure 2-2.— Illustration of densities for nonsymmetric labeling errors.

3. PERFORMANCE OF BAYES CLASSIFIER WITH AND WITHOUT MISLABELING

In this section, we analyze the performance of the Bayes classifier with and without imperfections in the labels.

3.1 TWO-CLASS BAYES CLASSIFIER PERFORMANCE WITH SYMMETRIC MISLABELING

For a symmetric model, we develop a unique relationship between the probability of errors of the two-class Bayes classifier with and without mislabeling errors. The Bayes decision rule is

$$\left.\begin{aligned} &\text{Decide } X \in \omega = 1 \text{ if } P(\omega = 1)p(X|\omega = 1) > P(\omega = 2)p(X|\omega = 2) \\ &\text{Decide } X \in \omega = 2 \text{ otherwise} \end{aligned}\right\} \quad (3\text{-}1)$$

The Bayes probability of error (P_e) is

$$P_e = \int \min[P(\omega = 1)p(X|\omega = 1), P(\omega = 2)p(X|\omega = 2)]dX \quad (3\text{-}2)$$

For any two positive real numbers A and B,

$$\min(A,B) = \frac{1}{2}(A + B) - \frac{1}{2}|A - B| \quad (3\text{-}3)$$

Using equations (3-2) and (3-3) yields

$$P_e = \frac{1}{2} - \frac{1}{2}\int \left|P(\omega = 1)p(X|\omega = 1) - P(\omega = 2)p(X|\omega = 2)\right| dX \quad (3\text{-}4)$$

Following the argument presented by equations (3-1) through (3-4), the probability of error of a two-class Bayes classifier with imperfect labels ($\hat{P}_e$) can be written as

$$\hat{P}_e = \frac{1}{2} - \frac{1}{2}\int \left|P(\hat{\omega} = 1)p(X|\hat{\omega} = 1) - P(\hat{\omega} = 2)p(X|\hat{\omega} = 2)\right| dX \quad (3\text{-}5)$$

From equations (2-12), (3-4), and (3-5), we obtain the following.

$$\hat{P}_e = \frac{1}{2} - \frac{1}{2}|2\beta - 1| \int \left| P(\omega = 1)p(X|\omega = 1) - P(\omega = 2)p(X|\omega = 2) \right| dX$$

$$= \frac{1}{2} - \frac{1}{2}|2\beta - 1|(1 - 2P_e)$$

$$= \frac{1}{2}(1 - |2\beta - 1|) + |2\beta - 1|P_e \qquad (3\text{-}6)$$

From equation (3-6), writing P_e in terms of $\hat{P}_e$,

$$P_e = \frac{1}{|2\beta - 1|}\hat{P}_e + \frac{1}{2}\left(1 - \frac{1}{|2\beta - 1|}\right) \qquad (3\text{-}7)$$

Equation (3-7) is graphically displayed for various values of β in figure 3-1.

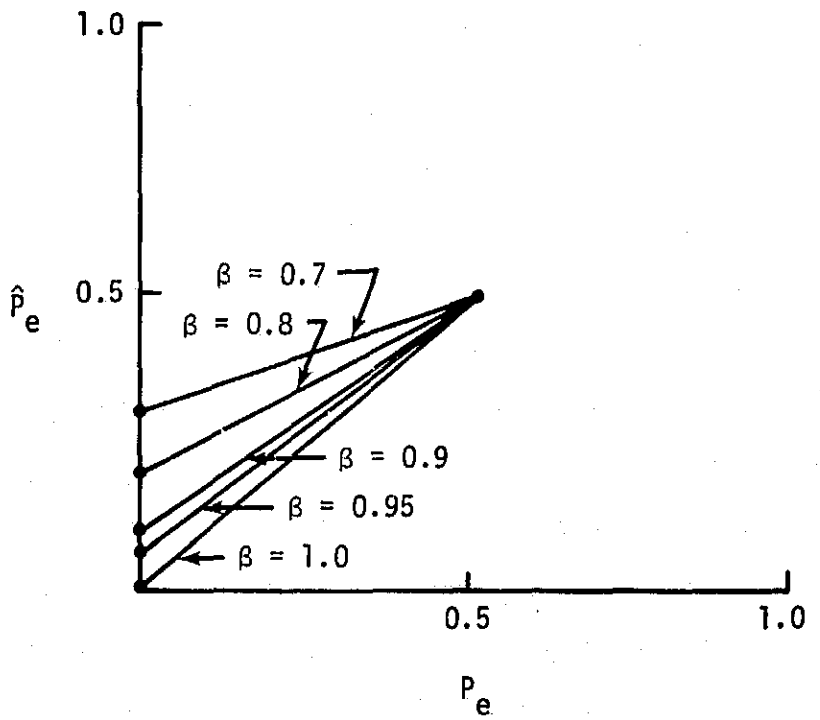

Figure 3-1.— Bayes risk for a symmetric model with and without labeling errors.

Figure 3-1 shows the increase in P_e because of labeling errors. When $P_e = 0.5$, the decision is random; hence, $\hat{P}_e$ is independent of β.

3.2 MULTICLASS BAYES ERROR UNDER A GENERAL MISLABELING MODEL

The Bayes risk (r) in classifying a pattern X, can be written as

$$r(X) = 1 - \max_i [p(\omega = i|X)] \tag{3-8}$$

$$\hat{r}(X) = 1 - \max_i [p(\hat{\omega} = i|X)] \tag{3-9}$$

Then the probability of error can be written as

$$P_e = E_{p(X)}[r(X)] \tag{3-10}$$

$$\hat{P}_e = E_{p(X)}[\hat{r}(X)] \tag{3-11}$$

where E is the expectation operator. If the imperfections are not symmetric, then the Bayes errors depend on the particular probability density functions of the patterns. However, we obtain bounds in the following manner.

3.2.1 LOWER BOUND

Let

$$K = \max_i [p(\hat{\omega} = i|X)] \tag{3-12}$$

$$\beta_{RSM} = \max_i \left(\sum_{j=1}^{M} \beta_{ji} \right) \tag{3-13}$$

where RSM = row sum maximum.

From equations (3-8), (3-9), (3-12), and (3-13), we obtain

$$\begin{aligned} \hat{r}(X) &= 1 - \max_i \left[\sum_{j=1}^{M} \beta_{ji} p(\omega = j|X) \right] \\ &\geq 1 - K \left[\max_i \left(\sum_{j=1}^{M} \beta_{ji} \right) \right] \\ &= 1 - K\beta_{RSM} \\ &= (1 - \beta_{RSM}) + \beta_{RSM} r(X) \end{aligned} \tag{3-14}$$

Taking the expectations on both sides of equation (3-14) with respect to p(X) obtains the desired inequality

$$\hat{P}_e \geq (1 - \beta_{RSM}) + \beta_{RSM} P_e \tag{3-15}$$

3.2.2 UPPER BOUND

Let

$$p(\omega = k|X) = \max_j [p(\omega = j|X)] \tag{3-16}$$

$$b_i = \min_j (\beta_{ji}) \tag{3-17}$$

$$\alpha = \min_k \max_i (\beta_{ki} - b_i) \tag{3-18}$$

where k is a scalar indicating class.

Consider

$$\max_i [p(\hat{\omega} = i|X)] = \max_i \left[\beta_{ki} p(\omega = k|X) + \sum_{\substack{j=1 \\ j \neq k}}^{M} \beta_{ji} p(\omega = j|X) \right]$$

$$\geq \max_i [(\beta_{ki} - b_i) p(\omega = k|X) + b_i]$$

$$\geq \alpha[1 - r(X)] \tag{3-19}$$

However,

$$\hat{r}(X) = 1 - \max_i [p(\hat{\omega} = i|X)] \leq (1 - \alpha) + \alpha r(X) \tag{3-20}$$

Taking the expectations on both sides of equation (3-20) with respect to p(X), we obtain

$$\hat{P}_e \leq (1 - \alpha) + \alpha P_e \tag{3-21}$$

3.3 ILLUSTRATION OF BOUNDS

To illustrate the upper and lower bounds, let M = 3. Consider a matrix of mislabeling probabilities β_{ji} as shown in figure 3-2. The various quantities

required to compute the bounds are also shown.

$$\beta_{ji} = \begin{array}{c} j \downarrow \\ \\ \\ \\ b_i = \min_j(\beta_{ji}) \end{array} \begin{array}{c} i \longrightarrow \\ \begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix} \\ \begin{bmatrix} 0.05 & 0.05 & 0.05 \end{bmatrix} \end{array} \quad \begin{array}{c} \sum_{j=1}^{M}\beta_{ji} \\ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ \\ \end{array} \quad \begin{array}{c} \max_i(\beta_{ki} - b_i) \\ \begin{bmatrix} 0.85 \\ 0.85 \\ 0.85 \end{bmatrix} \\ \\ \end{array}$$

$$\beta_{RSM} = 1, \ \alpha = 0.85$$

Figure 3-2.— Example illustrating upper and lower bounds.

Let $\hat{P}_e = 0.2$ and using the above probabilistic mislabeling model, the bounds on the true probability of error P_e without imperfections in the labels are given by

$$0.059 \leq P_e \leq 0.2 \tag{3-22}$$

4. PERFORMANCE OF MULTICATEGORY NEAREST NEIGHBOR CLASSIFIER UNDER A GENERAL MODEL OF MISLABELING

In the case of imperfectly labeled patterns, given the pattern X, the conditional nearest neighbor risk can be written as

$$\hat{r}_N(X) = p(\omega = 1|X) \sum_{\substack{j=1 \\ j\neq 1}}^{M} p(\hat{\omega} = j|X) + \cdots + p(\omega = M|X) \sum_{\substack{j=1 \\ j\neq M}}^{M} p(\hat{\omega} = j|X)$$

$$= 1 - \sum_{i=1}^{M} p(\omega = i|X)p(\hat{\omega} = i|X)$$

$$= 1 - \sum_{i=1}^{M}\sum_{j=1}^{M} \beta_{ji} p(\omega = i|X)p(\omega = j|X) \tag{4-1}$$

In the following subsections, we obtain bounds on the nearest neighbor error in terms of Bayes classifier error.

4.1 LOWER BOUND

Substituting equation (3-12) into (4-1) obtains the following result.

$$\begin{aligned}\hat{r}_N(X) &\geq 1 - K\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_{ji}P(\omega = j|X) \\ &= 1 - K\sum_{j=1}^{M}p(\omega = j|X)\sum_{i=1}^{M}\beta_{ji} \\ &= 1 - K \\ &= r(X) \end{aligned} \tag{4-2}$$

Taking expectations with respect to $p(X)$, on both sides of equation (4-2), results in

$$\hat{P}_{eN} \geq P_e \tag{4-3}$$

where $\hat{P}_{eN}$ is the nearest neighbor error.

4.2 UPPER BOUND

Let

$$p(\omega = k|X) = \max_i[p(\omega = i|X)] \tag{4-4}$$

$$\beta = \min_i(\beta_{ii}) \tag{4-5}$$

$$\beta_m = \min_{j,k}(\beta_{jk}) \tag{4-6}$$

Consider the following.

$$\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_{ji}p(\omega = i|X)p(\omega = j|X) = \sum_{i=1}^{M}\beta_{ii}p^2(\omega = i|X)$$

$$+ \sum_{i=1}^{M}p(\omega = i|X)\sum_{\substack{j=1\\ j\neq i}}^{M}\beta_{ji}p(\omega = j|X)$$

$$= \beta_{kk}p^2(\omega = k|X) + \sum_{\substack{i=1\\ i\neq k}}^{M}\beta_{ii}p^2(\omega = i|X)$$

$$+ p(\omega = k|X)\sum_{\substack{j=1\\ j\neq k}}^{M}\beta_{jk}p(\omega = j|X)$$

$$+ \sum_{\substack{i=1\\ i\neq k}}^{M}p(\omega = i|X)\sum_{\substack{j=1\\ j\neq i}}^{M}\beta_{ji}p(\omega = j|X) \qquad (4\text{-}7)$$

However,

$$\sum_{\substack{i=1\\ i\neq k}}^{M}p^2(\omega = i|X) \geq \frac{1}{M-1}\left[\sum_{\substack{i=1\\ i\neq k}}^{M}p(\omega = i|X)\right]^2$$

$$= \frac{1}{M-1}[1 - p(\omega = k|X)]^2 \qquad (4\text{-}8)$$

Now consider

$$p(\omega = k|X)\sum_{\substack{j=1\\ j\neq k}}^{M}\beta_{jk}p(\omega = j|X) \geq \beta_m p(\omega = k|X)[1 - p(\omega = k|X)] \qquad (4\text{-}9)$$

Substituting equations (4-8) and (4-9) into (4-7) results in

$$\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_{ji}p(\omega = i|X)p(\omega = j|X) \geq \beta[1 - r(X)]^2 + \frac{\beta}{M-1}r^2(X)$$

$$+ \beta_m r(X)[1 - r(X)] \qquad (4\text{-}10)$$

From equations (4-1) and (4-10), we obtain

$$\hat{r}_N(X) \leq 1 - \beta[1 - r(X)]^2 - \frac{\beta}{M-1}r^2(X) - \beta_m r(X)[1 - r(X)]$$

$$= (1 - \beta) + (2\beta - \beta_m)r(X) - \left(\frac{M}{M-1}\beta - \beta_m\right)r^2(X) \qquad (4\text{-}11)$$

But we have

$$E_{p(X)}[r^2(X)] = \mathrm{Var}[r(X)] + P_e^2$$

$$\geq P_e^2 \qquad (4\text{-}12)$$

$$\frac{M}{M-1}\beta - \beta_m \geq 0 \qquad (4\text{-}13)$$

Taking the expectations on both sides of equation (4-11) with respect to $p(X)$ and using equations (4-12) and (4-13) results in

$$\hat{P}_{eN} \leq (1 - \beta) + (2\beta - \beta_m)P_e - \left(\frac{M}{M-1}\beta - \beta_m\right)P_e^2 \qquad (4\text{-}14)$$

When $\beta = 1$ and $\beta_m = 0$, we have the perfectly labeled situation, in which case equation (4-14) becomes

$$\hat{P}_{eN} \leq 2P_e - \frac{M}{M-1}P_e^2 \qquad (4\text{-}15)$$

It is seen that equation (4-15) is identical to the nearest neighbor bound obtained by Cover and Hart [9].

5. DESIGN OF PATTERN CLASSIFIERS WITH IMPERFECTLY LABELED PATTERNS

In this section, we consider the problem of designing a classifier with imperfectly labeled training patterns. Once the amount of imperfections is known, this knowledge can be incorporated into the classifier training and results in improved performance.

5.1 INCREMENT ERROR CORRECTION CLASSIFIER (NONPARAMETRIC TRAINING)

Consider the case of two pattern classes. Assume a given set of training patterns $X_1(1), X_1(2), \cdots, X_1(N_1)$ and $X_{-1}(1), X_{-1}(2), \cdots, X_{-1}(N_{-1})$ from classes 1 and -1, with imperfect labels $\hat{\omega}_1(1), \cdots, \hat{\omega}_1(N_1)$; $\hat{\omega}_{-1}(1), \cdots, \hat{\omega}_{-1}(N_{-1})$. Let the perfect labels of these patterns be $\omega_1(1), \cdots, \omega_1(N_1)$; $\omega_{-1}(1), \cdots, \omega_{-1}(N_{-1})$. The imperfect and perfect labels take values of 1 and -1. The objective is to find a decision function d(X), such that

$$\left.\begin{aligned} d(X) > 0 \text{ when } X \in \omega_1 \\ d(X) < 0 \text{ when } X \in \omega_{-1} \end{aligned}\right\} \tag{5-1}$$

5.1.1 ALTERNATIVE REPRESENTATION OF IMPERFECTIONS

The imperfect labels can be modeled from perfect labels as follows.

$$\hat{\omega} = \omega\eta \tag{5-2}$$

where η = labeling noise and takes values of 1 and -1. Since ω takes 1 and -1, whenever ω differs from η we have an imperfect label. Recalling our previous model of imperfections, we have

$$\beta_{ij} = P(\hat{\omega} = j | \omega = i) \quad ; \quad i,j = \pm 1 \tag{5-3}$$

Let

$$\overline{\eta} = E(\eta) \tag{5-4}$$

and

$$P_i = P(\omega = i) \quad ; \quad i = \pm 1 \tag{5-5}$$

Since η takes values of 1 or -1, the average value of η, $\overline{\eta}$, can be written as follows.

$$
\begin{aligned}
\bar{\eta} &= E(\eta) \\
&= P(\eta = 1) - P(\eta = -1) \\
&= P(\eta = 1, \omega = 1) + P(\eta = 1, \omega = -1) - P(\eta = -1, \omega = 1) \\
&\quad - P(\eta = -1, \omega = -1) \\
&= P_1 P(\eta = 1 | \omega = 1) + P_{-1} P(\eta = 1 | \omega = -1) - P_1 P(\eta = -1 | \omega = 1) \\
&\quad - P_{-1} P(\eta = -1 | \omega = -1) \\
&= P_1 P(\hat{\omega} = 1 | \omega = 1) + P_{-1} P(\hat{\omega} = -1 | \omega = -1) - P_1 P(\hat{\omega} = -1 | \omega = 1) \\
&\quad - P_{-1} P(\hat{\omega} = 1 | \omega = -1) \\
&= P_1 \beta_{11} + P_{-1} \beta_{-1,-1} - P_1 \beta_{1,-1} - P_{-1} \beta_{-1,1} \\
&= (2\beta_{11} - 1) P_1 + (2\beta_{-1,-1} - 1) P_{-1} \qquad (5\text{-}6)
\end{aligned}
$$

Under a symmetric model, the above becomes

$$\bar{\eta} = 2\beta - 1 \qquad (5\text{-}7)$$

5.1.2 AN ERROR CORRECTION ALGORITHM WITH IMPERFECT LABELS

To obtain a linear approximation to d(X),

$$d(X) = X^T W \qquad (5\text{-}8)$$

where W is the weight vector. Let $d_0(X)$ be the unknown decision function and d*(X) be the optimal decision function. Suppose we set up a criterion

$$C(W) = E[\alpha(W,X)] \qquad (5\text{-}9)$$

where

$$\alpha(W,X) = W^T X \left\{ \mathrm{Sgn}(W^T X) - \mathrm{Sgn}[d_0(X)] \right\} \qquad (5\text{-}10)$$

Let W* be the value of W which minimizes C(W), then

$$C(W) \geq C(W^*) \qquad (5\text{-}11)$$

The corresponding optimal approximation $d^*(X)$ to $d_0(X)$ is given by

$$d^*(X) = X^T W^* \tag{5-12}$$

At the ℓth step of training, the weight vector $W(\ell)$ is updated using the steepest descent method.

$$W(\ell + 1) = W(\ell) - \nu(\ell) \frac{\partial C(W)}{\partial W}\Big|_{W = W(\ell)} \tag{5-13}$$

Since the gradient is not known, the above is approximated to

$$W(\ell + 1) = W(\ell) - \nu(\ell) g(W,X)\Big|_{X = X(\ell), W = W(\ell)} \tag{5-14}$$

where $X(\ell)$ is the training pattern at the ℓth step and

$$g(W,X) = \{\mathrm{Sgn}(W^T X) - \mathrm{Sgn}[d_0(X)]\}X \tag{5-15}$$

Since the perfect label $\mathrm{Sgn}[d_0(X)]$ is unknown, we replace $g(W,X)$ with $f[W,X,\hat{\omega}(X)]$ so that f is observable for any X and has the same expected value as g, where

$$f[W,X,\hat{\omega}(X)] = \frac{1}{\eta}[\bar{\eta}\,\mathrm{Sgn}(W^T X) - \hat{\omega}(X)]X \tag{5-16}$$

$$\begin{aligned} E\{f[W,X,\hat{\omega}(X)]|W\} &= E\left[\left(\frac{1}{\eta}[\bar{\eta}\,\mathrm{Sgn}(W^T X) - \hat{\omega}(X)]X + \{\mathrm{Sgn}[d_0(X)] - \mathrm{Sgn}[d_0(X)]\}X\right)|W\right] \\ &= E\left(\{\mathrm{Sgn}(W^T X) - \mathrm{Sgn}[d_0(X)]\}X|W\right) \\ &\quad + E\left(\left\{\mathrm{Sgn}[d_0(X)] - \frac{\eta}{\bar{\eta}}\mathrm{Sgn}[d_0(X)]\right\}|W\right) \\ &= E\left(\{\mathrm{Sgn}(W^T X) - \mathrm{Sgn}[d_0(X)]\}X|W\right) \\ &= E[g(W,X)|W] \end{aligned} \tag{5-17}$$

Hence, the error correction algorithm for updating the weight vector $W(\ell)$ at the ℓth step of training can be written as

$$W(\ell + 1) = W(\ell) - \nu(\ell)\frac{1}{\eta}\{\bar{\eta}\,\mathrm{Sgn}[X^T W(\ell)] - \hat{\omega}(X)\}X \tag{5-18}$$

where X is the training pattern at the ℓth step of training and $\overline{\eta}$ is given by equation (5-6). For the convergence of this algorithm, the conditions on $\nu(\ell)$ can be shown to be [3]:

$$\nu(\ell) \geq 0, \sum_{\ell=1}^{\infty} \nu(\ell) = \infty, \sum_{\ell=1}^{\infty} \nu^2(\ell) < \infty \tag{5-19}$$

5.2 BAYES CLASSIFIER (PARAMETRIC TRAINING)

Once the a priori probabilities and class conditional densities of the imperfectly labeled patterns and the mislabeling probabilities β_{ij}'s are estimated, equations (2-7) and (2-8) can be used to obtain the a priori probabilities and class conditional densities of the perfectly labeled patterns. Then, the following algorithm can be used to classify the patterns.

$$\text{Decide } X \in \omega = i \text{ if } P(\omega = i)p(X|\omega = i) > \max_{\substack{\ell \\ \ell=1,2,\cdots,M \\ \ell \neq i}} [P(\omega = \ell)p(X|\omega = \ell)] \tag{5-20}$$

5.3 THE CASE WHEN β_{ij}'S ARE UNKNOWN

In the case when β_{ij}'s are unknown, the scheme shown in figure 5-1 can be used to design the classifier.

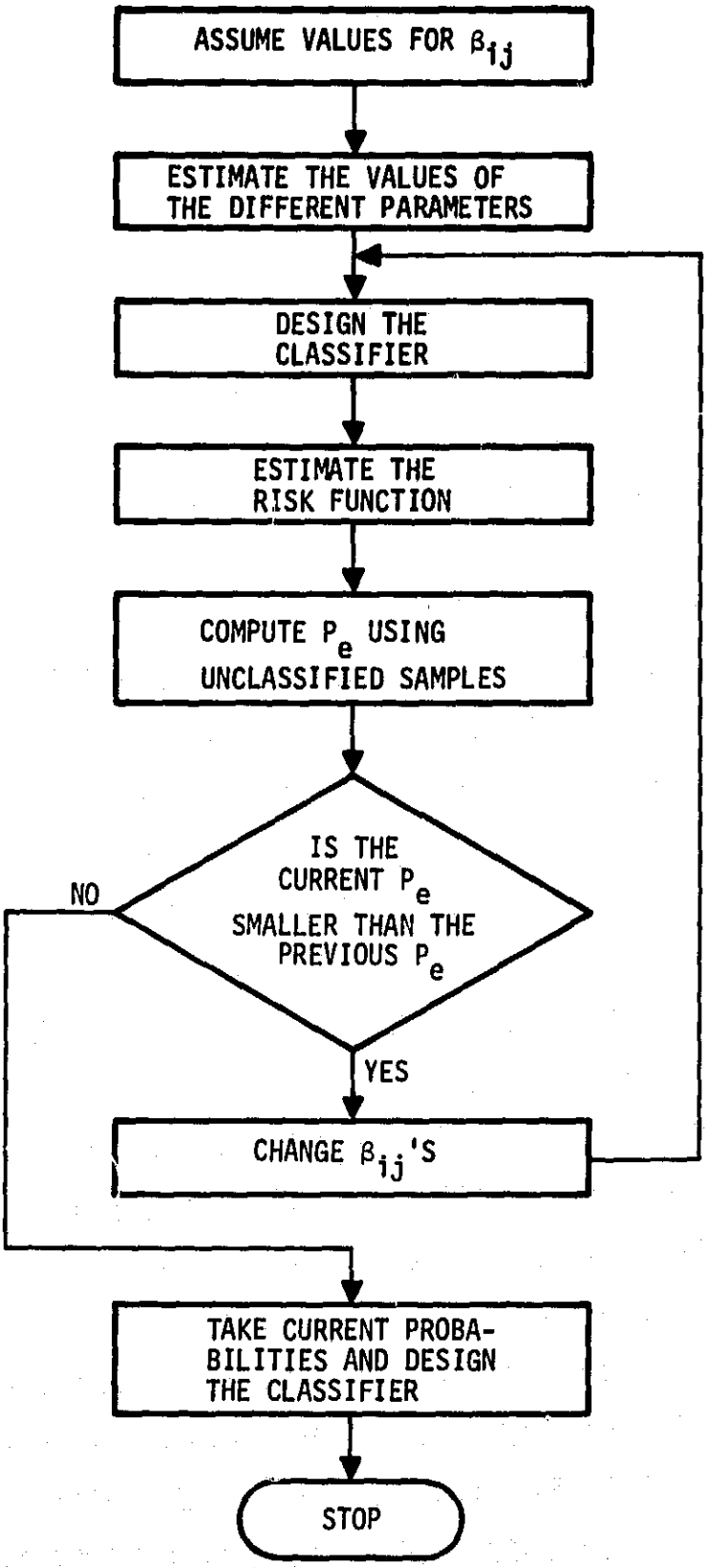

Figure 5-1.— Flow diagram for learning with imperfectly labeled patterns when β_{ij}'s are unknown.

6. CORRECTION OF IMPERFECT LABELS OF TRAINING PATTERNS

This section deals with methods of correcting imperfect labels of training patterns. Whenever a method is confident enough to show that the label is imperfect, we correct the label of the training pattern.

6.1 LABEL CORRECTION USING k-NEAREST NEIGHBOR DECISION RULE

The nearest neighbor decision rule can be used to correct imperfect labels in training patterns [4]. Suppose that n - 1 training patterns are processed with imperfect labels. When a pattern X_n with imperfect label $\hat{\omega}(X_n)$ is presented to the algorithm, a guess of the true label of X_n must be made by combining the information in $\hat{\omega}(X_n)$ and the information in the previously processed n - 1 training patterns. The new label of X_n, $\omega(X_n)$, is determined in the following manner.

Assume that two positive integers k and k' are given, where k is an odd integer and k' is an integer such that $k' \geq (k + 1)/2$. Using a distance metric d, the k-nearest neighbors to X_n among the training patterns X_1, X_2, $\cdots$, X_{n-1} are located. If at least k' of the nearest neighbors to X_n have the same value for their class labels, $\omega(X_n)$ is set to that value. Otherwise, $\omega(X_n)$ is set to the value of $\hat{\omega}(X_n)$, the label provided by the teacher. The process is repeated to obtain the label of pattern X_{n+1}.

The integer k' specifies the degree of confidence required in labeling the k-nearest neighbors of X_n before the label of the teacher of X_n is changed. At least k of the training patterns must be obtained before the beginning of the label correction process. Also, at least k - k' + 1 of the teacher's labels are accepted for each class before the label correction process is begun, in order to avoid the algorithm's labeling of all patterns into one class.

At the termination of the label correction process, unlabeled patterns can be classified using the k-nearest neighbor decision algorithm.

6.2 LABEL CORRECTION USING BAYES DECISION RULE

In this section, an algorithm is developed to correct the imperfections in the labels of the training patterns. Assuming two pattern classes and that the densities are Gaussian; i.e.,

$$p(X|\hat{\omega} = i) \sim N(\hat{M}_i, \hat{\Sigma}_i) \quad ; \quad i = 1,2 \tag{6-1}$$

where $\hat{M}_i$ is the mean and $\hat{\Sigma}_i$ is the covariance matrix of the patterns in the classes $\hat{\omega}_i$, $i = 1,2$. The Bayes decision rule uses the following criterion.

$$\left.\begin{array}{l} \text{Decide } X \in \hat{\omega} = 1 \text{ if } d(X) > 0 \\ \text{and} \qquad \text{Decide } X \in \hat{\omega} = 2 \text{ if } d(X) < 0 \end{array}\right\} \tag{6-2}$$

where

$$d(X) = \log \frac{P(\hat{\omega} = 1)p(X|\hat{\omega} = 1)}{P(\hat{\omega} = 2)p(X|\hat{\omega} = 2)} \tag{6-3}$$

The following scheme for correcting imperfections in the labels is proposed.

$$\left.\begin{array}{l} \text{Change the label of } X \text{ to } \omega = 1 \text{ if } d(X) > t_1 \\ \text{Change the label of } X \text{ to } \omega = 2 \text{ if } d(X) < -t_2 \\ \text{Do not change the label of } X \text{ if } -t_2 \leq d(X) \leq t_1 \end{array}\right\} \tag{6-4}$$

where t_1 and t_2 are the thresholds.

6.2.1 DISTRIBUTION OF d(X)

Assume that $[d(X)|X \in \hat{\omega} = i]$, $i = 1,2$, is a Gaussian random variable with mean m_i and variance σ_i, $i = 1,2$; i.e., $p[d(X)|X \in \hat{\omega} = i] = N(m_i, \sigma_i^2)$. Then,

the expressions for m_i and σ_i can be shown to be [10]:

$$
\begin{aligned}
m_1 &= \operatorname{tr}\left(I - \hat{\Sigma}_1\hat{\Sigma}_2^{-1}\right) - (\hat{M}_1 - \hat{M}_2)^T\hat{\Sigma}_2^{-1}(\hat{M}_1 - \hat{M}_2) + \ell n\left(\frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|}\right) \\
&\quad - 2\ell n\left[\frac{P(\hat{\omega} = 1)}{P(\hat{\omega} = 2)}\right] \\
m_2 &= \operatorname{tr}\left(\hat{\Sigma}_1^{-1}\hat{\Sigma}_2 - I\right) + (\hat{M}_1 - \hat{M}_2)^T\hat{\Sigma}_1^{-1}(\hat{M}_1 - \hat{M}_2) + \ell n\left(\frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|}\right) \\
&\quad - 2\ell n\left[\frac{P(\hat{\omega} = 1)}{P(\hat{\omega} = 2)}\right] \\
\sigma_1^2 &= 2\left\{\operatorname{tr}\left[\left(I - \hat{\Sigma}_2^{-1}\hat{\Sigma}_1\right)^2\right]\right. \\
&\quad \left. + 2(\hat{M}_1 - \hat{M}_2)^T\hat{\Sigma}_2^{-1}\hat{\Sigma}_1\hat{\Sigma}_2^{-1}(\hat{M}_1 - \hat{M}_2)\right\} \\
\sigma_2^2 &= 2\left[\operatorname{tr}\left(\hat{\Sigma}_1^{-1}\hat{\Sigma}_2 - I\right)^2\right. \\
&\quad \left. + 2(\hat{M}_1 - \hat{M}_2)^T\hat{\Sigma}_1^{-1}\hat{\Sigma}_2\hat{\Sigma}_1^{-1}(\hat{M}_1 - \hat{M}_2)\right]
\end{aligned}
\qquad (6\text{-}5)
$$

6.2.2 SELECTION OF THRESHOLDS t_1 AND t_2

We propose to select the thresholds t_1 and t_2 for correcting the imperfect labels by specifying the probability α that mislabeling will occur in the correction process.

$$
\begin{aligned}
\alpha &= P(\text{bad label}) \\
&= P(\omega = 1)P(\text{bad label} \mid X \in \omega = 1) + P(\omega = 2)P(\text{bad label} \mid X \in \omega = 2) \\
&= P(\omega = 1)P\{d(X) < -t_2 \mid X \in \omega = 1) + P(\omega = 2)P[d(X)] > t_1 \mid X \in \omega = 2\}
\end{aligned}
\qquad (6\text{-}6)
$$

Following the argument and assumptions similar to those of section 2, it can be shown that

$$P(\omega = 1)p[d(X)|\omega = 1] = \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}\left\{\beta_{22}P(\hat{\omega} = 1)p[d(X)|\hat{\omega} = 1] - \beta_{21}P(\hat{\omega} = 2)p[d(X)|\hat{\omega} = 2)]\right\} \tag{6-7}$$

$$P(\omega = 2)p[d(X)|\omega = 2] = \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}\left\{\beta_{11}P(\hat{\omega} = 2)p[d(X)|\hat{\omega} = 2] - \beta_{12}P(\hat{\omega} = 1)p[d(X)|\hat{\omega} = 1]\right\} \tag{6-8}$$

Then,

$$P(\omega = 1)p[d(X) < -t_2|X \in \omega = 1] = P(\omega = 1)\int_{-\infty}^{-t_2} p[d(X)|\omega = 1]d[d(X)]$$

$$= \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}\left\{\beta_{22}P(\hat{\omega} = 1)\int_{-\infty}^{-t_2} p[d(X)|\hat{\omega} = 1]d[d(X)] - \beta_{21}P(\hat{\omega} = 2)\int_{-\infty}^{-t_2} p[d(X)|\hat{\omega} = 2]d[d(X)]\right\}$$

$$= \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}\left[\beta_{22}P(\hat{\omega} = 1)\int_{-\infty}^{\frac{-t_2-m_1}{\sigma_1}} N(0,1)d\xi - \beta_{21}P(\hat{\omega} = 2)\int_{-\infty}^{\frac{-t_2-m_2}{\sigma_2}} N(0,1)d\xi\right] \tag{6-9}$$

where $N(0,1)\sim$ is a Gaussian density function with zero mean and unit variance. After an argument similar to equation (6-9), from equation (6-8) we obtain the following.

$$P(\omega = 2)P(\text{bad label}|X \in \omega_2) = P(\omega = 2)\int_{t_1}^{\infty} p[d(X)|X \in \omega_2]d[d(X)]$$

$$= \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}\left[\beta_{11}P(\hat{\omega} = 2)\int_{-\infty}^{\frac{-t_1+m_2}{\sigma_2}} N(0,1)d\xi\right.$$

$$\left. - \beta_{12}P(\hat{\omega} = 1)\int_{-\infty}^{\frac{-t_1+m_1}{\sigma_1}} N(0,1)d\xi\right] \qquad (6\text{-}10)$$

For a specified α, using equations (6-9) and (6-10) in (6-6), t_1 and t_2 can be determined by one-dimensional numerical integration, and imperfect labels can be corrected using the algorithm in equation (6-4).

7. ONE-DIMENSIONAL CLASSIFIER WITH IMPERFECTLY LABELED PATTERNS

Consider the training of a one-dimensional version of the increment error correction classifier considered in section 5.1. Let ω and $\hat{\omega}$ be the perfect and imperfect labels of training patterns that take values 1 or -1. Assuming a symmetric model for the imperfections, i.e.,

$$\left.\begin{aligned} P(\hat{\omega} = 1|\omega = 1) = \beta = P(\hat{\omega} = -1|\omega = -1) \\ \text{then} \qquad P(\hat{\omega} = -1|\omega = 1) = 1 - \beta = P(\hat{\omega} = 1|\omega = -1) \end{aligned}\right\} \qquad (7\text{-}1)$$

As in section 5.1, the imperfections in the labels can be considered in terms of a quantity η, which takes values 1 or -1.

$$\hat{\omega} = \omega\eta \qquad (7\text{-}2)$$

In section 5.1, $\overline{\eta} = E(\eta)$ is related to β as

$$\overline{\eta} = 2\beta - 1 \qquad (7\text{-}3)$$

The decision rule for one-dimensional patterns is

$$\left.\begin{aligned} &\text{Decide } x_n \in \omega = 1, \text{ if } x_n > k_n \\ &\text{Decide } x_n \in \omega = -1, \text{ if } x_n < k_n \end{aligned}\right\} \qquad (7\text{-}4)$$

where k_n is the threshold at the n*th* training step. If $\omega(x_n)$ is the label of x_n, the pattern at the n*th* training step, the usual training procedure is to adjust k_n according to the following relation.

$$k_{n+1} = k_n - [\omega(x_n) - \mathrm{Sgn}(x_n - k_n)]\Delta k_n \tag{7-5}$$

where Δk_n is the increment of adjustment. Since the label $\omega(x_n)$ is not known, we modify equation (7-5) according to (7-6).

$$k_{n+1} = k_n - \frac{\nu_n}{\bar{\eta}} [\hat{\omega}(x_n) - \bar{\eta}\mathrm{Sgn}(x_n - k_n)]\Delta k_n \tag{7-6}$$

Similar to the results shown in section 5.1, the conditions for convergence of equation (7-6) can be shown to be

$$\nu_n \geq 0, \sum_{n=1}^{\infty} \nu_n = \infty, \sum_{n=1}^{\infty} \nu_n^2 < \infty \tag{7-7}$$

This section contains the learning dynamics of this procedure; i.e., expressions for the learning curves. Let $f_1(x)$ and $f_{-1}(x)$ be the constituent densities of patterns in classes $\omega = 1$ and $\omega = -1$, respectively. Similarly, let $\hat{f}_1(x)$ and $\hat{f}_{-1}(x)$ be the constituent densities of the patterns in classes $\hat{\omega} = 1$ and $\hat{\omega} = -1$, respectively. Let

$$\left.\begin{aligned} P_1 &= \int_{-\infty}^{\infty} f_1(x)dx \text{ and } P_{-1} = \int_{-\infty}^{\infty} f_{-1}(x)dx \\ \hat{P}_1 &= \int_{-\infty}^{\infty} \hat{f}_1(x)dx \text{ and } \hat{P}_{-1} = \int_{-\infty}^{\infty} \hat{f}_{-1}(x)dx \end{aligned}\right\} \tag{7-8}$$

Then, the following probabilities can be easily obtained.

$$
\left.
\begin{aligned}
\Pr[\hat{\omega} = 1, \mathrm{Sgn}(x - k) = 1] &= \int_k^{\infty} \hat{f}_1(x)dx \\
&= \hat{P}_1 - \int_{-\infty}^{k} \hat{f}_1(x)dx \\
&= \hat{P}_1 - \hat{F}_1(k) \\
\Pr[\hat{\omega} = 1, \mathrm{Sgn}(x - k) = -1] &= \int_{-\infty}^{k} \hat{f}_1(x)dx \\
&= \hat{F}_1(k) \\
\Pr[\hat{\omega} = -1, \mathrm{Sgn}(x - k) = 1] &= \int_k^{\infty} \hat{f}_{-1}(x)dx \\
&= \hat{P}_{-1} - \hat{F}_{-1}(k) \\
\Pr[\hat{\omega} = -1, \mathrm{Sgn}(x - k) = -1] &= \int_{-\infty}^{k} \hat{f}_{-1}(x)dx \\
&= \hat{F}_{-1}(k)
\end{aligned}
\right\} \quad (7\text{-}9)
$$

Let $P_n(k)$ be the probability of occurrence of k at time instant n. Then the training algorithm, equation (7-6), may be described by the following difference equation.

$$
\begin{aligned}
P_{n+1}(k) = {} & P_n\left[k + \frac{\nu_n(1 - \bar{\eta})}{\bar{\eta}} \Delta k_n\right] \Pr[\hat{\omega} = 1, \mathrm{Sgn}(x - k) = 1] \\
& + P_n\left[k + \frac{\nu_n(1 + \bar{\eta})}{\bar{\eta}} \Delta k_n\right] \Pr[\hat{\omega} = 1, \mathrm{Sgn}(x - k) = -1] \\
& + P_n\left[k - \frac{\nu_n(1 + \bar{\eta})}{\bar{\eta}} \Delta k_n\right] \Pr[\hat{\omega} = -1, \mathrm{Sgn}(x - k) = 1] \\
& + P_n\left[k - \frac{\nu_n(1 - \bar{\eta})}{\bar{\eta}} \Delta k_n\right] \Pr[\hat{\omega} = -1, \mathrm{Sgn}(x - k) = -1] \quad (7\text{-}10)
\end{aligned}
$$

The substitution of equation (7-9) into (7-10) obtains

$$
\begin{aligned}
P_{n+1}(k) = {} & P_n\left[k + \frac{\nu_n(1-\bar{\eta})}{\bar{\eta}}\Delta k_n\right]\left\{\hat{P}_1 - \hat{F}_1\left[k + \frac{\nu_n(1-\bar{\eta})}{\bar{\eta}}\Delta k_n\right]\right\} \\
& + P_n\left[k + \frac{\nu_n(1+\bar{\eta})}{\bar{\eta}}\Delta k_n\right]\hat{F}_1\left[k + \frac{\nu_n(1+\bar{\eta})}{\bar{\eta}}\Delta k_n\right] \\
& + P_n\left[k - \frac{\nu_n(1+\bar{\eta})}{\bar{\eta}}\Delta k_n\right]\left\{\hat{P}_{-1} - \hat{F}_{-1}\left[k - \frac{\nu_n(1+\bar{\eta})}{\bar{\eta}}\Delta k_n\right]\right\} \\
& + P_n\left[k - \frac{\nu_n(1-\bar{\eta})}{\bar{\eta}}\Delta k_n\right]\hat{F}_{-1}\left[k - \frac{\nu_n(1-\bar{\eta})}{\bar{\eta}}\Delta k_n\right] \qquad (7\text{-}11)
\end{aligned}
$$

Now a differential equation describing the learning process can be obtained. Using a continuous approximation and rearranging equation (7-11), after subtracting both sides of it, $P(n\Delta t,k)$, we obtain

$$
\begin{aligned}
\frac{P(\overline{n+1}\,\Delta t,k) - P(n\Delta t,k)}{\Delta t} \times \Delta t = {} & \frac{P\left[k + \frac{\nu_n(1-\bar{\eta})}{\bar{\eta}}\Delta k_n, n\Delta t\right] - P(k,n\Delta t)}{\frac{2(1-\beta)}{2\beta-1}\nu_n\Delta k_n} \\
& \times \frac{2(1-\beta)}{2\beta-1}\nu_n\Delta k_n \\
& - \frac{\hat{P}_{-1}\left\{P\left[k + \frac{\nu_n(1-\bar{\eta})}{\bar{\eta}}\Delta k_n, n\Delta t\right] - P\left[k - \frac{\nu_n(1+\bar{\eta})}{\bar{\eta}}\Delta k_n, n\Delta t\right]\right\}}{\frac{2\nu_n\Delta k_n}{2\beta-1}} \\
& \times \frac{2\nu_n\Delta k_n}{2\beta-1} + \left\{P\left[k + \frac{\nu_n\Delta k_n(1+\bar{\eta})}{\bar{\eta}}, n\Delta t\right]\hat{F}_1\left[k + \frac{\nu_n\Delta k_n(1+\bar{\eta})}{\bar{\eta}}\right]\right. \\
& \left. - P\left[k + \frac{\nu_n\Delta k_n(1-\bar{\eta})}{\bar{\eta}}, n\Delta t\right]\hat{F}_1\left[k + \frac{\nu_n\Delta k_n(1-\bar{\eta})}{\bar{\eta}}\right]\right\} \times \frac{2\nu_n\Delta k_n}{2\nu_n\Delta k_n} \\
& + \left\{P\left[k - \frac{\nu_n(1-\bar{\eta})}{\bar{\eta}}\Delta k_n, n\Delta t\right]\hat{F}_{-1}\left[k - \frac{\nu_n(1-\bar{\eta})}{\bar{\eta}}\Delta k_n\right]\right. \\
& \left. - P\left[k - \frac{\nu_n(1+\bar{\eta})}{\bar{\eta}}\Delta k_n, n\Delta t\right]\hat{F}_{-1}\left[k - \frac{\nu_n(1+\bar{\eta})}{\bar{\eta}}\Delta k_n\right]\right\} \times \frac{2\nu_n\Delta k_n}{2\nu_n\Delta k_n}
\end{aligned}
$$

(7-12)

Letting $\Delta t \to 0$ and $\Delta k \to 0$, we get the following from equation (7-12).

$$\frac{\partial P}{\partial t} = \frac{\partial P}{\partial k}\nu(t)\frac{\partial k}{\partial t}\frac{2(1-\beta)}{2\beta-1} - \hat{P}_{-1}\frac{\partial P}{\partial k}\nu(t)\frac{\partial k}{\partial t}\frac{2}{2\beta-1}$$

$$+ \frac{\partial P\hat{F}_1}{\partial k}2\nu(t)\frac{\partial k}{\partial t} + \frac{\partial P\hat{F}_{-1}}{\partial k}2\nu(t)\frac{\partial k}{\partial t} \qquad (7\text{-}13)$$

Rewriting equation (7-13) yields

$$\frac{\partial P}{\partial t} = \nu(t)g(t)\frac{\partial P\hat{F}}{\partial k} \qquad (7\text{-}14)$$

where

$$\hat{F} = \frac{2(1-\beta)}{2\beta-1} - \hat{P}_{-1}\frac{2}{2\beta-1} + 2\hat{F}_1 + 2\hat{F}_{-1}$$

$$= 2\left[\hat{F}_1 + \hat{F}_{-1} + \frac{\hat{P}_1 - \beta}{2\beta-1}\right] \qquad (7\text{-}15)$$

$$g(t) = \frac{\partial k}{\partial t} \qquad (7\text{-}16)$$

The conditions on $\nu(t)$ and $g(t)$ for convergence become

$$\left.\begin{aligned} &\int_0^\infty g(t)dt = \infty \\ &0 < g(t) < \infty \\ &\int_0^\infty \nu(t)dt = \infty,\ \int_0^\infty \nu^2(t)dt < \infty,\ \nu(t) \geq 0 \end{aligned}\right\} \qquad (7\text{-}17)$$

The conditional probability of success (S), given k, is

$$S(k) = P[\omega = 1, Sgn(x-k) = 1] + P[\omega = -1, Sgn(x-k) = -1]$$

$$= \int_k^\infty f_1(x)dx + \int_{-\infty}^k f_{-1}(x)dx$$

$$= P_1 - \int_{-\infty}^k f_1(x)dx + \int_{-\infty}^k f_{-1}(x)dx$$

$$= P_1 + F_{-1}(k) - F_1(k) \qquad (7\text{-}18)$$

From equation (2-8), we have

$$P_1 = \frac{1}{2\beta - 1}\left[\beta\hat{P}_1 - (1 - \beta)\hat{P}_{-1}\right]$$

$$= \frac{1}{2\beta - 1}\left(\beta - \hat{P}_{-1}\right) \tag{7-19}$$

From equation (2-12), we have

$$F_{-1}(k) - F_1(k) = \frac{1}{2\beta - 1}\left[\hat{F}_{-1}(k) - \hat{F}_1(k)\right]$$

Thus, S(k) is given by

$$S(k) = \frac{1}{2\beta - 1}\left[\hat{F}_{-1}(k) - \hat{F}_1(k) + (\beta - \hat{P}_{-1})\right] \tag{7-20}$$

The success probability (Z) at any time instant t is defined as

$$Z(t) = \int_{-\infty}^{\infty} p(k,t)S(k)dk \tag{7-21}$$

and P satisfies the differential equation

$$\frac{\partial P}{\partial t} = \nu(t)g(t)\,\frac{\partial P\hat{F}}{\partial k} \tag{7-22}$$

The solution to this equation is given by [11]:

$$\left.\begin{aligned} y(t) &= \int_{t_0}^{t} \nu(u)g(u)du \\ y &= -\int_{k_0}^{V(y)} \frac{dv}{\hat{F}(v)} \\ p(k,t) &= \delta[k - V(y - y_0)] \end{aligned}\right\} \tag{7-23}$$

Then

$$Z(t) = S[V(y - y_0)] \tag{7-24}$$

Hence, Z(t) can be plotted as a function of time to study the learning characteristics of the training algorithm.

8. FEATURE SELECTION CRITERIA WITH IMPERFECTLY LABELED PATTERNS

Probabilistic distance measures are normally used in practice as a feature evaluation criterion for selecting best features. Of all the probabilistic distance measures, the Bhattacharyya distance is most frequently used, since it is easy to evaluate under the Gaussian assumption and its general relationship to the Bayes probability of error. In this section, we present a relationship between the Bayes probability of error and the Bhattacharyya distance with imperfectly labeled patterns under the symmetric model discussed in section 2.

Consider the case of two pattern classes. Let $\hat{P}_e$ and $\hat{\rho}$ be the Bayes probability of error and the Bhattacharyya coefficient with imperfectly labeled patterns. $\hat{\rho}$ is defined as

$$\hat{\rho} = \int [p(X|\hat{\omega} = 1)p(X|\hat{\omega} = 2)]^{1/2} dx \tag{8-1}$$

It is well known [12] that $\hat{P}_e$ and $\hat{\rho}$ are related as

$$\frac{1}{2}\left[1 - \sqrt{1 - P(\hat{\omega} = 1)P(\hat{\omega} = 2)\hat{\rho}^2}\right] \leq \hat{P}_e \leq \sqrt{P(\hat{\omega} = 1)P(\hat{\omega} = 2)}\hat{\rho} \tag{8-2}$$

From section 2, the relationship between the Bayes probability of error with $\hat{P}_e$ and without P_e imperfections is

$$P_e = \frac{\hat{P}_e}{|2\beta - 1|} - \frac{1}{2|2\beta - 1|}(1 - |2\beta - 1|) \tag{8-3}$$

From equations (8-2) and (8-3), the desired relationship is obtained.

$$\frac{1}{2|2\beta - 1|}\left[1 - \sqrt{1 - 4P(\hat{\omega} = 1)P(\hat{\omega} = 2)\hat{\rho}^2}\right] - \frac{1}{2}\frac{(1 - |2\beta - 1|)}{|2\beta - 1|} \leq P_e$$

$$\leq \frac{\sqrt{P(\hat{\omega} = 1)P(\hat{\omega} = 2)}}{|2\beta - 1|}\hat{\rho} - \frac{1}{2|2\beta - 1|}(1 - |2\beta - 1|) \tag{8-4}$$

Similarly, using the relations developed in section 2, other probabilistic distance measures can be studied [6,7,8].

9. REFERENCES

1. Duda, R. O.; and Singleton, R. C.: Training a Threshold Logic Unit With Imperfectly Classified Patterns. Presented at the WESCON Conv. (Los Angeles), Aug. 1964.

2. Whitney, A. W.; and Dwyer, S. J.: Performance and Implementation of the k-Nearest Neighbor Decision Rule With Incorrectly Identified Training Samples. Proc. 4th Annual Allerton Conf., Circuit and System Theory, pp. 96-106, 1966.

3. Kashyap, R. L.; and Blaydon, C. C.: Recovery of Functions From Noisy Measurements Taken at Randomly Selected Points. Proc. IEEE, vol. 54, pp. 1127-1128, 1966.

4. Gimlin, D. R.; and Ferrell, D. R.: A k-k' Error Correcting Procedure for Nonparametric Imperfectly Supervised Learning. IEEE Trans. Systems, Man and Cybernetics, pp. 304-306, May 1974.

5. Shanmugam, K.; and Breipohl, A. M.: An Error Correcting Procedure for Learning With an Imperfect Teacher. IEEE Trans. Systems, Man and Cybernetics, pp. 223-229, July 1971.

6. Chittineni, C. B.: On the Selection of Effective Features From the Imperfectly Labeled Patterns. Int. J. Computer and Information Sciences, vol. 2, no. 2, pp. 103-114, 1973.

7. Chittineni, C. B.: On Feature Extraction From Imperfectly Labeled Patterns. IEEE Trans. Systems, Man and Cybernetics, pp. 290-292, May 1973.

8. Chittineni, C. B.: On the Application of Probabilistic Distance Measures for the Extraction of Features From Imperfectly Labeled Patterns. Proc. 6th Annual Princeton Conf. on Information Sciences and Systems, Mar. 1972.

9. Cover, T. M.; and Hart, P. E.: Nearest Neighbor Pattern Classification. IEEE Trans. Systems, Information Theory, vol. IT-13, pp. 21-27, Jan. 1967.

10. Wilks, S.: Mathematical Statistics. John Wiley & Sons (N.Y.), 1963.

11. Sklansky, J.; and Bershad, N. J.: A Study of Time-Varying Threshold Learning. Tech. Report TP-68-2, School of Engineering, Univ. of Calif. (Irvine), Sept. 1968.

12. Kailath, T.: The Divergence and Bhattacharyya Distance Measures in Signal Selection. IEEE Trans. Comm. Tech., vol. 15, pp. 52-60, Jan. 1967.