General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

Produced by the NASA Center for Aerospace Information (CASI)
TECHNICAL MEMORANDUM

LEARNING WITH IMPERFECTLY LABELED PATTERNS

By

C. B. Chittineni

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

April 1979

LEC-13068
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. INTRODUCTION.</td>
<td>1</td>
</tr>
<tr>
<td>2. A MODEL FOR LABELING IMPERFECTIONS.</td>
<td>2</td>
</tr>
<tr>
<td>3. PERFORMANCE OF BAYES CLASSIFIER WITH AND WITHOUT MISLABELING</td>
<td>5</td>
</tr>
<tr>
<td>3.1 TWO-CLASS BAYES CLASSIFIER PERFORMANCE WITH SYMMETRIC MISLABELING.</td>
<td>5</td>
</tr>
<tr>
<td>3.2 MULTICLASS BAYES ERROR UNDER A GENERAL MISLABELING MODEL</td>
<td>7</td>
</tr>
<tr>
<td>3.2.1 LOWER BOUND.</td>
<td>7</td>
</tr>
<tr>
<td>3.2.2 UPPER BOUND.</td>
<td>8</td>
</tr>
<tr>
<td>3.3 ILLUSTRATION OF BOUNDS</td>
<td>8</td>
</tr>
<tr>
<td>4. PERFORMANCE OF MULTICATEGORY NEAREST NEIGHBOR CLASSIFIER UNDER A GENERAL MODEL OF MISLABELING.</td>
<td>9</td>
</tr>
<tr>
<td>4.1 LOWER BOUND.</td>
<td>10</td>
</tr>
<tr>
<td>4.2 UPPER BOUND.</td>
<td>10</td>
</tr>
<tr>
<td>5. DESIGN OF PATTERN CLASSIFIERS WITH IMPERFECTLY Labeled PATTERNS.</td>
<td>12</td>
</tr>
<tr>
<td>5.1 INCREMENT ERROR CORRECTION CLASSIFIER (NONPARAMETRIC TRAINING).</td>
<td>13</td>
</tr>
<tr>
<td>5.1.1 ALTERNATIVE REPRESENTATION OF IMPERFECTIONS.</td>
<td>13</td>
</tr>
<tr>
<td>5.1.2 AN ERROR CORRECTION ALGORITHM WITH IMPERFECT LABELS.</td>
<td>14</td>
</tr>
<tr>
<td>5.2 BAYES CLASSIFIER (PARAMETRIC TRAINING)</td>
<td>16</td>
</tr>
<tr>
<td>5.3 THE CASE WHEN $b_{ij}$'S ARE UNKNOWN.</td>
<td>16</td>
</tr>
<tr>
<td>6. CORRECTION OF IMPERFECT LABELS OF TRAINING PATTERNS</td>
<td>18</td>
</tr>
<tr>
<td>6.1 LABEL CORRECTION USING k-NEAREST NEIGHBOR DECISION RULE.</td>
<td>18</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.2 LABEL CORRECTION USING BAYES DECISION RULE</td>
<td>19</td>
</tr>
<tr>
<td>6.2.1 DISTRIBUTION OF d(X)</td>
<td>19</td>
</tr>
<tr>
<td>6.2.2 SELECTION OF THRESHOLDS t₁ AND t₂</td>
<td>20</td>
</tr>
<tr>
<td>7. ONE-DIMENSIONAL CLASSIFIER WITH IMPERFEKTLY LABELED PATTERNS</td>
<td>22</td>
</tr>
<tr>
<td>8. FEATURES SELECTION CRITERIA WITH IMPERFEKTLY LABELED PATTERNS</td>
<td>28</td>
</tr>
<tr>
<td>9. REFERENCES</td>
<td>29</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>2-1</td>
<td>Illustration of densities for symmetric labeling errors</td>
</tr>
<tr>
<td>2-2</td>
<td>Illustration of densities for nonsymmetric labeling errors</td>
</tr>
<tr>
<td>3-1</td>
<td>Bayes risk for a symmetric model with and without labeling errors</td>
</tr>
<tr>
<td>3-2</td>
<td>Example illustrating upper and lower bounds</td>
</tr>
<tr>
<td>5-1</td>
<td>Flow diagram for learning with imperfectly labeled patterns when $\beta_{ij}$'s are unknown</td>
</tr>
</tbody>
</table>
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.


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 \( \omega \) 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

\[ R_{ji} = P(m = i|\omega = j) \quad i, j = 1, 2, \cdots, M \]  \hspace{1cm} (2-1)

where \( i \) and \( j \) indicate class, we have the constraint

\[ \sum_{i=1}^{M} R_{ji} = 1 \]  \hspace{1cm} (2-2)

Assume further that

\[ p(X|\hat{\omega}, \omega = j) = p(X|\omega) \]  \hspace{1cm} (2-3)

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

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

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

\[ = \frac{1}{p(\hat{\omega} = i)} \sum_{j=1}^{M} R_{ji} p(\omega = j)p(X|\omega = j) \]  \hspace{1cm} (2-4)

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

\[ p(\hat{\omega} = i|X) = \sum_{j=1}^{M} R_{ji} p(\omega = j|X) \]  \hspace{1cm} (2-5)
Similarly, the a priori probabilities are related as

\[
P(\hat{\omega} = i) = \sum_{j=1}^{M} \beta_{ij}^* P(\omega = j) \tag{2-6}
\]

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

\[
\begin{align*}
P(\omega = 1)p(X\mid\omega = 1) &= \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}} \left[ \beta_{22} P(\hat{\omega} = 1)p(X\mid\hat{\omega} = 1) \\
- \beta_{21} P(\hat{\omega} = 2)p(X\mid\hat{\omega} = 2) \right] \\
\end{align*}
\]

\[
\begin{align*}
P(\omega = 2)p(X\mid\omega = 2) &= \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}} \left[ \beta_{11} P(\hat{\omega} = 2)p(X\mid\hat{\omega} = 2) \\
- \beta_{12} P(\hat{\omega} = 1)p(X\mid\hat{\omega} = 1) \right] \\
\end{align*}
\]  

(2-7)

Similarly, for the a priori and a posteriori probabilities,

\[
P(\omega = 1) = \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}} \left[ \beta_{jj} P(\hat{\omega} = i) - \beta_{ij}^* P(\hat{\omega} = j) \right] \quad ; \quad i,j = 1,2 \tag{2-8}
\]

For a symmetric case, when

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

then

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

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

\[
\begin{align*}
[P(\omega = 1)p(X\mid\omega = 1) - P(\omega = 2)p(X\mid\omega = 2)] &= \frac{1}{2\beta - 1} \left[ P(\hat{\omega} = 1)p(X\mid\hat{\omega} = 1) - P(\hat{\omega} = 2)p(X\mid\hat{\omega} = 2) \right] \\
\end{align*}
\]  

(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

\[
\begin{align*}
\text{Decide } X \in \omega = 1 & \text{ if } P(\omega = 1)p(X|\omega = 1) > P(\omega = 2)p(X|\omega = 2) \\
\text{Decide } X \in \omega = 2 & \text{ otherwise }
\end{align*}
\]

(3-1)

The Bayes probability of error ($P_e$) is

\[
P_e = \int \min[P(\omega = 1)p(X|\omega = 1), P(\omega = 2)p(X|\omega = 2)]dX
\]

(3-2)

For any two positive real numbers $A$ and $B$,

\[
\min(A, B) = \frac{1}{2}(A + B) - \frac{1}{2}|A - B|
\]

(3-3)

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

\[
P_e = \frac{1}{2} - \frac{1}{2} \int \left| P(\omega = 1)p(X|\omega = 1) - P(\omega = 2)p(X|\omega = 2) \right| dX
\]

(3-4)

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

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

(3-5)

From equations (2-12), (3-4), and (3-5), we obtain the following.
\[ \hat{P}_e = \frac{1}{2} - \frac{1}{2}|2\beta - 1| \int |P(\omega = 1)p(X|\omega = 1) - P(\omega = 2)p(X|\omega = 2)| dX \]

\[ = \frac{1}{2} - \frac{1}{2}|2\beta - 1|(1 - 2\hat{P}_e) \]

\[ = \frac{1}{2}(1 - |2\beta - 1|) + |2\beta - 1|\hat{P}_e \quad (3-6) \]

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

\[ P_e = \frac{1}{|2\beta - 1|} \hat{P}_e + \frac{1}{2} \left( 1 - \frac{1}{|2\beta - 1|} \right) \quad (3-7) \]

Equation (3-7) is graphically displayed for various values of \( \beta \) 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 \( \beta \).
### 3.2 Multiclass Bayes Error Under a General Mislabeling Model

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

\[
r(X) = 1 - \max_i \left[ p(\omega = i | X) \right] \quad (3-8)
\]

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

Then the probability of error can be written as

\[
P_e = E[p(X)[r(X)] \quad (3-10)
\]

\[
\hat{P}_e = E[p(X)[\hat{r}(X)] \quad (3-11)
\]

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

\[
K = \max_i \left[ p(\hat{\omega} = i | X) \right] \quad (3-12)
\]

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

where RSM = row sum maximum.

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

\[
\hat{r}(X) = 1 - \max_i \left[ \sum_{j=1}^{M} \beta_{ji} p(\omega = j | X) \right] \geq 1 - K \left( \max_j \left( \sum_{i=1}^{M} \beta_{ji} \right) \right) = 1 - K \beta_{RSM} = (1 - \beta_{RSM}) + \beta_{RSM} \hat{r}(X) \quad (3-14)
\]
Taking the expectations on both sides of equation (3-14) with respect to \( p(X) \) obtains the desired inequality

\[
\hat{P}_e \geq (1 - \beta_{\text{RSM}}) + \beta_{\text{RSM}} P_e
\]  

(3-15)

### 3.2.2 UPPER BOUND

Let

\[
p(\omega = k|X) = \max_j p(\omega = j|X)
\]  

(3-16)

\[
b_i = \min_j (\beta_{kj})
\]  

(3-17)

\[
\alpha = \min_k \max_i (\beta_{ki} - b_i)
\]  

(3-18)

where \( k \) is a scalar indicating class.

Consider

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

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

\[
\geq \alpha [1 - r(X)]
\]  

(3-19)

However,

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

(3-20)

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

\[
\hat{P}_e \leq (1 - \alpha) + \alpha P_e
\]  

(3-21)

### 3.3 ILLUSTRATION OF BOUNDS

To illustrate the upper and lower bounds, let \( M = 3 \). Consider a matrix of mislabeling probabilities \( \beta_{ji} \) as shown in figure 3-2. The various quantities
required to compute the bounds are also shown.

\[
\begin{align*}
\beta_{ji} &= \begin{bmatrix}
0.9 & 0.05 & 0.05 \\
0.05 & 0.9 & 0.05 \\
0.05 & 0.05 & 0.9
\end{bmatrix} \\
\max_{i} (\beta_{ki} - b_i) &= \begin{bmatrix}
0.85 \\
1 \\
1
\end{bmatrix} \\
b_i &= \min_{j} (\beta_{ji}) \\
&= 0.05
\end{align*}
\]

\[\beta_{RSM} = 1, \alpha = 0.85\]

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

\[0.059 < P_e < 0.2 \quad (3-22)\]

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

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

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

\[= 1 - \sum_{i=1}^{M} \sum_{j=1}^{M} \beta_{ji} p(\omega = i|X)p(\omega = j|X) \quad (4-1)\]
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.

\[ \hat{r}_N(X) \geq 1 - K \sum_{i=1}^{M} \sum_{j=1}^{M} \beta_{ij} p(w = j|X) \]

\[ = 1 - K \sum_{j=1}^{M} p(w = j|X) \sum_{i=1}^{M} \beta_{ij} \]

\[ = 1 - K \]

\[ = r(X) \quad (4-2) \]

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

\[ \hat{P}_{eN} \geq P_e \quad (4-3) \]

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

4.2 UPPER BOUND
Let

\[ p(w = k|X) = \max_i [p(w = i|X)] \quad (4-4) \]

\[ \beta = \min_i (\beta_{ii}) \quad (4-5) \]

\[ \beta_m = \min_{j,k} (\beta_{jk}) \quad (4-6) \]

Consider the following.
\[
\sum_{i=1}^{M} \sum_{j=1}^{M} \beta_{ji} p(\omega = i | X)p(\omega = j | X) = \sum_{i=1}^{M} \beta_{ii} p^2(\omega = i | X)
\]

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

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

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

\[
+ \sum_{i=1, i \neq k}^{M} p(\omega = i | X) \sum_{j=1, j \neq i}^{M} \beta_{ji} p(\omega = j | X)
\]

(4-7)

However,

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

\[
= \frac{1}{M - 1} [1 - p(\omega = k | X)]^2
\]

(4-8)

Now consider

\[
p(\omega = k | X) \sum_{j=1, j \neq k}^{M} \beta_{jk} p(\omega = j | X) \geq \beta_m p(\omega = k | X) [1 - p(\omega = k | X)]
\]

(4-9)

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

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

\[
+ \beta_m r(X) [1 - r(X)]
\]

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

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

\[ = (1 - \beta) + (2\beta - \beta_m) r(X) - \left( \frac{M}{M - 1} \beta - \beta_m \right) r^2(X) \]  \hspace{1cm} (4-11)

But we have

\[ E_p (X)[r^2(X)] = \text{Var}[r(X)] + p_e^2 \geq p_e^2 \] \hspace{1cm} (4-12)

\[ \frac{M}{M - 1} \beta - \beta_m \geq 0 \] \hspace{1cm} (4-13)

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

\[ \hat{p}_{e_N} \leq (1 - \beta) + (2\beta - \beta_m) p_e - \left( \frac{M}{M - 1} \beta - \beta_m \right) p_e^2 \] \hspace{1cm} (4-14)

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

\[ \hat{p}_{e_N} \leq 2p_e - \frac{M}{M - 1} p_e^2 \] \hspace{1cm} (4-15)

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 \( X_1(1), X_1(2), \ldots, X_1(N_1) \) and \( X_{-1}(1), X_{-1}(2), \ldots, X_{-1}(N_{-1}) \) from classes 1 and -1, with imperfect labels \( \hat{w}_1(1), \ldots, \hat{w}_1(N_1); \hat{w}_{-1}(1), \ldots, \hat{w}_{-1}(N_{-1}) \). Let the perfect labels of these patterns be \( w_1(1), \ldots, w_1(N_1); w_{-1}(1), \ldots, w_{-1}(N_{-1}) \). The imperfect and perfect labels take values of 1 and -1. The objective is to find a decision function \( d(X) \), such that

\[
\begin{align*}
\text{d}(X) > 0 \text{ when } X \in w_1 \\
\text{d}(X) < 0 \text{ when } X \in w_{-1}
\end{align*}
\]

5.1.1 ALTERNATIVE REPRESENTATION OF IMPERFECTIONS

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

\[
\hat{\omega} = \omega \eta
\]

where \( \eta \) = labeling noise and takes values of 1 and -1. Since \( \omega \) takes 1 and -1, whenever \( \omega \) differs from \( \eta \) we have an imperfect label. Recalling our previous model of imperfections, we have

\[
\beta_{ij} = P(\hat{\omega} = j | \omega = i) \quad i, j = \pm 1
\]

Let

\[
\bar{\eta} = E(\eta)
\]

and

\[
P_i = P(\omega = i) \quad i = \pm 1
\]

Since \( \eta \) takes values of 1 or -1, the average value of \( \eta, \bar{\eta} \), can be written as follows.
\[ \overline{n} = E(\eta) \]
\[ = P(\eta = 1) - P(\eta = -1) \]
\[ = P(\eta = 1, \omega = 1) + P(\eta = 1, \omega = -1) - P(\eta = -1, \omega = 1) \]
\[ - P(\eta = -1, \omega = -1) \]
\[ = P_1P(\eta = 1 | \omega = 1) + P_{-1}P(\eta = 1 | \omega = -1) - P_1P(\eta = -1 | \omega = 1) \]
\[ - P_{-1}P(\eta = -1 | \omega = -1) \]
\[ = P_1P(\hat{\omega} = 1 | \omega = 1) + P_{-1}P(\hat{\omega} = -1 | \omega = -1) - P_1P(\hat{\omega} = -1 | \omega = 1) \]
\[ - P_{-1}P(\hat{\omega} = 1 | \omega = -1) \]
\[ = P_1\beta_{1,1} + P_{-1}\beta_{-1,-1} - P_1\beta_{1,-1} - P_{-1}\beta_{-1,1} \]
\[ = (2\beta_{1,1} - 1)P_1 + (2\beta_{-1,-1} - 1)P_{-1} \quad (5-6) \]

Under a symmetric model, the above becomes

\[ \overline{n} = 2\beta - 1 \quad (5-7) \]

5.1.2 AN ERROR CORRECTION ALGORITHM WITH IMPERFECT LABELS

To obtain a linear approximation to \( d(X) \),

\[ d(X) = X^T W \quad (5-8) \]

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

\[ C(W) = E[\alpha(W,X)] \quad (5-9) \]

where

\[ \alpha(W,X) = W^T X \{ \text{Sgn}(W^T X) - \text{Sgn}[d_0(X)] \} \quad (5-10) \]

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

\[ C(W) \geq C(W^*) \quad (5-11) \]
The corresponding optimal approximation $d^*(X)$ to $d_0(X)$ is given by

$$d^*(X) = X^T W^* \quad (5-12)$$

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

$$W(\lambda + 1) = W(\lambda) - \nu(\lambda) \frac{\partial C(W)}{\partial W} \quad (5-13)$$

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

$$W(\lambda + 1) = W(\lambda) - \nu(\lambda) g(W,X) \quad (5-14)$$

where $X(\lambda)$ is the training pattern at the $\lambda$th step and

$$g(W,X) = \{Sgn(W^TX) - Sgn[d_0(X)]\}_X \quad (5-15)$$

Since the perfect label $Sgn[d_0(X)]$ 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

$$f[W,X,\hat{\omega}(X)] = \frac{1}{n} \left[ \eta Sgn(W^TX) - \hat{\omega}(X) \right]_X \quad (5-16)$$

$$E[f[W,X,\hat{\omega}(X)]|W] = E \left[ \left( \frac{1}{n} \left[ \eta Sgn(W^TX) - \hat{\omega}(X) \right]_X + \{Sgn[d_0(X)] - Sgn[d_0(X)]\}_X \right) | W \right]$$

$$= E \left\{ \{Sgn(W^TX) - Sgn[d_0(X)]\}_X | W \right\}$$

$$+ E \left( \left\{ \{Sgn[d_0(X)] - \frac{1}{n} Sgn[d_0(X)] \}_X \right\} | W \right)$$

$$= E \left( \{Sgn(W^TX) - Sgn[d_0(X)]\}_X | W \right)$$

$$= E[g(W,X)|W] \quad (5-17)$$

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

$$W(\lambda + 1) = W(\lambda) - \nu(\lambda) \frac{1}{n} \{\eta Sgn[X^TW(\lambda)] - \hat{\omega}(X)\}_X \quad (5-18)$$
where $X$ is the training pattern at the $\lambda$th step of training and $\Pi$ is given by equation (5-6). For the convergence of this algorithm, the conditions on $v(\lambda)$ can be shown to be [3]:

$$v(\lambda) \geq 0, \sum_{\lambda=1}^{\infty} v(\lambda) = \infty, \sum_{\lambda=1}^{\infty} v^2(\lambda) < \infty \quad (5-19)$$

5.2 BAYES CLASSIFIER (PARAMETRIC TRAINING)

Once the a priori probabilities and class conditional densities of the imperfectly labeled patterns and the mislabeling probabilities $\beta_{ij}$'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.

Decide $X \in \omega = i$ if $P(\omega = i)p(X|\omega = i) > \max_{\lambda}[P(\omega = \lambda)p(X|\omega = \lambda)] \quad (5-20)$

5.3 THE CASE WHEN $\beta_{ij}$'S ARE UNKNOWN

In the case when $\beta_{ij}$'s are unknown, the scheme shown in figure 5-1 can be used to design the classifier.
ASSUME VALUES FOR $\beta_{ij}$

ESTIMATE THE VALUES OF THE DIFFERENT PARAMETERS

DESIGN THE CLASSIFIER

ESTIMATE THE RISK FUNCTION

COMPUTE $P_e$ USING UNCLASSIFIED SAMPLES

IS THE CURRENT $P_e$ SMALLER THAN THE PREVIOUS $P_e$?

NO

YES

CHANGE $\beta_{ij}$'S

TAKE CURRENT PROBABILITIES AND DESIGN THE CLASSIFIER

STOP

Figure 5-1.— Flow diagram for learning with imperfectly labeled patterns when $\beta_{ij}$'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 \( X_1, X_2, \ldots, X_{n-1} \) 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.,

\[ p(X|\omega = i) \sim N(\hat{M}_i, \hat{S}_i) \quad ; \quad i = 1, 2 \quad (6-1) \]

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

\[
\begin{align*}
\text{Decide } X \in \hat{\omega} = 1 \text{ if } d(X) > 0 \\
\text{Decide } X \in \hat{\omega} = 2 \text{ if } d(X) < 0 
\end{align*}
\quad (6-2)
\]

where

\[
d(X) = \log \frac{P(\omega = 1)p(X|\omega = 1)}{P(\omega = 2)p(X|\omega = 2)}
\quad (6-3)
\]

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

\[
\begin{align*}
\text{Change the label of } X \text{ to } \omega = 1 \text{ if } d(X) > t_1 \\
\text{Change the label of } X \text{ to } \omega = 2 \text{ if } d(X) < -t_2 \\
\text{Do not change the label of } X \text{ if } -t_2 \leq d(X) \leq t_1
\end{align*}
\quad (6-4)
\]

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 \( \sigma_i \), \( i = 1, 2; \) i.e., \( p[d(X)|X \in \hat{\omega} = i] = N(m_i, \sigma_i^2) \). Then,
the expressions for $m_i$ and $\sigma_i$ can be shown to be [10]:

\[
\begin{align*}
m_1 &= \text{tr}(I - \hat{\Sigma}_2 \hat{\Sigma}_1^{-1}) - (\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}_1^{-1} (\hat{\mu}_1 - \hat{\mu}_2) + 2\ln \left( \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} \right) \\
&\quad - 2\ln \left[ \frac{P(\hat{\omega} = 1)}{P(\hat{\omega} = 2)} \right] \\
m_2 &= \text{tr}(\hat{\Sigma}_1 \hat{\Sigma}_2^{-1} - I) + (\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}_2^{-1} (\hat{\mu}_1 - \hat{\mu}_2) + 2\ln \left( \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} \right) \\
&\quad - 2\ln \left[ \frac{P(\hat{\omega} = 1)}{P(\hat{\omega} = 2)} \right] \\
\sigma_i^2 &= 2\left\{ \text{tr} \left[ (I - \hat{\Sigma}_2 \hat{\Sigma}_1^{-1})^2 \right] \\
&\quad + 2(\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}_2^{-1} \hat{\Sigma}_1 \hat{\Sigma}_2^{-1} (\hat{\mu}_1 - \hat{\mu}_2) \right\}
\end{align*}
\]  

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 $\alpha$ that mislabeling will occur in the correction process.

\[
\alpha = P(\text{bad label})
\]

\[
= P(\omega = 1)P(\text{bad label} | X \in \omega = 1) + P(\omega = 2)P(\text{bad label} | X \in \omega = 2)
\]

\[
= P(\omega = 1)P(d(X) < -t_2 | X \in \omega = 1) + P(\omega = 2)P[d(X)] > t_1 | X \in \omega = 2)
\]

(6-6)
Following the argument and assumptions similar to those of section 2, it can be shown that

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

(6-7)

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

(6-8)

Then,

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

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

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

(6-9)

where \(N(0,1)\) 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.
\[ P(\omega = 2)P(\text{bad label}|X \in \omega_2) = P(\omega = 2) \int_{t_1}^{\infty} p[d(X)|X \in \omega_2]d[d(X)] \]

\[ = \frac{1}{\beta_{11}^2 \beta_{22} - \beta_{12}^2 \beta_{21}} \left[ \beta_{11} P(\hat{\omega} = 2) \int_{-t_1 + m_1}^{t_1 + m_1} N(0,1)d\xi \right. \]

\[ \left. - \beta_{12} P(\hat{\omega} = 1) \int_{-t_1 + m_1}^{t_1 + m_1} N(0,1)d\xi \right] \]  

For a specified \( \alpha \), 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 \( \omega \) 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.,

\[ P(\hat{\omega} = 1|\omega = 1) = \beta = P(\hat{\omega} = -1|\omega = -1) \]

then

\[ P(\hat{\omega} = -1|\omega = 1) = 1 - \beta = P(\hat{\omega} = 1|\omega = -1) \]  

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

\[ \hat{\omega} = \omega \eta \]  

In section 5.1, \( \bar{\eta} = E(\eta) \) is related to \( \beta \) as

\[ \bar{\eta} = 2\beta - 1 \]  

The decision rule for one-dimensional patterns is

\[ \text{Decide } x_n \in \omega = 1, \text{ if } x_n > k_n \]

\[ \text{Decide } x_n \in \omega = -1, \text{ if } x_n < k_n \]
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.

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

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

\[
k_{n+1} = k_n - \frac{\nu_n}{n} [\hat{\omega}(x_n) - \text{Sgn}(x_n - k_n)] \Delta k_n
\]  
(7-6)

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

\[
\nu_n > 0, \quad \sum_{n=1}^{\infty} \nu_n = \infty, \quad \sum_{n=1}^{\infty} \nu_n^2 < \infty
\]  
(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

\[
P_1 = \int_{-\infty}^{\infty} f_1(x) \, dx \quad \text{and} \quad P_{-1} = \int_{-\infty}^{\infty} f_{-1}(x) \, dx
\]

\[
\hat{P}_1 = \int_{-\infty}^{\infty} \hat{f}_1(x) \, dx \quad \text{and} \quad \hat{P}_{-1} = \int_{-\infty}^{\infty} \hat{f}_{-1}(x) \, dx
\]

Then, the following probabilities can be easily obtained.
\[
\begin{align*}
\Pr[\hat{\omega} = 1, \text{Sgn}(x - k) = 1] &= \int_{k}^{\infty} \hat{f}_1(x)\,dx \\
&= \hat{p}_1 - \int_{-\infty}^{k} \hat{f}_1(x)\,dx \\
&= \hat{p}_1 - \hat{f}_1(k) \\
\Pr[\hat{\omega} = 1, \text{Sgn}(x - k) = -1] &= \int_{-\infty}^{k} \hat{f}_{-1}(x)\,dx \\
&= \hat{f}_{-1}(k) \\
\Pr[\hat{\omega} = -1, \text{Sgn}(x - k) = 1] &= \int_{k}^{\infty} \hat{f}_{-1}(x)\,dx \\
&= \hat{p}_{-1} - \hat{f}_{-1}(k) \\
\Pr[\hat{\omega} = -1, \text{Sgn}(x - k) = -1] &= \int_{-\infty}^{k} \hat{f}_1(x)\,dx \\
&= \hat{f}_1(k)
\end{align*}
\] (7-9)

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.

\[
\begin{align*}
p_{n+1}(k) &= p_n\left[k + \frac{v_n(1 - \bar{n})}{\bar{n}} \Delta k_n\right]\Pr[\hat{\omega} = 1, \text{Sgn}(x - k) = 1] \\
&+ p_n\left[k + \frac{v_n(1 + \bar{n})}{\bar{n}} \Delta k_n\right]\Pr[\hat{\omega} = 1, \text{Sgn}(x - k) = -1] \\
&+ p_n\left[k - \frac{v_n(1 + \bar{n})}{\bar{n}} \Delta k_n\right]\Pr[\hat{\omega} = -1, \text{Sgn}(x - k) = 1] \\
&+ p_n\left[k - \frac{v_n(1 - \bar{n})}{\bar{n}} \Delta k_n\right]\Pr[\hat{\omega} = -1, \text{Sgn}(x - k) = -1] \quad (7-10)
\end{align*}
\]
The substitution of equation (7-9) into (7-10) obtains

\[ p_{n+1}(k) = p_n \left[ k + \frac{\nu_n(1 - \bar{n})}{\bar{n}} \Delta k_n \right] \left\{ \hat{\beta}_1 - \hat{\beta}_1 \left[ k + \frac{\nu_n(1 - \bar{n})}{\bar{n}} \Delta k_n \right] \right\} \\
+ p_n \left[ k + \frac{\nu_n(1 + \bar{n})}{\bar{n}} \Delta k_n \right] \hat{\beta}_1 \left[ k + \frac{\nu_n(1 + \bar{n})}{\bar{n}} \Delta k_n \right] \\
+ p_n \left[ k - \frac{\nu_n(1 + \bar{n})}{\bar{n}} \Delta k_n \right] \left\{ \hat{\beta}_1 - \hat{\beta}_1 \left[ k - \frac{\nu_n(1 + \bar{n})}{\bar{n}} \Delta k_n \right] \right\} \\
+ p_n \left[ k - \frac{\nu_n(1 - \bar{n})}{\bar{n}} \Delta k_n \right] \hat{\beta}_1 \left[ k - \frac{\nu_n(1 - \bar{n})}{\bar{n}} \Delta k_n \right] \tag{7-11} \]

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

\[ \frac{P(n + 1\Delta t, k) - P(n\Delta t, k)}{\Delta t} = \frac{p \left[ k + \frac{\nu_n(1 - \bar{n})}{\bar{n}} \Delta k_n, n\Delta t \right] - P(k, n\Delta t)}{2} \frac{2(1 - \beta)}{2\beta - 1} \frac{\nu_n \Delta k_n}{\Delta t} \]

\[ \times \left\{ \hat{\beta}_1 \left[ k + \frac{\nu_n(1 - \bar{n})}{\bar{n}} \Delta k_n, n\Delta t \right] - \hat{\beta}_1 \left[ k + \frac{\nu_n(1 + \bar{n})}{\bar{n}} \Delta k_n, n\Delta t \right] \right\} \\
\times \frac{2\nu_n \Delta k_n}{2\beta - 1} + \left\{ p \left[ k + \frac{\nu_n \Delta k_n(1 + \bar{n})}{\bar{n}}, n\Delta t \right] \hat{\beta}_1 \left[ k + \frac{\nu_n \Delta k_n(1 + \bar{n})}{\bar{n}} \right] \\
- P \left[ k + \frac{\nu_n \Delta k_n(1 - \bar{n})}{\bar{n}}, n\Delta t \right] \hat{\beta}_1 \left[ k + \frac{\nu_n \Delta k_n(1 - \bar{n})}{\bar{n}} \right] \right\} \times \frac{2\nu_n \Delta k_n}{2\beta - 1} \\
+ \left\{ P \left[ k - \frac{\nu_n(1 - \bar{n})}{\bar{n}} - \Delta k_n, n\Delta t \right] \hat{\beta}_1 \left[ k - \frac{\nu_n(1 - \bar{n})}{\bar{n}} \Delta k_n \right] \right\} \times \frac{2\nu_n \Delta k_n}{2\beta - 1} \\
- P \left[ k - \frac{\nu_n(1 + \bar{n})}{\bar{n}} - \Delta k_n, n\Delta t \right] \hat{\beta}_1 \left[ k - \frac{\nu_n(1 + \bar{n})}{\bar{n}} \Delta k_n \right] \times \frac{2\nu_n \Delta k_n}{2\beta - 1} \tag{7-12} \]
Letting $\Delta t \to 0$ and $\Delta k \to 0$, we get the following from equation (7-12).

\[
\frac{\partial P}{\partial t} = \frac{\partial P}{\partial k} \frac{\partial v(t)}{\partial t} \frac{2(1 - \beta)}{2 \beta - 1} - \hat{P}_{-1} \frac{\partial P}{\partial k} \frac{\partial v(t)}{\partial t} \frac{2}{2 \beta - 1} \\
+ \frac{\partial P \hat{F}_1}{\partial k} 2v(t) \frac{\partial k}{\partial t} + \frac{\partial P \hat{F}_{-1}}{\partial k} 2v(t) \frac{\partial k}{\partial t}
\]  

(7-13)

Rewriting equation (7-13) yields

\[
\frac{\partial P}{\partial t} = v(t)g(t) \frac{\partial P \hat{F}}{\partial k}
\]  

(7-14)

where

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

(7-15)

\[
g(t) = \frac{\partial k}{\partial t}
\]  

(7-16)

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

\[
\begin{align*}
\int_0^\infty g(t)dt &= \infty \\
0 < g(t) < \infty
\end{align*}
\]  

(7-17)

\[
\int_0^\infty v(t)dt = \infty, \int_0^\infty v^2(t)dt < \infty, v(t) \geq 0
\]

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

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

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

(7-18)

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

\[
= P_1 + F_{-1}(k) - F_1(k)
\]
From equation (2-8), we have

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

\[ = \frac{1}{2^\beta - 1} (\beta - \hat{P}_{-1}) \]  

(7-19)

From equation (2-12), we have

\[ F_1(k) - F_1(k) = \frac{1}{2^\beta - 1} [\hat{F}_1(k) - \hat{F}_{-1}(k)] \]

Thus, \( S(k) \) is given by

\[ S(k) = \frac{1}{2^\beta - 1} [\hat{F}_1(k) - \hat{F}_{-1}(k) + (\beta - \hat{P}_{-1})] \]  

(7-20)

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

\[ Z(t) = \int_{-\infty}^{\infty} p(k,t)S(k)dk \]  

(7-21)

and \( P \) satisfies the differential equation

\[ \frac{dP}{dt} = v(t)g(t) \frac{d\hat{P}}{dk} \]  

(7-22)

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

\[ y(t) = \int_{t_0}^{t} v(u)g(u)du \]

\[ y = -\int_{k_0}^{V(y)} \frac{dv}{\hat{F}(v)} \]

\[ p(k,t) = \delta[k - V(y - y_0)] \]

Then

\[ Z(t) = S[V(y - y_0)] \]  

(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 \( \hat{\beta} \) be the Bayes probability of error and the Bhattacharyya coefficient with imperfectly labeled patterns. \( \hat{\beta} \) is defined as

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

It is well known [12] that \( \hat{P}_e \) and \( \hat{\beta} \) are related as

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

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.

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

\[
\leq \sqrt{P(\hat{\omega} = 1)P(\hat{\omega} = 2)\hat{\beta}} - \frac{1}{2|2\beta - 1|} \frac{1}{2|2\beta - 1|} (1 - |2\beta - 1|) \tag{8-4}
\]
Similarly, using the relations developed in section 2, other probabilistic distance measures can be studied [6,7,8].

9. REFERENCES


