Error probability in decision functions for character recognition*

by J. T. CHU and J. C. CHUEH**
University of Pennsylvania
Philadelphia, Pennsylvania

INTRODUCTION

In this paper, we investigate the evaluation and reduction of error probability, when statistical decision functions are used for computer character recognition. Suppose that the given alphabet consists of m characters, $\Theta_1, \Theta_2, \ldots, \Theta_m$, and that a character Θ is to be identified by the observed value of a random vector $X = (X_1, X_2, \ldots, X_n)$, where each X_k is associated with a certain feature of Θ. Let p_i be the probability that $\Theta = \Theta_i$, $i = 1,2, \ldots, m$; and $f_i(x)$, where $x = (x_1, x_2, \ldots, x_n)$ is a real vector, be the pdf (probability density function) of X given that $\Theta = \Theta_i$. In order to minimize the probability of error, i.e., incorrect recognition, it is well known [3] that Bayes decision functions should be used. Namely, one identifies Θ as Θ_i, if the observed value x of X is in

$$T_i = \{x: p_i f_i(x) = \max p_j f_j(x), j = 1,2, \ldots, m\}. \tag{1}$$

In case x belongs to more than one T_i, Θ may be identified as the one corresponding to that T_i with the smallest subscript i.

When p_i and $f_i(x)$ are all given, a Bayes decision function is simple to apply, since all that one has to do is to observe X and compare for different i the value of $p_i f_i(x)$. On the other hand, the corresponding error probability is generally difficult to evaluate. Furthermore, the X_i's and n which depend on the types and number of features used for recognition, are usually not given at the start. One would then like to know how they should be chosen so as to reduce the error probability of the recognition system below a certain level. These problems of evaluating and reducing the error probability are obviously of importance in applications. However, general solutions to such problems have been so far lacking.

*Research supported in whole by U.S. Public Health Service Grant GMO9054 from the National Institute of General Medical Sciences.

**Now with the Bell Telephone Laboratories, Holmdel, New Jersey.

In this paper, we shall see that some solutions of a general nature can be obtained by the use of the Bayes majority decision functions defined in Section 3. Upper bounds for the error probability are derived in terms of the "differences" between pairs of $f_i(x)$ and $f_{ik}(x_k)$ respectively, where $f_{ik}(x_k)$ is the pdf of X_k given $\Theta = \Theta_i$. From there, we obtain the main result that if a sufficient number of features is used in a recognition system, and corresponding to each and every feature X_k, the $f_{ik}(x_k)$ have positive "differences" among themselves, then the error probability of the Bayes decision function can be made arbitrarily small. Hence, to set up a character recognition system, the following procedure may be considered:

(a) Select a set of features having the largest possible "differences" among the corresponding pdfs.

(b) Determine the number of features to be used by the requirement on error probability and/or cost consideration.

(c) Use Bayes decision functions to identify the characters.

On the other hand, for a given recognition system, the upper bounds mentioned above may be used as conservative approximations to the error probability.

The details are presented in separate sections. In Section 2, we define the "difference" between $f_i(x)$ and $f_j(x)$ and obtain relations between "difference" and error probability. In Section 3, we introduce the majority decision functions from which an upper bound for the error probability is derived. In Section 4, we discuss various kinds of applications and give illustrative examples where the pdfs are binomial and normal respectively. Some numerical comparisons are also made.

A special case

Consider first the case where the alphabet contains only two characters, i.e., $\Theta = \Theta_1$ or Θ_2. Then the Bayes decision function given in (1) recognizes Θ as Θ_i if $x \in S_i$, $i = 1,2$, where $S_2 = S'$, the complement of S_1, and

$$S_1 = \{x: p_1 f_1(x) \geqslant p_2 f_2(x)\}. \qquad (2)$$

The corresponding error probability is given by

$$P_e = \int_{S_1} p_2 f_2(x)\, dx + \int_{S_2} p_1 f_1(x)\, dx.$$

To simplify the notations, we shall from now on write f_i instead of $f_i(x)$ and delete "dx" from integration, whenever there is no danger of confusion.

Intuitively, it seems clear that the more different f_1 and f_2 are, the less P_e should be. In the following, we shall see that this is indeed the case.

Theorem 1. If $\int | p_1 f_1 - p_2 f_2 | \geqslant \delta$, then $P_e \leqslant (1-\delta)/2$, and equalities correspond.

Proof. By assumption, $\sum_{i=1}^{2} \int S_i p_i f_i - P_e \geqslant \delta$. Since $\sum_{i=1}^{2}$ $+ P_e = 1$, the assertions follow immediately.

Theorem 2. If $\int |f_1 - f_2| \geqslant 2\,\delta$, then $P_e \leqslant 1/2 - \delta/4$.

Proof. Let $R_1 = \{x: f_1(x) \geqslant f_2(x)\}$ and R_1, the complement of R_1. Assume that $p_1 \geqslant p_2$. Then $1 - P_e = \int_{S_1} p_1 f_1 + \int_{S_2} p_2 f_2 \geqslant \int_{R_1} p_1 f_1 + \int_{S_1 R'_1} p_1 f_1 + \int_{S_2} p_1 f_1$.

By assumption, $\int_{R_1} (f_1 - f_2) \geqslant \delta$, hence $\int_{R_1} p_1 f_1 \geqslant p_1\,\delta + \int_{R_1} p_1 f_2 \geqslant p_1\,\delta + \int_{R_1} p_2 f_2$. Furthermore, $\int_{S_1 R'_1} p_1 f_1 \geqslant \int_{S_1 R'_1} p_2 f_2$. It follows that $1 - P_e \geqslant p_1\,\delta + P_e$. Hence, $P_e \leqslant 1/2 - p_1\,\delta/2 \leqslant 1/2 - \delta/4$. In a similar way, we show that the theorem holds for $p_1 \leqslant p_2$.

The integral $\int | p_1 f_1 - p_2 f_2 |$ may be viewed as a weighted difference between f_1 and f_2. Theorem 1 says that if the weighted difference is δ, then $P_e = (1 - \delta)/2$. Furthermore, in order to reduce P_e below a given level α, one must select X such that the corresponding weighted difference is at least $1 - 2\,\alpha$. The ideal case is that the weighted difference is 1, since then $P_e = 0$. The integral in Theorem 2 may be viewed as an unweighted difference between f_1 and f_2 and plays a similar role. But the result of Theorem 2 is somewhat weaker in that equalities do not correspond, that $\delta = 1$ does not imply $P_e = 0$, and that for $p_1 = p_2 = 1/2$, Theorem 1 provides a better upper bound.

We have just seen that the evaluation and reduction of error probability depend very much on the differences between the pdfs of X. However, these differences are in general not easy to obtain. In the next section, we shall derive an upper bound for the error probability in terms of the differences between the pdfs of each and every X_k which should be much easier to find.

Majority decision functions

By the use of the Bayes majority decision functions defined below, upper bounds for the error probability P_e can be derived in terms of the differences between the pdfs $F_{ik}(X_k)$. For simplicity, we shall assume that the X_k's are statistically independent. However, similar results can also be obtained for dependent random variables, provided that the Central Limit Theorem holds. One type of dependence encountered in practice is the so-called M-dependence, i.e., X_r and X_s are independent if $s - r > M$. For reference, we cite ([1], p. 14) and [5].

In general, a decision function for character recognition is a function $d(x)$ which maps every x into one of the Θ_i's. If $d(x) = \Theta_i$, it means that if x is the observed value of X, then the decision is that $\Theta = \Theta_i$, i.e., Θ is recognized as Θ_i. For the case where the alphabet consists of only two characters, say Θ_1 and Θ_2, we define a majority decision function as follows:

Definition 1. Let $d(x) = (d_1(x_1), \ldots, d_{2n+1}(x_{2n+1}))$, i.e., the kth component decision depends only on the observed value x_k of X_k; and $d_k(x_k) = \Theta_1$ or Θ_2, $k = 1, \ldots, 2n+1$. Then $d(x)$ is called a majority decision function if it follows the decision of the majority. (Note that we use 2n+1 to avoid the minor complication caused by 2n.)

Let $d(x)$ be a majority decision function defined above and

$$S_{ik} = \{x_k: d_k(x_k) = \Theta_i\}, \text{ and } \alpha_{ik} = \int_{S_{jk}} f_{ik}(x_k)\, dx_k, \qquad (3)$$

where $i \neq j$, $i, j = 1, 2$, and $k = 1, 2, \ldots, 2n+1$.

Theorem 3. Let $u_k = 0$ or 1, $k = 1, \ldots, 2n+1$, and Σ^* denote the summation (of a function of the u_i's) over all u_k such that $\sum_{k=1}^{2n+1} u_k \geqslant n+1$. Then, the error probability associated with the majority decision function $d(x)$ in (3) is given by

$$P_e(d) = \sum_{i=1}^{2} \left[p_i\, \Sigma^* \prod_{k=1}^{2n+1} \alpha_{ik}^{u_k} (1 - \alpha_{ik})^{1-u_k} \right] \qquad (4)$$

$$\sim \sum_{i=1}^{2} p_i\, \Phi\left(n+1, \sum_{k=1}^{2n+1} \alpha_{ik}, \sum_{k=1}^{2n+1} \alpha_{ik}(1 - \alpha_{ik})\right) \qquad (5)$$

where

$$\Phi(x, \xi, \sigma^2) = \int_x^{\infty} (2\pi\sigma^2)^{-1/2}\, e^{-(y-\xi)^2/2\sigma^2}\, dy, \qquad (6)$$

and $a \sim b$ means that a and b are approximately equal to each other if n is large.

Proof. Define random variables

$$U_k = 0,\ V_k = 1, \text{ if } d_k(x_k) = \Theta_1 \text{ and}$$
$$U_k = 1,\ V_k = 0, \text{ if } d_k(x_k) = \Theta_2.$$

Let u_k and v_k be the observed values of U_k and V_k
2n+1

respectively. Then by definition, $d(x) = \Theta_1$ if $\sum_{k=1} v_k \geqslant n+1$; and $d(x) = \Theta_2$, if $\sum_{k=1}^{2n+1} u_k \geqslant n+1$. Since $P(U_k = u_k | \Theta = \Theta_1) = \alpha_{1k}(1-\alpha_{1k})^{1-u_k}$, where $P(A|B)$ is the probability of A given B, we have $P_e(d) = P(d(x) = \Theta_2, \Theta = \Theta_1) + P(d(x) = \Theta_1, \Theta = \Theta_2) = p_1 P\left(\sum_{k=1}^{2n+1} U_k \geqslant n+1 \,|\, \Theta = \Theta_1\right) + p_2 P\left(\sum_{k=1}^{2n+1} V_k \geqslant n+1 \,|\, \Theta = \Theta_2\right) = \sum_{l=1}^{2}\left[p_i \Sigma^* \prod_{k=1}^{2n+1} \alpha_{ik}^{u_k} (1-\alpha_{ik})^{1-u_k}\right]$.

By the Central Limit Theorem for the sum of random variables that are independently but not necessarily identically distributed ([4], 215-218), we see that for large n, $P_e(d)$ may be approximated by (5).

Corollary if $\alpha_{ik} = \beta_i$, $i = 1, 2$, $k = 1, \ldots, 2n+1$, then

$$P_e(d) = \sum_{i=1}^{2}\left[p_i \sum_{K=n+1}^{2n+1} \binom{2n+1}{k} \beta_i^k \, (1-\beta_i)^{2n+1-k}\right]$$

$$\sim \sum_{i=1}^{2} p_i \, \Phi(n+1, (2n+1)\beta_i, (2n+1)\beta_i(1-\beta_i)) \quad (7)$$

From Theorem 3 and its Corollary, we see that if the α_{ik}'s are known, it will not be difficult to evaluate $P_e(d)$. The special case where $\alpha_{ik} = \beta_i$ is easy to handle, since each $\sum_{k=n+1}^{2n+1}$ in (7) is a cumulative binomial distribution and tables are available for computing its value.[7] In general, for small n, we use (4) to obtain $P_e(d)$, since the value of a Σ^* can be found by direct tabulation. For large n, we use the approximation in (5), where the value of a Φ also may be found from tables.[6]

The remaining problem is then how to find α_{ik}. From (3), it is obvious that α_{ik} depends on S_{ik} and S_{jk}. One type of S_{ik} is the following.

Definition 2. A Bayes majority decision function is a majority decision function such that for every $k = 1, 2, \ldots, 2n+1$, the sets S_{ik} in (3) are given by

$$S^1_k = \{x_k: q_{1k} f_{1k}(x_k) \geqslant q' \ f_{2k}(x_k)\}, \quad \text{and } S_{2k} = S', \quad (8)$$

where $q_{1k}, q_{2k} \geqslant 0$, and $q_{1k} + q_{2k} = 1$.

Note that the q_{ik}'s may be different from p_i, and are not necessarily the same for different k's. For given q_{ik}, the corresponding α_{ik} are not difficult to find in most applications. This is because f_{1k} and f_{2k} often have the same functional form; consequently, the sets S_{ik} are easy to handle. (See examples in Section 4.) A type of q_{ik}, known as the least favorable distribution, ([2] p. 154), is of specific interest to us. For each k, q_{1k} and q_{2k} are said to be the least favorable distribution of Θ with respect to $f_{1k}(x_k)$ and $f_{2k}(x_k)$ if $\alpha_{1k} = \alpha_{2k}$.

Theorem 4. Let $d(x)$ be the Bayes majority decision function such that for each $k = 1, \ldots, 2n+1$, q_{1k} and q_{2k} are the least favorable distribution of Θ. Suppose that for every k, $\int | f_{1k} - f_{2k} | \geqslant 2\,\delta > 0$. Then for large n,

$$P_e(d) \leqslant \Phi\,(n+1, (2n+1)\varepsilon, (2n+1)\,\varepsilon\,(1-\varepsilon)), \quad (9)$$

where $\varepsilon = 1/2 - \delta/4$; consequently, $P_e(d) \to 0$ as $n \to \infty$.

Proof. From Theorem 2, we see that $\alpha_{1k} = \alpha_{2k} \leqslant \varepsilon$. Now the function $y = x(1-x)$ increases as x increases from 0 to 1/2. Hence, an upper bound is obtained if the α_{ik}'s in (5) are replaced by ε. But $\Phi(n+1, (2n+1)\varepsilon, (2n+1)\varepsilon(1-\varepsilon)) \sim \Phi(a\, n^{1/2}, 0, 1)$ where $a > 0$. Since the latter tends to 0 as $n \to \infty$, we see that $P_e(d) \to 0$.

Applications and examples

For the general case where the alphabet consists of $\Theta_1, \ldots, \Theta_m$, it is well known[3] that if the Bales decision function defined in (1) is used, the corresponding error probability is

$$P_e = \sum_{i<j}\left[\int_{T_i} p_j f_j + \int_{T_j} p_i f_i\right] \leqslant \sum_{i<j} P_e(i,j), \quad (10)$$

where

$$P_e(i,j) = \int_{S_{ij}} p_j f_j + \int_{S'_{ij}} p_i f_i,$$

and $S_{ij} = \{x: p_i f_i(x) \geqslant p_j f_j(x)\}$. If for every pair i and j, $P_e(i,j) \to 0$ as $n \to \infty$, then $P_e \to 0$ as $n \to \infty$. From Theorem 4, we have

Theorem 5. If for all $i \neq j$; $i,j = 1, \ldots, m$; and $k = 1, 2, \ldots, 2n+1$, $\int |f_{ik} - f_{jk}| \geqslant 2\,\delta > 0$, then $P_e \leqslant \binom{m}{2} \Delta$, where Δ is the bound given in (9); consequently $P_e \to 0$ as $n \to \infty$.

The following are some applications of the results that we have so far obtained for the design of a character recognition system.

(a) Feature Selection. The difference between $f_{ik}(x_k)$ and $f_{jk}(x_k)$ depends, among other things, the type of X_k that is selected. In dealing with an alphabet consisting of Θ_1 and Θ_2 only, it is obvious that we should first rank the X_k's into a sequence $X_1, X_2, \ldots$, in descending order of $\int | f_{1k} - f_{2k} |$, and select the X_k's one by one from the beginning of the sequence. In general, we suggest that the X_k's be ranked according to $\min \int |f_{ik} - f_{jk}|$, for $i \neq j$; and $i,j = 1, \ldots, m$.

(b) Error Reduction. To reduce P_e below a required level α, one way is to select an $X = (X_1, \ldots, X_{2n+1})$ such that $\binom{m}{2}\Delta \leqslant \alpha$, where Δ is the bound given in (9). In case where $\Delta_1 = \min_{i \neq j} \int |p_i f_i - p_j f_j|$ of Theorem 1 or $\Delta_2 = \min_{i \neq j} \int |f_i - f_j|$ of Theorem 2 is not difficult to obtain, we may also select X such that $\binom{m}{2}\Delta_1$ or $\binom{m}{2}\Delta_2 \leqslant \alpha$. On the other hand, if cost is of primary importance, then the following method may be used. Suppose that a loss c is incurred whenever an error is made and that c_k is the unit cost associated with X_k. Then, the optimal n is the one which minimizes $\binom{m}{2}\Delta c + \sum_{k=1}^{2n+1} c_k$.

(c) Decision Functions. In should be emphasized that after X is chosen and the corresponding $f_i(x)$ are found, then the Bayes decision function as defined in (1) rather than the majority decision functions, should be used for actual recognition. This is because Bayes decision function minimizes the error probability. The actual value of the error probability is not known, but we know that it is below the required level α, and may be conservatively estimated by the various upper bounds.

The following are some illustrative examples:

(a) Binomial Distributions. The use of binary random variables in character recognition is quite common. For example, $X_k = 0$ and 1 may indicate that the k^{th} "region" of a character is black and white. respectively. The corresponding probability density function given $\Theta = \Theta_i$ is

$$f_{ik}(x_k) = \Theta_{ik}^{x_k}(1-\Theta_{ik})^{1-x_k},\ x_k = 0, 1, 0 < \Theta_{ik} < 1.$$

The difference between f_{ik} and f_{jk} is

$$\sum_{x_k=0}^{1} |\Theta_{ik}^{x_k}(1-\Theta_{ik})^{1-x_k} - \Theta_{jk}^{x_k}(1-\Theta_{jk})^{1-x_k}| = 2|\Theta_{ik} - \Theta_{jk}|.$$

Therefore, to select X_k, a simple criterion is min $\{|\Theta_{ik} - \Theta_{jk}|, i\neq j, i,j=1,\ldots,m\}$. The probabilities α_{ik} and α_{jk} can also be found, but they do not provide as clear a picture as the differences do; hence, will not be discussed.

(b) Normal Distributions. Consider the case where the X_k's have normal distributions, i.e.,

$$f_{ik}(x_k) = (2\pi\sigma_{ik}^2)^{-1/2}\, e^{-(x_k - \mu_{ik})^2/2\sigma_{ik}^2},\quad i=1,\ldots,m;\ \text{and}\ k = 1,\ldots,2n+1. \quad (11)$$

For simplicity of notation, we shall omit the subscript k unless there is confusion. It is easy to verify that for $i = 1, 2$ only, the set S_1 of (8) is that of all x for which

$$x^2\left(\frac{1}{\sigma^2_1}-\frac{1}{\sigma^2_2}\right) + 2x\left(\frac{\mu_2}{\sigma^2_2}-\frac{\mu_1}{\sigma^2_1}\right) + \frac{\mu^2_1}{\sigma^2_1}-\frac{\mu^2_2}{\sigma^2_2} - 2\log\frac{q_1\sigma_2}{q_2\sigma_1} \leqslant 0. \quad (12)$$

Let a and b be the solutions of the quadratic equation corresponding to (12). Then, S_1 is either the set $\{x: a < x < b\}$ or $\{x: x < a \text{ or } x > b\}$. Hence, α_1 and α_2 of (3) and the corresponding $P_e(d)$ may be found.

Now suppose that $\sigma_1 = \sigma_2 = \sigma$. Then (12) can be simplified and if $\alpha = \alpha_1 + \alpha_2$, then

$$\alpha=\Phi\ (w/2+q/w)+1-\Phi\ (-w/2+q/w), \quad (13)$$

where $w = |\mu_2 - \mu_1|/\sigma$, $q = \log q_1/q_2$, and $\Phi(x) = \Phi(x,0,1)$ of (6). Furthermore, $d\alpha/dw \leqslant 0$, if and only if $-(1 + e^q)/2 \leqslant q(e^q - 1)/w^2$. The latter inequality holds for all w and q, since the right and left hand sides are respectively non-negative and non-positive. Hence, $d\alpha/dw \leqslant 0$, and is a decreasing function of w. If q_1 and q_2 are the least favorable distribution, we know that $\alpha_1 = \alpha_2 = \alpha$. Therefore, the larger w is, the smaller α_1, α_2, and the corresponding $P_e(d)$ are. This suggests that in order to reduce P_e, one should choose those X_k for which $|\mu_{2k} - \mu_{1k}|/\sigma_k$ are large.

Finally, consider the special case of (11) where $\mu_{ik} = \mu_i$, $\sigma_{ik}^2 = \sigma^2$, and $p_i = ½$, for all $i = 1,2$ and $k = 1,\ldots, 2n+1$. From (13), $\alpha_{ik} = \Phi(w/2)$ for all i and k. Hence, (7) may be used to compute $P_e(d)$. Now it is easy to see that the exact value of the error probability P_e is $\Phi(w(2n+1)^{1/2}/2)$. For comparison, we give the following table where $w = 1$.

Table: Error probability and upper bound

n	P_e	Upper Bound
1	.3086	.3086
3	.1933	.2269
5	.1318	.1747
11	.0486	.0883
31	.0027	.0127
51	.0002	.0010
101	.0000	.0000

Note that proportionally the upper bounds are not close to the actual probabilities. But this is to be expected, since the bounds are valid for any kind of distribution.

REFERENCES

1 S BERNSTEIN
Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantites dépendantes
Mathematische Annalen 97 1 1927

2 D BLACKWELL M A GIRSHICK
Theory of games and statistical decisions
Wiley New York 1954

3 J T CHU
Optimal decision functions for computer character recognition
Jour Assn Comp Mach 12 213 1965

4 H CRAMÉR
Mathematical methods of statistics
Princeton University Press 1946

5 W HOEFFDING H ROBBINS
The central limit theorem for dependent random variables
Duke Math Jour 15 773 1948

6 *Tables of probability functions*
National Bureau of Standards Washington DC 1950

7 *Tables of the binomial probability distribution*
National Bureau of Standards 1950