ON THE DIVERGENCE BETWEEN TWO DISTRIBUTIONS AND THE PROBABILITY OF MISCLASSIFICATION OF SEVERAL DECISION RULES

Godfried T. Toussaint

School of Computer Science
McGill University
Montreal, Quebec
CANADA

Abstract

Sharp lower bounds are derived for the divergence between two distributions and the probabilities of misclassification of three decision rules. The three decision rules considered are the optimal Bayes rule, the nearest neighbour rule, and the "proportional prediction" randomized decision rule. It is shown that the randomized rule yields a probability of misclassification equal to the asymptotic nearest neighbour error rate. The bound between the Bayes error rate and the divergence is more general than the Kullback bound and, unlike the latter, is distribution free. The bounds are used to obtain sharp inequalities between measures of probabilistic dependence between features and classes in the multi-class pattern recognition problem. The bounds lead to sharp inequalities between the divergence and various information and distance measures found in the literature. Finally, the divergence is related to the least-mean-square-error design criterion in pattern recognition.

1. Introduction

Consider the two-category pattern classification problem. Let $P(X/C_i)$ denote the class-conditional probability density function of the feature vextor X conditioned on the class C_i, i=1,2. The Bhattacharyya coefficient and the divergence are defined, respectively, by

$$\rho = \int \sqrt{P(X/C_1)\, P(X/C_2)}\; dX \qquad (1)$$

and

$$J = \int [P(X/C_1) - P(X/C_2)] \log \left[\frac{P(X/C_1)}{P(X/C_2)}\right] dX. \qquad (2)$$

These measures are well known in the pattern recognition literature [1] and are useful for feature selection when the underlying distributions are Gaussian because they are much easier to evaluate than the error probability.

Let class C_i occur with a priori probability π_i, i=1,2, $\pi_1 + \pi_2 = 1$. It is useful to define more general measures than (1) and (2) above, as follows:

$$\rho(\pi_1, \pi_2) = \sqrt{\pi_1 \pi_2} \int \sqrt{P(X/C_1)\, P(X/C_2)}\; dX$$

$$= \int P(X) \sqrt{P(C_1/X)\, P(C_2/X)}\; dX$$

$$= \int P(X)\; \rho(X)\; dX, \qquad (3)$$

and

$$J(\pi_1,\pi_2) = \int [\pi_1 P(X/C_1) - \pi_2 P(X/C_2)] \log \frac{\pi_1 P(X/C_1)}{\pi_2 P(X/C_2)}\, dX$$

$$= \int P(X)\, [P(C_1/X) - P(C_2/X)] \log \frac{P(C_1/X)}{P(C_2/X)}\, dX$$

$$= \int P(X)\, J(X)\, dX. \qquad (4)$$

It follows that $\rho(1/2,1/2) = \rho/2$ and $J(1/2,1/2) = J/2$. In (3) and (4) P(X) is the mixture distribution and is given by

$$P(X) = P(X/C_1)\,\pi_1 + P(X/C_2)\,\pi_2 .$$

In this paper, the divergence is related to the probabilities of misclassification of three well known decision rules. These relationships are important when one would like to know what performance can be expected from a decision rule when features have been selected using the divergence. The first decision rule is the optimal Bayes rule. Given a feature vector X from some unknown pattern P, P is classified as belonging to class C_i if $P(C_i/X) > P(C_j/X)$, i=1,2, i≠j. This rule gives the minimum possible probability of misclassification [2] which is given by

$$P_e = \int \min_i [\, P(X/C_i)\, \pi_i \,]\; dX, \quad i=1,2$$

$$= \int P(X) \min [\, P(C_1/X), P(C_2/X) \,]\; dX$$

$$= \int P(X)\, P_e(X)\; dX . \qquad (5)$$

The second decision rule considered here is the nearest neighbour rule (NN-rule). Let $\{X,\Theta\} = \{X_1,\Theta_1; X_2,\Theta_2; \ldots ; X_N,\Theta_N\}$ be the set of N pattern samples available, where X_i and Θ_i denote, respectively, the feature vector or measurement information and the label or classification information of the ith pattern sample. It is assumed that each Θ_i associated with X_i is the correct label, i.e., the pattern samples have been correctly pre-classified. Let $(X_n,\Theta_n) \in \{X,\Theta\}$ to be the sample nearest to the unknown X. P is then classified as belonging to the class associated with the label Θ_n. Cover and Hart [3] have shown that as $N \to \infty$ the asymptotic nearest neighbour error rate, denoted by R, is given by

$$R = \int [2P(X/C_1)\ \pi_1\ P(X/C_2)\ \pi_2\ /\ P(X)]\ dX$$
$$= \int P(X)\ [2P(C_1/X)\ P(C_2/X]\ dX$$
$$= \int P(X)\ R(X)\ dX\ . \qquad (6)$$

The third decision rule under investigation is the randomized decision rule. Let the class-conditional distributions be known as in the deterministic Bayes rule. Given a feature vector X from some unknown pattern P, P is classified as belonging to Class C_i, i=1,2, by a flip of a biased coin which indicates C_i with probability $P(C_i/X)$. This type of decision rule tends to produce a distribution of classifications more similar to the original distribution than does the deterministic Bayes rule and is also known as proportional prediction [4]. The probability of misclassification using this rule, denoted by R for reasons that will become apparent, can be derived as follows. For any given value of X, C_1 occurs with probability $P(C_1/X)$ and it is decided to belong to class C_2 with probability $1 - P(C_1/X)$. Similarly, C_2 occurs with probability $P(C_2/X)$ and it is decided to belong to class C_1 with probability $1 - P(C_2/X)$. Hence, for a given value of X the resulting probability of misclassification is given by

$$R(X) = P(C_2/X)\ [\ 1 - P(C_2/X)\ ] + P(C_1/X)\ [1-P(C_1/X)]$$
$$= 2P(C_1/X)\ P(C_2/X) \qquad (7)$$

Taking the expected value of (7) with respect to P(X) yields

$$R = \int P(X)\ R(X)\ dX\ , \qquad (8)$$

which is the same as the asymptotic nearest neighbour error rate of (6). This equivalence between the NN-rule and the proportional prediction randomized rule (PPR-rule) has not made its appearance in the literature and provides added insight into the deterministic NN-rule. When an unknown X is far from the decision boundary into the region for Class C_i the NN-rule will almost always choose class C_i unless a maverick is near X. In the PPR-rule mavericks are explained by the fact that $P(C_i/X)$ is hardly ever equal to one. On the other hand, when X lies around the decision boundary it is likely to have nearest neighbours of either class. In terms of the PPR-rule one chooses C_i with probability $P(C_i/X)$ which is close to 0.5 when X is close to the decision boundary.

In this paper a generalized version of the inequality of Hoeffding and Wolfowitz [5] is derived. Using this inequality lower bounds are derived for P_e and R in terms of J. It is shown that the bounds for P_e are tighter than existing bounds. In addition, sharper lower bounds are derived for the divergence J in terms of P_e and R. Their relation to the Kullback bound is discussed. The bounds are applied to dependence measures between features and classes, equivocation measures, and distance measures found in the literature. The divergence is finally also related to the least-mean-square-error design criterion.

2. An Inequality Between $J(\pi_1,\pi_2)$ and $\rho(\pi_1,\pi_2)$

The divergence between two distributions occurring with a priori probabilities π_1 and π_2 can be written as

$$J(\pi_1,\pi_2) = -\pi_1 E_1 \left\{ \log \frac{P(X/C_2)\ \pi_2}{P(X/C_1)\ \pi_1} \right\} - \pi_2 E_2 \left\{ \log \frac{P(X/C_1)\ \pi_1}{P(X/C_2)\ \pi_2} \right\} \qquad (9)$$

where E_i denotes expected value with respect to $P(X/C_i)$. Since log x is a convex upward function ($\cap$), Jensen's inequality applied to (9) gives

$$J(\pi_1,\pi_2) \geq -2\pi_1 \log E_1\left\{ \frac{P(X/C_2)\ \pi_2}{P(X/C_1)\ \pi_1} \right\} - 2\pi_2 \log E_2 \left\{ \frac{P(X/C_1)\ \pi_1}{P(X/C_2)\ \pi_2} \right\},$$

which, in turn, yields

$$J(\pi_1,\pi_2) \geq -2\pi_1 \log [\ \rho(\pi_1,\pi_2)/\ \pi_1\ ] - 2\pi_2 \log [\ \rho(\pi_1,\pi_2)/\ \pi_2\ ]\ . \qquad (10)$$

Expanding (10) and recombining terms yields the desired result given by

$$J(\pi_1,\pi_2) \geq -2\ [\ H(\pi) + \log \rho(\pi_1,\pi_2)\ ]\ , \qquad (11)$$

where $H(\pi)$ is the entropy function given by

$$H(\pi) = -\pi_1 \log \pi_1 - \pi_2 \log \pi_2\ . \qquad (12)$$

When $\pi_1 = \pi_2 = 1/2$, $H(\pi) = \log 2$ and (11) reduces to

$$J \geq -4 \log \rho \qquad (13)$$

which is a well known inequality due to Hoeffding and Wolfowitz [5]. Hence (11) is a generalization of (13) to take into account the a priori probabilities.

3. Lower Bounds for R and J

Although there exists a lower bound on R(X) in terms of J(X) no bounds are available, in the literature, between R and J. Horibe [6] showed that

$$R \geq 2\ [\ \rho(\pi_1,\pi_2)\ ]^2\ ,$$

from which it follows that

$$\log \rho(\pi_1,\pi_2) \leq \log \sqrt{R/2}\ . \qquad (14)$$

Substituting (14) into (11) yields

$$R \geq \exp [-2 H(\pi) - J(\pi_1, \pi_2)] \tag{15}$$

For $\pi_1 = \pi_2 = 1/2$, (15) reduces to

$$R \geq (1/2) \exp [-J/2] , \tag{16}$$

where the equality holds for both $J=o$ and $J=\infty$.

Chitti Babu [7] showed that

$$R(X) \geq (1/2) [1 - J (X)/2] . \tag{17}$$

Taking expected values of both sides of (17) yields a second lower bound on R in terms of J as shown below.

$$R \geq (1/2) [1 - J(\pi_1, \pi_2)/2] \tag{18}$$

For $\pi_1 = \pi_2 = 1/2$, (18) reduces to

$$R \geq 1/2 - J/8 , \tag{19}$$

where the equality holds when $J=o$. Both (16) and (19) are illustrated in Fig. 1. It is observed that, for $J \leq 3.2$, (19) is the sharper bound. On the other hand, for $J \geq 3.2$, (16) is sharper. A lower bound on J in terms of R that is sharper than both (16) and (19) is given in (20) and (21).

$$J(\pi_1, \pi_2) \geq \sqrt{1-2R} \log \left[\frac{1 + \sqrt{1 - 2R}}{1 - \sqrt{1 - 2R}} \right] \tag{20}$$

For $\pi_1 = \pi_2 = 1/2$, (20) reduces to

$$J \geq 2 \sqrt{1 - 2R} \log \left[\frac{1 + \sqrt{1 - 2R}}{1 - \sqrt{1 - 2R}} \right] \tag{21}$$

These bounds follow from a result in section 4 and, hence, their derivation is deferred to that section. Inequality (21) is also illustrated in Fig. 1. Although (21) gives the sharpest inequality it has the disadvantage that it cannot be solved for R as a function of J, which is a more useful form since J is what is to be computed explicitly rather than R. For a proof that (21) is sharper than both (16) and (19) see Appendix A.

It should be pointed out that the bounds given by (15), (16), (18), (19), (20) and (21) are not important from the practical point of view when R is interpreted as the asymptotic nearest-neighbour error rate because we want to compute J only for the case of Gaussian distributions - a situation in which we would not use the nearest-neighbour rule. However, the bounds are useful when R is interpreted as the error rate of the proportional-prediction randomized decision rule - a parametric rule which has knowledge of the underlying distributions. From the theoretical point of view these bounds are very important and lead to some of the results in sections 7 - 9.

4. Lower Bounds for P_e and J

There exists in the literature a lower bound on P_e in terms of J. It appears to have been derived first by Kailath [8] and has since appeared in a number of papers on feature selection and texts on pattern recognition [9]. It is given by

$$P_e \geq (1/4) \exp [-J/2] \tag{22}$$

where the equality holds of $J = \infty$. This bound is illustrated in Fig. 2 where it is seen that it is a loose bound.

Cover and Hart [3] have shown that

$$R \leq 2P_e(1-P_e) \tag{23}$$

and, hence, that

$$R \leq 2P_e . \tag{24}$$

Substituting (24) into (15) yields

$$P_e \geq \exp [- 2 H(\pi) - J(\pi_1, \pi_2)] . \tag{25}$$

For $\pi_1 = \pi_2 = 1/2$, (25) reduces to (22) and, hence, (25) can be considered as a generalization of Kailath's bound. Similarly, substituting (24) into (18) yields

$$P_e \geq (1/4) [1 - J(\pi_1, \pi_2)/2] . \tag{26}$$

For $\pi_1 = \pi_2 = 1/2$, (26) reduces to

$$P_e \geq 1/4 - J/16 , \tag{27}$$

which is illustrated in Fig. 2. For $J \geq 4$ it is useless, but for $J \leq 3.2$ it is sharper than Kailath's bound and, hance, improves the latter when used in conjunction with it.

Tighter bounds than (25) and (26) can be obtained by using (23) rather than (24). Substituting (23) into (15) yields

$$P_e (1 - P_e) \geq \exp [- 2 H(\pi) - J(\pi_1, \pi_2)] . \tag{28}$$

Solving (28) for P_e yields

$$P_e \geq (1/2) - (1/2) \sqrt{1-4 \exp[-2 H(\pi) - J(\pi_1,\pi_2)]} . \tag{29}$$

For $\pi_1, = \pi_2 = 1/2$, (29) reduces to

$$P_e \geq (1/2) - (1/2) \sqrt{1 - \exp (-J/2)} , \tag{30}$$

where the equality holds for both $J=0$ and $J=\infty$. Similarly, substituting (23) into (18) and solving for P_e yields

$$P_e \geq (1/2) - \sqrt{J(\pi_1, \pi_2)/8} , \tag{31}$$

which for equal a priori probabilities reduces to

$$P_e \geq (1/2) - (1/4) \sqrt{J} , \tag{32}$$

where the equality holds when $J=0$. Bounds (30) and (32) are also illustrated in Fig. 2 which shows that (32) is sharper than (30) for $J \leq 3.2$.

A lower bound on J in terms of P_e which is sharper than all the above bounds can be derived as follows. From (4) it follows that

$$J(X) = [P(C_1/X) - P(C_2/X] \log \left[\frac{P(C_1/X)}{P(C_2/X)} \right] . \tag{33}$$

Also, from (5) it follows that

$$P_e(X) = \min [P(C_1/X), P(C_2/X)] . \tag{34}$$

The crucial step in the derivation of the bound is the realization that, since J(X) is symmetrical with respect to $P(C_1/X)$ and $P(C_2/X)$, it can be expressed in terms of $P_e(X)$ as

$$J(X) = [2 P_e(X) - 1] \log \left[\frac{P_e(X)}{1 - P_e(X)} \right] . \tag{35}$$

Now consider the function

$$f(x) = (2x-1) \log (x/1-x))$$

in the interval $0 \leq x \leq 1/2$. The first derivative of f(x) with respect to x is given by

$$\frac{df(x)}{dx} = \frac{2x-1}{1-x} + \frac{2x-1}{x} + 2 \log (x/(1-x)) . \tag{36}$$

The second derivative is given by

$$\frac{d^2f(x)}{dx^2} = \frac{1-2x}{x^2} + \frac{4}{x} + \frac{4}{1-x} - \frac{1-2x}{(1-x)^2} . \tag{37}$$

It can easily be shown that (37) is non negative. Hence f(x) and (35) are convex downward $\bigcup$ functions. For a convex downward function Jensen's inequality is given by

$$E \{ f(x) \} \geq f(E \{ x \}) , \tag{38}$$

where E denotes expected value. Taking the expected value of both sides of (35) with respect to P(X) and using (38) yields the desired bound given by

$$J(\pi_1, \pi_2) \geq (2 P_e - 1) \log [P_e/(1-P_e)] . \tag{39}$$

For equal a priori probabilities (39) reduces to

$$J \geq 2(2 P_e - 1) \log [P_e/(1-P_e)] , \tag{40}$$

where the equality holds for both $P_e=0$ and $P_e=1/2$. As illustrated in Fig. 2, (40) is the sharpest inequality between J and P_e but has the disadvantage that it cannot be solved for P_e as a function of J. For a proof that (40) is sharper than (30) and (32) see Appendix B.

From (6) it follows that

$$R(X) = 2 P(C_1/X) P(C_2/X) , \tag{41}$$

which can be written as

$$R(X) = 2 P_e(X) [1 - P_e(X)] , \tag{42}$$

where $P_e(X)$ is given by (34). Solving the $P_e(X)$ yields

$$P_e(X) = (1/2) - \sqrt{(1/4) - R(X)/2} . \tag{43}$$

It can easily be shown that (43) is a convex downward function of R(X). Taking expected values, with respect to P(X), of both sides of (43) yields, using Jensen's inequality,

$$P_e \geq (1/2) - \sqrt{(1/4) - R/2} . \tag{44}$$

Substituting (44) into (39) yields (20).

5. Relation to Kullback Bounds

The Kullback-Liebler numbers [10] are given by

$$I(1,2) = \int P(X/C_1) \log \left[\frac{P(X/C_1)}{P(X/C_2)} \right] dX \tag{45}$$

and

$$I(2,1) = \int P(X/C_2) \log \left[\frac{P(X/C_2)}{P(X/C_1)} \right] dX . \tag{46}$$

Let P_{ei} denote the probability of misclassification given class C_i, i=1,2, where $P_e = \pi_1 P_{e1} + \pi_2 P_{e2}$. The Kullback bounds are given by [8], [10], [11],

$$I(1,2) \geq P_{e1} \log [P_{e1}/(1-P_{e2})] + (1-P_{e1}) \log [(1-P_{e1})/P_{e2}] , \tag{47}$$

and

$$I(2,1) \geq P_{e2} \log [P_{e2}/(1-P_{e1})] + (1-P_{e2}) \log [(1-P_{e2})/P_{e1}] . \tag{48}$$

When the distributions are such that

$$\int_{\Omega_{X/C_1}} P(X/C_2)\, dX = \int_{\Omega_{X/C_2}} P(X/C_1)\, dX \tag{49}$$

where $\Omega_{X/C_i} \in \{ \Omega_X : P(X/C_i) > P(X/C_j) \}$,

$i,j=1,2,\ i \neq j$, and Ω_X

is the entire feature space, then $P_{e1}=P_{e2}=P_e$. For example, (49) is true for Gaussian distributions with equal covariance matrices. Adding (47) and (48), substituting $P_{e1}=P_{e2}=P_e$, and using the fact that $I(1,2) + I(2,1) = J$, yields

$$J \geq 2(2P_e -1) \log [P_e/(1-P_e)] , \tag{50}$$

which is a special case of (39). It is nice to know that assumption (49) is not needed and that (50) actually holds in general.

6. Application to Dependence Measures

Consider the M-class problem. A measure of the dependence between features and classes can be obtained by measuring the distance, in some sense, between the joint probability distribution P(X,C) and the product of the marginals P(X)P(C). Vilmansen [12] considers various measures of probabilistic dependence in this way and relates them to the probability of misclassification P_e. Two measures considered in [12] are the Kolmogorov dependence, first proposed by Hoeffding [13] and, given by

$$D_K(X,C) = \sum_{i=1}^{M} \int |P(X,C_i) - P(X) \pi_i| \, dX , \tag{51}$$

and the Joshi dependence, first proposed as a measure of channel capacity, Joshi [14], and, given by

$$D_J(X,C) = \sum_{i=1}^{M} \int [P(X,C_i) - P(X) \pi_i] \log \left[\frac{P(X,C_i)}{P(X) \pi_i} \right] dX . \tag{52}$$

The bounds between J and P_e, for the 2-class problem derived in section 4, can be used to form sharp inequalities between the above dependence measures for the M-class problem.

It can easily be shown [15] that, for equal a priori probabilities,

$$P_e = (1/2) - V/4 , \tag{53}$$

where V is the Kolmogorov variational distance given by

$$V = \int | P(X/C_1) - P(X/C_2) | \; dX . \tag{54}$$

Substituting (53) into (31), (32) and (40) yields, respectively,

$$V \leq 2 \; [1 - \exp(-J/2)]^{1/2}, \tag{55}$$

$$V \leq (J)^{1/2}, \tag{56}$$

and

$$J \geq V \log \left(\frac{2+V}{2-V} \right) . \tag{57}$$

Realizing that $D_K(X,C)$ are distance measures between two distributions in a continuous-discrete space of dimensionality one greater than the dimensionality of X, allows one to write (55)-(57), respectively, in the following way.

$$D_J(X,C) \geq D_K(X,C) \log \left[\frac{2 + D_K(X,C)}{2 - D_K(X,C)} \right] \tag{58}$$

$$D_K(X,C) \leq 2 \{1 - \exp [- D_J(X,C)/2] \}^{1/2} \tag{59}$$

$$D_K(X,C) \leq [D_J(X,C)]^{1/2} \tag{60}$$

The Kolmogorov dependence $D_K(X,C)$ can also be related to the expected divergence $\overline{J}$ which is given by

$$\overline{J} = \sum_{i=1}^{M} \sum_{j=1}^{M} \pi_i \pi_j J_{ij} , \tag{61}$$

where J_{ij} is the divergence between $P(X/C_i)$ and $P(X/C_j)$. It was shown in [16] that

$$\overline{J} = 2 D_J(X,C) . \tag{62}$$

This relation supports Vilmansen's conjecture [17] that there is a close relationship between the dependence of features and classes and the distance between class-conditional distributions. Substituting (62) into (58), (59) and (60) yields, respectively,

$$\overline{J} \geq 2 D_K(X,C) \log \left[\frac{2 + D_K(X,C)}{2 - D_K(X,C)} \right], \tag{63}$$

$$D_K(X,C) \leq 2 [1 - \exp(-\overline{J}/4)]^{1/2}, \tag{64}$$

and

$$D_K(X,C) \leq (\overline{J}/2)^{1/2}, \tag{65}$$

where the equality holds when classes and features are independent.

One measure of dependence not considered in [12] can be developed from the asymptotic nearest neighbour error rate R. For equal a priori probabilities R is given by

$$R = \int \frac{P(X/C_1) P(X/C_2)}{P(X/C_1) + P(X/C_2)} \; dX , \tag{66}$$

which, in a sense, measures the distance between $P(X/C_1)$ and $P(X/C_2)$. Hence, a new measure of dependence can be defined as

$$D_R(X,C) = \sum_{i=1}^{M} \int \frac{P(X,C_i) \, P(X) \, \pi_i}{P(X,C_i) + P(X) \, \pi_i} \; dX . \tag{67}$$

Furthermore, from the fact that [3]

$$P_e \leq R \leq 2 P_e (1 - P_e) ,$$

using (53) and similar arguments as above, it follows that

$$(1/2) - D_K(X,C)/4 \leq D_R(X,C) \leq (1/2) - (1/8) [D_K(X,C)]^2 , \tag{68}$$

where the equalities hold for $D_K(X,C) = 0$, i.e., when the features and classes are independent.

7. Application to Equivocation Measures

Shannon's measure of equivocation is the most well known and, for the 2-class problem, is given by

$$H(C/X) = - \int P(X) \sum_{i=1}^{2} P(C_i/X) \log P(C_i/X) \; dX . \tag{69}$$

Not as well known is Vajda's quadratic equivocation [18] given by

$$Q(C/X) = - \int P(X) \sum_{i=1}^{2} P(C_i/X) [P(C_i/X) - 1] \; dX$$

$$= 1 - \int P(X) \sum_{i=1}^{2} [P(C_i/X)]^2 \; dX . \tag{70}$$

Recently, Toussaint [16], [19], [20] proposed a family of equivocation measures given by

$$M_k(C/X) = \int P(X) \sum_{i=1}^{2} | P(C_i/X) - 1/2 |^{k^*} \; dX , \tag{71}$$

where $k^* = 2(k+1)/(2k+1)$ and $k=0,1,2,\ldots$. Of particular interest here is $M_0(C/X)$ given by

$$M_0(C/X) = \int P(X) \sum_{i=1}^{2} [P(C_i/X) - 1/2]^2 \; dX . \tag{72}$$

It was shown in [16] that $M_0(C/X)$ is related to the asymptotic nearest neighbour error rate by the relation

$$R = (1/2) - M_0(C/X) . \tag{73}$$

It also follows that

$$M_0(C/X) = 1 - Q(C/X) \tag{74}$$

and

$$R = Q(C/X) . \tag{75}$$

Hence, the information measure Q(C/X), which is obtained by approximating log x by (x-1) in Shannon's logarithmic equivocation, is also a distance measure (the harmonic mean between $P(X,C_1)$ and $P(X,C_2)$) as well as the asymptotic nearest neighbour error rate, and the probability of error of the proportional prediction randomized decision rule. Since $\log x \leq x-1$, it follows that R is bounded above by Shannon's equivocation, i.e.,

$$R \leq H(C/X) . \tag{76}$$

Substituting (74) and (75) into the bounds on R in section 3 gives sharp inequalities between the divergence J and the various equivocation measures. For example, substituting (75) into (15), (18), and (20) yields, respectively,

$$Q(C/X) \geq 2 \exp[- 2 H(\pi) - J(\pi_1, \pi_2)] , \quad (77)$$

$$Q(C/X) \geq (1/2) [1 - J(\pi_1, \pi_2)/2] , \quad (78)$$

and

$$J(\pi_1,\pi_2) \geq \sqrt{1-2 Q(C/X)} \log \left[\frac{1+\sqrt{1 - 2 Q(C/X)}}{1-\sqrt{1 - 2 Q(C/X)}} \right] . \quad (79)$$

8. Application to Distance Measures

Ito [21] proposed a family of distance measures, called the Q-function, given by

$$Q_n = (1/2) - (1/2) \int P(X) [P(C_1/X) - P(C_2/X)] \cdot [P(C_1/X) - P(C_2/X)]^{n^*} dX , \quad (80)$$

where $n^* = 1/(2n+1)$ and n is a natural number. Of particular interest is Q_0 given by

$$Q_0 = (1/2) - d/2$$

where

$$d = \int P(X) [P(C_1/X) - P(C_2/X)]^2 dX . \quad (81)$$

Ito [21] showed that

$$Q_0 = R , \quad (82)$$

$$Q_\infty = P_e , \quad (83)$$

and

$$Q_{n+1} \leq Q_n . \quad (84)$$

Substituting these results into the lower bounds for R relates the Q-function to the divergence.

Lissack and Fu [22] have investigated feature selection and estimation of misclassification using the separability measure

$$J_\alpha = \int P(X) \mid P(C_1/X) - P(C_2/X) \mid^\alpha dX \quad (85)$$

for $\alpha > 0$. It can easily be shown that

$$P_e = (1/2) - J_1/2 , \quad (86)$$

and

$$R = (1/2) - J_2/2 . \quad (87)$$

Hence, substituting (86) and (87) into the results of sections 3 and 4 relates J_α to the divergence J.

Devijver [23], [24] has recently done a lot of work on the so-called Bayesian distance given by

$$B(C/X) = \int P(X) \sum_{i=1}^{2} [P(C_i/X)]^2 dX . \quad (88)$$

It is obvious that

$$B(C/X) = 1 - R . \quad (89)$$

Hence, using the results of section 3 yields sharp inequalities between the Bayesian distance and the divergence. For example, letting B denote B(C/X) , to simplify notation, and substituting (89) into (21) yields

$$J \geq 2 \sqrt{2 B - 1} \log \left[\frac{1 +\sqrt{2 B - 1}}{1 -\sqrt{2 B - 1}} \right] . \quad (90)$$

9. Concluding Remarks

It has been shown that the probability of misclassification of the proportional-prediction randomized decision rule is equivalent to the error rate of the deterministic nearest neighbour rule, asymptotically. Previously, no bounds were available for R and the divergence J. In this paper better lower bounds are given for R and J. The tightest bound is given by (21). However, for feature evaluation using J, (16) and (19) are more useful. Letting

$$\underline{R}_1 = (1/2) \exp[-J/2] ,$$

and

$$\underline{R}_2 = (1/2) - J/8 ,$$

the best lower bound recommended for future use is

$$R \geq \max [\underline{R}_1 , \underline{R}_2] .$$

Similar comments hold true for P_e. For Gaussian distributions, an upper bound on P_e in terms of J is available and is given by [25]

$$P_e \leq (1/2) (J/4)^{-1/4} . \quad (91)$$

Letting

$$L_1 = (1/2) - (1/2) \sqrt{1 - \exp(-J/2)}$$

and

$$L_2 = (1/2) -(1/4) \sqrt{J} ,$$

from (30) and (32) , the best available lower bound to complement (91) above is given by

$$P_e \geq \max [L_1 , L_2] ,$$

which is greatly superior to the previous available bound, (22).

A final comment is in order as regards the well known least-mean-square-error (LMSE) design criterion [26] which has received a great deal of attention in the pattern recognition literature. Devijver [27] has shown that for a certain class of risk functions the LMSE criterion is equal to R. Under these conditions, (21) shows that minimizing the LMSE is equivalent to maximizing a lower bound on the divergence J.

Appendix A

Equation (19) can be written in the form

$$J \geq 4(1 - 2R) . \quad (A1)$$

To show that (21) is sharper than (19) it must be proved that

$$2 \sqrt{1 - 2 R} \log \left[\frac{1 + \sqrt{1 - 2 R}}{1 - \sqrt{1 - 2 R}} \right] \geq 4(1-2R). \quad (A2)$$

Making use of the transformation $[1-2R]^{1/2} = x$, $0 \leq x \leq 1$, it must follow that

$$2x \log [(1+x)/(1-x)] \geq 4x^2 ,$$

which in turn yields

$$\log [(1+x)/(1-x)] \geq 2x . \quad (A3)$$

It is known that

$$\log [(1+x)/(1-x)] = 2 \sum_{k=1}^{\infty} [1/(2k-1)] x^{2k-1} \quad (A4)$$

for $x^2 < 1$. Since $x \geq 0$, all terms in (A4) are non negative and it follows that for k=1 (A4) reduces to (A3), proving the result.

Equation (16) can be written in the form

$$J \geq -2 \log (2R). \quad \text{(A5)}$$

To show that (21) is sharper than (16) it must hold true that

$$2x \log [(1+x)/(1-x)] \geq -2 \log (1 - x^2), \quad \text{(A6)}$$

where x is as above. Making use of the transformation $x = 2y - 1$, $1/2 \leq y \leq 1$, it must hold true that

$$2(2y-1) \log [y/(1-y)] \geq -2 \log [4y(1-y)]. \quad \text{(A7)}$$

Expanding (A7) and recombining terms results in

$$H(y,1-y) \leq \log 2, \quad \text{(A8)}$$

where $H(y,1-y)$ is the entropy function. The maximum of $H(y,1-y)$ occurs for $y=1/2$ and is given by log 2, thus proving the desired result.

Appendix B

Equation (30) can be written in the form

$$J \geq -2 \log [4 P_e (1 - P_e)]. \quad \text{(B1)}$$

To prove that (40) is sharper than (30) it must be shown that

$$2(2P_e-1) \log [P_e/(1-P_e)] \geq -2 \log [4 P_e (1-P_e)],$$

which is of the same form as (A7), thus proving the result.

Equation (32) can be written in the form

$$J \geq 4(1 - 2P_e)^2. \quad \text{(B2)}$$

To prove that (40) is sharper than (32) it must be shown that

$$\log [(1-x)/x] \geq 2(1-2x), \quad \text{(B3)}$$

for $0 \leq x \leq 1/2$. Using the transformation $x = 1/(z+1)$, $1 \leq z \leq \infty$, it must be shown that

$$\log z \geq 2[(z-1)/(z+1)]. \quad \text{(B4)}$$

It is known that for $z > 0$

$$\log z = 2 \sum_{k=1}^{\infty} [1/(2k-1)][(z-1)/(z+1)]^{2k-1}. \quad \text{(B5)}$$

For $z \geq 1$, all terms in (B5) are non negative. Hence, for k=1 (B5) reduces to (B4), thus proving the result.

REFERENCES

[1] P.H.Swain and R.C.King, "Two effective feature selection criteria for multispectral remote sensing, "Proc. First International Joint Conf. on Pattern Recognition, Nov. 1973, pp.536-540.

[2] R.O.Duda and P.E.Hart, Pattern Classification and Scene Analysis, John Wiley, 1973, pp. 10-22.

[3] T.M.Cover and P.E.Hart, "Nearest neighbour pattern classification", IEEE Trans. Information Theory, Vol. IT-13, Jan.1967, pp. 21-27.

[4] L.A.Goodman and W.H.Kruskal, "Measures of association for cross classifications", J.A.S.A., Dec. 1954, pp. 732-763.

[5] W.Hoeffding and J.Wolfowitz, "Distinguishability of sets of distributions," Ann.Math.Stat., Vol. 29, 1958, pp. 700-718.

[6] Y.Horibe, "On zero error probability of binary decisions," IEEE Trans. Information Theory, Vol. IT-16, May 1970, pp. 347-348.

[7] C.Chitti Babu, "On divergence and probability of error in pattern recognition," Proc. IEEE, June 1973, pp.798-799.

[8] T.Kailath, "The divergence and Bhattacharyya distance measures in signal selection," IEEE Trans. Communication Technology, Vol. COM-15, Feb. 1967, pp. 52-60.

[9] C.H.Chen, Statistical Pattern Recognition, Hayden Book Co., Inc., 1973, Chapter 4.

[10] S.Kullback, Information Theory and Statistics, Dover Publications, Inc., New York, 1968.

[11] J.Ziv and M.Zakai, "On functionals satisfying a data-processing theorem," IEEE Trans. Information Theory, Vol. IT-19, May, 1973, pp. 275 - 283.

[12] T.R.Vilmansen, "Feature evaluation with measures of probabilistic dependence," IEEE Trans. Computers, Vol. C-22, April 1973, pp. 381-388.

[13] W.Hoeffding, "Stochastishe abhangigkeir und funktionaler zusammenhang," Skand. Aktuarietidskr Vol. 25, 1942, pp. 200-227.

[14] A.K.Joshi, "A note on a certain theorem stated by Kullback," IEEE Trans. Information Theory, Vol. IT-10, Jan. 1964, pp. 93-94.

[15] G.T.Toussaint, "Comments on the divergence and Bhattacharyya distance measures in signal selection," IEEE Trans. Communication Technology, Vol. COM-20, June 1972, p. 485.

[16] G.T.Toussaint, "Feature evaluation criteria and contextual decoding algorithms in statistical pattern recognition," Ph.D. thesis, University of British Columbia, 1972.

[17] T.R.Vilmansen, "On dependence and discrimination in pattern recognition," IEEE Trans. Computers, Vol. C-21, Sept. 1971, pp. 1029-1031.

[18] I.Vajda, "A contribution to the informational analysis of pattern," in Methodologies of Pattern Recognition, Ed., S.Watanabe, Academic Press, 1969, pp. 509-519.

[19] G.T.Toussaint, "A certainty measure for feature evaluation in pattern recognition," Proc. Fifth Hawaii International Conf. on Systems Sciences, Jan. 1972, pp. 37-39.

[20] G.T.Toussaint, "Distance measures as measures of certainty and their application to statistical pattern recognition," Conf. on Theoretical and Applied Statistics and Data Analysis, Queen's University, Kingston, Ontario, June 4-6, 1973. Also, Canadian Journal of Statistics, Vol. 1, No. 1, 1973, abstract, p. 134.

[21] T.Ito, "On approximate error bounds in pattern recognition," Systems, Computers, Controls, Vol. 4, Feb. 1973, pp. 85-92.

[22] T.Lissack and K.S.Fu, "A separability measure for feature selection and error estimation in pattern recognition," Technical Report No. TR-EE 72-15, May 1972, School of Electrical Engineering, Purdue University.

[23] P.A.Devijver, "The Bayesian distance. A new concept in statistical decision theory," Proc. 1972 IEEE Conf. on Decision and Control, pp. 543-544.

[24] P.A.Devijver, "On a new class of bounds on Bayes

risk in multi-hypothesis pattern recognition," IEEE Trans. Computers, Vol. C-23, January 1974 pp. 70-80.

[25] T.T.Kadota and L.A.Shep, "On the best finite set of linear observables for discriminating two Gaussian signals," IEEE Trans. Information Theory, Vol. IT-13, April 1967, pp. 278-284.

[26] S.S.Yau and J.M.Garnet, "Least-mean-square approach to pattern classification," in Frontiers of Pattern Recognition, M.S.Watanabe, Ed., Academic Press, 1972, pp. 575-587.

[27] P.A.Devijver, "Relationships between statistical risks and the least-mean-square-error design criterion in pattern recognition," Proc. First International Joint Conference on Pattern Recognition, Oct. 30-Nov 1, 1973, pp. 139-148.

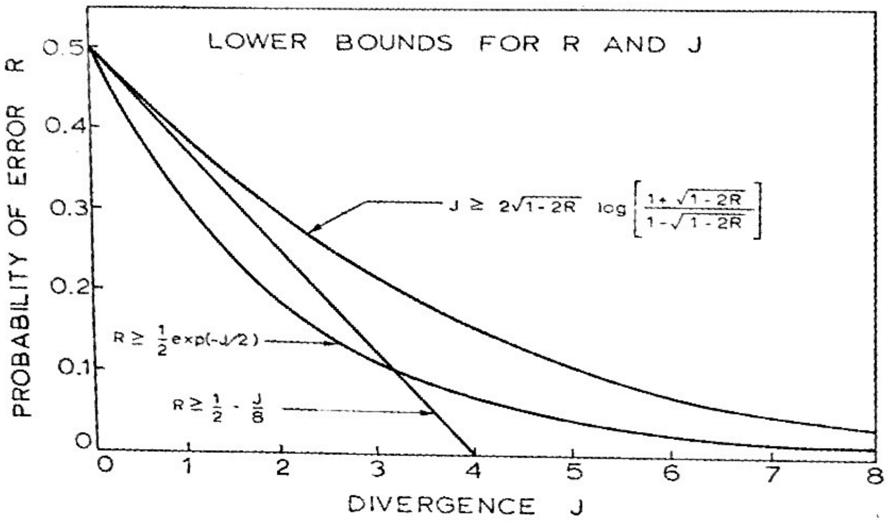

Figure 1

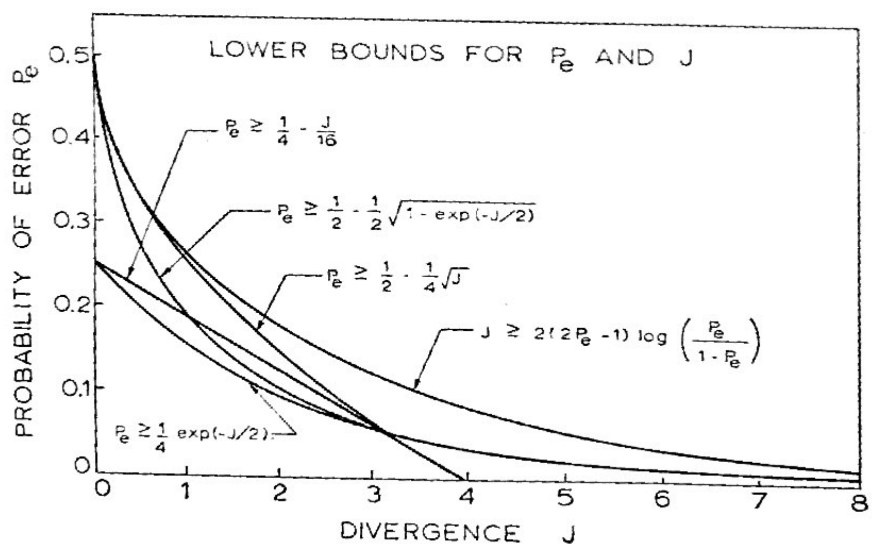

Figure 2