Pattern Recognition, Vol. 21, No. 2, pp. 183–188, 1988.

EFFICIENCY OF LEARNING WITH IMPERFECT SUPERVISION

T. KRISHNAN
Indian Statistical Institute, 203 Barrackpore Trunk Rd, Calcutta, 700 035, India

(*Received* 16 *September* 1986; *in revised form* 20 *March* 1987)

Abstract—We consider the problem of supervision errors in training samples in two-group discriminant analysis based on normal distributions. Using a model for training sample misclassification, we derive Efron's Asymptotic Relative Efficiency (ARE) of the discriminant function estimated under this model, relative to the case when classification is perfect. We tabulate this ARE for certain values of the Mahalanobis distance between the groups and for various levels of supervision errors. We show that training samples are useful even if prone to a certain amount of misclassification. Our formulae and tables give, for a training sample prone to a certain amount of error, sample size equivalent to that of one error-free training sample as well as that of an unsupervised sample, the equivalence being in terms of estimation efficiency.

Normal discrimination   Supervision error   Asymptotic relative efficiency

1. INTRODUCTION

In many applications of pattern recognition, classifying a training sample is expensive and difficult and is subject to error; some examples of this situation are remote sensing[1,2,3] and medical diagnosis.[7,8,12] In a problem of remote sensing of crop patterns, the training samples may be visually classified and hence may be prone to error; in a medical diagnosis problem, the training samples may be classified by experts on the basis of the same vector variable x as the one used for learning and hence may be prone to error. In the first case, the errors in supervision may be presumed to be independent of the observable x and in the second dependent on x. These two situations are called respectively random and nonrandom misclassification.[1,7,8]

Studies have been made on the effect of supervision errors on the estimates of the Bayes discriminant function for the special case of two multivariate normal populations with a common covariance matrix. Lachenbruch[7] started these studies using a random misclassification model studying the means and variances of conditional error rates. He used a combination of theoretical and Monte Carlo studies. McLachlan[9] studied conditional error rates using their asymptotic expansions for the case where one group does not get misclassified. Lachenbruch[8] studied by Monte Carlo methods the conditional error rates under two types of nonrandom misclassification models where the probability of misclassification depended on the vector variable x through the distance of x from its group mean. Chhikara and McKeon[1] used models more general than those of Lachenbruch[7,8] and derived the asymptotic distribution of the discriminant boundary and the asymptotic means and variances of conditional error rates and of the average error rate; they pointed out the adverse effects of misclassification on the boundary and on these error rates. Michalek and Tripathi[10] used Efron's Asymptotic Relative Efficiency (ARE) and studied the effect on the estimate of the discriminant function. All these authors are concerned with the effect of supervision errors when the discriminant function and other parameters are estimated using methods suitable for the perfectly supervised case. The general conclusion is that under nonrandom misclassification, the true error rates are only slightly affected and apparent error rates considerably affected and give an optimistic picture. Under random misclassification, the estimates of the discriminant function and the true error rates are biased and maximum likelihood estimates converge to false values and the efficiency of the discriminant function estimates decreases.

Chittineni[2,3] considers random misclassification and compares misclassified and correctly classified situations for the general case using Bayes and nearest neighbour classifiers. He also develops nonparametric learning techniques and methods for correction of supervision errors.

Thus, if training sample misclassification occurs in such a way that observations falling in the "doubtful" region are more prone to misclassification, then ignoring supervision errors does not lead to much harm. However, if training sample misclassification occurs at random, then ignoring supervision errors does distort the learning process. Thus it is clear that, especially in the case of random misclassification, it would be helpful to have methods of estimating the discriminant function taking such misclassification into account. This is quite a different objective from those of most of the papers discussed above.

In Katre and Krishnan,(6) we have discussed the problem of maximum likelihood estimation under a random misclassification model and have developed the EM algorithm,(4) for the case of two multinormal populations with a common dispersion matrix. In this paper, our object is to examine, for the same case, the extent to which a training sample subject to misclassification with constant probability α, is useful in learning the parameters. This extent of usefulness will also depend on other parameters, the Mahalanobis distance Δ between the two populations and the prior probabilities of the two groups. Values of α from 0 to $\frac{1}{2}$ will cover a range of situations; $\alpha = 0$ implies perfect supervision (the usual supervised learning) and $\alpha = \frac{1}{2}$ implies no supervision at all (the usual unsupervised case). In other words, we examine the relative sample sizes required under these types of supervision, to achieve the same error rate for the discriminant function when maximum likelihood estimates are used.

We consider the case of two p-variate normal populations with a common covariance matrix $N_p(\mu_0, \Sigma)$ and $N_p(\mu_1, \Sigma)$ occurring in proportions $\eta_0 (=\eta)$ and $\eta_1 (1-\eta)$ respectively. The Bayes rule here is to use

$$\beta_0 + \beta' x \tag{1.1}$$

where

$$\beta_0 = \log(\eta_1 | \eta_0) - \tfrac{1}{2}(\mu_1' \Sigma^{-1} \mu_1 - \mu_0' \Sigma^{-1} \mu_0)$$
$$\beta = \Sigma^{-1}(\mu_1 - \mu_0). \tag{1.2}$$

The Bayes rule is the one with the least error rate. The Asymptotic Error Rate (AER) of a procedure based on estimates $\begin{pmatrix} a_0 \\ a \end{pmatrix}_n$ of vector $\begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}$ from a sample of size n is defined to be the limiting value (as $n \to \infty$) of the additional error rate of $\begin{pmatrix} a_0 \\ a \end{pmatrix}$ over the Bayes error. This AER will naturally depend on the nature of the procedure and the parameters. For supervised and unsupervised procedures it will be different and for the same parameters, the unsupervised procedure will have a larger AER. The ratio of the AER of two procedures gives the Asymptotic Relative Efficiency of one procedure with respect to the other. When several procedures less efficient than the supervised one are considered, such as unsupervised, combined supervised and unsupervised or error-prone supervised, the supervised procedure may be used as the basis of comparison. This leads to Efron's(5) Asymptotic Relative Efficiency (ARE). These ideas are defined precisely in the next section.

O'Neill(11) adopts this approach in studying the efficiency of a procedure where the training sample has a proportion γ of unsupervised samples and $(1-\gamma)$ of (correctly) supervised samples. Thus our work can be regarded as an extension of O'Neill's work. We derive a formula for ARE of the error-prone supervision relative to a correctly supervised scheme similar to O'Neill's(11) formula (3.1). In fact, we use similar techniques too; these techniques require the computation of the information matrix of the logistic regression estimators of β_0, β while computing the ARE of maximum likelihood estimators.

2. ASYMPTOTIC RELATIVE EFFICIENCY

Since error rates of discriminant rules based on β_0, β or its estimates are invariant under linear transformation on x, we assume the canonical form for (μ_0, Σ) and (μ_1, Σ) to be $\left(\frac{\Delta}{2} e_1, I_p\right)$ and $\left(-\frac{\Delta}{2} e_1, I_p\right)$ where Δ is the Mahalanobis distance between the two groups, e_1 is the vector $(1, 0, \ldots, 0)$ and I_p is the $p \times p$ identity matrix; this canonical form can be obtained by a linear transformation on x. Let $(a_0, a)_n$ denote the estimate of (β_0, β) based on a sample of n by a certain procedure and let $\text{ER}(a_0, a)_n$ denote the error rate on using $(a_0, a)_n$ for (β_0, β) in (1.1). Let $\lambda = \log(\pi_1 | \pi_0)$.

Efron(5) shows that if

$$\sqrt{n}\left[\begin{pmatrix} a_0 \\ a \end{pmatrix}_n - \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}\right] \xrightarrow{L} N_{p+1}(0, M)$$

then

$$n\,[\text{ER}(a_0, a)_n - \text{ER}(\beta_0, \beta)] \xrightarrow{L}$$
$$\frac{\pi_1}{2\Delta}\phi\left(\frac{\Delta}{2} - \frac{\lambda}{\Delta}\right)\left[z_0^2 - \left(\frac{2\lambda}{\Delta}\right) z_0 z_1 + \left(\frac{\lambda}{\Delta}\right)^2 z_1^2 + z_2^2 + \ldots + z_p^2\right]$$

where $\xrightarrow{L}$ means convergence in law (distribution), $z = (z_0, z_1, \ldots, z_p) \sim N_{p+1}(0, M)$, ϕ is standard normal density function, and 0 the $(p+1)$-null vector. The Asymptotic Error Rate of a procedure with estimates $(a_0, a)_n$ is then defined to be the expectation of the limit above, which is equal to

$$\pi_1 \frac{\phi\left(\frac{\Delta}{2} - \frac{\lambda}{\Delta}\right)}{2\Delta}\left[m_{00} - \left(\frac{2\lambda}{\Delta}\right) m_{01} + \left(\frac{\lambda}{\Delta}\right)^2 m_{11} + m_{22} + \ldots + m_{pp}\right]$$

when $((m_{ij})) = M$. This is denoted for convenience by $\text{AER}(a_0, a)$. Then the Asymptotic Relative Efficiency (ARE) of a procedure with $(c_0, c)_n$ with respect to a procedure yielding estimate $(b_0, b)_n$ is

$$\text{Eff}_p = \text{AER}\,(b_0, b)/\text{AER}\,(c_0, c). \tag{2.1}$$

In order to compute this efficiency for error-prone initial samples relative to a supervised sample we need the matrices M for these cases for the maximum likelihood estimates. This is done by computing the information matrix of β_0, β and inverting it; Efron has already computed this for the supervised case as

$$I_c = [\eta_0 \, \eta_1] \begin{bmatrix} H & 0 \\ 0 & (1 + \Delta^2 \eta_0 \eta_1)^{-1} I_{p-1} \end{bmatrix}$$

where

$$H^{-1} = \begin{bmatrix} 1 + \Delta^2/4 & -(\eta_0 - \eta_1)\Delta/2 \\ -(\eta_0 - \eta_1)\Delta/2 & 1 + 2\eta_0\eta_1\Delta^2 \end{bmatrix}. \qquad (2.2)$$

3. EFFICIENCY OF ERROR-PRONE SCHEME

We denote by

x: p-dimensional observation vector:
z: the group (0 or 1) to which the observed unit is allotted.

The observation thus has the form (x, z). In our formulation z is a random variable. Let

y: actual class to which an observation belongs (unknown or unobserved).

Let w: $1 - 2z$ (w is 1 or -1 as z is 0 or 1).

Thus z may not be the same as y and

$$P(z \neq y \mid y) = \alpha$$

according to our model; further z and x are independent given y, by our model. In our model, the probability of misclassification of group 1 into group 2 is the same as that of misclassifying group 2 into group 1.

We have n observations (x_j, z_j), $j = 1, 2, \ldots, n$. Denoting by $f_i(x)$, the multivariate normal density N_p (μ_i, Σ), $i = 0, 1$, we write the density of (x_i, z_i) in the two groups 0 and 1 as

$$f_0(x, z) = f_0(x)\alpha^z(1-\alpha)^{1-z};$$
$$f_1(x, z) = f_1(x)(1-\alpha)^z\alpha^{1-z}.$$

We assume that random sampling was done from the mixture

$$\eta_0 f_0(x, z) + \eta_1 f_1(x, z).$$

To compute the information matrix, we need full, various marginal and conditional likelihoods. We use l to denote log-likelihood; the arguments of l and the conditioning symbol $|$ indicate which likelihood is being considered.

$$l(x, z, y) = \log\{[\eta_1 f_1(x, z)]^y [\eta_0 f_0(x, z)]^{1-y}\}$$
$$l(x, y) = \log\{[\eta_1 f_1(x)]^y [\eta_0 \, y_0(x)]^{1-y}\}$$
$$l(x, z) = \log[\eta_1 f_1(x, z) + \eta_0 f_0(x, z)]$$
$$l(x) = \log[\eta_1 f_1(x) + \eta_0 f_0(x)]$$
$$l(y \mid x, z) = \log\{[\eta_1(x, z)]^y [\eta_0(x, z)]^{1-y}\}$$

where $\eta_0(x, z) = 1 - \eta_1(x, z)$

$$= \frac{\eta_0 f_0(x, z)}{\eta_0 f_0(x, z) + \eta_1 f_1(x, z)} = \frac{1}{1 + \dfrac{\eta_1 f_1(x)}{\eta_0 f_0(x)} \dfrac{(1-\alpha)^z \alpha^{1-z}}{\alpha^z (1-\alpha)^{1-z}}}$$

$$= \frac{1}{1 + \dfrac{(1-\alpha)^z \alpha^{1-z}}{\alpha^z (1-\alpha)^{1-z}} e^{\beta_0 + \beta' x}}$$

$$= 1/[1 + \exp(\beta_0 + \beta' x + \delta w)]$$

where $\delta = \log[\alpha/(1-\alpha)]$.

Finally $l(y \mid x) = \log\{[\eta_1(x)]^y [\eta_0(x)]^{1-y}\}$,

where $\eta_0(x) = 1/[1 + \exp(\beta_0 + \beta' x)] = 1 - \eta_1(x)$.

Maximum likelihood estimation of the parameters under the given observational structure involves maximising

$$\sum_{i=1}^{n} l(x_i, z_i).$$

This situation can be regarded as similar to estimation under an unclassified observational structure, but with the addition of one more dimension z. Thus the likelihoods $l(x, z)$ and $l(x)$ are analogues both referring to unsupervised initial samples. Just like O'Neill's[11] study of efficiency of unsupervised initial samples compared to supervised initial samples, required the likelihoods $l(x, y)$ and $l(y \mid x)$, in our study also we need the analogous likelihoods $l(x, z, y)$ and $l(y \mid x, z)$. Note that when y is actually known z does not provide any information about the parameters η_0, μ_0, μ_1, Σ but only about α.

As in O'Neill[11], we denote the information matrices based on a single observation for the parameters β_0, β by I with suffixes C, UC and LR to denote Classified, Unclassified and Logistic Regression log-likelihoods. Further, we use I with superfix * if the observation used is (x, z) and without if x. From the log-likelihoods $l(x, y, z)$ and $l(x, z)$ it is immediately seen that $I_C = I_C^*$. Further, I_{UC}^* corresponds to the case of error-prone initial samples.

We reparametrise η_1, μ_1, μ_0, Σ, α as follows:

$$u = \eta_1 \mu_1 + \eta_0 \mu_0$$
$$R = \Sigma + \eta_0 \eta_1 (\mu_1 - \mu_0)(\mu_1 - \mu_0)'$$
$$\beta_0, \ \beta \text{ and } \delta.$$

Let A, B, C denote information matrices of $(u, R, \beta_0, \beta, \delta)$ based on $l(x, y, z)$, $l(x, z)$ and $l(y \mid x, z)$ respectively. Then since

$$l(x, y, z) = l(x, z) + l(y \mid x, z)$$

and A, B, C are expection matrices

$$A = B + C.$$

These are different from I_C^*, I_{UC}^* and I_{LR}^* which are information matrices for β_0, β only. In the Appendix, we derive a formula for I_{UC}^* as follows:

$$\text{Let } A_i(\eta_1, \Delta) = \int_{-\infty}^{\infty} \frac{e^{-\Delta 2/8} x^i \phi(x) dx}{\eta_1 e^{\Delta x/2} + \eta_0 e^{-\Delta x/2}}, \quad i = 0, 1, 2$$

$\phi(x)$: standard normal density.

$a_0 = \eta_0(1-\alpha) + \eta_1\alpha$; $a_1 = \eta_0\alpha + \eta_1(1-\alpha)$

$p_0 = \eta_1\alpha/a_0$; $p_1 = \eta_1(1-\alpha)/a_1$

$F_i = [A_i(p_0, \Delta)/a_0] + [A_i(p_1, \Delta)/a_1]$, $i = 0, 1, 2$

$B_i = [A_i(p_0, \Delta)/a_0] - [A_i(p_1, \Delta)/a_1]$, $i = 0, 1$.

$$D = \alpha(1-\alpha)\begin{bmatrix} F_0 + [\eta_1\eta_0 B_0^2/(1-\eta_1\eta_0F_0)]F_1 + [\eta_1\eta_0B_0B_1/(1-\eta_1\eta_0F_0)] \\ F_1 + [\eta_1\eta_0B_0B_1/(1-\eta_1\eta_0F_0)]F_2 + [\eta_1\eta_0B_1^2/(1-\eta_1\eta_0F_0)] \end{bmatrix}.$$

Thus $I_{UC}{}^* = I_C - \eta_1\eta_0 D$,

where $I_C{}^* = I_C$ is given in (2.2).

For the case $\alpha = \frac{1}{2}$, we have $a_0 = a_1 = \frac{1}{2}$; $p_0 = p_1 = \eta_1$ and the F_i's are exactly Efron's A_i; $B_i = 0$ for $i = 0, 1$. This makes the D matrix the same as the information matrix for β_0, β under logistic regression yielding the result that

$$I_{UC}{}^* = I_C{}^* - I_{LR} = I_C - I_{LR}$$

which when compared to O'Neill's[11] Lemma 3 that $I_{UC} = I_C - I_{LR}$ shows that this case is like a completely unsupervised scheme and the z-values are totally useless; this indeed, is evident, since in such a case, the initial samples are merely classified with equal probability in each group.

Substituting these in the formula (2.1), we can write the ARE as

$$\text{Eff}_p(\lambda, \Delta, \alpha) = \frac{q(\lambda, \Delta, \alpha)\,\text{Eff}_1(\lambda, \Delta, \alpha) + (p-1)\,\text{Eff}_\infty(\lambda, \Delta, \alpha)}{q(\lambda, \Delta, \alpha) + p - 1}$$

where

$$q(\lambda, \Delta, \alpha) = (1, -\lambda/\Delta)\,(H-D)^{-1}\,(1, -\lambda/\Delta)'$$
$$(1-\alpha(1-\alpha)F_0(1+\eta_1\eta_0\Delta^2))/[1+\eta_1\eta_0\Delta^2].$$

For $p = 1$ and $p \to \infty$, Eff_p becomes

$$\text{Eff}_1(\lambda, \Delta, \alpha) = (1, -\lambda/\Delta)$$
$$H^{-1}(1, -\lambda/\Delta)'/(1, -\lambda/\Delta)(H-D)^{-1}\left(1, \frac{-\lambda}{\Delta}\right)'$$

$$\text{Eff}_\infty(\lambda, \Delta, \alpha) = 1 - \alpha(1-\alpha)F_0(1+\eta_1\eta_0\Delta^2)$$

which are called Intercept (β_0) and Angle (β) Efficiencies respectively by Efron.

4. VALUES OF EFFICIENCY AND DISCUSSION

Table 1 gives Eff_1 and Eff_∞ for $\pi_1 = 0.5, 0.667, 0.9$, $\Delta = 2.0\,(0.5)\,4.0$ and $\alpha = 0.01, 0.05, 0.20, 0.35, 0.50$. We also give values of γ, the proportion of unclassified observations at which the same efficiency is obtained by O'Neill's formula. This gives an idea of the usefulness of error-prone supervision. Values of Δ from 2 to 4 were considered by Efron[5] and O'Neill[11] as statistically most interesting in the sense that enough separation exists between the populations to be able to discriminate effectively and not too much to make formal discriminant analysis unnecessary.

The interpretation of our table is as follows: for instance, if the misclassification rate is 5, $\pi_1 = 0.5$, $\Delta = 2.5$, the asymptotic relative efficiency of an initial sample is 0.756. From O'Neill's formula, for $\pi_1 = 0.5$, $\Delta = 2.5$, such an efficiency is attained when the initial sample of the same size contains $\gamma = 0.31$ proportion of unsupervised observations. Thus for these parameter values if we have an initial sample of 100 units, subject to a misclassification rate of 0.05, it is like having a correctly supervised sample of 76 or an initial sample of 100 consisting of a mixture of 69 correctly classified units and 31 unclassified units.

Thus our formula and table put into perspective the relative amount of information contained in supervised, unsupervised, mixed and error-prone supervision schemes. If in a situation, the costs of unclassified, error-prone, and correctly classified schemes are known, and are say 0.25, 0.5 and 1 respectively, then the unit cost efficiency are respectively 0.84, 1.5 and 1 for the three schemes for these parameter values, and the error-prone scheme is to be preferred.

As the groups are better separated, the efficiency of error-prone observations increase. As η_1 goes away from $\frac{1}{2}$, the efficiency decreases. As the distance between the groups increases, the equivalent γ generally increases but only slightly. Indeed as α increases, the efficiency decreases and the equivalent γ increases to 100 at $\alpha = 0.5$. Values of γ are affected only a little by values of Δ and η_1, more by Δ than by η_1.

Clearly, if $\alpha = 0.5$, then the supervision is useless and it is ARE-equivalent to $\gamma = 1$, which is precisely what happens with our formula.

For a misclassification rate of 10%, the ARE ranges from 0.93 to 0.96 and the equivalent γ between 7 and 12.

The corresponding figures for 5% misclassification rate are 74–90% and 30–45%; 20% misclassification rate are 35–80% and 74–79%; 35% misclassification rate are 18–70% and 94–95%; and 50% misclassification rate are 13–75% and 100%.

SUMMARY

Discriminant analysis is usually carried out assuming that the training samples are classified deterministically and correctly. Recently, there has been interest in discriminant analysis with incorrectly classified training samples, motivated by examples from remote sensing and medical diagnosis. In this paper, we consider the case of supervision errors occurring at random with a constant probability in a two-group discriminant problem with normal distributions. We compute Efron's Asymptotic

Table 1. Asymptotic relative efficiency of normal discrimination based on an initial sample with misclassification probability α and equivalent proportion γ of unclassified observations

η_1	Δ	α	EFF_1	EQUI γ	EFF_∞	EQUI γ	η_1	Δ	α	EFF_1	EQUI γ	EFF_∞	EQUI γ
0.5	2.0	0.01	0.933	0.07	0.933	0.07		3.5	0.01	0.942	0.11	0.944	0.12
		0.05	0.744	0.28	0.744	0.28			0.05	0.825	0.35	0.836	0.35
		0.20	0.347	0.73	0.347	0.73			0.20	0.616	0.76	0.644	0.76
		0.35	0.159	0.94	0.159	0.94			0.35	0.523	0.94	0.559	0.94
		0.50	0.109	1.00	0.109	1.00			0.50	0.495	1.00	0.533	1.00
	2.5	0.01	0.930	0.09	0.930	0.09		4.0	0.01	0.955	0.13	0.958	0.13
		0.05	0.756	0.31	0.756	0.31			0.05	0.871	0.36	0.880	0.36
		0.20	0.417	0.74	0.417	0.74			0.20	0.726	0.77	0.748	0.77
		0.35	0.262	0.94	0.262	0.94			0.35	0.662	0.95	0.690	0.95
		0.50	0.214	1.00	0.214	1.00			0.50	0.642	1.00	0.672	1.00
	3.0	0.01	0.934	0.10	0.934	0.10	0.90	2.0	0.01	0.944	0.07	0.845	0.19
		0.05	0.787	0.33	0.787	0.33			0.05	0.765	0.28	0.666	0.42
		0.20	0.518	0.75	0.518	0.75			0.20	0.342	0.75	0.370	0.79
		0.35	0.396	0.94	0.396	0.94			0.35	0.127	0.95	0.239	0.95
		0.50	0.359	1.00	0.359	1.00			0.50	0.059	1.00	0.199	1.00
	3.5	0.01	0.943	0.12	0.943	0.12		2.5	0.01	0.927	0.09	0.896	0.16
		0.05	0.830	0.35	0.830	0.35			0.05	0.740	0.32	0.746	0.38
		0.20	0.630	0.76	0.630	0.76			0.20	0.364	0.76	0.488	0.77
		0.35	0.541	0.94	0.541	0.94			0.35	0.189	0.95	0.374	0.95
		0.50	0.514	1.00	0.514	1.00			0.50	0.136	1.00	0.340	1.00
	4.0	0.01	0.956	0.13	0.956	0.13		3.0	0.01	0.921	0.11	0.927	0.13
		0.05	0.876	0.36	0.876	0.36			0.05	0.751	0.34	0.809	0.37
		0.20	0.737	0.77	0.737	0.77			0.20	0.438	0.76	0.606	0.77
		0.35	0.676	0.95	0.676	0.95			0.35	0.300	0.95	0.516	0.95
		0.50	0.657	1.00	0.657	1.00			0.50	0.254	1.00	0.489	1.00
0.667	2.0	0.01	0.935	0.07	0.910	0.10		3.5	0.01	0.928	0.12	0.948	0.12
		0.05	0.745	0.28	0.728	0.31			0.05	0.789	0.36	0.861	0.37
		0.20	0.339	0.72	0.353	0.74			0.20	0.544	0.77	0.714	0.77
		0.35	0.145	0.93	0.176	0.94			0.35	0.435	0.95	0.649	0.95
		0.50	0.085	1.00	0.121	1.00			0.50	0.402	1.00	0.630	1.00
	2.5	0.01	0.930	0.09	0.923	0.10		4.0	0.01	0.942	0.13	0.964	0.14
		0.05	0.753	0.31	0.755	0.32			0.05	0.838	0.37	0.905	0.38
		0.20	0.404	0.74	0.431	0.75			0.20	0.660	0.77	0.805	0.78
		0.35	0.243	0.94	0.283	0.94			0.35	0.582	0.95	0.761	0.95
		0.50	0.193	1.00	0.238	1.00			0.50	0.558	1.00	0.749	1.00
	3.0	0.01	0.933	0.10	0.932	0.11							
		0.05	0.782	0.33	0.792	0.34							
		0.20	0.502	0.75	0.533	0.76							
		0.35	0.376	0.94	0.417	0.94							
		0.50	0.338	1.00	0.382	1.00							

Relative Efficiency (ARE) of the estimator of the linear discriminant function in this case relative to the correctly supervised case. This ARE gives an idea of the amount of information contained in an error-prone training sample relative to an error-free training sample. We present formulae for ARE and values of ARE for certain ranges of the parameters. It is found that moderately error-prone training samples are still fairly useful and efficient and should be used under appropriate models to estimate the discriminant function.

REFERENCES

1. R. S. Chhikara and J. McKeon, Linear discriminant analysis with misallocation in training samples, *J. Am. statist. Ass.* 79, 899–906 (1984).
2. C. B. Chittineni, Learning with imperfectly labelled patterns, *Pattern Recognition* **12**, 221–281 (1980).
3. C. B. Chittineni, Estimation of probabilities of label imperfections and correction of mislabels, *Pattern Recognition* 13, 257–268 (1981).
4. A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. R. statist. Soc.* **39B**, 1–38 (1977).
5. B. Efron, The efficiency of logistic regression compared to normal discriminant analysis, *J. Am. statist. Ass.* **70**, 892–898 (1975).
6. U. Katre and T. Krishnan, Pattern recognition with an imperfect supervisor, unpublished (1987).
7. P. A. Lachenbruch, Discriminant functions when the initial samples are misclassified, *Technometrics* **8**, 657–662 (1966).

8. P. A. Lachenbruch, Discriminant functions when the initial samples are misclassified. II. Nonrandom misclassification models, *Technometrics* **16**, 419–424 (1974).

9. G. S. McLachlan, Asymptotic results for discriminant analysis when the initial samples are misclassified, *Technometrics* **14**, 415–422 (1972).

10. J. E. Michalek and R. C. Tripathi, The effect of errors in diagnosis and measurement on the estimation of the probability of an event, *J. Am. statist. Ass.* **75**, 713–721 (1980).

11. T. J. O'Neill, Normal discrimination with unclassified observations, *J. Am. statist. Ass.* **73**, 821–826 (1978).

12. D. M. Titterington, Updating a diagnostic system using unconfirmed cases, *Appl. Statistics* **25**, 238–247 (1976).

APPENDIX

FORMULA FOR EFFICIENCY OF ERROR-PRONE INITIAL SAMPLES

Let the matrix A be partitioned corresponding to (u, R), (β_0, β) and δ as

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

and similarly B and C. It is easily checked that

$$A = \begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & 0 \\ 0 & 0 & A_{33} \end{bmatrix}. \quad C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & C_{22} & C_{23} \\ 0 & C_{32} & C_{33} \end{bmatrix}.$$

Further, as already observed $A = B + C$.

We partition the inverses of A, B, C in a similar manner and denote the blocks by A^{11}, A^{12} etc. From above

$$B = \begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22}-C_{22} & C_{23} \\ 0 & C_{32} & A_{33}-C_{33} \end{bmatrix}.$$

Further, $l(x, y, z)$ breaks up into two factors one involving δ, z and y only and another y, x and parameters other than δ, yielding easily

$$A_{13} = E\left(-\frac{\partial^2(x, y, z)}{\partial \delta^2}\right)\alpha(1-\alpha).$$

Now, $I_C{}^* = I_C = (A^{22})^{-1}$

$$= A_{22} - A_{21}A_{11}^{-1}A_{12} = B_{22} + C_{22} - B_{21}B_{11}^{-1}B_{12}.$$

But $I_{UC}{}^* = (B^{22})^{-1}$

$$= B_{22} - (B_{21} B_{23}) \begin{bmatrix} B_{11} & B_{13} \\ B_{31} & B_{33} \end{bmatrix}^{-1} \begin{pmatrix} B_{12} \\ B_{32} \end{pmatrix}$$

$$= B_{22} - (B_{21} C_{23}) \begin{bmatrix} B_{11} & 0 \\ 0 & B_{33} \end{bmatrix}^{-1} \begin{pmatrix} B_{12} \\ C_{32} \end{pmatrix}$$

$$= B_{22} - B_{21}B_{11}^{-1}B_{12} - C_{23}B_{33}^{-1}C_{32}$$

$$= B_{22} - B_{21}B_{11}^{-1}B_{12} - C_{23}[\alpha(1-\alpha) - C_{33}]^{-1}C_{32}.$$

Hence $I_C = I_{UC}{}^* + C_{22} + C_{23}[\alpha(1-\alpha) - C_{33}]^{-1}C_{32}$. (A.1)

Following Efron[5], we make a linear transformation on the x-variable to make $N_p(\mu_0, \Sigma)$ and $N_p(\mu_1, \Sigma)$ respectively $N_p\left(-\frac{\Delta}{2}e_1, I\right)$ and $N_p\left(\frac{\Delta}{2}e_1, I\right)$ where $e_1 = (1, 0, \ldots, 0)$ and Δ the Mahalanobis distance between the two groups based on x. Then $\beta_0 = \lambda$, $\beta' = \Delta e_1$. The matrix $I_C{}^* = I_C$ was computed by Efron[5] as (2.2). Thus to compute $I_{UC}{}^*$ from (A.1) for use in (2.1) for AER we need to compute matrix C. For this, we follow the technique of Efron[5] for his Lemma 3 (pp. 895–896), which makes essential use of the exponential family form of $l(y|x, z)$. Here it is convenient to use w rather than z; then the parametrisation turns out directly in terms of δ. Then from the theory of exponential families

$$C = \lim_{n\to\infty}\frac{1}{n}\mathrm{Cov}_{\beta_0,\beta}T$$

where $T = \sum_{j=1}^{n}\begin{Bmatrix} 1 \\ x_j \\ w_j \end{Bmatrix} y_j$

and C is given by

$$\int_{R^p}\sum_{w=-1,+1}\begin{Bmatrix} 1 \\ x \\ w \end{Bmatrix}(1\ x'\ w)\eta_1(x, w)\eta_0(x, w)\,\mathrm{d}F(x, w)$$

where $\mathrm{d}F$ is the mixture density $\eta_0 f_0(x, z) + \eta_1 f_1(x, z)$. This now yields a formula analogous to (3.16) of Efron.[5] We thus obtain

$$\begin{bmatrix} C_{22} & C_{23} \\ C_{32} & C_{33} \end{bmatrix} = \eta_1\eta_0\alpha(1-\alpha)\left[\begin{array}{cccccc:c} F_0 & F_1 & 0 & \ldots & 0 & & B_0 \\ F_1 & F_2 & 0 & \ldots & 0 & & B \\ 0 & 0 & A_0 & \ldots & 0 & & . \\ 0 & 0 & 0 & \ldots & & F_0 & 0 \\ B_0 & B_1 & 0 & \ldots & 0 & & F \end{array}\right.$$

giving $C_{22} + C_{23}[\alpha(1-\alpha) - C_{33}]^{-1}C_{32}$ as D defined earli
Thus $I_{UC}{}^* = I_C - \eta_1\eta_0 D$.

About the Author—T. KRISHNAN received his Master's degree in Mathematics from Madras University in 1958 and Master's degree in Statistics and the Ph.D. degree from the Indian Statistical Institute in 1965 and 1968, respectively. He has been in the Indian Statistical Institute since then and is now an Associate Professor. He has held visiting positions in the University of Southampton and the University of Western Australia. His research interests are in Pattern Recognition, Biostatistics and Psychometry.