Research Journal of Mathematical and Statistical Sciences _____ISSN 2320–6047
Vol. **1(8),** 28-38, September **(2013)** *Res. J. Mathematical & Statistical Sci.*

Review Paper

A Review of Classification Models Using Discrete Variables

Onyeagu S.I. and Osuji G.A.[1], Ekezie D.D. and Ogbonna C.J.
Department of Statistics, Nnamdi Azikiwe University, Awka NIGERIA
Department of Statistics, Imo State University, Owerri, NIGERIA
Department of Statistics, Federal University of Technology Owerri NIGERIA

Available online at: www.isca.in
Received 21st August 2013, revised 28th August 2013, accepted 6th September 2013

Abstract

This paper is a review of the classification models used for discrete variables. Nine classification procedures for binary variables are discussed and some of them evaluated at each of 118 configurations of the sampling experiments. The results obtained ranked the procedures as follows: Optimal, first order Bahadur, LDF, Second Order, Full, Distance, Nrule.

Keywords: Classification models, orthogonal polynomials, likelihood ratio, discrete variables, misclassification, multinomial rule.

Introduction

The problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several categories or population groups based on the measurements made on the individual. So the problem is that of assigning item(s) into one of k, $k \geq 2$ known populations assuming the item(s) actually belong to one of the populations. Our interest is in deriving a rule that can be used to optimally assign an item to one of the populations. The optimality criterion is to minimize the risk associated with the rule. We shall be concerned with $k = 2$ population classification problems.

In some problems, fairly complete information is available about the distribution of the measurement vector X_{rx1} in the two groups. In this case, we may use this information and treat the problem as if the distributions are known. In most cases, information about the distribution of X_{rx1} comes from a relatively small sample from the populations, and slightly different procedures are used.

When all the parameters of the populations are known, the error of misclassification called the optimum error rate is evaluated directly from the given probability man function. If the parameters are unknown, the probabilities of misclassification, based on estimates of unknown parameters are obtained by simulation.

The Problem

Let π_1 *and* π_2 be two populations available with infinite number of individual objects. Let there be r characteristics of interest with corresponding measurement variables $X_1, X_2, X_3 ...., X_r$, where $r \geq 1$. Let the measurement vector of an individual in π_1 be $\underline{X_1} = \left(X_{11} \quad X_{12\ldots} \quad X_{1r}\right)'$ and in $\underline{X_2} = \left(X_{21} \quad X_{22\ldots} \quad X_{2r}\right)'$. Supposing we find an object "o" with measurement vector $X_0 = \left(X_{01}, X_{02}, X_{or}\right)'$ outside π_1 *and* π_2 and which must belong to either π_1 *or* π_2. The problem is how to assign 0 to π_1 *or* π_2 such that the risk or expected cost or probability of error is a minimum. The measurement vector $X_{r\times1}$ can be discrete, or continuous or a mixture of discrete and continuous variables. We are interested in $X_{r\times1}$ whose arguments are discrete and to be more precise Bernoulli. The case of continuous measurement vector X has been studied extensively and the case of mixed variables (discrete and continuous) is yet to be studied in detail.

In this inferential setting, the researcher can commit one of the following errors. An object from π_1 may be misclassified into π_2. Likewise, an object from π_2 may be misclassified into π_1 If misclassification occurs, a loss would be suffered. Let C(i / j) be the

cost of misclassifying an object π_j int o π_i. For the two population setting, we have that $C(2/1)$ means cost of misclassifying and object into π_2 given that it is from π_1.

$C(1/2)$ is the cost of misclassifying an object into π_1 given that it is from π_2. The relative magnitude of the loss $L(j,i) = C(i/j)$ depends on the case in question; for example failure to detect an early cancer in a patient is costlier than stating that a patient has cancer and discovering otherwise.

Classification Procedures

Full Multinomial Rule: The full multinomial rule is based on estimating the probability mass function in population π_i denoted by $f_i(x)$ with the minimum variance unbiased estimators

$$f_i(x) = \frac{n_i(x)}{n_i}, \quad i = 1, 2$$

where $n_i(x)$ is the number of individuals in a sample of size n_i from the population having response pattern X. The classification rule is;

Classify an item with response pattern X into π_i if $q_1 \frac{n_1(x)}{n_1} > q_2 \frac{n_2(x)}{n_2}$

and to π_2 if $q_1 \frac{n_1(x)}{n_1} < q_2 \frac{n_2(x)}{n_2}$

and with probability $1/2$ if $q_1 \frac{n_1(x)}{n_1} = q_2 \frac{n_2(x)}{n_2}$

Pires and Bronco (2004) noted as pointed out by Dillon and Goldstein (1978) that one of the undesirable properties of the Full Multinomial Rule is the way it treats zero frequencies. If $n_1(x) = 0 \; and \; n_2(x) \neq 0,$ a new observation with vector X will be allocated to π_2, irrespective of the sample sizes n_1 and n_2.

The Dillon-Goldstein Rule: Dillon and Goldstein (1978) proposed the following rule as a result of the problem arising from zero frequency. The rule called the D-rule is based on Matusita's distribution distance using the notation;
$n_i(x) = n_{ij}$ if x belong to state j.

The rule is classify item into π_1 if $\frac{[n_{2j}(n_{1j}+1)]^{1/2} + \sum_{k \neq j} (n_{1k} n_{2k})^{1/2}}{[n_{1j}(n_{2j}+1)]^{1/2} + \sum_{k \neq j} (n_{1k} n_{2k})^{1/2}} < [\frac{n_2(n_1+1)}{n_1(n_2+1)}]^{1/2}$

and to π_2 if $\frac{[n_{2j}(n_{1j}+1)]^{1/2} + \sum_{k \neq j} (n_{1k} n_{2k})^{1/2}}{[n_{1j}(n_{2j}+1)]^{1/2} + \sum_{k \neq j} (n_{1k} n_{2k})^{1/2}} > [\frac{n_2(n_1+1)}{n_1(n_2+1)}]^{1/2}$

randomly classify if $\frac{[n_{2j}(n_{1j}+1)]^{1/2} + \sum_{k \neq j} (n_{1k} n_{2k})^{1/2}}{[n_{1j}(n_{2j}+1)]^{1/2} + \sum_{k \neq j} (n_{1k} n_{2k})^{1/2}} = [\frac{n_2(n_1+1)}{n_1(n_2+1)}]^{1/2}$

Note that if $n_1 = n_2$, the D-rule reduces to the Full Multinomial Rule.
For $n_1 < n_2$ and $n_{1j} = 0$ and $n_{2j} > 0$ the rule becomes

Classify X into π_1 if $\sqrt{n_{2j}} < \left[\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2} - 1\right]\sum_{k\neq j}(n_{1k}n_{2k})^{1/2}$

and to π_2 if $\sqrt{n_{2j}} > \left[\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2} - 1\right]\sum_{k\neq j}(n_{1k}n_{2k})^{1/2}$

and randomly classify if $\sqrt{n_{2j}} = \left[\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2} - 1\right]\sum_{k\neq j}(n_{1k}n_{2k})^{1/2}$

However, suppose $n_2 > n_1$ and $n_{2j}=0$ but $n_{ij} > 0$ we shall classify item with response pattern X into π_1 if

$$\left[1-\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2}\right]\sum_{k\neq j}(n_{1k}n_{2k})^{1/2} < \left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2}\sqrt{n_{2j}}$$

and to π_2 if $\left[1-\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2}\right]\sum_{k\neq j}(n_{1k}n_{2k})^{1/2} > \left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2}\sqrt{n_{2j}}$

randomly classify if $\left[1-\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2}\right]\sum_{k\neq j}(n_{1k}n_{2k})^{1/2} = \left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2}\sqrt{n_{2j}}$

The Likelihood Ratio Criterion Rule (L-Rule): Considering a generalized ratio test for the hypothesis

$H_0 : X, X_{11},...,X_{1n_1} \sim f_1(x)\ and\ X_{21},...,X_{2n_2} \sim f_2(x)$

against $H_1 : X_{11},...,X_{1n_1} \sim f_1(x)\ and\ X, X_{21},...,X_{2n_2} \sim f_2(x)$

As was proposed by Anderson (1982), Pires and Bronco found that the likelihood ratio criterion also handles the problem of zero frequency. For the multinomial model, they proposed a test statistics that is a function of X and is given by

$$L(x) = \frac{\left(1+1/n_1(x)\right)^{n_1(x)}(n_1(x)+1)}{\left(1+1/n_2(x)\right)^{n_2(x)}(n_2(x)+1)} \times \frac{\left(1+\frac{1}{n_2}\right)^{n_2}(n_2+1)}{\left(1+\frac{1}{n_1}\right)^{n_1}(n_1+1)}$$

This rule fails to take account of several factors that may be important in practice. These factors are the deferential prior-probabilities of observing individuals from the two populations and differential cost incurred by misclassification and a-prior probabilities and where if $n_1(x) = 0$ and $n_2(x) = 0$, the classification rule becomes

Classify item with response pattern into $\pi_1\ of\ L(x) > 1\ and$ into π_2 if $L(x) < 1$. For $n_1 = n_2$, this rule falls back to the Full Multinomial Rule.

Procedure Based on the Linear Discriminant Function: The linear discriminant function for discrete variables is given by

$$\hat{L}\left[X_j(1)\right] = \sum_j\sum_k \left(\hat{P}_{2j} - \hat{P}_{ij}\right)S^{kj}X_k - \frac{1}{2}\sum_j\sum_k \left(\hat{P}_{2j} - \hat{P}_{ij}\right)S^{kj}\left(\hat{P}_{2k} + \hat{P}_{1k}\right)$$

where S^{kj} are the elements of the inverse of the pooled sample covariance matrix, $\hat{P}_{ij}\ and\ \hat{P}_{2j}$ are the estimates of the sample means in $\pi_1\ and\ \pi_2$ respectively. The classification rule obtained using this estimation is;

Classify an item with response pattern X into π_1 if $\sum_j\sum_k\left(\hat{P}_{2j}-\hat{P}_{ij}\right)S^{kj}X_k - \frac{1}{2}\sum_j\sum_k\left(\hat{P}_{2j}-\hat{P}_{ij}\right)S^{kj}\left(\hat{P}_{2k}+\hat{P}_{1k}\right) > 0$ and to π_2 otherwise.

First Order Bahadur Procedure: The first order Bahadur procedure is based on the Bahadur model

$$f_i(x) = \prod_{j=1}^{r} P_{ij}^{x_j}(1-P_{ij})^{1-x_j}[1+\sum_{j<k}\ell_i(jk)Z_{ij}Z_{ik}+\sum_{j<k<i}\ell_i(jki)Z_{ij}Z_{ik}Z_{ii} +\ldots+\ell_i(1,2,\ldots,r)Z_{i1}Z_{i2}\ldots Z_{ir}$$

where $f_i(x)$ is the probabilities that response pattern X is observed in the i^{th} population

$$P_{ij} = E_i(x_j)\ i=1,2,\ j=1,2,\ldots,r$$

$$Z_{ij}=\frac{x_j - p_{ij}}{\sqrt{p_{ij}(1-p_{ij})}}$$

$$\ell_i(jk) = E(Z_{jk}Z_{ik})$$

$$\ell_i(jkl) = E(Z_{ij}Z_{ik}Z_{il})$$

$$e_i(1,2,\ldots r) = E(Z_{i1}, Z_{i2},\ldots, Z_{ir})$$

$$e_{ij} = \sum_{sj}\frac{n_i(x)}{n_i}$$

where S_j in the set of all pattern X with $X_j = 1$.

The first order Bahadur is obtained by omitting all correlation terms $\ell_i(jk)\ \ell_i(jkl)\ldots\ell_i(1,2,\ldots r)$. The classification rule is obtained using the likelihood ratio $f_i(x)\big/f_i(x)$.

Second Order Bahadur Procedure

The second order Bahadur procedure is obtained when $P_{ij}\ and\ P_i(jk)$ terms are retained and all other correlation terms omitted. The classification is obtained by taking the likelihood ratio

$$\hat{f}_1(x)\big/\hat{f}_2(x)$$

Where $\hat{f}_i(x) = \prod_{j=1}^{r}\hat{P}_{ij}^{x_j}(1-P_{ij})^{1-x_j}[1+\sum_{j<k}\hat{e}_j(jk)\hat{Z}_{ij}\hat{Z}_{ik}]$

where $\hat{e}_i(jk) = \dfrac{\sum_{sjk}\frac{n_i(x)}{n_i} - \hat{P}_{ij}\hat{P}_{ik}}{\sqrt{\hat{P}_{ij}(1-\hat{P}_{ij})\hat{P}_{ijk}(1-\hat{P}_{ijk})}}$

$$\hat{Z}_{ij} = \frac{n_{ij}(x) - \hat{P}_{ij}}{\sqrt{\hat{P}_{ij}(1-\hat{P}_{ij})}}$$

where $n_{ij}(x)$ is number of response pattern X from the ith population with X_j=1

Nearest Negbour Procedure

Hills was concerned with the problem of estimating likelihood ratios for Bernoulli variables. As there are 2^k possible outcomes, the distribution may be considered multinomial, although in some cases it is possible to represent the distribution more economically. He proposed the used of nearest neighbour procedures to help overcome the problem of small (or zero) cell frequencies. For the Bernoulli case, a cell may be represented by the corresponding pattern of zeros and ones for example with k = 3 variable, the possible patterns are

000 001 010 011
100 101 110 111

The near neighbour of order 1 is one that differs from the pattern in only one variable. The near neighbours of 010 are011, 000, and 110

If the cell count for the jth cell is n_{ij}, then the nearest neighbour procedure assigns the observation to π_1 if

$$\frac{\left(n_{1j}+\sum_{A} n_{1j}\right)/n_1}{\left(n_{2j}+\sum_{A} n_{2j}\right)/n_2} > \frac{p_2}{p_1}$$

where A is the set of neighbour of state j. Hills comments that the estimate of the likelihood ratio has less sampling variability than the simple method using cell frequencies.

The Optimal Classification Rule

Let $\pi_1 \; and \; \pi_2$ be two multivariate Bernoulli populations.

Let $C(i/j)$ be the cost of misclassifying an item with measurement X from π_j into π_i and let q_i be the prior prob. of π_c where $i = 1,2 \; with \; q_1 + q_2 = 1$ and prob. mass function $f_i(x)$ in π_i where $i = 1,2$. Supposing we assign an item with measurement $\underline{X}$ to π_i if is in some region $R \subset R^r$ and to π_2 if X is in some region $R_1 \subset R^r$ where $R^r = R_1 \cup R_2 \; and \; R_1 \cap R_2 = 0$

The expected cost of misclassification is given by $ECM = C(2/1)q_1 \sum_{R_1} f(X/\pi_1) + C(1/2)q_2 \sum_{R_2} f(X/\pi_2)$

where $\sum_{R_2} f(X/\pi_1) = P[Classifying \text{ into } \pi_2/\pi_1] = P(2/1)$

$\sum_{R_1} f(X/\pi_1) = P[Classifying \text{ into } \pi_1/\pi_2] = P(1/2)$

The optimal classification rule is one that partitions R^r such that ECM is a minimum

$ECM = C(2/1)q_1\left[1 - \sum_{R_1} f(X/\pi_1)\right] + C[1/2]q_2 \sum_{R_2} f(X/\pi_2)$

Since $\sum_{R_1} f(X/\pi_1) + \sum_{R_2} f(X/\pi_1) = 1$

$\sum_{R_2} f(X/\pi_1) = 1 - \sum_{R_1} f(X/\pi_1)$

$ECM = C(2/1)q_1 + \sum_{R_1} [C(1/2)q_2 f(X/\pi_2) - C(1/2)q_1 f(X/\pi_1)]$

ECM is minimized if the second term is minimized. ECM is minimized, it R_1 is chosen such that

$C(1/2)q_2 f(X/\pi_2) - C(2/1)q_1 f(X/\pi_1) \le 0$

$$R_1 = \left[X / \frac{f(X/\pi_1)}{f(X/\pi_2)} \ge \frac{c(1/2)q_2}{c(2/1)q_1} \right]$$

Therefore the optimal classification rule with respect to minimization of the expected cost of misclassification is given by: classify object with measurement X into π_1 if $\frac{f(X/\pi_1)}{f(X/\pi_2)} \ge \frac{c(1/2)q_2}{c(2/1)q_1}$ otherwise, classify into π_2 ,

Without loss of generality, we can assume that $C(1/2) = C(2/1)$. Then minimization of ECM becomes minimization of the probability of misclassification P(MC).

The optimal rule reduces to classify an item with measurement X into π_1 if $\frac{f(X/\pi_1)}{f(X/\pi_2)} \geq 1$ otherwise classify into π_2

Since X is multivariate Bernoulli with $P_{ij} = 1 - q_{ij} > 0\ i = 1,2\ j = 1,2,....r$ the optimal rule is classify an object with measurement X into π_1 if $\sum_{j=1}^{r} Xj \ln\left(\frac{P_{ij}}{q_{ij}} . \frac{q_{2j}}{P_{2j}}\right) > \sum_{j=1}^{r} \ln \frac{q_{ij}}{qij}$

otherwise classify into π_2

Evaluating the Probability of Misclassification for the Optimal Classification Rule: For the optimal classification rule, we obtained the probabilities of misclassification for two cases.

Case 1 Known Parameters

i. General case where $P_i = (P_{i1}\ P_{i2}\ ...P_{ir})$, ii. Special case where $P_i = (P_i\ P_i\ ...P_i)$ With assumption that $P_1 < P_2$, iii. Special case 1b with additional assumption that $P_1 = \theta P_2 \quad 0 < \theta < 1$

Case 2

i. General case $P_i = (P_{i1}\ P_{i2}...P_{ir})$, ii. We estimate $P_1\ and\ P_2$ by taking samples of size from $\pi_1\ and\ \pi_2$ respectively, iii. Special case where $P_i = (P_i\ P_i...P_i)$ with the assumption that $P_{1i} < P_{2i}$ we also estimate $P_1\ and\ P_2$, iv. Special case 2b with $P_1 = \theta P_2\ P_1 < P_2\ 0 < \theta < 1$. We take training samples of size $n_2\ from\ \pi_2$ and estimate P_2 for fixed value of θ, $P_1 = \theta P_2$.

For case 1b $P(MC) = \frac{1}{2}\left[1 + B_{(r,P_2)}\left(\frac{r \ln\left(\frac{q_2}{q_1}\right)}{\ln\left(\frac{P_1}{P_2} . \frac{q_2}{q_1}\right)}\right) - B_{(r,P_2)} \frac{r \ln\left(\frac{q_2}{q_1}\right)}{\ln\left(\frac{P_1}{P_2} . \frac{q_2}{q_1}\right)}\right]$

$P(MC) = \frac{1}{2}[1 + B(r, P_2\ \lambda) - B(r, P_1\ \lambda)]$

where $B(r, p)(X) = \sum_{y=0}\left(\frac{r}{y}\right) P^y (1-P)^{r-y}$ given by

For case (1c) the probability of misclassification is $P(MC) = \frac{1}{2}\left[1 + B_{(r,P_2)} \frac{r \ln\left(\frac{1-P_2}{1-\theta P_2}\right)}{\ln\left(\frac{1-P_2}{1-\theta P_2}\right)} - B_{(r,\theta P_2)}\left(\frac{r \ln\left(\frac{1-P_2}{1-\theta P_2}\right)}{\ln\left(\theta \frac{1-P_2}{1-\theta P_2}\right)}\right)\right]$

$P(MC) = \frac{1}{2}[1 + B(r, P_2\ \lambda) - B(r, P_1\ \lambda)]$

For cases 2b and 2c the formula remains the same except that the parameters are estimated by their MLE estimates.

Application with Life Data: The data used in this example was collected at the University of Ilorin Teaching Hospital. The data is made of the following categories of heart disease patients. i. Heart failure, ii. Hypertensive heart failure, iii. Hypertensive heart failure with stroke in evolution, iv. Congestive heart failure, v. Cardiovascular accident

There are two populations, i. Those who survived the attack, ii. Those who died

Three variables were used Systolic blood pressure. i. Diastolic blood pressure, ii. Heart rate

π_1 has 131 patients, π_2 has 83 patients

The systolic blood pressure is normal if it is less than 140mmHg ie 90-140. The diastolic blood pressure is normal if it is less than 90mmHg ie 60-90. The heart rate is normal if it is less than 100 beats per minute ie (60-100) with the above information the dichotomized the measurements.

Variable 1: Systolic Blood =1 if it is less than 140mmHg Pressure = 0 otherwise

Variable 2: diastolic = 1 if it is less than 90mmHg = 0 otherwise

Variable 3: Heart rate = 1 if it is less than 100 beats per minute = 0 otherwise

Let $X = (X_{1,}\ X_2\ \ X_3)$ denote the total response to the measurements and this leads to the following 2^r response patterns.
000, 100, 010, 110, 001, 101, 011, 111

We selected the first 50 patients from each group and computed a classification rule. The frequency distribution is seen below.

STATE			**Survivable Group**	**Non Survivable Group**
x_1	**x_2**	**x_3**	**Frequency**	**Frequency**
0	0	0	4	3
1	0	0	1	1
0	1	0	1	3
1	1	0	3	12
0	0	1	22	15
1	0	1	2	1
0	1	1	2	3
1	1	1	15	12

The population parameters are not known so they are estimated by the MLE $\hat{P}_y = \sum_{k=1}^{n_1} \frac{X_{ikj}}{n_i} = \frac{n_i(xj)}{n_i}$

$$\hat{P}_1 = \left(\frac{21}{50}\frac{21}{50}\frac{41}{50}\right) = (.42\,.\,42\,.\,82)$$

$$\hat{P}_2 = \left(\frac{26}{50}\frac{30}{50}\frac{31}{50}\right) = (.52\,.\,6\,.\,62)$$

Using these estimates we obtained the classification rule. Classify the item with response pattern X into π_1 if

$-0.0402816\ x_1 - 0.728238\ x_2 + 1.026799\ x_3 > 0.1864082$ and to π_2 otherwise. The response patterns were classified as follows.

Response pattern	**Classification**
000	π_2
100	π_2
010	π_2
110	π_2
001	π_1
101	π_1
011	π_1
111	π_2

$$\hat{P}(1/2) = P\left[x = (001)(101)(011) / \hat{P}_2 = (.552\,.\,6.\,62)\right]$$

$$= \hat{q}_{21}\,\hat{q}_{22}\,\hat{p}_{23} + \hat{p}_{21}\,\hat{q}_{22}\,\hat{p}_{23} + \hat{q}_{21}\,\hat{p}_{22}\,\hat{p}_{23}$$

$$= .48\,x.40\,x.62 + .52\,x.4\,x.62 + .48\,x.6\,x.62 = 0.42656$$

$$\hat{P}(2/1) = P\left[x = (000)(100)(010)(110\,(111)) / \hat{P}_1 = (.42\,.42\,.82)\right]$$

$$= \hat{q}_{11}\,\hat{q}_{12}\,\hat{q}_{13} + \hat{p}_{11}\,\hat{q}_{12}\,\hat{q}_{13} + \hat{q}_{11}\,\hat{P}_{12}\,\hat{q}_{13} + \hat{P}_{11}\,\hat{P}_{12}\,\hat{q}_{13} + \hat{P}_{11}\,\hat{P}_{12}\,\hat{P}_{13}$$

$$= .58\,x.58\,x.18 + .42\,x.58\,x.18\,.58\,x.42\,x.18 + .42\,x.42\,x.18 + .42\,x.42\,x.82 = .32\,.4648$$

$$error\ \ rate = \frac{.42656 + .324648}{2} = 0.3756$$

Supposing we want to use the Full Multinomial rule for classification, then, the response patterns are classified as follows

STATE			Survivable Group	Non Survivable Group	
x_1	x_2	x_3	Frequency	Frequency	Allocation
0	0	0	4	3	π_1
1	0	0	1	1	π_1
0	1	0	1	3	π_2
1	1	0	3	12	π_2
0	0	1	22	15	π_1
1	0	1	2	1	π_1
0	1	1	2	3	π_2
1	1	1	15	12	π_1

$$\hat{P}(1/2) = p[(000), (100), (001), (101), (111)/\ \hat{p} = (0.52, 0.6, 0.62)]$$

$$= q_{21}q_{22}q_{23} + p_{21}q_{22}q_{23} + q_{21}q_{22}p_{23} + p_{21}q_{22}p_{23} + p_{21}p_{22}p_{23}$$

$$= 0.48 \times 0.4 \times 0.38 + 0.52 \times 0.4 \times 0.38 + 0.48 \times 0.4 \times 0.62 + 0.52 \times 0.4 \times 0.62 + 0.52 \times 0.6 \times 0.62$$

$$= 0.07296 + 0.07904 + 0.11904 + 0.12896 + 0.19344 = 0.59344$$

$$\hat{P}(2/1) = p[(010), (1`10), (011)/\ \hat{p} = (0.42, 0.42, 0.82)]$$

$$= \hat{q}_{11}\hat{p}_{12}\hat{q}_{13} + \hat{p}_{11}\hat{p}_{12}\hat{q}_{13} + \hat{q}_{11}\hat{p}_{12}\hat{p}_{13}$$

$$= 0.58 \times 0.42 \times 0.18 + 0.42 \times 0.42 \times 0.18 + 0.58 \times 0.42 \times 0.82 = 0.27535$$

$$Error\ \ Rate = (0.27535 + 0.59344)/2 = 0.4344$$

Procedure Based on Orthogonal Polynomials

This section focuses on two approaches that utilize orthogonal functions to affect classification in both approaches; the multivariate binary density at the point X is expressed as a linear combination of orthogonal polynomials. The approaches are the Martin-Bradley Model and the Kronmal-OH-Tarter Model. Due to the fact that both utilize orthogonal functions to affect classification, we are going to present only the Kronmal-OH-Tarter Model. Suppose $X=(X_1,X_2,\ldots,X_r)$ is a multinomial binary vector with an associated sample space consisting of 2^r points. Let these points be numbered by a binary index p and consider the orthogonal functions

$$\Psi_p(x) = (-1)^{x'_p}$$

where $X'_p = \sum_j x_j p_j$

The orthogonal of these functions follows by observing that

$$\sum_x \Psi_p(x)\Psi_j(x) = \sum_x (-1)^{2x'_p} = 2^r \ \ for\ \ p = j$$

$\sum_x (-1)^{x'(p+j)} = 0 \quad for \quad p \neq j$

Kronmal and Tarter (1968) represent the probability associated with a point X by $f(x) = 2^{-r} \sum_p d_p \Psi_p(x)$

where d_j is a parameter associated or determined by noting that

$\Psi_j(x) f(x) = 2^{-r} \sum_p d_p \Psi_p(x) \Psi_j(x)$

and hence, $\sum_x \Psi_j(x) f(x) = 2^{-r} \sum_p d_p \sum_x \Psi_p(x) \Psi_j(x) = d_j = E[\Psi_j(x)]$

Based on a random sample of size n, let n(x) be the frequency of the state defined by X. Then, the maximum likelihood estimate of f(x) can be written as; $\hat{f}(x) = 2^{-r} \sum_p d_p \Psi_j(x)$

where $\hat{d}_r = \sum_x \Psi_j(x) n(x) \Big/ n$

Kronmal and Tarter have shown using mean summed squared error $E \sum_x (f(x) - \hat{f}(x))^2$

As a criterion of fit that the increase in error due to inclusion of the rth term, namely, d_r, in the representation of f(x) is given by

$\frac{2^{-r}\left[1 - (n+1) d_p^2\right]}{n}$

and is estimated by $\frac{2^{-r}\left[2 - (n+1) \hat{d}_p^2\right]}{n-1}$

It follows therefore that inclusion of $\hat{d}_r \; in \; f(x)$ leads to a decrease in error if the above equation is negative, that is if $\hat{d}_r > \frac{2}{n+1}$

Assuming equal prior probabilities and denoting by f_1 and f_2, the underlying multinomial densities respectively associated with $\pi_1 \; and \; \pi_2$, it follows that the optimal classification rule using the representation discussed above is to clarify x into $\pi_1 (\pi_2)$ if

$\sum_p \Psi_p(x) d_{1.p} > (<) \sum_p \Psi_p(x) d_{2.p}$ and randomly assign otherwise where the parameter sets $\{ d_{i.p} \}$ are associated with $\pi_1 \; and \; \pi_2$ respectively.

If all parameters are estimated, the sample-based rule is simply the rule given above with the parameter set d_{ip} replaced by their estimates $\hat{d}_{i.p}$, $i = 1,2.$

Sampling Experiments and Results

We have compared some of these procedures. Included in the list are the optimal rule, full multinomial, first, second and LDF procedures. The simulation experiments are based on population characterized by three, four and five variables. In general, a simulation experiment is characterized by the values assigned to the input parameters $P_{1j}$and P_{2j}. In addition to mean structures characterized by marginal probabilities P_{1j} and P_{2j}, we consider structures determined by the differences

$d = (p_{2j} - p_{1j}) \geq 0, \; j = 1, ..., r$

To make the study reasonable in size, we take $d \leq 0.4$.

On the whole 21 population pairs given rise to 118 configurations are formed specifying values for the means P_{ij}. Seven population pairs are based on three variables; eight are based on 4 variables and six on five variables. The five classification procedures are evaluated at each of the 118 configurations of n, r and d. The 118 configurations of n, r, and d are all possible combination of n=40, 60, 100, 140, 200, 300, 400, 600, 700, 800, 900, 1000, r=3, 4, 5 and d=0.1, 0.2, 03 and 0.4.

A training data set of size n is generated via IMSL where n_1= n/2 observations are sampled from π_1 which has multivariate Bernoulli distribution with input parameter P_1 and n_2 = n/2 observations from π_2 which is multivariate Bernoulli with input parameter P_2, j=1, 2, …, r. These samples are used to consider the rule for each procedure and estimate the probability of misclassification for each procedure is obtained by the plug-in rule or the confusion matrix in the sense of the full multivariate.

The likelihood ratios are used to define classification rules. The plug-in estimates of error rates are determined for each of the classification rules.

Step 1 and 2 are repeated 1000 times and the mean plug-in error and variances for the 1000 trials are recorded. The method of estimation used here is called the re-substitution method. The simulation experiments have been implemented using the IMSL. The number of iteration used for each of the configurations of n, r and d is 1000. The entire simulation required approximately 1008hrs of CPU time.

The result of just one configuration is displayed here $P_1 = (.3, .3, .3)'$ $P_2 = (.5, .5, .5)$

Sample Size	Optimal	Full	First	Second	LDF
50	0.341097	0.325212	0.341117	0.328162	0.340302
70	0.347421	0.334872	0.347272	0.337122	0.346458
100	0.349955	0.341089	0.34995	0.34958	0.34958
150	0.353234	0.346947	0.352976	0.352877	0.352877
400	0.357489	0.35587	0.357321	0.357302	0.357302

P(MC)= 0.358

Sample Size	Optimal	Full	First	Second	LDF
50	0.016903	0.03278	0.01688	0.02983	0.01769
70	0.010588	0.02312	0.01072	0.02087	0.01154
100	0.008045	0.01691	0.00805	0.01516	0.00842
150	0.00476	0.01105	0.00502	0.00988	0.00512
400	0.00057	0.00293	0.00067	0.0024	0.00069

Classification Rule	Performance
Optimal	1
First	2
LDF	3
Second	4
Full	5

Conclusion

We observed several marginal trends. We observed the good performance of the optimal classification rule and the first order and LDF procedures for small ratios of ln(n/r). Even when $\ln\left(n/r\right) \leq 3$ these three procedures yielded good approximations to the exact probability of misclassification P(MC) whereas the full multinomial and the second order Bahadur procedures falter. The performance of the LDF is close to that of the first order for decreasing ratios of $\ln\left(n/r\right)$ and is not notably worse at all the four d values for $\ln\left(n/r\right) < 4.60$. The performance of the optimal, first and LDF procedures are little affected by the magnitude of d within the range of the design. The first and LDF recorded large variances for small values of n. The accuracy of the optimal classification rule improved as the size of n is increased. In terms of minimum variance, the optimal classification rule performed better than all the procedures. From the analysis, the procedures can be ranked as follows; Optimal, First Order Bahadur, LDF, Second Order Bahadur, Full Multinomial.

Osuji compared seven classification procedures namely the optimal rule, first and second order Bahadur procedures, LDF, Full Multinomial, distance and nearest neighbour procedures. He concluded that the classification rules that can be considered to be good are the optimal rule, Full Multinomial, Distance and Second Order Bahadur procedures. The following table from Osuji illustrates the result;

Frequencies		
Distance	**Classification Rule**	**1 2 3 4**
0.1		OPT OPT OPT FULL/DISTANCE
0.2		OPT FULL OPT
0.3		OPT FIRST
0.4		OPT

References

1. Cochran W.G. and Hopkins C.E., Some classification problems with multivariate qualitative data, *Biometrics,* **17,** 10-32 **(1961)**

2. Gilbert E.S., on discriminant using qualitative variables, *J. Ame. Stat. Ass.,* 63, 1399 **(1968)**

3. Glick N., Sample base multinomial classification, *Biometrics,* 29, 241-256 **(1973)**

4. Goldstein M. and Dillon W.R., Discrete discriminant analysis. John Wiley and Sons Inc. New York **(1978)**

5. Goldstein and Rabinowitz, Selection of variables for the two group multinomial classification, *J .Ame. Stat. Ass*., **70,** 776-781 **(1975)**

6. Goldstein M. and Wolf E., On the problem of bias in multinomial classification, *Biometrics,* **33,** 325-331 **(1977)**

7. Hoel M. and Paterson C., A solution to the problem of optimum classification, *Ann Math Stat.,* **20**, 433-438 **(1949)**

8. Hills M., Discriminant and Allocation with discrete data, *J. Royal Statistical Society,* **C16,** 237-250 **(1967)**

9. Mclachlan R., Discriminant Analysis and Statistical Pattern Recognition, John Wiley and Sons Inc, New York **(1992)**