Ann. Inst. Statist. Math.
31 (1979), Part A, 1-7

ON THE EMPIRICAL BAYES APPROACH TO CLASSIFICATION IN THE CASE OF DISCRETE MULTIVARIATE DISTRIBUTION HAVING ONLY FINITE MASS POINTS

HIROSI HUDIMOTO

(Received Feb. 21, 1978; revised Oct. 24, 1978)

1. Introduction

In [2], the empirical Bayes approach to classification problems has been considered for the case that a population π is divided into r mutually exclusive sub-groups $\pi_1, \cdots, \pi_r$ and that proportions w_i's of individuals of π belonging to π_i's are unknown. The similar approach is dealt with in a special setting of classification for the case of $r=2$.

Suppose to be known that each individual in the given population π belongs to one of two mutually exclusive groups π_1 and π_2. Our purpose is to classify each of n individuals randomly drawn from π to either π_1 or π_2 as correctly as possible. An application for the case of classification based on individuals responses to a battery of m dichotomous items is given in the last section. Assume that observations are obtained from any m-variate distribution that a random vector can take only a finite number of distinct points.

Let w_1 and w_2 be unknown proportions of individuals of π belonging to π_1 and π_2, respectively. We shall regard $w=(w_1, w_2)$ as an unknown prior distribution which represents the chance that an individual randomly drawn from π belongs to π_1 or π_2.

2. Preliminary consideration

Let $f_i(x)$ be the joint probability function of the m-variate x which can take only s distinct points $x_1, \cdots, x_s$, in π_i, $i=1, 2$. For the preliminary consideration to show the basic idea of our procedure, we assume that $f_1(x)$ and $f_2(x)$ are known. Consider the likelihood ratio $L(x)=f_2(x)/f_1(x)$, and denote the ξth of $L(x)$ arranged in ascending order by $L_{(\xi)}$, that is,

$$0 \leqq L_{(1)} \leqq \cdots \leqq L_{(\xi)} \leqq \cdots \leqq L_{(s)} \leqq \infty . \tag{2.1}$$

Let $g_i(\xi)$ be $f_i(x)$ arranged in the rank order of $L_{(\xi)}$, and let $G_i(y)$ be defined by $G_i(y)=\sum_{\xi\leq y} g_i(\xi)$.

Suppose now that a random sample of the size n is obtained from π. Observed vectors included in the sample from π are transformed into the rank order based on $L_{(\xi)}$. Let $(y_1,\cdots,y_n)$ be the sample from π transformed into the rank order defined above.

Our purpose is to classify each of individuals contained in the sample to either π_1 or π_2 as correctly as possible. Firstly, the unknown prior distribution $w=(w_1, w_2)$ is estimated from the sample. Secondly, the empirical Bayes decision rule is made for our classification problem. Take the statistic

$$\hat{p}_i=\frac{1}{n}\sum_{\nu=1}^{n}\left\{G_i(y_\nu)-\frac{1}{2}g_i(y_\nu)\right\}, \qquad i=1,2. \tag{2.2}$$

Then, for unbiased and consistent estimates of w_1 and w_2, we have

$$\hat{w}_1=\frac{1}{\Delta_{12}}\left\{\frac{1}{2}-\hat{p}_2\right\} \quad \text{and} \quad \hat{w}_2=\frac{1}{\Delta_{12}}\left\{\hat{p}_1-\frac{1}{2}\right\}, \tag{2.3}$$

where Δ_{12} is

$$\Delta_{12}=\frac{1}{2}\sum_{\xi=1}^{s}\{G_1(\xi)g_2(\xi)-G_2(\xi)g_1(\xi)\}. \tag{2.4}$$

It is obvious that $\Delta_{ij}=-\Delta_{ji}$ and $\Delta_{ii}=0$, $i,j=1,2$. The expectation of $G_i(y)-(1/2)g_i(y)$ with respect to $g_j(\xi)$ is

$$\mathcal{E}_{g_j}\left\{G_i(y)-\frac{1}{2}g_i(y)\right\}=\frac{1}{2}+\Delta_{ij}, \qquad i,j=1,2. \tag{2.5}$$

Thus, we have

$$\mathcal{E}_{g^n}\{\hat{p}_1-\hat{p}_2\}=\Delta_{12}, \tag{2.6}$$

where $\mathcal{E}_{g^n}$ denotes the expectation with respect to the joint distribution of $y_1,\cdots,y_n$ which have the common probability function given by

$$g(\xi)=w_1g_1(\xi)+w_2g_2(\xi). \tag{2.7}$$

Possibly it may be that either of $\hat{w}_i$'s $(i=1,2)$ comes to a negative value, because $\hat{w}_i$'s are unbiased estimates of w_i's. Such a happening will be more probable in the case that either one of w_i's is fairly small. Then, we have defined the ordering based on $L_{(\xi)}$ in order to get a greater Δ_{12}. The formula (2.6) intuitively shows our aim for the above described situation.

Let $L(j|i)$ be the loss incurred if a decision is made to classify him as coming from π_j when the individual is actually from π_i. It is assumed that $0<L(j|i)<\infty$ if $i\neq j$ and $L(i|i)=0$, $i,j=1,2$.

Consider

(2.8) $$D_{\hat{w}}(x)=L(1|2)f_2(x)\left(\hat{p}_1-\frac{1}{2}\right)-L(2|1)f_1(x)\left(\frac{1}{2}-\hat{p}_2\right) .$$

Then, if $\Delta_{12}>0$, we make a decision rule

(2.9) $$\delta_{\hat{w}}(x)=\begin{cases} a_1 & \text{if } D_{\hat{w}}(x)\leqq 0 , \\ a_2 & \text{if } D_{\hat{w}}(x)>0 , \end{cases}$$

where a_i indicates to classify the individual with x to π_i, $i=1, 2$, and $\hat{w}$ means $\hat{w}=(\hat{w}_1, \hat{w}_2)$.

Denote by $B(w, \delta_{\hat{w}})$ the expected risk of $\delta_{\hat{w}}(x)$ with respect to $w=(w_1, w_2)$. The rule of the form (2.9) with $(1/2-\hat{p}_2)$ and $(\hat{p}_1-1/2)$ in (2.8) replaced by w_1 and w_2 is a Bayes decision rule with respect to w, and $B(w)=B(w, \delta_w)$ is the corresponding Bayes risk. Then, it can be shown that

(2.10) $$\mathcal{E}_{g^n}B(w, \delta_{\hat{w}})\rightarrow B(w) \qquad \text{as } n\rightarrow\infty .$$

Robbins, [3], has defined by (2.10) "asymptotic optimality" of an estimated Bayes decision rule. Thus, we have the following: *If $\Delta_{12}>0$, the decision rule given by* (2.9) *is asymptotically optimal relative to $w=(w_1, w_2)$.*

About each of n individuals contained in the sample, the same rule as (2.9) can be written as follows:

($*$) *Classify an individual with x having the rank (ξ) to π_1 if $(\xi)\leqq(\xi_0)$ or to π_2 if otherwise, where (ξ_0) is determined by*

$$L_{(\xi_0)}\leqq L(2|1)\left(\frac{1}{2}-\hat{p}_2\right)\Big/L(1|2)\left(\hat{p}_1-\frac{1}{2}\right)<L_{(\xi_0+1)} .$$

3. A procedure for the case that $f_1(x)$ and $f_2(x)$ are unknown

In the case that $f_1(x)$ and $f_2(x)$ are not completely known, we shall assume that past observations randomly obtained from π_1 and π_2 are available, respectively.

Let $(x_1^{(1)},\cdots, x_{n_1}^{(1)})$ and $(x_1^{(2)},\cdots, x_{n_2}^{(2)})$ be those samples obtained from π_1 and π_2, respectively. Define the relative frequencies $\hat{f}_{n_1}(x)$ and $\hat{f}_{n_2}(x)$ by

(3.1) $$\hat{f}_{n_i}(x)=\frac{1}{n_i}\sum_{\mu=1}^{n_i} n(x, x_\mu^{(i)}) , \qquad i=1, 2 ,$$

and

$$n(x, x_\mu^{(i)}) = \begin{cases} 1 & \text{if } x = x_\mu^{(i)} , \\ 0 & \text{if } x \neq x_\mu^{(i)} . \end{cases}$$

Consider $\hat{L}_{n_1,n_2}(x) = \hat{f}_{n_2}(x)/\hat{f}_{n_1}(x)$ instead of the likelihood ratio $L(x)$ in the Section 2 and denote the ξth of $\hat{L}_{n_1,n_2}(x)$ arranged in ascending order by $\hat{L}_{(\xi)}(n_1, n_2)$.

For logical convenience, assume to be

$$0 < L_{(1)} < \cdots < L_{(\xi)} < \cdots < L_{(s)} < \infty \tag{3.2}$$

for the ordered relation given in (2.1). It can be shown that $\hat{L}_{n_1,n_2}(x)$ converges to $L(x)$ in probability for any fixed x which can take the points $x_1, \cdots, x_s$, for $\hat{f}_{n_i}(x)$ converges to $f_i(x)$ with probability one and $f_i(x) > 0$ by (3.2), $i = 1, 2$. Thus, we can find out a number $n_0 = n_0(\varepsilon)$ such that

$$|\hat{L}_{(y_\mu)}(n_1, n_2) - L_{(y_\mu)}| \leqq \varepsilon \qquad \text{for } n_i \geqq n_0, \ i = 1, 2; \ y_\mu = 1, \cdots, s , \tag{3.3}$$

in the above-mentioned sense, when we take $\varepsilon > 0$ as

$$\varepsilon = \frac{1}{3} \min_{\xi = 2, \cdots, s} (L_{(\xi)} - L_{(\xi - 1)}) . \tag{3.4}$$

From (3.3), we have

$$L_{(y_\mu)} - L_{(y_\nu)} - 2\varepsilon \leqq \hat{L}_{(y_\mu)}(n_1, n_2) - \hat{L}_{(y_\nu)}(n_1, n_2) \leqq L_{(y_\mu)} - L_{(y_\nu)} + 2\varepsilon . \tag{3.5}$$

Then, (3.5) means that if $L_{(y_\mu)} > L_{(y_\nu)}$, $\hat{L}_{(y_\mu)}(n_1, n_2) > \hat{L}_{(y_\nu)}(n_1, n_2)$ and if $\hat{L}_{(y_\mu)}(n_1, n_2) > \hat{L}_{(y_\nu)}(n_1, n_2)$, $L_{(y_\mu)} > L_{(y_\nu)}$. Thus, we can obtain the following:

For sufficiently large n_1 and n_2, the rank order obtained from $\hat{L}_{(\xi)}(n_1, n_2)$ based on $\hat{L}_{n_1,n_2}(x)$ tends to coincide with the rank order obtained from (3.2) based on $L(x)$.

Now, let $(y_1^{(1)}, \cdots, y_{n_1}^{(1)})$ and $(y_1^{(2)}, \cdots, y_{n_2}^{(2)})$ be $(x_1^{(1)}, \cdots, x_{n_1}^{(1)})$ and $(x_1^{(2)}, \cdots, x_{n_2}^{(2)})$ transformed into the rank order based on $\hat{L}_{(\xi)}(n_1, n_2)$. Suppose that a new random sample of the size n is obtained from π. Every observation from π is transformed into a rank as mentioned above from a vector. Let $(y_1, \cdots, y_n)$ be the new sample transformed into the rank order.

Define $Z(y_\mu^{(i)}, y_\nu)$, $z(y_\mu^{(i)}, y_\nu)$ and $\hat{p}_i$ by

$$\begin{aligned} Z(y_\mu^{(i)}, y_\nu) &= \begin{cases} 1 & \text{if } y_\mu^{(i)} \leqq y_\nu , \\ 0 & \text{if } y_\mu^{(i)} > y_\nu , \end{cases} \\ z(y_\mu^{(i)}, y_\nu) &= \begin{cases} 1 & \text{if } y_\mu^{(i)} = y_\nu , \\ 0 & \text{if } y_\mu^{(i)} \neq y_\nu , \ \mu = 1, \cdots, n_i ; \ \nu = 1, \cdots, n , \end{cases} \end{aligned} \tag{3.6}$$

and

$$\hat{\hat{p}}_i=\frac{1}{nn_i}\sum_{\nu=1}^{n}\sum_{\mu=1}^{n_i}\left\{Z(y_\mu^{(i)}, y_\nu)-\frac{1}{2}z(y_\mu^{(i)}, y_\nu)\right\}, \qquad i=1, 2. \tag{3.7}$$

Let $\hat{g}_{n_i}(y)$ and $\tilde{g}_i(y)$ be $\hat{f}_{n_i}(x)$ and $f_i(x)$ arranged in the rank order based on $\hat{L}_{(\xi)}(n_1, n_2)$, and let $\hat{G}_{n_i}(y)$ and $\tilde{G}_i(y)$ be defined by $\hat{G}_{n_i}(y)=\sum_{\xi\leq y}\hat{g}_{n_i}(\xi)$ and $\tilde{G}_i(y)=\sum_{\xi\leq y}\tilde{g}_i(\xi)$, $i=1, 2$, respectively. Define $\tilde{\varDelta}_{ij}$ by

$$\tilde{\varDelta}_{ij}=\frac{1}{2}\sum_{\xi=1}^{s}\{\hat{G}_{n_i}(\xi)\tilde{g}_j(\xi)-\tilde{G}_j(\xi)\hat{g}_{n_i}(\xi)\}, \qquad i, j=1, 2. \tag{3.8}$$

Then, the conditional expectations of $\hat{\hat{p}}_i$'s, $i=1, 2$, with respect to the joint distribution of $y_1, \cdots, y_n$ which have the common probability function $\tilde{g}(\xi)=w_1\tilde{g}_1(\xi)+w_2\tilde{g}_2(\xi)$ are

$$\mathcal{E}_{\tilde{g}^n}\hat{\hat{p}}_1=\frac{1}{2}+\tilde{\varDelta}_{11}+w_2(\tilde{\varDelta}_{12}-\tilde{\varDelta}_{11}),$$

and

$$\mathcal{E}_{\tilde{g}^n}\hat{\hat{p}}_2=\frac{1}{2}+\tilde{\varDelta}_{22}+w_1(\tilde{\varDelta}_{21}-\tilde{\varDelta}_{22}),$$

under the condition that observed values of $(y_1^{(1)}, \cdots, y_{n_1}^{(1)})$ and $(y_1^{(2)}, \cdots, y_{n_2}^{(2)})$ have been obtained. Thus, we have

$$p\lim_{\substack{n_1\to\infty\\ n_2\to\infty}}\mathcal{E}_{\tilde{g}^n}\left(\hat{\hat{p}}_1-\frac{1}{2}\right)=w_2\varDelta_{12},$$

$$p\lim_{\substack{n_1\to\infty\\ n_2\to\infty}}\mathcal{E}_{\tilde{g}^n}\left(\frac{1}{2}-\hat{\hat{p}}_2\right)=w_1\varDelta_{12},$$

and

$$p\lim_{\substack{n_1\to\infty\\ n_2\to\infty}}\mathcal{E}_{\tilde{g}^n}(\hat{\hat{p}}_1-\hat{\hat{p}}_2)=\varDelta_{12},$$

since $p\lim_{\substack{n_1\to\infty\\ n_2\to\infty}}\tilde{\varDelta}_{ii}=0$, $i=1, 2$, $p\lim_{\substack{n_1\to\infty\\ n_2\to\infty}}\tilde{\varDelta}_{12}=\varDelta_{12}$ and $p\lim_{\substack{n_1\to\infty\\ n_2\to\infty}}\tilde{\varDelta}_{21}=-\varDelta_{12}$.

Therefore, *a decision rule as given in* (*) *with* $\hat{p}_1$ *and* $\hat{p}_2$ *replaced by* $\hat{\hat{p}}_1$ *and* $\hat{\hat{p}}_2$ *is adequate for our purpose.*

4. An approximation in the case of dichotomous response patterns

The classification procedure based on response patterns of individuals to m dichotomous items is considered for the case that a population π is composed of two mutually exclusive sub-groups π_1 and π_2, from the viewpoint of empirical Bayes approach.

Let $x=(e_1,\cdots,e_m)$ denote the total response to the given battery of items, where $e_k=1$ if the response on the kth item is "positive" and $e_k=0$ if otherwise, $k=1,\cdots,m$. Then, s in the preceding sections is $s=2^m$ in this case. Let $f_1(x)$ and $f_2(x)$ be also the probability functions of x in π_1 and π_2, respectively. In this case, the representation of $f_i(x)$ given by Bahadur [1], is as follows:

$$f_i(x)=\left(\prod_{k=1}^{m}\alpha_k^{(i)e_k}(1-\alpha_k^{(i)})^{1-e_k}\right)\cdot\varphi_i(x)\ ,\qquad i=1,2\ , \tag{4.1}$$

and

$$\varphi_i(x)=1+\sum_{k_1<k_2}r_{k_1k_2}^{(i)}z_{k_1}^{(i)}z_{k_2}^{(i)}+\cdots+r_{12\cdots m}^{(i)}z_1^{(i)}z_2^{(i)}\cdots z_m^{(i)}\ ,$$

where $\alpha_k^{(i)}=\mathrm{P}(e_k=1\,|\,\pi_i)$, $z_k^{(i)}=(e_k-\alpha_k^{(i)})/\sqrt{\alpha_k^{(i)}(1-\alpha_k^{(i)})}$ and $\mathcal{E}_{f_i}z_{k_1}^{(i)}\cdots z_{k_l}^{(i)}=r_{k_1\cdots k_l}^{(i)}$.

Bahadur [1], pointed out that the optimum solution based on $L(x)$ requires knowledge of the probability distribution of response patterns in each group, but this is a strong requirement if m is large, since both $f_1(x)$ and $f_2(x)$ are distributions with 2^m-1 parameters. Then, he has given certain approximations to $l(x)=\log L(x)$ and to error curve attainable with $l(x)$.

But, in actual applications of classification problems, it may be rather unusual that $f_1(x)$ and $f_2(x)$ are completely known. In such cases, if respective past observations from π_1 and π_2 are available, the procedure as discussed in Section 3 may be applicable. However, the procedure based on $\hat{L}_{n_1,n_2}(x)$ may need fairly large sample sizes n_1 and n_2 for obtaining a stable result if m is large. Then, we shall also use the approximation for $l(x)$ proposed by Bahadur to transform x into a rank order.

For $f_i(x)$'s in (4.1), $l(x)$ is

$$\begin{aligned} l(x)&=\log L(x)=\log f_2(x)-\log f_1(x)\\ &=\sum_{k=1}^{m}(A_k+B_ke_k)+(\log\varphi_2(x)-\log\varphi_1(x))\ , \end{aligned} \tag{4.2}$$

where A_k and B_k are $A_k=\log((1-\alpha_k^{(2)})/(1-\alpha_k^{(1)}))$ and $B_k=[(\alpha_k^{(2)}/(1-\alpha_k^{(2)}))\cdot((1-\alpha_k^{(1)})/\alpha_k^{(1)})]$. In this case, the simplest approximation for $l(x)$ is to replace $\log\varphi_i$ by φ_i-1, $i=1,2$, giving

$$\begin{aligned} \tilde{l}(x)=\sum_{k=1}^{m}(A_k+B_ke_k)&+\sum_{k_1<k_2}(r_{k_1k_2}^{(2)}z_{k_1}^{(2)}z_{k_2}^{(2)}-r_{k_1k_2}^{(1)}z_{k_1}^{(1)}z_{k_2}^{(1)})+\cdots\\ &+(r_{1\cdots m}^{(2)}z_1^{(2)}\cdots z_m^{(2)}-r_{1\cdots m}^{(1)}z_1^{(1)}\cdots z_m^{(1)})\ . \end{aligned} \tag{4.3}$$

Consider the case that observed response vectors $x_1^{(i)},\cdots,x_{n_i}^{(i)}$ regarded as surely coming from π_i, $i=1, 2$, are available, as $f_i(x)$ is unknown. The parameters $\alpha_k^{(i)}$, $r_{k_1 k_2}^{(i)}, \cdots$, and $r_{12\cdots m}^{(i)}$ included in $\tilde{l}(x)$ can be estimated from $(x_1^{(i)}, \cdots, x_{n_i}^{(i)})$. An estimate $\hat{l}_{n_1, n_2}(x)$ of $\tilde{l}(x)$ is obtained by means of replacing the parameters included in $\tilde{l}(x)$ by their estimates. Consider the rank order based on $\hat{l}_{n_1, n_2}(x)$, and denote the ξth of $\hat{l}_{n_1, n_2}(x)$ arranged in ascending order by $\hat{l}_{(\xi)}(n_1, n_2)$. Putting to use the same notations as in the Section 3, $(x_1^{(i)}, \cdots, x_{n_i}^{(i)})$ is transformed to the rank order $(y_1^{(i)}, \cdots, y_{n_i}^{(i)})$ based on $\hat{l}_{(\xi)}(n_1, n_2)$.

Suppose that a new random sample $(x_1, \cdots, x_n)$ is obtained from π in order to classify each of individuals included in this sample based on their response patterns. Let $(y_1, \cdots, y_n)$ be also the transformed sample to above-mentioned rank order from $(x_1, \cdots, x_n)$.

Then, the statistics of the form given by (3.7) can be obtained from $(y_1^{(1)}, \cdots, y_{n_1}^{(1)})$, $(y_1^{(2)}, \cdots, y_{n_2}^{(2)})$ and $(y_1, \cdots, y_n)$, and the procedure in order to make the decision rule can be carried out in the same fashion as in the Section 3.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Bahadur, R. R. (1961). On classification based on responses to n dichotomous items, *Studies in Item Analysis and Prediction*, Stanford University Press, 169-176.

[2] Hudimoto, H. (1976). On the empirical Bayes approach to classification problems, *Essays in Probability and Statistics*, 697-707.

[3] Robbins, H. (1964). The empirical Bayes approach to statistical decision problems, *Ann. Math. Statist.*, **35**, 1-20.

[4] Van Ryzin, J. and Susarla, V. (1977). On the empirical Bayes approach to multiple decision problems, *Ann. Statist.*, **5**, 172-181.