ON A DISTRIBUTION-FREE TWO-WAY CLASSIFICATION

HIROSI HUDIMOTO

1. Introduction

Suppose that $X^{(1)}$ and $X^{(2)}$ are random variables with continuous distribution functions $F_1(x^{(1)})$ and $F_2(x^{(2)})$, respectively. When $F_2(x^{(2)})$ is completely known and $F_1(x^{(1)})$ can be estimated from a random sample, Z. W. Birnbaum [3] proposed the statistic

$$\hat{p}=\int_{-\infty}^{\infty}\hat{F}_1(t)dF_2(t) \tag{1.1}$$

as an unbiased estimator of

$$p=P\{X^{(1)}<X^{(2)}\}=\int_{-\infty}^{\infty}F_1(t)dF_2(t)\,, \tag{1.2}$$

where $\hat{F}_1(t)$ is the empirical distribution function of $F_1(t)$. The statistic proposed by H. B. Mann and D. R. Whitney [6] is applicable as an unbiased estimator of p in the case when both $F_1(x^{(1)})$ and $F_2(x^{(2)})$ are unknown and can be estimated from samples. In this paper, we shall consider the use of such statistics for classifying a sample $(X_1, \cdots, X_n)$ into either of two populations $\prod_1$ and $\prod_2$ characterized by $F_1(x^{(1)})$ and $F_2(x^{(2)})$, respectively, and give an evaluation of their success rates. In section 2, we treat the case where $F_1(x^{(1)})$ and $F_2(x^{(2)})$ are known and in section 3, the case where both are unknown.

A fairly general consideration for classifying an individual into one of two or more populations has been presented by T. W. Anderson [1], [2], and C. R. Rao [7]. The particular case of two p-variate normal populations with equal covariance matrices has been fully discussed by A. Wald [10], T. W. Anderson [2], C. R. Rao [8], R. Sitgreaves [9] and the others. However, approaches in these papers cannot be applied to the non-parametric case where distributions are not specified but must be estimated from samples. Our method will be useful for such a case. Although it is assumed that $F_1(t)>F_2(t)$ for all t in the following sections, the method is applicable so far as it can be expected that $\Delta=\int_{-\infty}^{\infty}[F_1(t)-$

$F_2(t)]d\frac{F_1(t)+F_2(t)}{2} \neq 0$ and may be favorable in cases where Δ is fairly different from zero.

2. The case of known distribution functions

When $(X_1, \cdots, X_n)$ is a random sample of size n which has come from either of two populations $\prod_1$ and $\prod_2$, we wish to classify it correctly. Suppose $\prod_1$ and $\prod_2$ are characterized by univariate distribution functions $F_1(x^{(1)})$ and $F_2(x^{(2)})$, and $X^{(1)}$ and $X^{(2)}$ stand for their random variables, respectively. Then, it will be desirable for our procedure to have higher probability of correct classification.

As a preliminary for the next section, we shall first deal with the case where the distribution functions $F_1(x^{(1)})$ and $F_2(x^{(2)})$ are completely known, and assume that $F_1(x^{(1)})$ and $F_2(x^{(2)})$ are continuous and $F_1(t) > F_2(t)$ for all real t, i.e. $X^{(2)}$ is stochastically larger than $X^{(1)}$, for convenience. Consider statistics

$$\hat{p}_1 = \frac{1}{n}\sum_{k=1}^{n} F_1(X_k)\,, \tag{2.1}$$

and

$$\hat{p}_2 = 1 - \frac{1}{n}\sum_{k=1}^{n} F_2(X_k)\,. \tag{2.2}$$

Their expectations are

$$\mathscr{E}(\hat{p}_1 \mid \prod_1) = \int_{-\infty}^{\infty} F_1(t)dF_1(t) = \frac{1}{2}\,, \tag{2.3}$$

and

$$\begin{aligned}\mathscr{E}(\hat{p}_2 \mid \prod_1) &= 1 - \int_{-\infty}^{\infty} F_2(t)dF_1(t) \\ &= \int_{-\infty}^{\infty} F_1(t)dF_2(t) = p\,,\end{aligned} \tag{2.4}$$

where $\mathscr{E}(\hat{p}_u \mid \prod_v)$ denotes the conditional expectation of $\hat{p}_u$ under the condition that $(X_1, \cdots, X_n)$ has come from $\prod_v$ $(u=1, 2$ and $v=1, 2)$. Similarly, we have

$$\mathscr{E}(\hat{p}_1 \mid \prod_2) = \int_{-\infty}^{\infty} F_1(t)dF_2(t) = p\,, \tag{2.5}$$

and

$$\mathscr{E}(\hat{p}_2 \mid \prod_2) = 1 - \int_{-\infty}^{\infty} F_2(t) dF_2(t) = \frac{1}{2} . \tag{2.6}$$

The statistic of this kind was proposed by Z. W. Birnbaum [3]. p must be larger than 1/2 because of the assumption that $X^{(2)}$ is stochastically larger than $X^{(1)}$. Thus, we shall take the following as our procedure.

Decision rule 1. If the continuous distribution functions $F_1(x^{(1)})$ and $F_2(x^{(2)})$ are completely known and $F_1(t) > F_2(t)$ for all real t, classify $(X_1, \cdots, X_n)$ into $\prod_1$ or $\prod_2$ according to $\hat{p}_1 < \hat{p}_2$ or $\hat{p}_1 > \hat{p}_2$, respectively, where $\hat{p}_1$ and $\hat{p}_2$ are given by (2.1) and (2.2).

Let $H(x) = \frac{1}{2}\{F_1(X) + F_2(X)\}$. Then, $\hat{p}_1 \gtreqless \hat{p}_2$ is equivalent to $\frac{1}{n}\sum_{k=1}^{n} H(X_k) \gtreqless \frac{1}{2}$. The conditional expectations of $\frac{1}{n}\sum_{k=1}^{n} H(X_k)$ are

$$\mathscr{E}\left(\frac{1}{n}\sum_{k=1}^{n} H(X_k) \mid \prod_1\right) = \frac{1}{2}\left\{1 - \left(p - \frac{1}{2}\right)\right\} = \frac{1}{2} - \frac{\Delta}{2} \tag{2.7}$$

and

$$\mathscr{E}\left(\frac{1}{n}\sum_{k=1}^{n} H(X_k) \mid \prod_2\right) = \frac{1}{2}\left\{1 + \left(p - \frac{1}{2}\right)\right\} = \frac{1}{2} + \frac{\Delta}{2} , \tag{2.8}$$

where $\Delta = p - \frac{1}{2}$. Thus, if $\frac{1}{n}\sum_{k=1}^{n}\left[H(X_k) - \mathscr{E}\{H(X_k) \mid \prod_1\}\right]$ is smaller than $\frac{\Delta}{2}$, $\frac{1}{n}\sum_{k=1}^{n} H(X_k)$ will be smaller than $\frac{1}{2}$ when $(X_1, \cdots, X_n)$ comes from $\prod_1$, and if $\frac{1}{n}\sum_{k=1}^{n}[\mathscr{E}\{H(X_k) \mid \prod_2\} - H(X_k)]$ is smaller than $\frac{\Delta}{2}$, $\frac{1}{n}\sum_{k=1}^{n} H(X_k)$ will be larger than $\frac{1}{2}$ when $(X_1, \cdots, X_n)$ comes from $\prod_2$.

We shall use above relations to evaluate the success rate of the rule 1. W. Hoeffding [5] has given the certain probability inequalities for sums of bounded random variables. Since $H(X_1), \cdots, H(X_n)$ are mutually independent and $0 \leqq H(X_k) \leqq 1$ $(k = 1, \cdots, n)$, theorem 1 of his paper [5] gives us

$$P\left\{\frac{1}{n}\sum_{k=1}^{n} H(X_k) < \frac{1}{2} \mid \prod_1\right\} \geqq 1 - e^{-2n(\Delta/2)^2} \tag{2.9}$$

and

$$P\left\{\frac{1}{n}\sum_{k=1}^{n} H(X_k) > \frac{1}{2} \mid \prod_2\right\} \geqq 1 - e^{-2n(\Delta/2)^2} , \tag{2.10}$$

where $P\{\cdots \mid \prod_v\}$ denotes the conditional probability under the condition that each $X_1, \cdots, X_n$ have come from $\prod_v$ $(v=1, 2)$. Thus, we obtain the following:

The success rate of the rule 1 is greater than or equal to $1-\exp\left(-\frac{n\Delta^2}{2}\right)$,

where $\Delta=\int_{-\infty}^{\infty} F_1(t)dF_2(t)-\frac{1}{2}$ *and* n *is the size of sample to be classified.*

It will be found that this result is useful in the next section.

3. The case of unknown distribution functions

In this section, we shall consider the case where the distribution functions $F_1(x^{(1)})$ and $F_2(x^{(2)})$ are not known, but random samples of sizes n_1 and n_2 obtained from these populations are available. When $(X_1^{(1)}, \cdots, X_{n_1}^{(1)})$ and $(X_1^{(2)}, \cdots, X_{n_2}^{(2)})$ are the random samples from $\prod_1$ and $\prod_2$, respectively, and $(X_1, \cdots, X_n)$ is a random sample to be classified into either of them, we shall use the statistics $\hat{\hat{p}}_1$ and $\hat{\hat{p}}_2$ instead of $\hat{p}_1$ and $\hat{p}_2$:

$$\hat{\hat{p}}_1=\frac{U_{n_1 n}}{n_1 n} \quad \text{and} \quad \hat{\hat{p}}_2=\frac{U_{n n_2}}{n n_2}, \tag{3.1}$$

where $U_{n_1 n}$ is the number of pairs $(X_i^{(1)}, X_k)$ such that $X_i^{(1)}<X_k$, and $U_{n n_2}$ is the number of pairs $(X_k, X_j^{(2)})$ such that $X_k<X_j^{(2)}$, $i=1, \cdots, n_1$, $j=1, \cdots, n_2$ and $k=1, \cdots, n$, respectively. Such a statistic U has been proposed by Mann and Whitney [6] for the two-sample problem. In this case, we have also

$$\mathscr{E}(\hat{\hat{p}}_1 \mid \prod_1)=\frac{1}{2} \quad \text{and} \quad \mathscr{E}(\hat{\hat{p}}_2 \mid \prod_1)=p \tag{3.2}$$

and

$$\mathscr{E}(\hat{\hat{p}}_1 \mid \prod_2)=p \quad \text{and} \quad \mathscr{E}(\hat{\hat{p}}_2 \mid \prod_2)=\frac{1}{2}. \tag{3.3}$$

Thus, a procedure analogous to the rule 1 in the previous section can be used.

Decision rule 2. If it can be assumed that $F_1(t)>F_2(t)$ for all real t, classify $(X_1, \cdots, X_n)$ into $\prod_1$ or $\prod_2$ according to $\hat{\hat{p}}_1<\hat{\hat{p}}_2$ or $\hat{\hat{p}}_1>\hat{\hat{p}}_2$, where $\hat{\hat{p}}_1$ and $\hat{\hat{p}}_2$ are given by (3.1).

We shall assume here that $n \leqq n_1$ and $n \leqq n_2$. Let $g(X_{i(k)}^{(1)}, X_k)$ and $g(X_{j(k)}^{(2)}, X_k)$ be two sets of random variables defined as

(3.4)
$$g(X^{(1)}_{i(k)}, X_k)=\begin{cases}\frac{1}{2} & \text{if } X^{(1)}_{i(k)}<X_k\\ 0 & \text{if } X^{(1)}_{i(k)}\geqq X_k\end{cases}$$

$$g(X^{(2)}_{j(k)}, X_k)=\begin{cases}\frac{1}{2} & \text{if } X^{(2)}_{j(k)}<X_k\\ 0 & \text{if } X^{(2)}_{j(k)}\geqq X_k\end{cases}$$

$$(k=1, \cdots, n;\ i(k)=1, \cdots, n_1 \text{ and } j(k)=1, \cdots, n_2)$$

and let

(3.5)
$$V(X^{(1)}_{i(1)}, \cdots, X^{(1)}_{i(n)};\ X^{(2)}_{j(1)}, \cdots, X^{(2)}_{j(n)})$$
$$=\frac{1}{n}\sum_{k=1}^{n}\{g(X^{(1)}_{i(k)}, X_k)+g(X^{(2)}_{j(k)}, X_k)\}$$

$$(k=1, \cdots, n,\ i(1)\neq \cdots \neq i(n) \text{ and } j(1)\neq \cdots \neq j(n)).$$

Then, we have also

(3.6)
$$\mathscr{E}\left\{V(X^{(1)}_{i(1)}, \cdots, X^{(1)}_{i(n)}, X^{(2)}_{j(1)}, \cdots, X^{(2)}_{j(n)}) \mid \prod_1\right\}=\frac{1}{2}-\frac{\Delta}{2}$$

and

(3.7)
$$\mathscr{E}\left\{V(X_{i(1)}, \cdots, X_{i(n)}, X_{j(1)}, \cdots, X_{j(n)}) \mid \prod_2\right\}=\frac{1}{2}+\frac{\Delta}{2}.$$

Consider

(3.8)
$$V=\frac{(n!)^2}{n_1!\, n_2!}\sum_{i(1)\neq \cdots \neq i(n)}^{{}_{n_1}P_n}\ \sum_{j(1)\neq \cdots \neq j(n)}^{{}_{n_2}P_n} V(X^{(1)}_{i(1)}, \cdots, X^{(1)}_{i(n)}, X^{(2)}_{j(1)}, \cdots, X^{(2)}_{j(n)}),$$

where sums $\sum_{i(1)\neq \cdots \neq i(n)}^{{}_{n_1}P_n}$ and $\sum_{j(1)\neq \cdots \neq j(n)}^{{}_{n_2}P_n}$ are taken over all n-tuples $\{i(1), \cdots, i(n)\}$ of distinct positive integers $\leqq n_1$ and all n-tuples $\{j(1), \cdots, j(n)\}$ of distinct positive integers $\leqq n_2$.

Then V will become a statistic corresponding to $\frac{1}{n}\sum_{k=1}^{n} H(X_k)$ in the previous section and $\hat{\hat{p}}_1 \lessgtr \hat{\hat{p}}_2$ is equivalent to $V \lessgtr \frac{1}{2}$ except possibly for $X_k=X^{(1)}_{i(k)}$ and $X_k=X^{(2)}_{j(k)}$ $(k=1, \cdots, n,\ i(k)=1, \cdots, n_1$ and $j(k)=1, \cdots, n_2)$. Thus, the situation is similar as before and the result of section 5 in Hoeffding's paper [5] is applicable in this case, too.

If we rewrite (3.8) as

$$V=\alpha_1 V_1+\cdots+\alpha_N V_N,$$

where $\alpha_1=\cdots=\alpha_N=\dfrac{(n!)^2}{n_1!\,n_2!}$ and $N=\dfrac{n_1!\,n_2!}{(n!)^2}$,

we have

$$P\left\{V\geqq\frac{1}{2}\,\middle|\,\Pi_1\right\}=P\left\{V-\mathscr{E}_{\Pi_1}V\geqq\frac{\Delta}{2}\,\middle|\,\Pi_1\right\} \tag{3.9}$$

$$\leqq\mathscr{E}\left[\exp\left\{h\left(V-\mathscr{E}_{\Pi_1}V-\frac{\Delta}{2}\right)\right\}\,\middle|\,\Pi_1\right]$$

$$=\mathscr{E}\left[\exp\left\{h\sum_{i=1}^{N}\alpha_i\left(V_i-\mathscr{E}_{\Pi_1}V_i-\frac{\Delta}{2}\right)\right\}\,\middle|\,\Pi_1\right]$$

for an arbitrary positive constant h. Since the exponential function $\exp\{\ \}$ is convex,

$$P\left\{V\geqq\frac{1}{2}\,\middle|\,\Pi_1\right\}\leqq\sum_{i=1}^{N}\alpha_i\,\mathscr{E}\left[\exp\left\{h\left(V_i-\mathscr{E}_{\Pi_1}V_i-\frac{\Delta}{2}\right)\right\}\,\middle|\,\Pi_1\right]. \tag{3.10}$$

Since $g(X^{(1)}_{i(1)}, X_1)+g(X^{(2)}_{j(1)}, X_1), \cdots, g(X^{(1)}_{i(n)}, X_n)+g(X^{(2)}_{j(n)}, X_n)$ are mutually independent and $0\leqq g(X^{(1)}_{i(k)}, X_k)+g(X^{(2)}_{j(k)}, X_k)\leqq 1$, it can be seen from [5] that the minimum of $\mathscr{E}\left[\exp\left\{h\left(V_i-\mathscr{E}_{\Pi_1}V_i-\dfrac{\Delta}{2}\right)\right\}\,\middle|\,\Pi_1\right]$ with respect to $h(>0)$ is not larger than $\exp\left(-\dfrac{n\Delta^2}{2}\right)$. Thus,

$$P\left\{V<\frac{1}{2}\,\middle|\,\Pi_1\right\}\geqq 1-e^{-n\Delta^2/2},$$

and similarly

$$P\left\{V>\frac{1}{2}\,\middle|\,\Pi_2\right\}\geqq 1-e^{-n\Delta^2/2}.$$

Summarizing the above, we have the following:

The success rate of the rule 2 is greater than or equal to $1-\exp\left(-\dfrac{n\Delta^2}{2}\right)$, *if it can be assumed that* $F_1(t)>F_2(t)$ *for all real* t, *and* $\Delta=p-\dfrac{1}{2}$ *is given.*

Although the above assumptions are not always fulfilled in actual

applications, Δ can be estimated from samples $(X_1^{(1)}, \cdots, X_{n_1}^{(1)})$ and $(X_1^{(2)}, \cdots, X_{n_2}^{(2)})$, and we may remove the assumption of stochastic orderings, if Δ is fairly larger than zero.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley and Sons, New York, 1958.

[2] T. W. Anderson, "Classification by multivariate analysis," *Psychometrika*, 16 (1951), 31-50.

[3] Z. W. Birnbaum, "On a use of Mann-Whitney statistic," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Univ. of California Press, I (1956) 13-17.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, 7 (1936), 179-188.

[5] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Amer. Statist. Assoc.*, 58 (1963), 13-30.

[6] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, 18 (1947), 50-60.

[7] C. R. Rao, *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons, New York, 1952.

[8] C. R. Rao, "On a general theory of discrimination when the information on alternative hypotheses is based on samples," *Ann. Math. Statist.*, 25 (1954) 651-669.

[9] R. Sitgreaves, "On the distribution of two random matrices used in classification procedures," *Ann. Math. Statist.*, 23 (1952), 263-270.

[10] A. Wald, "On a statistical problem arising in the classification of an individual into one of two groups," *Ann. Math. Statist.*, 15 (1944) 145-163.