

# A Note on the Classification of Observation Data

By Hirojiro AOYAMA

(Received April 30, 1950)

1. Introduction. Recently R. v. Mises treated the problem about the classification of data as follows. 1) Let an aggregate be given, upon each element of which a trial is done and consequently one definite real number  $x$  is obtained. And let each element of this aggregate belong to one class of  $n$  classes which are characterized by  $n$  density functions  $f_1(x), f_2(x), \dots, f_n(x)$ . Then having a value  $x$  as the result of observation, there arises the problem "To which class will this  $x$  belong"? As an application of this problem Hayashi recently published an interesting article about the parole prediction. 2) In this note we shall show that we can get similar results as Mises did from some different point of view.

2. Case when there are two classes. We consider now of the certain infinite population where two classes exist, being mixed, which are characterized respectively by the two unimodal density functions  $f_1(x)$  and  $f_2(x)$  as follows

$$\int_{-\infty}^{\infty} f_1(x) dx = 1, \quad \int_{-\infty}^{\infty} f_2(x) dx = 1, \quad (1)$$

$$\int_{-\infty}^{\infty} \{k_1 f_1(x) + k_2 f_2(x)\} dx = 1, \quad (2)$$

$k_1$  and  $k_2$  being two constants which are proportional rate in the population satisfying  $k_1 + k_2 = 1$  and are not necessarily known previously.

When we have an observation data  $x$ , we may select the  $x_0$  so as to decide that  $x$  should belong to 1st class when  $x > x_0$  and to 2nd class when  $x < x_0$ . Then arises the problem at what value  $x_0$  the reliability will become maximum. Representing the reliability  $P$  by the degree of true judging, we may put

$$P = \int_{-\infty}^{x_0} k_2 f_2(x) dx + \int_{x_0}^{\infty} k_1 f_1(x) dx \quad (3)$$

We may first treat the case when  $f_1(x)$  and  $f_2(x)$  are both normal density function whose means are  $m_1$  and  $m_2$ , and whose standard deviations are  $\sigma_1$  and  $\sigma_2$  respectively and when  $k_1$  and  $k_2$  are known previously. For the maximum reliability we get from  $\partial P / \partial x_0 = 0$

$$\frac{k_1}{\sigma_1} e^{-\frac{(x_0 - m_1)^2}{2\sigma_1^2}} = \frac{k_2}{\sigma_2} e^{-\frac{(x_0 - m_2)^2}{2\sigma_2^2}},$$

that is

$$(\sigma_2^2 - \sigma_1^2)x_1^2 - 2x_1(m_1\sigma_2^2 - m_2\sigma_1^2) + \left(m_1^2\sigma_2^2 - m_2^2\sigma_1^2 - 2\sigma_1^2\sigma_2^2 \log \frac{k_1\sigma_2}{k_2\sigma_1}\right) = 0. \quad (4)$$

When  $m_1 > m_2$ , we may take  $x_0 > m_2$  of two roots. When  $\sigma_1 = \sigma_2 = \sigma$  we get

$$x_0 = \frac{1}{2} \left\{ (m_1 + m_2) - \frac{2\sigma^2}{m_1 - m_2} \log \frac{k_1}{k_2} \right\} \quad (5)$$

For every case the reliability is computed from (3). If  $k_1$  and  $k_2$  are not known previously, we decide  $x_0, k_1$  so as to make (3) maximum.

Then putting

$$P = \frac{k_2}{\sqrt{2\pi}\sigma_2} \int_{-\infty}^{x_0} e^{-\frac{(x-m_2)^2}{2\sigma_2^2}} dx + \frac{k_1}{\sqrt{2\pi}\sigma_1} \int_{x_0}^{\infty} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}} dx \quad (6)$$

where  $k_1 + k_2 = 1$ , we have from  $\partial P / \partial k_1 = 0$

$$\frac{1}{\sigma_1} \int_{x_0}^{\infty} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}} dx = \frac{1}{\sigma_2} \int_{-\infty}^{x_0} e^{-\frac{(x-m_2)^2}{2\sigma_2^2}} dx. \quad (7)$$

Then from  $\partial P / \partial x_0 = 0$  we have also

$$\frac{f_1(x_0)}{f_2(x_0)} = \frac{k_2}{k_1} \quad (8)$$

For the general density function we get instead of (7)

$$\int_{x_0}^{\infty} f_1(x) dx = \int_{-\infty}^{x_0} f_2(x) dx \quad (9)$$

3. Case when there are  $n$  classes. Similarly to the previous section we may take  $n$  classes which are characterized by the unimodal density functions  $f_1(x), f_2(x), \dots, f_n(x)$  respectively and which exist mixed each other in the population. Let the reliability be

$$P = k_1 \int_{-\infty}^{r_1} f_1(x) dx + k_2 \int_{r_1}^{r_2} f_2(x) dx + \dots + k_n \int_{r_{n-1}}^{\infty} f_n(x) dx \quad (10)$$

where

$$k_1 + k_2 + \dots + k_n = 1. \quad (11)$$

Making  $P$  maximum, we have similarly as before

$$\int_{-\infty}^{r_1} f_1(x) dx = \int_{r_1}^{r_2} f_2(x) dx = \dots = \int_{r_{n-1}}^{\infty} f_n(x) dx \quad (12)$$

These  $x_1, x_2, \dots, x_{n-1}$  are the points of division of  $n$  classes. And we have at each point of division

$$\frac{f_\nu(x)}{f_{\nu+1}(x)} = \frac{k_{\nu+1}}{k_\nu} \quad (\nu = 1, 2, \dots, n-1). \quad (13)$$

If we know the value  $k_1, k_2, \dots, k_n$  previously, we can compute directly the reliability from (10). In another case we first decide the  $x_\nu$  from (12), then decide  $k_\nu$  from (11) and (13).

4. General case. When  $n$  classes are characterized by the  $m$  variate unimodal density functions  $f_1(x_1, x_2, \dots, x_m), f_2(x_1, x_2, \dots, x_m), \dots, f_n(x_1, x_2, \dots, x_m)$  respectively, we put the reliability  $P$  as follows

$$P = k_1 \int_{R_1} f_1 dR_1 + k_2 \int_{R_2} f_2 dR_2 + \dots + k_n \int_{R_n} f_n dR_n \quad (14)$$

where  $k_1 + k_2 + \dots + k_n = 1$ , and  $R_\nu$  is a certain region in  $m$ -dimensional space, and  $R_1 + R_2 + \dots + R_n$  equal to the whole space. From  $\partial P / \partial k_\nu = 0$  we can deduce immediately

$$\int_{R_1} f_1 dR_1 = \int_{R_2} f_2 dR_2 = \dots = \int_{R_n} f_n dR_n \quad (15)$$

and if we put the increment  $\Delta P$  of  $P$  equal to zero for the infinitesimal increment  $\Delta R$  of some region  $R_\mu$  and the same decrement of adjacent region  $R_\nu$ , we have

$$\frac{f_\nu(x_1, x_2, \dots, x_m)}{f_\mu(x_1, x_2, \dots, x_m)} = \frac{k_\mu}{k_\nu}. \quad (16)$$

So we first decide the regions  $R_\nu$  from (15) and then compute  $P$  from (16) and  $k_1 + k_2 + \dots + k_n = 1$ .

### References :

- 1) R. v. Mises: On the classification of observation data into distinct groups, *Annals of Mathe. Statis.* vol. XVI, No. 1. 1945.
- 2) Chikio Hayashi: On an application of the statistical method in Parole Prediction. Institute of Case Work.

*Institute of Statistical Mathematics  
Sangenjaya Laboratory*