STATISTICAL INFERENCE APPLIED TO CLASSIFICATORY PROBLEMS

By C. RADHAKRISHNA RAO
Statistical Laboratory, Calcutta

PART II: THE PROBLEM OF SELECTING INDIVIDUALS FOR VARIOUS DUTIES IN A SPECIFIED RATIO

1. Introduction

In a discussion of the author's paper on the 'Utilization of multiple measurements in problems of biological classification' read before the Royal Statistical Society in 1948, Mr. Patrick Slater posed the following problem.

"Could Mr. Rao give a rule of procedure to be followed when out of an undistributed population with parameters $X_1, X_2, \ldots, X_m$ and $[M]$, n_a have to be selected for duty A, n_b for duty B etc., given the means for each duty? This supposes that the whole population must be distributed among the duties and the quota and the desired standards for each duty are laid down."

In reply the author has given a complete solution without proof. The object of this paper is to study this problem in its generality, supply the necessary proofs and suggest some practical methods of arriving at the desired classification. Although the main results are deducible from the general theory of decision functions as developed by Prof. Abraham Wald in his book on 'Statistical decision functions' and in his lectures at the Indian Statistical Institute, the computational procedure presents some difficulty. The two lemmas given in the next section are useful for this purpose. It is also shown that the solution derived by the method of maximum likelihood possesses some important properties.

2. Two lemmas

Consider an array of elements

$$\begin{array}{cccc} a_{11} & a_{21} & \ldots & a_{p1} \\ a_{12} & a_{22} & \ldots & a_{p2} \\ . & . & \ldots & . \\ a_{1n} & a_{2n} & \ldots & a_{pn} \end{array}$$

consisting of p columns and n rows. Let P denote the product and S the sum of n elements chosen one from each row such that the total number of elements coming from the first column is equal to a specified value n_1, from the second n_2 and so on from the p-th column n_p. Obviously $n \geq p$ if no n is zero and $n = n_1 + \ldots + n_p$.

Lemma 1 : *If the elements a_{ij} are not negative and if there exist quantities $\lambda_1, \lambda_2, \ldots, \lambda_p$ such that each element a_{ik} of n_i elements chosen from the i-th column satisfies the relationships*

$$\lambda_i a_{ik} \geqslant \lambda_j a_{jk} \quad \text{for } j = 1, 2, \ldots, p \qquad \ldots \quad (2.1.1)$$

and if similar reationships are satisfied for all $i = 1,2,\ldots, p$ with the same set of λ's, then for this choice of $n_1, n_2, \ldots, n_p$ elements the product P defined above is a maximum.

To prove this consider any other choice of n_i elements

$$a_{1r}, \quad a_{1s}, \ldots$$

from the i-th column. Remembering that the second subscript refers to the row number, the elements from the rth, sth,... rows occuring in the former selection may be represented by

$$a_{br}, \quad a_{cs}, \ldots$$

from which by definition it follows that the product

$$a_{br}\ a_{cs} \ldots$$

is not less than

$$\frac{\lambda_1}{\lambda_b}\ a_{1r}\ \frac{\lambda_1}{\lambda_c}\ a_{1s} \ldots$$

The λ's are positive by definition so that division by a λ does not change the inequality sign. Considering all the groups of elements in the second selection we find

$$\Pi\ a_{br}\ a_{cs} \ldots \geqslant \Pi_1 \frac{\lambda_1}{\lambda_b} \frac{\lambda_1}{\lambda_c} \ldots a_{1r} a_{1s} \ldots$$

$$\geqslant \Pi_1\ a_{1r}\ a_{1s} \ldots$$

since $$\Pi_1 \frac{\lambda_1}{\lambda_b} \frac{\lambda_1}{\lambda_c} \ \ldots \ = 1$$

all the λ's in the numerator cancelling with those in the denominator in the final product.

Corollary 1.1 : *If the object is to maximise the product without the restriction on the number of elements coming from each column then, obviously, the best procedure is to choose the biggest element from each row.*

Corollary 1.2. : *The product P will be a minimum if the inequality relationship (2.1.1) is reversed.*

Corollary 1.3 : *If $p=2$ the method described in lemma 1 reduces to evaluating the ratios*

$$\frac{a_{11}}{a_{21}}, \frac{a_{12}}{a_{22}}, \ldots, \frac{a_{1n}}{a_{2n}}$$

and arranging them in descending order of magnitude and choosing the elements in the numerator from the first n_1 ratios and the elements in the denominator from the second n_2 ratios.

Lemma 2 : *If there exist quantities $\mu_1, \mu_2, \ldots, \mu_p$ such that each element a_{ik} of the n_i elements chosen from the ith column satisfies the relationships*

$$a_{ik}+\mu_i \geqslant a_{jk}+\mu_j \text{ for all } j = 1,2,\ldots,p$$

and if similar relationships hold for all $i = 1,2,\ldots,p$ with the same set of μ's, then for this choice of $n_1,\ldots,n_p$ elements the sum S defined above is a maximum.

The result follows from lemma 1 by considering $exp\,(a_{ik})$ and maximising the product. The existence of μ's leads to the existence of positive quantities λ's used in lemma 1. The sum S is minimised when the reverse relationships hold good.

3. Problems of two groups

Let us consider two groups characterised by probability densities $f_1(x|\theta_1)$ and $f_2(x|\theta_2)$ where x stands for all the available set of measurements and θ for all the parameters. These may be simply denoted by $f_1(x)$ and $f_2(x)$.

Two types of problems arise. Firstly a sample of size n_1+n_2 may be drawn from a group with two populations mixed in the ratio $\pi_1 : \pi_2$ and a division of the sample in two groups of n_1 and n_2 individuals is required. Secondly two samples of sizes n_1 and n_2 drawn independently from the first and second groups may get mixed. In this case what is the best method of separating the individuals belonging to two different groups? The difference between the first and second problems is that in the latter every sample is known to consist of n_1 individuals from the first group and n_2 from the second while in the former no such information is available, the sample being drawn at random from a mixed population.

3.1. *Solution to the first problem*

Let $x_1,\ldots,x_n$ represent the measurements on n individuals. As observed earlier x_i will stand for all the available set of measurement on the ith individual. The probability of the set is

$$\prod_i [\pi_1 f_1(x_i)+\pi_2 f_2(x_i)]$$

Consider the following set of functions

$$\delta_i(x_1,\ldots,x_n),\ \delta'_i(x_1,\ldots,x_n);\ i = 1,\ldots,n$$

which can be represented simply as δ_i, δ'_i satisfying the conditions

$$\delta_i = 0 \text{ or } 1, \qquad \delta_i + \delta'_i = 1$$

and

$$\sum_1^n \delta_i = n_1, \qquad \sum_1^n \delta'_i = n_2$$

where n_1 and n_2 are the specified numbers. If the individual with measurments x_i is assigned to the first group when $\delta_i = 1$ and to the second when $\delta'_i = 1$ then the above set of functions constitute a decision rule for the problem (Wald, 1950). The problem is then to construct the above functions such that the expected risk associated with this decision rule is a minimum. To calculate the expected risk we need know as a datum of the problem the loss of assigning an individual of one group to another. If r_{12} represents the loss in assigning an individual of the first to the second and r_{21} in the other case then the quantity to be minimised is the expected value of

$$\sum_1^n \frac{r_{21}\delta_i \pi_2 f_2(x_i) + r_{12}\delta'_i \pi_1 f_1(x_i)}{\pi_1 f_1(x_i) + \pi_2 f_2(x_i)}$$

$$= \sum_1^n (\delta_i a_{i1} + \delta'_i a_{i2}) \qquad \ldots \quad (3.1.1)$$

where $a_{i1} = r_{21}\pi_2 f_2/(\pi_1 f_1 + \pi_2 f_2)$, $a_{i2} = r_{12}\pi_1 f_1/(\pi_1 f_1 + \pi_2 f_2)$

The expected loss will be a minimum if δ_i and δ'_i are chosen such that the expression (3.1.1) has the least value. The problem is the same as that treated in lemma 2 leading to the solution

$$\begin{aligned} \delta_i &= 1 \quad \text{if} \quad a_{i1} + \mu_1 \leqslant a_{i2} + \mu_2 \\ &= 0 \quad \text{if} \quad a_{i1} + \mu_1 \geqslant a_{i2} + \mu_2 \end{aligned}$$

where μ_1 and μ_2 are suitably chosen to satisfy the condition $\Sigma\delta_i = n_1$. Now $a_{i1} - a_{i2} \leqslant \mu_2 - \mu_1$ implies that

$$\frac{r_{21}\pi_2 f_2(x_i) - r_{12}\pi_1 f_1(x_i)}{\pi_1 f_1(x_i) + \pi_2 f_2(x_i)} \leqslant$$

or

$$\frac{f_1(x_i)}{f_2(x_i)} \geqslant \lambda$$

so that the decision rule reduces to the evaluation of the likelihood ratios $\lambda_r = f_1(x_r)/f_2(x_r)$ and assigning, all the individuals with highest n_1 values of of the ratios to the first group and the rest to the second. Or if the differences $s_i = \log f_1(x_i) - \log f_2(x_i)$ are first calculated then the highest n_1 values will form the basis for selection to the first group. Fortunately the decision rule is independent of the apriori probabilities and also the loss function.

Corresponding to every decision rule we can set up a density function of the observations by considering all individuals assigned to the first group as having been drawn at random from the first group and similarly for the second. By using lemma 1 we find that the best decision rule found above maximises the corresponding probability density. As shown below this forms the basis on which the solution of the second problem depends.

3.2. *Solution to the second problem*

In this problem the mixture is known to consist of n_1 individuals drawn from the first group and n_2 from the second. The observations $x_1, x_2, \ldots, x_n$ could have arisen in ${}^nC_{n_1}$ ways, any subset of n_1 observations belonging to the first group. The probability density of the observations is equal to the sum of the densities associated with ${}^nC_{n_1}$ ways of splitting the sample. If $x_a, x_b, \ldots$ and $x_p, x_q, \ldots$ represents a division into two groups of sizes n_1 and n_2 then the probability density of the observations can be written as

$$P(x) = \Sigma f_1(x_a) f_1(x_b) \ldots f_2(x_p) f_2(x_q) \ldots$$

where the sumation is over ${}^nC_{n_1}$ such terms. Corresponding to any one of ${}^nC_{n_1}$ possible decision rules the loss relative to the given set of observations is $1/P(x)$ times

$$\Sigma l(a,b,\ldots,p,q,\ldots) f_1(x_a) f_1(x_b) \ldots f_2(x_p) f_2(x_q) \ldots \qquad \ldots \quad (3.2.1)$$

where $l(a,b,\ldots,p,q,\ldots)$ is the loss incurred in adopting an assigned decision rule when in fact $x_a, x_b, \ldots$ come from the first group and $x_p, x_q \ldots$ from the second. The loss will generally be a function of the number of wrong classifications only. That decision rule for which the expression (3.2.1) is the least is the best in the sense that it minimises the expected loss. The solution depends on the evaluation of the expression (3.2.1) for each of the ${}^nC_{n_1}$ decision rules and this makes its application a little difficult in practice even when n is small.

As an alternative we may try to minimise the maximum loss incurred by following a decision rule. If we suppose that the loss function is proportional to the number of wrong classifications then the maximum loss occurs when all the individuals in the smaller group are wrongly classified. With this assumption it can be shown that the division of the sample corresponding to the maximum probability density supplies the best possible solution to the problem. This solution may be referred to as the maximum likelihood solution; we consider the ${}^nC_{n_1}$ ways of splitting the sample as associated with ${}^nC_{n_1}$ different hypotheses concerning the individuals in the sample and choose that hypothesis which has the maximum likelihood.

To prove the property referred above, consider any other decision rule leading to a division of the sample

$$x_{a_1} x_{a_2} \ldots, x_{b_1} x_{b_2} \ldots \text{ and } x_{c_1} x_{c_2} \ldots, x_{d_1} x_{d_2} \ldots$$

into two groups of sizes n_1 and n_2 and compare with the division

$$x_{a1}x_{a2}\ldots x_{r1}x_{r2}\ldots \text{ and } x_{b1}x_{b2}\ldots x_{s1}x_{s2}\ldots$$

associated with the maximum density. The measurements classified in the same way by the two decision rules are represented by $x_{a1}, x_{a2}\ldots$ for the first group and by $x_{s1}, x_{s2}, \ldots$ for the second group. By definition

$$\frac{f_1(x_{r1})}{f_2(x_{r1})} \geq \frac{f_1(x_{b1})}{f_2(x_{b1})} \qquad \ldots \quad (3.2.2)$$

and the same is true for the product of a number of ratios involving x_r and the product of the same number of ratios involving x_b.

Let $n_2 \leq n_1$ without loss of generality. In this case maximum loss occurs, by following the first decision rule when

$$x_{r1}x_{r2}, \ldots, x_{s1}x_{s2}\ldots \qquad \ldots \quad (3.2.3)$$

arise from the first group in which case a subset n_2 out of

$$x_{a1}x_{a2}\ldots, x_{b1}x_{b2}\ldots \qquad \ldots \quad (3.2.4)$$

arise from the second group. Let this subset be

$$x_{a1}x_{a2}, \ldots, x_{bp}x_{bq}\ldots \qquad \ldots \quad (3.2.5)$$

By replacing the subscript c by b we obtain the corresponding situation for the proposed maximum likelihood decision rule. The difference between the two probability densities associated with maximum errors for the two rules is, apart from a common multiplier, equal to

$$f_1(x_{rp})f_1(x_{r_1})\ldots f_2(x_{np})f_2(x_{bq})\ldots - f_1(x_{bp})f_1(x_{bq})\ldots f_2(x_{rp})f_2(x_{r_1})\ldots$$

which is not less than zero according to result (3.2.2). By considering all subsets of n_2 observations out of (3.2.4) we exhaust all possible ways in which the observations leading to a maximum error according to the first rule can arise. To each such case there is a corresponding division leading to the maximum error for the proposed rule. But this division leads to a smaller probability density. The total chance of maximum error relative to the given set of observations is thus a minimum for the maximum likelihood decision rule.

4. THE PROBLEM OF THREE GROUPS

As in the case of two groups we consider two situations, firstly when the sample consists of n individuals observed at random from a mixed population and secondly when the sample is a mixture of n_1 individuals drawn from the first group, n_2 from the second and n_3 from the third. The problem in either case is to select n_1 individuals for the first group, n_2 for the second and n_3 for the third, where $n_1+n_2+n_3 = n$.

Let $f_1(x)$, $f_2(x)$ and $f_3(x)$ represent the probability densities of x for the three groups and π_1, π_2, π_3 the proportions of mixture in the general population. The loss in assigning a person to the i-th group when, in fact, he belongs to the j-th group is denoted by r_{ji}. The aposteriori risks in assigning an individual with measurements x_i to the first, second and third groups are respectively equal to

$$a_{11} = \frac{\pi_2 f_2(x_1) r_{21} + \pi_3 f_3(x_1) r_{31}}{\pi_1 f_1(x_1) + \pi_2 f_2(x_1) + \pi_3 f_3(x_1)}$$

$$a_{21} = \frac{\pi_1 f_1(x_1) r_{12} + \pi_3 f_3(x_1) r_{32}}{\pi_1 f_1(x_1) + \pi_2 f_2(x_1) + \pi_3 f_3(x_1)}$$

and

$$a_{31} = \frac{\pi_1 f_1(x_1) r_{13} + \pi_2 f_2(x_1) r_{23}}{\pi_1 f_1(x_1) + \pi_2 f_2(x_1) + \pi_3 f_3(x_1)}$$

Consider a set of functions

$$\delta_i = 0 \text{ or } 1,\ \delta'_i = 0 \text{ or } 1,\ \delta''_i = 0 \text{ or } 1,\ \delta_i + \delta'_i + \delta''_i = 1$$

such that

$$\Sigma \delta_i = n_1,\ \Sigma \delta'_i = n_2,\ \Sigma \delta''_i = n_3$$

They define a decision rule if the individual with measurements x_i is assigned to the first group when $\delta_i = 1$ to the second when $\delta'_i = 1$ and to the third when $\delta''_i = 1$. The aposteriori risk for such a selection procedure is

$$\sum_1^n (\delta_i a_{11} + \delta' a_{21} + \delta''_i a_{31}) \qquad \ldots \quad (4.1)$$

The best decision rule is one which minimises the above expression. This is exactly the problem solved in lemma 2. The best solution is

$$\begin{aligned}
\delta_i &= 1 \quad \text{if} \quad a_{11} + \mu_1 \leqslant a_{21} + \mu_2,\ a_{11} + \mu_1 \leqslant a_{31} + \mu_3 \\
\delta'_i &= 1 \quad \text{if} \quad a_{21} + \mu_2 \leqslant a_{11} + \mu_1,\ a_{21} + \mu_2 \leqslant a_{31} + \mu_3 \\
\delta''_i &= 1 \quad \text{if} \quad a_{31} + \mu_3 \leqslant a_{11} + \mu_1,\ a_{31} + \mu_3 \leqslant a_{21} + \mu_2
\end{aligned}$$

where μ_1, μ_2 and μ_3 are determined such that $\Sigma \delta_i = n_1$, $\Sigma \delta'_i = n_2$ and $\Sigma \delta''_i = n_3$. As it stands the problem of determination of μ_1, μ_2, μ_3 appears to be complicated. There is a geometrical device which is helpful in the solution of the problem. In higher dimensional cases involving four or more groups the geometrical method cannot be applied. For three groups we replace a_{11}, a_{21}, a_{31} by two coordinates

$$X_i = a_{11} - a_{21},\quad Y_i = a_{11} - a_{31}$$

and represent the n points (X_i, Y_i) on a two dimensional chart with rectangular axes. The problem is to determine a point (X_0, Y_0) on this chart such that the regions formed by the lines $X = X_0$, $Y = Y_0$ and $Y - Y_0 = X - X_0$ contain the requisite number of points. This can be done by moving three thin rods OX', OY', OZ' fixed at the point

O as shown in the figure below, with OX' and OY' parallel to the X and Y axes and arrive at the required division by trial and error. It will help in this process if the numbers falling in the three regions are recorded for a few positions of the frame with the frame marked on the chart.

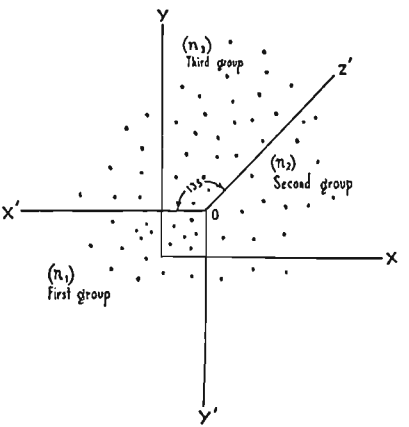

Figure 1

To solve the second problem when the sample consists of a mixture of n_1 individuals drawn from the first group, n_2 from the second and n_3 from the third, one can set up the total risk relative to the given set of observations for each of $n!/(n_1!n_2!n_3!)$ possible decision rules and choose that rule for which the risk is a minimum. This is very difficult in practice so that a simplified procedure is needed. As before we may choose that decision rule which leads to a division of the sample with the maximum probability density. This possesses an important property that the probability of the maximum number of wrong classifications for any one group is as small as possible.

The method of arriving at the required division is first to obtain the quantities

$$a_{1i} = \log f_1(x_i),\ a_{2i} = \log f_2(x_i),\ a_{3i} = \log f_3(x_i)$$
$$i = 1, 2, \ldots, n$$

and plot the points

$$X_i = a_{2i} - a_{1i},\quad Y_i = a_{3i} - a_{1i}$$

and proceed geometrically as in figure 1. For this division the probability density will be a maximum.

5. On a problem of optimum selection

Birnbaum and Chapman (1950) considered a problem of selecting candidates on the basis of p admission scores $y_1, y_2, \ldots y_p$. The object is to select those whose performance is expected to be better in the final test. The offered solution does not refer to a case where the scores of a number of individuals N have been observed but to a hypothetical set of individuals offering for the admission test. The former problem is often met with because the question asked is whom out of a number of individuals whose scores are available should be admitted. Let the scores of N individuals be be represented by

$$\begin{array}{cccc} y_{11}, & y_{21}, & \ldots, & y_{p1} \\ y_{12}, & y_{22}, & \ldots, & y_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ y_{1N}, & y_{2N} & \ldots, & y_{pN} \end{array}$$

To answer this problem we need the expected performance in the final test of an individual with the initial scores $y_{11}, \ldots, y_{p1}$. Let this expected performance be

$$x_1 = \phi(y_{11}, y_{21}, \ldots, y_{p1})$$

which actually stands for the regression equation of the final performance on the initial scores. The regression function which may be of any complicated type supplies us with the expected performances $x_1, x_2, \ldots, x_N$ of the candidates and these latter scores form the basis for selection. The regression function can be estimated on the basis of the previous information. No assumption of multivariate normality is needed.

For instance if k, a given number of seats, are available then the best plan is to admit k candidates corresponding to the k largest x's because this maximises the expected performance under the condition that k have to be chosen.

A second alternative may be to admit as many as possible with the restriction that the expected average performance is not less than an assigned number x_0. The best plan is then to arrange the x scores in a decreasing order and find the cumulative averages from the top and admit all those for whom the cumulative average is greater than or equal to x_0. Obviously under such a selection procedure the maximum number is admitted subject to the condition that the expected average performance of the selected candidates is not less than x_0.

If the restriction is that the average performance of the chosen candidates should exceed a given value x_0 with a probability greater than β, then again we start with the highest score of x and go on adding the others in the decreasing order of x till the required probability remains greater than β. The calculations are not however, simple.

If we consider a hypothetical set of candidates, a situation which may arise when the statistician is asked to give a uniform rule for independent recruitment at various places without specifying the numbers to be selected from each place, then what is needed is the determination of a score x_0 leading to the selection of all individuals with expected x score (calculated on the basis of the admission y score) greater than or equl to x_0. For this the distribution of expected score x as a function of y has to be studied. Let this be $f(x)$. If the criterion is that the maximum number of candidates have to be admitted subject to the condition that the expected average performance is not less than x_0 then x_0 is determined from the formula

$$\int_{x_0}^{\infty} x f(x)\, dx = x_0 \int_{x_0}^{\infty} f(x) dx$$

in which case the expected proportion admitted is

$$\int_{x_0}^{\infty} f(x)\, dx$$

References

Birnbaum, Z. W. and Chapman, D. G. (1950): On optimum selections from multinormal populations *Ann. Math. Stat.* 21, 443.

Rao, C. R. (1948): The utilization of multiple measurements in problems of biological classification *J.R.S.S.*, Series B, 10, 159.

Wald, A. (1950) : Statistical decision functions. John Wiley & Sons, New York.

Paper received : January, 1951.

This treatment of the problem then presupposes that the individual agent has stable probability estimates of finding someone to trade with, of that person being costly to bargain with, of that party being inclined to cheat on the terms of the agreement. In the case of externalities this implies that some unwanted side effects remain because of the uncertainty associated with undertaking the transaction that would eliminate them. The original problem was how to perform the market basket analysis which attempted to find all the interesting relationships between products bought in a given context. Association rule mining has been applied to LMS in order to reveal which contents students tend to access together, or which combination of tools they use.Â Subgroup discovery is usually seen as being different from classification, as it addresses different goals.