2群判別で判別保留を考慮した総期待費用最小化モデル

01000170　青山学院大学　阿部 俊一　ABE Shun-ichi

1　はじめに

医学検査などでは，陽性(+)，陰性(−)のほかに，しばしば疑陽性(±)の判定が用いられている．このことに対応して2群判別に判別保留領域を追加し，結果的に3個の判別領域を設定するモデルを考察する*.

2　確率モデル

群G_iに属する個体(標本)の出現確率を$P_i>0$，その個体に関するp個の変量のベクトル$x=(x_1,x_2,\cdots,x_p)$の確率密度関数を$f_i(x)$ $(i=1,2)$とし，$P_1+P_2=1$とする．ここで，群G_iに属する個体を$G_j(i\neq j)$に属するものと誤判別したときとその個体を判別保留としたときの1個体当たり期待費用をそれぞれ$C(j|i)$，$C(0|i)$とし，更に，

(1)　$C(j|i)>C(0|i)>0 \qquad (i,j=1,\ 2;\ i\neq j)$

とする．各個体に対する起こり得る誤判別と判別保留に基づく期待費用の総和は

(2)
$$\begin{aligned}C&=P_1C(2|1)\int_{x\in D_2}f_1(x)dx+P_2C(1|2)\int_{x\in D_1}f_2(x)dx\\&+\int_{x\in D_0}\{P_1C(0|1)+P_2C(0|2)f_2(x)\}dx\end{aligned}$$

であり，このCの値を最小にするように集合D_0，D_1，D_2を決定することがこの研究の目的である．ここに，データxを持つ個体をG_iに属するものと判別することを$d(x)=i$ $(i=1,2)$とし，またその個体を判別保留とすることを$d(x)=0$と表すものとし，

(3) $D_i=\{x\mid d(x)=i\}$　$(i=0,1,2)$，

$D_0\cup D_1\cup D_2=R^p$ (p次元ユークリッド空間)，

$D_0\cap D_1=D_1\cap D_2=D_0\cap D_2=\phi$ (空集合)

とする．

以上の前提の下に以下の記号を定義する：

(4)　$\rho(x)=\dfrac{P_1f_1(x)}{P_2f_2(x)}$，　$\rho_0=\dfrac{C(1|2)}{C(2|1)}$，

$$\rho_1=\frac{C(1|2)-C(0|2)}{C(0|1)},\quad \rho_2=\frac{C(0|2)}{C(2|1)-C(0|1)};$$

(5)　$m=\left\{\dfrac{C(2|1)}{C(0|1)}-1\right\}\left\{\dfrac{C(1|2)}{C(0|2)}-1\right\}.$

3　ρ_0，ρ_1，ρ_2とmの関係

上の(4)と(5)から

$$\begin{cases}\rho_1-\rho_0=(m-1)C(0|2)/C(2|1),\\ \rho_1-\rho_2=(m-1)\rho_2,\\ \rho_0-\rho_2=(m-1)\rho_2C(0|1)/C(2|1)\end{cases}$$

が得られる．これらから，2つの同等関係式：

(6)　$\rho_1<\rho_0<\rho_2 \Longleftrightarrow m<1,$

(7)　$\rho_1>\rho_0>\rho_2 \Longleftrightarrow m>1,$

が得られる．

4　総期待費用を最小にする判別

任意の個体に対する観測データのベクトルxが与えられた時，

(8)　$x\in D_i\,(i=1,2)$ではなく，　$x\in D_0$とした方が，総期待費用Cが小さくなるための条件は

$$\begin{aligned}&P_1C(0|1)f_1(x)+P_0C(0|2)f_2(x)\\&<P_iC(j|i)f_i(x)\qquad (i,j=1,2;\ i\neq j),\end{aligned}$$

すなわち，

$$\rho_2<\rho(x)<\rho_1;$$

(9)　$x\in D_0$でも$x\in D_2$でもなく，　$x\in D_1$とした方が，総期待費用Cが小さくなるための条件は

$$\begin{aligned}&P_2C(1|2)f_2(x)<P_1C(2|1)f_1(x),\\&P_2C(1|2)f_2(x)<P_1C(0|1)f_1(x)+P_2C(0|2)f_2(x),\end{aligned}$$

すなわち，

$$\rho(x)>\rho_0,\rho(x)>\rho_1;$$

(10)同様に，　$x\in D_0$でも$x\in D_1$でもなく，　$x\in D_2$

*本稿の内容は，すでに10年以上筆者のﾌｧｲﾙに保管されていたものであるが，文献[1]に触発されて公表を思い立ったものである。

とした方が，C が小さくなるための条件は

$\rho(x) > \rho_0, \rho(x) > \rho_2$;

となる．(8)，(9)，(10)から次の 3 つの命題が得られる．

命題１：$m<1$ ならば，(6)より，$\rho_1 < \rho_0 < \rho_2$ となり，(8)の D_0 は空集合 $D_0 = \phi$ であり，(9)，(10)から最適な D_1，D_2 は

$D_1^0 = \{x \mid \rho(x) > \rho_0\}$，$D_2^0 = \{x \mid \rho(x) < \rho_0\}$

となる．特に，$\rho(x) = \rho_0$ を満足する $x = x_0$ は D_0，D_1 のいずれに含めても C は変わらないが，誤判別確率をなるべく小さくする意味から，

(11)

$$\begin{cases} \rho(x_0) > 1 \text{ならば} D_2^0 \text{を} D_2^0 = \{x \mid \rho(x) \le \rho_0\} \text{に修正；} \\ \rho(x_0) < 1 \text{ならば} D_1^0 \text{を} D_1^0 = \{x \mid \rho(x) \ge \rho_0\} \text{に修正；} \\ \rho(x_0) = 1 \text{の時} x_0 \text{は} D_1^0, D_2^0 \text{のいずれに入れてもよい．} \end{cases}$$

命題２：$m>1$ ならば，(7)より $\rho_1 > \rho_0 > \rho_2$ となり，(8)より $D_0^0 = \{x \mid \rho_2 < \rho(x) < \rho_1\}$，(9)より $D_1^0 = \{x \mid \rho(x) > \rho_1\}$，(10)より $D_2^0 = \{x \mid \rho(x) < \rho_2\}$ となる．特に，$\rho(x) = \rho_1$ 及び $\rho(x) = \rho_2$ を満足する x は D_0，D_1，D_2 のいずれに含めても C は変わらないが，上の(11)と類似の修正を加える方がよい．

命題３：$m=1$ ならば，3節から，$\rho_0 = \rho_1 = \rho_2$ となるから，(8)より $D_0^0 = \phi$，(9)より $D_1^0 = \{x \mid \rho(x) > \rho_0\}$，(10)より $D_2^0 = \{x \mid \rho(x) < \rho_0\}$ とするほか，$\rho(x) = \rho_0$ を満足する $x = x_0$ の取り扱いについては命題１と同様にする．

命題１で“$m<1$”を“$m \le 1$”に修正すれば，命題３は命題１に統合することができる．

5　数値計算例

数値例：２群 G_1，G_2 の判別問題で，$P_1 = 2/5$，$P_2 = 3/5$，$C(2|1) = 3$，$C(0|1) = C(0|2) = 1/2$，$C(1|2) = 1$，G_i に属する個体に対するデータの $x = (x_1, x_2)$ の確率密度関数を

$f_i(x) = (2\pi|\Sigma|)^{-1}\exp\{-Q_i(x)/2\}$,

$Q_i(x) = (x-\mu_i)\Sigma^{-1}(x-\mu_i)'$, $(i = 1,\ 2)$ とし，

$$\Sigma = \begin{bmatrix} 94.4071 & 54.6949 \\ 54.6949 & 121.9243 \end{bmatrix},$$

$\mu_1 = (74.7,\ 62.6)$,　$\mu_2 = (52.4,\ 49.1)$

とする時，誤判別と判別保留の総期待費用を最小にする判別領域を求めよ．また $x_* = (65,45)$ はどの領域に入るか判別せよ．

(解)：$\rho(x) = P_1 f_1(x) / P_2 f_2(x)$

$$= q\exp[a(x-\overline{\mu})'],$$

$$a = (a_1, a_2) = \Sigma^{-1}(\mu_1 - \mu_2)'$$
$$= \begin{bmatrix} 0.014312 & -0.006420 \\ -0.006420 & 0.011082 \end{bmatrix}\begin{bmatrix} 22.3 \\ 13.5 \end{bmatrix}$$
$$= (0.2325,\ 0.006441),$$

$\overline{\mu} = (\mu_1 + \mu_2)/2 = (63.55,\ 55.85)$,

$q = P_1 / P_2 = 2/3$;

$$Z(x) = \ln\rho(x) = a(x-\overline{\mu})' + \ln q$$
$$= 0.2325(x_1 - 63.55)$$
$$+ 0.006441(x_2 - 55.85) - 0.4055;$$

$\rho_0 = 1/3$，$\rho_1 = 1$，$\rho_2 = 1/5$，

$\theta_0 = \ln\rho_0 = -1.0986$，$\theta_1 = \ln\rho_1 = 0$，

$\theta_2 = \ln\rho_2 = -1.6094$；$m = 5$；

$D_0 = \{x \mid -1.6094 \le Z(x) \le 0\}$,

$D_1 = \{x \mid Z(x) > 0\}$,

$D_2 = \{x \mid Z(x) < -1.6094\}$;

$Z(x_*) = -0.1383$ より $x_* \in D_0$．

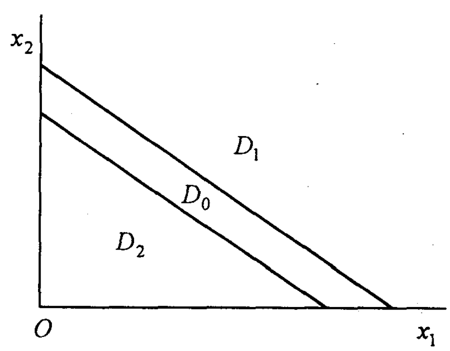

図　上の数値例における D_0，D_1，D_2 の概念図

文献[1]：末松俊幸ほか２名：“相対効率分析法と目標計画法から見た３種の判別分析法の比較・考察，” 1997年度日本OR学会春季研究発表会アブストラクト集．pp.122-123.