実用的重判別法について

統計数理研究所　駒　澤　　勉

(1980年1月　受付)

A Utilitarian Method of Multiple Discriminant Analysis

Tsutomu Komazawa
(The Institute of Statistical Mathematics)

This method aids the after processing of data analysis for discriminant analysis of Hayashi's quantification method or multiple discriminant method by Cooley and Johnes. Our thinking does not discriminate all groups quickly on one dimensional space. Because we have better discriminate two groups that we have united by our method, as many cases may take efficient effect from practical problems on processing of data. So we explain the processing of integration and discrimination with examples. We think this processing method may be used widely on multiple discriminant method in reality.

0. まえがき

この分析法は相関比を判別基準にした，数量化第II類（林，[2]）や重判別分析（Cooley & Johnes, [4]）の解析法に事後的解析の処理過程を追加したものである．一般に，この種の分析に当っては，本来，多次元空間上で判別すべきところを，最も判別効果のある座標軸（最大相関比に対する軸）を求めて，その軸上に各集団の分布を射影したもので判別を行っている．ところで，実際の問題では多くの場合，各集団の分布が離散するようきれいにずれていない．それらは傾向として，特定の集団がどちらか軸の一方の極に寄り，他の集団は反対側の極に，お互いに重なる度合が大きい一つの分布と考えられることが，大半である．そのため，一つの座標軸で一度に全集団を判別するのでなく，2つの集団に統合して判別した方が効率がよい場合が多い．そこで，その統合の仕方と判別の手順について考えてみた．また，その時の判別基準の相関比とミニマックス的中解の関係についても整理してみた．

1. 2集団（グループ）の相関比とミニマックス的中率

数量化理論第II類や重判別分析のデータ解析の方法論は相関比 η^2 を判別の基準測度として解析計算を行っている．しかし，実際上は直感的にもわかり易い的中率を得られた結果の分布から計算し直して，判別の基準測度に利用し分析を行っている．そこで，こゝでは相関比 η^2 とミニマックス的中率の関係について以後の分析上，利用するので述べておく．

いま，2集団の分布は $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ なる正規分布に従うものとし，確率密度関数を $f_1(Y), f_2(Y)$ とする．ただし μ_t は平均値，σ_t^2 は分散 $(t=1,2)$ である．よく知られているように，ミニマックス的中解は次式を解くことで求まる．

$$(1)\quad \frac{1}{\sqrt{2\pi}\,\sigma_1}\int_{-\infty}^{\alpha}\exp\left[-\frac{(Y-\mu_1)^2}{2\sigma_1^2}\right]dY=\frac{1}{\sqrt{2\pi}\,\sigma_2}\int_{\alpha}^{+\infty}\exp\left[-\frac{(Y-\mu_2)^2}{2\sigma_2^2}\right]dY$$

判別の分点 α は

$$\alpha = \frac{\mu_1 \sigma_2 + \mu_2 \sigma_1}{\sigma_1 + \sigma_2}$$

として求まる．

ところで，ミニマックス的中率 p は次式で求めることが出来る．

(2) $$p = \frac{1}{\sqrt{2\pi}} \int_{\alpha'}^{\infty} \exp\left[-\frac{t^2}{2}\right] dt$$

ただし，$\mu_2 \geqq \mu_1$，$t = \dfrac{Y-\mu_2}{\sigma_2}$，$\alpha' = \dfrac{\mu_1-\mu_2}{\sigma_1+\sigma_2}$

この α' を，ここでは区分点と呼ぶことにする．実際には的中率 p の計算は (2) 式を変形した (2)′ 式を利用すればよい．

(2)′ $$p = \frac{1}{\sqrt{2\pi}} \int_{0}^{|\alpha'|} \exp\left(-\frac{t^2}{2}\right) dt + 0.5$$

(2)′ 式の第1項の値は正規分布の表があれば直接それから読みとることで簡単にミニマックス的中率 p を求めることができる．

そこで，(2) 式によるミニマックス的中率 p を求める際の区分点 α' と相関比 η^2 の関係を示しておこう．いま，全分散 σ^2，外分散 $\sigma_B{}^2$ とすると相関比 η^2 は次のように表現できる．

(3) $$\eta^2 = \frac{\sigma_B{}^2}{\sigma^2} = \frac{\pi_1 \pi_2 (\mu_2-\mu_1)^2}{\sigma^2}$$

ただし，π_t, $(t=1,2)$ は集団 t に属する割合を示す．

即ち，相関比 η^2 と区分点 α' の関係は

$$\eta^2 = \frac{\pi_1 \pi_2 (\mu_2-\mu_1)^2}{\sigma^2} = \frac{\pi_1 \pi_2 (\sigma_1+\sigma_2)^2}{\sigma^2} \left(\frac{\mu_2-\mu_1}{\sigma_1+\sigma_2}\right)^2$$

$$= \frac{\pi_1 \pi_2 (\sigma_1+\sigma_2)^2}{\sigma^2} \alpha'^2$$

(4) $$\therefore \quad \alpha' = \frac{\sigma}{\sigma_1+\sigma_2} \sqrt{\frac{\eta^2}{\pi_1 \pi_2}}$$

である．実際の解析では全平均値 $\mu=0$，全分散 $\sigma^2=1$ と正規化された計算結果が出ていることが多い．その際は

(5) $$\alpha' = \frac{1}{\sigma_1+\sigma_2} \sqrt{\frac{\eta^2}{\pi_1 \pi_2}}$$

であり，区分点 α' を集団 1,2 に属する割合 π_1, π_2，標準偏差 σ_1, σ_2 と相関比 η^2 で示すことができる．

2. 2集団への統合法

統合の手順に入る前に，数量化理論第II類，または重判別分析の計算で求まった各値についての記号表示の定義をしておく．

集団を t，個体を i，座標軸を u で表示する．u は各軸の相関比の大きさの順番に対応する．座標軸 u において，集団 t における個体 i の数量を $y_{it}{}^{(u)}$，平均値を $m_t{}^{(u)}$，標準偏差を $s_t{}^{(u)}$，相関比を $\eta_u{}^2$，ミニマックス的中率を p_u，集団数を T とする．

i. 2集団への統合手順（その 1）

(1)　最大相関比 η_u^2 $(u=1)$ に対する集団別平均値を大小順に並べる．

(2)　隣りあう集団間の距離 d_τ を計算する．

$$d_\tau = m_\tau^{(u)} - m_{\tau+1}^{(u)} \qquad (m_\tau^{(u)} \geqq m_{\tau+1}^{(u)},\ \tau = 1, \cdots, G$$

G の初期値は $T-u$ とする．)

次に，d_τ が最小である集団番号 τ を求める．

$$\tau = \min \quad (d_1, d_2, \cdots, d_\tau, \cdots, d_G)$$

この τ を使って τ 集団と $\tau+1$ 集団を統合する．また，この統合された集団の平均値を求め，$\tau+2$ 番目の集団以後の順番を1番前へずらす．$(\tau+2 \to \tau+1, \tau+3 \to \tau+2, \cdots)$ さらに，$G \to G-1$ とし，この (2) の操作を繰り返し，$G=1$ になるまで行う．なお，統合作業中において統合された集団の認識番号を記録しておく．

次に，$\tau=1$，または $\tau=T-u$ のどちらか端の単独集団が統合結果において2集団のうちの一方の構成になった時は，手順 ii. はとばす．

ii.　2集団への統合手順（その 2）

(1)　i. 同様に集団別平均値を大小順に並べる．集団の平均が最小と最大なる集団と相隣る集団の平均値の距離 d を求める．

$$d_{\min} = d_G - d_{G-1}$$
$$d_{\max} = d_1 - d_2$$

ただし，　$(d_1 \geqq d_2 \geqq \cdots \geqq d_G)$

$$G = T-u$$

次に，$d_{\max}$ と $d_{\min}$ の大小を調べる．もし $d_{\max}$ が大きければ $\tau=1$ の集団を，$d_{\min}$ が大きければ $\tau=T-u$ の集団を特定集団に定め，他は一緒にまとめて混合集団とする

iii.　統合した2集団での相関比と的中率

手順 i., ii. で2集団化した構成の内容が異なった場合には互々の相関比を計算し，その値が大きい方を採用する．次に，採用された2集団のミニマックス的中点，および的中率を (1), (2) 式から求める．

このように，数量化理論第 II 類や重判別分析によって求まった最大相関比 η_1^2 に対して，判別を行ったら，第二番目に大きい相関比 η_2^2 で混合集団だけについて手順 i., ii., iii. を行う．同様に，相関比 $\eta_3^2, \eta_4^2, \cdots$ について順次判別を行う．

3. 計　算　例

数量化理論第 II 類の計算結果（表 1）を用い判別分析の事後処理を行う．ただし，表1の各

表1　集団別各値

集団 t	統計量	第1軸	第2軸	第3軸	個体数 n_t
集団 1	m_1 s_1	−0.3273 0.7261	0.6548 0.5622	0.0214 1.2320	14
集団 2	m_2 s_2	−0.5680 0.9666	−0.4493 1.2240	0.1559 0.8795	15
集団 3	m_3 s_3	0.2353 0.7722	−0.1712 0.8618	−0.4641 0.8484	12
集団 4	m_4 s_4	0.9345 0.7881	−0.0338 0.7854	0.2664 0.7844	11
—	η^2	0.319	0.180	0.071	—

注：　m_t は平均値，s_t は標準偏差 $(t=1, 2, 3, 4)$

軸の数量は全平均0，全分散1に正規化されている．

1) 最大相関比 η_1^2 の座標軸での判別

統合法（その1）の作業手順

作業 (1)	平均値の大小順	② −0.5680	① −0.3273	③ 0.2353	④ 0.9345
(2)	集団間の距離	0.2407	0.5626	0.6992	
(3)	統合集団の平均値		(②①) −0.4518	③ 0.2353	④ 0.9345
(4)	集団間の距離		0.6871	0.6992	
(5)	統合集団の平均値			(②①③) −0.2507	④ 0.9345
	標準偏差			0.8345	0.7881
	サンプル数			41	11

注： t, $t=1,2,3,4$ ははじめの集団番号． ＊＊＊＊ は集団距離最小を示す．

統合された2集団の一方が単独の集団構成になっているので統合法（その2）の作業手順はとばす．次に集団4とその他混合集団のミニマックス的中率，相関比などを計算する．

$$\text{判別の分点}\quad \alpha = \frac{s_2 m_1 + s_1 m_2}{s_1 + s_2} \fallingdotseq 0.358$$

$$\text{区分点}\quad \alpha' = \frac{m_1 - m_2}{s_1 + s_2} \fallingdotseq -0.7304$$

$$\text{ミニマックス的中率}\quad p = \frac{1}{\sqrt{2\pi}} \int_0^{-0.7304} \exp\left(-\frac{t^2}{2}\right) dt + 0.5 \fallingdotseq 0.767$$

$$\text{相関比}\quad \eta^2 = \pi_1 \pi_2 (m_1 - m_2)^2 \fallingdotseq 0.234$$

ただし，m_t, s_t, π_t, $(t=1,2)$ は2集団に統合された時の平均値，標準偏差，集団に属する割合を示す．

2) 第二に大きい相関比 η_2^2 の座標軸での判別（集団4は除く）

作業 (1)	平均値の大小順	② −0.4493	③ −0.1712	① 0.6548
(2)	集団間の距離	0.2781	0.8260	
(3)	統合集団の平均値		(② ③) −0.3257	① 0.6548
	標準偏差		1.0782	0.5622
	サンプル数		27	14

判別の分点 $\alpha \fallingdotseq 0.318$

区分点 $\alpha' \fallingdotseq -0.597$

ミニマックス的中率 $p \fallingdotseq 0.726$

$$\text{相関比}\quad \eta^2=\frac{\pi_1\pi_2(m_1-m_2)^2}{S^2}\fallingdotseq 0.197$$

ただし，第二番目に大きい相関比 η_2^2 以後の軸での新しい2集団に対する相関比の計算に際しては，一般に表1が全平均0，全分散1の結果表として出されているから，いくつかの集団を除いた時には，全平均 m ($\fallingdotseq 0.0091$) と全分散 S^2 ($\fallingdotseq 1.0935$) を再計算して相関比 η^2 を求める．

$$\text{全平均}\quad m=\frac{1}{n}\sum_{t=1}^{G}n_t m_t^{(u)}\qquad n=\sum_{t=1}^{G}n_t$$

$$\text{全分散}\quad S^2=\frac{1}{n}\sum_{t=1}^{G}n_t(S_t^{(u)2}+m_t^{(u)2})-m^2$$

（ただし，G はいくつかの集団を除いたあとの集団数，n は全サンプル数，n_t は集団 t のサンプル数）

3)　相関比 η_3^2 の座標軸での判別（集団 1, 4 を除く）残りの集団は2つであるから的中率，相関比の計算だけ．

判別の分点　$\alpha\fallingdotseq -0.159$
$\alpha'\fallingdotseq -0.358$
$p\fallingdotseq 0.640$
$\eta^2=0.112$　　$(m\fallingdotseq -0.119,\ s^2\fallingdotseq 0.844)$

次に，数量化計算で求まっていた各判別軸に対する個体数量と事後処理で求まった判別の分点から機械的に的中率を求めて見る．その結果を表2に掲げておく．

表2　的中頻度表

的中点	第1判別軸 0.358		第2判別軸 0.318		第3判別軸 −0.159	
集団 1	[11]	3	4	[10]	7	7
集団 2	13	2	[11]	4	6	[9]
集団 3	[9]	3	[8]	4	[8]	4
集団 4	1	[10]	7	4	2	9
的中率	0.827 (43/53)		0.707 (29/41)		0.630 (17/27)	

4. おわりに

この種の判別分析では要因分析的な面も多分にあるので，2つの集団にまとめて分析を進めることは各判別軸の解釈，各変量の重みの解釈などを容易にする上で有意義と思われる．一般に，コンピュータによる一貫処理計算は各集団別分布の結果の出力までゞある．的中率の計算は，多くの場合，その分布表を見て手作業で求めているのが実状と思われる．その点，従来の判別の手順にこの事後処理を附加することで，個体数量の各集団別分布を正規分布と仮定はしているが，判別の分点，ミニマックス的中率などを簡単かつ機械的に計算できる利点があるので，実際のデータ解析処理に大いに役立つものと思われる．

参 考 文 献

[1] 千野貞子 (1963) 数量化による予測の的中率と相関比との関係について，統計数理研究所彙報 第11巻第1号，7-24.

[2] 林 知己夫，樋口伊佐夫，駒澤 勉 (1970) 情報処理と統計数理，産業図書，72-79.

[3] Hayashi, C. (1952) On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view, *Ann. Inst. Math., Statist.* **3**, 69–98.

[4] Cooley, W.W. and Johnes, P.R. (1962) *Multivariate Procedures for the Behavioral Sciences*, John Wiley, New York.