

判定土类的逐步判别方法

莫 治 雄

(广西农垦职工大学)

逐步判别适用于事物属性的两类及两类以上的多类别判别,广泛应用于地质、气象、医学、经济、农业等领域^{〔1-2〕},^①,但在国内土壤学界应用还不多。本文根据红壤、黄壤、黄棕壤的一些剖面分析数据,将逐步判别分析方法应用于判定土壤类型归属方面作些尝试。

红壤、黄壤和黄棕壤都是形成于热带亚热带生物气候条件下的土壤类型,具有不同程度的富铝化作用。黄棕壤兼具有红、黄壤的某些特点^{〔3-4〕}。土体中粘粒部分的硅铝率虽然能反映土壤富铝化作用的强弱,但并不能据此区分这些土壤类型。有人根据土壤胶体硅、铁、铝含量及相应的分子率共21个指标,对红、黄壤进行了两类判别分析,收到了较为满意的效果^{〔5〕}。然而,用于判别多变量时,则需要许多测定数据,而且计算的工作量也大为增加。本文所采用的逐步判别方法,能在提供判别分类的变量中,挑选出对类属判别影响显著的那些变量来建立判别函数,把对判别函数影响不显著的那些变量剔除掉,从而克服了一般判别分析方法的缺点。同时,应用逐步判别分析,还可以进行多类别的判别,比之两类判别分析的应用范围更广。

本文选取了40个土壤剖面(其中红壤16个,黄壤15个,黄棕壤9个),每个剖面选取A、B、C三个发生层次中的 SiO_2 、 Fe_2O_3 、 Al_2O_3 含量和 $\text{SiO}_2/\text{Al}_2\text{O}_3$ 、 $\text{SiO}_2/\text{Fe}_2\text{O}_3$ 、 $\text{SiO}_2/\text{R}_2\text{O}_3$ 、 $\text{Al}_2\text{O}_3/\text{Fe}_2\text{O}_3$ 比值共21个指标作为判别变量,进行逐步判别分析。另选10个剖面作为待判样品,用来验证所建立的判别函数的判别效果(原始数据从略)。本方法用BASIC语言编写程序,全部数据在IBM PC/XT计算机上实现计算。

一、基本原理

逐步判别分析是在多类(组)判别分析基础上发展起来的,二者的原理基本相同。前者所不同之处在于它是在正态母体的假设下,采用逐步筛选方法,把有显著影响的变量引入判别函数,把影响不显著的变量剔除掉。

设有 N 个已知样品、 P 个判别变量, N 个样品已预分为 G 组(类),第 i 组的样品数为 n_i

$$N = \sum_{i=1}^G n_i$$

各组样品都是相互独立的正态随机向量,且各组的协方差矩阵一样,即

$$(X_{gk1}, X_{gk2}, \dots, X_{gkp}) \sim N(\mu_g, \Sigma)$$

$$g = 1, 2, \dots, G; k = 1, 2, \dots, n_p$$

其中 X_{gki} 为 g 组中第 k 个样品的第 j 个变量。

① 黄嘉佑编,气象统计预报试用教材,101—105页。

按贝树斯准则, 即假设各总体 $A_g (g = 1, 2, \dots, G)$ 的概率分布密度函数 $f_g(X)$ 为已知, 先验概率 q_g 已知, 且错分损失相等, 经过逐步计算, 选取了 L 个对判别分类影响显著的变量, 则可建立一组判别函数

$$Y_g = \ln q_g + C_{g0} + \sum_{i \in \text{已引入变量}} C_{gi} X_i \quad (1)$$

$$g = 1, 2, \dots, G$$

式中 C_{gi} 称为判别系数(计算见后述), q_g 为第 g 组的先验概率, 可采用样品频率代替

$$q_g = n_g / N \quad (2)$$

用判别函数判别某一样品 X_0 的类别时, 把实测得的变量 $X_i (i \in \text{已引入变量})$ 的值代入(1)式, 求得 G 个 Y 值, 若

$$Y_h(X_0) = \max\{Y_g(X_0)\} \quad (3)$$

$$1 \leq g \leq G$$

则把 X_0 判归 h 类。其归属各类的后验概率用下式计算:

$$P(g/X_0) = e^{Y_g} / \sum_{g=1}^G e^{Y_g} \quad (4)$$

$$g = 1, 2, \dots, G$$

二、方法步骤

1. 求各个变量的均值和每一组内各变量的均值

第 g 组中第 J 个变量的均值

$$\bar{X}_{gJ} = \frac{1}{n_g} \sum_{k=1}^{n_g} X_{gk_j} \quad (5)$$

第 J 个变量的总均值

$$\bar{X}_J = \frac{1}{N} \sum_{g=1}^G \sum_{k=1}^{n_g} X_{gk_j} \quad (6)$$

2. 计算组内离差矩阵 $W = (w_{ij})$ 和总的离差矩阵 $T = (t_{ij})$

$$W_{ij} = \sum_{g=1}^G \sum_{k=1}^{n_g} (X_{gk_i} - \bar{X}_{gi})(X_{gk_j} - \bar{X}_{gj}) \quad (7)$$

$$t_{ij} = \sum_{g=1}^G \sum_{k=1}^{n_g} (X_{gk_i} - \bar{X}_i)(X_{gk_j} - \bar{X}_j) \quad (8)$$

$$i, j = 1, 2, \dots, P$$

3. 逐步计算

引入或删除一个变量为逐步判别的一步。手工计算时, 前三步可只考虑引入变量, 不考虑剔除变量。但在计算机上计算时, 为统一格式, 形式上每一步都考虑剔除。记初始组内离

差矩阵为 \$(w_{ij}^{(0)})\$, 初始总的离差矩阵为 \$(t_{ij}^{(0)})\$, 逐步判别在计算上采用“求解求逆紧凑变换法”, 设已进行了 \$t\$ 步, 引入了 \$L\$ 个变量, 则第 \$l+1\$ 步无论是引入或删除变量 \$x_r\$, 都进行变换:

$$w_{ij}^{(l+1)} = \begin{cases} w_{rr}^{(l)} / w_{rr}^{(l)} & (i=r, j=r) \\ w_{ij}^{(l)} - w_{ir}^{(l)} w_{rj}^{(l)} / w_{rr}^{(l)} & (i \neq r, j \neq r) \\ 1 / w_{rr}^{(l)} & (i=r, j=r) \\ -w_{ir}^{(l)} / w_{rr}^{(l)} & (i \neq r, j=r) \end{cases} \quad (9)$$

$$t_{ij}^{(l+1)} = \begin{cases} t_{rr}^{(l)} / t_{rr}^{(l)} & (i=r, j=r) \\ t_{ij}^{(l)} - t_{ir}^{(l)} t_{rj}^{(l)} / t_{rr}^{(l)} & (i \neq r, j \neq r) \\ 1 / t_{rr}^{(l)} & (i=r, j=r) \\ -t_{ir}^{(l)} / t_{rr}^{(l)} & (i \neq r, j=r) \end{cases} \quad (10)$$

第 \$l+1\$ 步时变量 \$x_i\$ 的判别能力按如下公式计算:

$$\text{对于未引入变量 } V_i = w_{ii}^{(l)} / t_{ii}^{(l)} \quad (11)$$

$$\text{对于已引入变量 } V_i = t_{ii}^{(l)} / w_{ii}^{(l)} \quad (12)$$

\$V_i\$ 越小, 则其对应的 \$F\$ 检验值越大, 变量的判别能力越强。因此, 对于每一步, 或是在已引入变量中以 \$V_i\$ 最大者作为待剔除变量; 或是在未引入变量中以 \$V_i\$ 最小者作为待引入变量, 如果

$$V_r = \max\{V_i\} \quad i \in L \quad (13)$$

则计算剔除变量的 \$F\$ 检验统计量 (\$F_{\text{剔}}\$)

$$F_{\text{剔}} = \frac{1 - V_r}{V_r} \cdot \frac{N - G - (L - 1)}{G - 1} \quad (14)$$

若 \$F_{\text{剔}} \leq F_2\$ (预先指定的剔除变量的临界值), 则将变量 \$x_r\$ 从判别函数中剔除出去, 对组内离差矩阵和总的离差矩阵施行变换 (9), (10); 如果 \$F_{\text{剔}} > F_2\$, 则考虑将

$$V_r = \min\{V_i\} \quad i \in \text{未引入变量} \quad (15)$$

对应的变量 \$x_r\$ 引入, 计算引入变量的 \$F\$ 检验统计量 (\$F_{\text{引}}\$)

$$F_{\text{引}} = \frac{1 - V_r}{V_r} \cdot \frac{N - G - L}{G - 1} \quad (16)$$

如果 \$F_{\text{引}} > F_1\$ (预先指定的引入变量的临界值), 则将变量 \$x_r\$ 引入判别函数, 对组内离差矩阵和总的离差矩阵施行变换 (9), (10); 如果 \$F_{\text{引}} \leq F_1\$, 表明已既不能剔除也不能引入变量, 逐步计算结束。

4. 建立判别函数, 检验判别效果

设最终已引入了 \$L\$ 个变量, 则第 \$g\$ 组的判别函数为:

$$Y_g = \ln q_g + C_g o + \sum_{i \in L} C_{gi} \bar{X}_i \quad g = 1, 2, \dots, G$$

其中 \$q_g\$ 为第 \$g\$ 组的先验概率, 可由 (2) 式计算, 而判别系数

$$C_{gj} = (N - G) \sum_{i \in I_g} W_{ji} \bar{X}_{gj} \quad (17)$$

$$j \in L; g = 1, 2, \dots, G$$

$\bar{X}_{gj}$ 为第 g 组第 j 个变量的均值,

$$C_{go} = -\frac{1}{2} \sum_{i \in L} C_{gi} \bar{X}_{gi} \quad g = 1, 2, \dots, G \quad (18)$$

引入的 L 个变量对所有 G 组的判别效果, 用 X^2 检验:

$$X^2 = -[N - 1 - (L + G)/2] \ln U_{(L)} \quad (19)$$

其自由度为 $L(G - 1)$, 其中 $U(b)$ 为威尔克斯统计量 (Wilks), 设逐次引入或剔除的变量序号为 $r_0, r_1, \dots, r_{L-1}$ 则

$$U_{(L)} = \frac{w_{r_0 0}^{(0)}}{t_{r_0 0}^{(0)}} \cdot \frac{w_{r_1 1}^{(1)}}{t_{r_1 1}^{(1)}} \cdots \frac{w_{r_{L-1} L-1}^{(L-1)}}{t_{r_{L-1} L-1}^{(L-1)}} \quad (20)$$

对策 e 和 f 两类的判别效果用 F 检验:

$$F_{ef} = \frac{(N - G - L + 1)n_e n_f}{L(N - G)(n_e + n_f)} D_{ef}^2 \quad (21)$$

$$\text{其中 } D_{ef}^2 = \sum_{i \in L} (C_{ie} - C_{if})(\bar{X}_{ie} - \bar{X}_{if}) \quad (22)$$

F_{ef} 的自由度 $v_1 = L; v_2 = N - G - L + 1$ 。

若 $X^2 > X_\alpha^2, F_{ef} > F_\alpha$ 则判别效果显著, α 为信度。

5. 判别应用

对于任一样品 $X = (X_1, X_2, \dots, X_p)$, 无论是已知分类的样品, 还是待判样品, 可用

(1) 式计算判别函数进行回判或判别, 如果

$$Y_h = \max\{Y_g\} \quad (23)$$

$$1 \leq g \leq G$$

则将该样品判属 h 类。

后验概率可用 (4) 式计算。

三、结果和讨论

1. 利用 40 个土壤剖面的原始数据, 经上述计算步骤算得各变量的分组均值和总均值列于表 1, 各组已引入变量的判别系数列于表 2。利用 (1) 式和 (2) 式及表 2 的数据, 可建立三个判别函数:

$$Y_1 = -0.9809 + C_{10} + \sum C_{1i} X_i$$

$$Y_2 = -0.9163 + C_{20} + \sum C_{2i} X_i$$

$$Y_3 = -1.4917 + C_{30} + \sum C_{3i} X_i$$

$i \in \text{已引入变量}$

已引入的 14 个变量 ($X_1, X_3, X_{5-7}, X_{9-11}, X_{14}, X_{16}, X_{18-21}$) 对所有 3 组的判别效果用 X^2 检验。根据 (20) 式和 (19) 式, 计算得

$$X^2 = 92.91$$

表1

各个变量的分组均值和总均值

项目	组别	土层	SiO ₂	Fe ₂ O ₃	Al ₂ O ₃	SiO ₂ /Al ₂ O ₃	SiO ₂ /Fe ₂ O ₃	SiO ₂ /R ₂ O ₃	Al ₂ O ₃ /Fe ₂ O ₃
分 组 均 值	第一组	A	44.94	10.93	34.35	2.31	12.23	1.91	5.55
		B	41.68	12.59	34.58	2.08	9.96	1.68	4.88
		C	41.63	13.10	34.63	2.09	9.60	1.63	4.72
	第二组	A	34.63	14.63	31.66	1.87	6.90	1.45	3.61
		B	33.61	15.47	32.37	1.80	6.61	1.39	3.57
		C	34.31	15.35	32.23	1.85	6.80	1.43	3.53
	第三组	A	47.93	8.17	25.40	3.23	19.61	2.69	6.02
		B	42.05	12.99	26.55	2.76	11.62	2.11	4.16
		C	41.30	11.63	23.24	2.57	10.65	2.03	4.11
总 均 值		A	41.27	11.63	31.26	2.34	11.76	1.90	4.83
		B	39.67	13.23	31.89	2.12	8.99	1.66	4.20
		C	38.76	12.63	32.32	2.10	8.71	1.63	4.12

说明: (1) 第1组: 黄壤, 1-15号样品。第2组: 红壤, 16-31号样品; 第3组: 黄棕壤, 32-40号样品;

(2) A层的7个变量依次用 $X_1, X_2, \dots, X_7$ 代表; B层的用 $X_8, X_9, \dots, X_{14}$ 代表, C层的用 $X_{15}, X_{16}, \dots, X_{21}$ 代表。

表2 各组已引入变量的判别系数

组别 (C)	1	2	3
C_{g0}	-8802.4454	-8262.4729	-8621.4766
C_{g1}	-65.0217	-63.2497	-65.4924
C_{g3}	110.8056	106.4249	110.2611
C_{g5}	9.4571	10.0170	11.7780
C_{g8}	1279.1674	1226.4961	1273.0939
C_{g7}	-47.9740	-47.7022	-52.3510
C_{g9}	67.4809	65.8771	66.3521
C_{g10}	-120.9903	-117.0028	-119.5559
C_{g11}	-745.5449	-718.8484	-732.8790
C_{g14}	201.8548	195.2938	197.5939
C_{g16}	425.9210	413.4751	420.9760
C_{g18}	-6625.2464	-6461.9053	-6527.5688
C_{g19}	-1050.0425	-1020.8833	-1041.0265
C_{g20}	16173.3419	15746.9252	15975.4452
C_{g21}	2014.2739	1956.6308	1998.0973

样品判归黄棕壤类。除了44号样品外, 其余49个样品的判别分类均与原分类相符, 符合率达98%。

逐步判别分析结果表明, 该方法用于判定土壤类属是可行的, 结果也令人满意。应用逐步判别分析使土壤类型的归属判定建立在数值分析基础上, 有助于土壤分类的定量化研究。

2. 经逐步判别, 引入了14个变量, 比原来减少了三分之一。由于引入的变量都是对判别分类有显著影响的, 因此, 提高了所建立的判别函数的稳定性和判别结果的可靠性。对那些被剔除的变量, 在实际工作中就可以不测定或不计算, 给工作带来很大方便, 减少许多工作量。

查信度 $\alpha = 0.001$ 时, 自由度 $14(3-1) = 8$ 的 X^2_α 为 56.89。 $X^2 > X^2_\alpha$, 故判别效果达极显著水平。

对于两类之间的判别效果用F检验。根据(22)和(21)式, 计算得 F_{12}, F_{13}, F_2 分别为 8.42、4.38、6.05。查 $\alpha = 0.01$, 自由度 $v_1 = 14, v_2 = 24$ 的 F_α 为 2.94。 F_{12}, F_{13}, F_2 都大于 $F_{0.01}$, 故判别效果达极显著水平。

从原始样品的回判和待判样品的判别结果看 (见表3和表4), 只有44号样品与原分组(类)不符。分析其原因是, 该样品的 $SiO_2/Al_2O_3, SiO_2/Fe_2O_3$ 和 SiO_2/R_2O_3 比值都远远大于红壤的平均值, 也大于黄壤的平均值, 而与黄棕壤的平均值接近, 故把该

表3

每个原始样品的回判结果

样品号	原组号	新组号	后验概率	样品号	原组号	新组号	后验概率
1	1	1	1.0000	21	2	2	1.0000
2	1	1	0.8807	22	2	2	1.0000
3	1	1	1.0000	23	2	2	0.9703
4	1	1	0.9998	24	2	2	0.9998
5	1	1	1.0000	25	2	2	1.0000
6	1	1	0.9769	26	2	2	1.0000
7	1	1	0.9396	27	2	2	0.9994
8	1	1	1.0000	28	2	2	0.9999
9	1	1	0.9960	29	2	2	0.9977
10	1	1	0.9992	30	2	2	1.0000
11	1	1	0.9995	31	2	2	1.0000
12	1	1	0.9981	32	3	3	0.7433
13	1	1	1.0000	33	3	3	0.9995
14	1	1	1.0000	34	3	3	1.0000
15	1	1	0.9995	35	3	3	1.0000
16	2	2	0.9869	36	3	3	0.9796
17	2	2	1.0000	37	3	3	1.0000
18	2	2	1.0000	38	3	3	0.9991
19	2	2	1.0000	39	3	3	0.9257
20	2	2	0.9930	40	3	3	0.8639

表4 每个待判样品的判别结果

样品号	计算分组	后验概率
41	1	1.0000
42	1	1.0000
43	2	0.9995
44	3	0.8968
45	2	0.9987
46	2	0.9998
47	2	0.9997
48	2	1.0000
49	2	0.9985
50	3	0.8488

3. 对于新取回的样品, 只要把分析数据代入上述三个判别函数, 计算得3个Y值根据(23)式即可判定该样品的归属。

参 考 文 献

〔1〕罗碧玉等编, 微机用多元统计分析软件, 287—363页, 四川科学技术出版社, 1986。

〔2〕黄正南编, 医用多因素分析及计算机程序, 76—99页, 湖南科学技术出版社, 1986。

〔3〕中国科学院南京土壤研究所编, 中国土壤, 495—537页, 科学出版社, 1978。

〔4〕李天杰等编, 土壤地理学, 151—158页, 高等教育出版社, 1983。

〔5〕杨艳生, 利用土壤胶体硅、铁、铝含量判定红壤和黄壤的类型归属, 土壤, 13(1):23—29, 1981。