

数理统计在化探中的应用简介 (七)

地 文

四、判别分析

1、问题的提出

表1和表2为某地矽卡岩型铜矿上14个已验证异常中Cu、Ag、Bi的几何平均值。表1为7个见矿异常,表2为7个未见矿异常。现在,要对当地一些未验证的异常作出判断:哪些可能见矿,哪些可能不见矿。例如,有一异常各元素的几何平均值(γ/g)为Cu880, Ag1.41, Bi34.4。如果利用单个元素分别对比,很难判断它属哪一类异常。利用判别分析,就可以帮助解决这类问题。因为判别分析可以提供一个界限,如未知异常落于该界限以上,属见矿异常,反之则属不见矿异常。此外如铁帽评价、岩体评价和构造评价等,均属此类问题。

2、几个基本概念

从上列实际问题中,抽象出一些一般性的提法:有两大类母体A及B,研究了A、B中若干个体(n_a 、 n_b),对每个个体测定了若干标志($x_1, x_2, \dots, x_p$)。现在要以这些标志的测定为基础,对另一些新的个体进行分类,判断它属A或B。由于各标志的测定

表1

异常号	X_{a1} (Cu)	X_{a2} (Ag)	X_{a3} (Bi)	R值
叶花香	1 380	0.08	8.9	0.248
	2 800	0.17	10.1	0.299
	3 3550	0.14	10	0.498
	4 224	0.14	6	0.389
石头壳	3500	0.70	6	0.701
铜井	500	1.70	20	0.031
赤马山	500	0.50	3	0.497

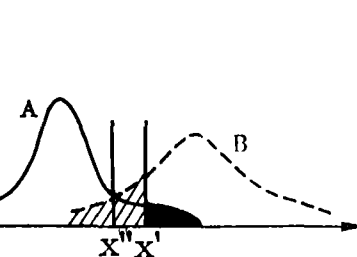


图1

值为统计量,有一个概率分布,所以据此进行分类时不可能百分之百正确,而只能要求尽可能少错些。如果把本来是属A的个体判别成属B,称为第一类错误,反之即称为第二类错误。这两类错误都应尽量避免。设某标志在不同母体的频率分布如图1。如果将判别界限划在 x' 处,发生第一类错误的概率 α 等于 x' 右边曲线A下面的面积(涂黑部份),发生第二类错误的概率 β 为 x' 左边曲线B下的面积(阴影部份)。 x' 取不同值时, α 、 β 相应变化。 x' 位置的选择是人为的,要根据两类错误的危害性大小来定。如第一类错误危害性更大,判别界线适当右移,反之则向左移。如果两条频率分布曲线完全重合,

表2

异常号	X_{b1} (Cu)	X_{b2} (Ag)	X_{b3} (Bi)	R值
I	177	0.95	11.5	0.03
II	143	0.64	11.5	0.014
V	215	0.55	12.5	0.048
VI	92	0.30	10.9	-0.016
VII	87	0.25	10	-0.007
小刘胜	1000	0.20	10	0.315
大刘胜	600	0.50	30	-0.001

就不能用这个标志判别A和B。如曲线A与B分得很开,用统计方法进行判别显然是多余的。然而这是很少遇到的两种极端情况。更常见的情况是A、B有部分重叠。此时如同时考虑多个标志,重叠现象可以大大改善,从而提高判

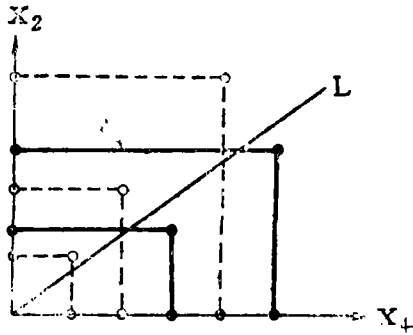


图 2

别能力。从图 2 可以看出这一点。以 x_1 、 x_2 分别代表两种不同的标志, 黑点表示含矿岩体, 小圈表示无矿岩体。在 x_1 、 x_2 轴上 (一度空间) 都找不出一个能把黑白点分开的界限, 但在作成 x_1-x_2 平面 (二度空间) 的散点图时, 就可以找到一条直线 L , 能把它们分开。当然这是大大简单化了的情况, 有时重叠现象很复杂, 仅用两种标志仍不能分开, 就要用第三个标志, 成为三度空间中分布的点, 用一个平面来区分它们。推而广之, 必要时可引入 K 个变量, 在 K 维空间中用一个超平面来区分代表不同母体的点子。当然在实际应用时并非变量越多越好, 因为这不仅会大大加重计算工作量, 而且有些变量的引入可能反使重叠现象加剧, 不利于判别。

综上所述, 我们可以找到一种判别函数 R , 它是由各种变量组成的一个方程式, (一次或二次的) 可据此算出 R 值, 如算出的 R 值大于某一临界值, 则属 A, 反之则属 B。下面介绍一种最简单而实用的判别函数。

3、线性判别函数

在下列三个条件下, 判别函数呈简单的线性公式: ①变量服从多维正态分布; ②多维正态分布的协

方差矩阵相等; ③两类错误的概率取得相等。只要 $n_a = n_b$, 化探资料一般能近似地满足这些条件。这时线性判别函数为:

$$R = \lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_k x_k$$

要建立这个方程, 必须求出 R 和 λ_i 。根据上述三个条件, 可导出下列计算公式:

$$\lambda_i = \frac{D_i}{D} \dots \dots (2),$$

$$R = \frac{1}{2D_i} \sum_{i=1}^K D_i (\bar{x}_{ai} + \bar{x}_{bi}) \dots \dots (3)$$

式中 D 为由 A、B 协方差矩阵之和构成的行列式, D_i 为 D 中等 i 列被 $d_i = \bar{x}_{ai} - \bar{x}_{bi}$ 所取代而构成的行列式。 $\bar{x}_{ai}$, $\bar{x}_{bi}$ 分别表示 A 和 B 的第 i 个标志的平均值。

以实例说明如何计算 D , 以至 λ 。我们从表 1, 表 2 出发。

第一步是进行对数转换, 使其成正态分布。得表 3 的左上部 (其中 Ag 的含量先乘以 100, 再换成对数, 以便计算)。

然后按公式计算方差和 ($i=j$ 时), 以及协方差和 ($i \neq j$) 时:

$$S_{ij} = \sum (x_{ai} - \bar{x}_{ai})(x_{aj} - \bar{x}_{aj})$$

$$= \sum x_{ai}x_{aj} - n\bar{x}_{ai}\bar{x}_{aj}$$

例如:

$$s_{11} = \sum x_{a1}^2 - n\bar{x}_{a1}^2 = 60.30 - 58.95 = 1.35$$

算出诸 S_{ij} 值后, 列成矩阵, 得:

表 3

	X_{a1} (Cu)	X_{a2} (Ag)	X_{a3} (Bi)	X_{a1}^2	X_{a2}^2	X_{a3}^2	$X_{a1}X_{a2}$	$X_{a1}X_{a3}$	$X_{a2}X_{a3}$
见 矿 异常 (A)	2.58	0.90	0.95	6.66	0.81	0.90	2.32	2.45	0.86
	2.90	1.23	1.00	8.41	1.51	1.00	3.57	2.90	1.23
	3.55	1.15	1.00	12.60	1.32	1.00	4.08	3.55	1.15
	2.35	1.15	0.79	5.52	1.32	0.62	2.70	1.86	0.91
	3.54	1.85	0.79	12.53	3.42	0.62	6.55	2.98	1.46
	2.70	2.23	1.30	7.29	4.97	1.69	6.02	3.51	2.90
	2.70	1.70	0.48	7.29	2.89	0.23	4.59	1.30	0.82
Σ	20.32	10.21	6.31	60.30	16.24	6.06	29.84	18.36	9.32
$\Sigma X_{ai}X_{aj}$	58.95	14.88	5.68	$n\bar{x}_{ai}\bar{x}_{aj}$			29.63	18.31	9.20

$$S_A = \begin{vmatrix} 1.35 & 0.21 & 0.05 \\ & 1.36 & 0.12 \\ & & 0.37 \end{vmatrix}$$

按同样方法处理未见矿异常, 得

$$S_B = \begin{vmatrix} 0.93 & -0.10 & 0.19 \\ & 0.33 & 0.08 \\ & & 0.17 \end{vmatrix}$$

二者相加(按位置对应地相加), 得:

$$S = \begin{vmatrix} 2.28 & 0.11 & 0.24 \\ & 1.69 & 0.20 \\ & & 0.54 \end{vmatrix}$$

由于 $S_{ij} = S_{ji}$, 所以按对角线对称的方式把左下部分填满, 即得行列式 D :

$$D = \begin{vmatrix} 2.28 & 0.11 & 0.24 \\ 0.11 & 1.69 & 0.20 \\ 0.24 & 0.20 & 0.54 \end{vmatrix} = 1.896$$

再将不同母体间各变量平均值之差算出来: $d_i = \bar{X}_{ai} - \bar{X}_{bi}$, 得 $d_1 = 5.56$, $d_2 = -1.70$, $d_3 = -2.05$ 。把它们排成一列, 用以代替行列式中的第 i 列, 即得 D_i 。例如:

$$D_1 = \begin{vmatrix} 5.56 & 0.11 & 0.24 \\ -1.70 & 1.69 & 0.20 \\ -2.05 & 0.20 & 0.54 \end{vmatrix} = 0.566$$

由此求出 $\lambda_1 = \frac{D_1}{D} = \frac{0.566}{1.896} = 0.298$ 。同法可

求出 $\lambda_2 = -0.062$, $\lambda_3 = -0.489$ 。至此, 已得判别方程式:

$$R = 0.298x_1 - 0.062x_2 - 0.489x_3 \dots \dots (4)$$

然后求 R 的临界值:

$$R_0 = \lambda_1 \cdot \frac{\sum X_{a1} + \sum X_{b1}}{n_a + n_b} + \lambda_2 \cdot \frac{\sum X_{a2} + \sum X_{b2}}{n_a + n_b} + \lambda_3 \cdot \frac{\sum X_{a3} + \sum X_{b3}}{n_a + n_b} + \dots \dots$$

(即将两个母体各变量的混合平均值代入式4中去)。可算出本例中 $R_0 = 0.2$ 。

将各个异常之 x_1, x_2, x_3 值代入式中, 可算出相应的 R 值(见表1, 2)。表1第

六个见矿异常 $R = 0.031$, 小于临界值(0.2), 据此将作出错误判断, 定为不见矿异常, 属第一类错误。另外表2第六个不见矿异常 $R = 0.315$, 大于临界值, 据此所作判断(定为见矿异常)属第二类错误。这两类错误的概率近似地为 $\alpha = \beta = \frac{1}{7}$ 。

对于本章开头列出的未知异常, 按式(4)算出其 R 值为 -0.007 , 小于临界值0.2, 故应定为不含矿异常。

4、几点说明

1. 标志的选择要少而精。尽量选那些在不同母体中平均值相差较大而各自的离散程度较小的元素。选择时可计算各变量之“贡献”, 舍去贡献小者。算法如下:

先算出综合距离函数

$$\Delta^2 = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3 \quad (\text{如本例中 } \Delta^2 = 0.277), \text{ 再分别求各元素之贡献,}$$

$$D_{xi} = \frac{\lambda_i d_i}{\Delta^2} \times 100, \quad (\text{如本例中Cu之贡献} = 60\%, \text{Ag为} 4\%, \text{Bi为} 36\%)$$

在本例中只用Cu、Bi两个变量即可, 但如只用Ag、Bi则不能得到满意的结果。

2. 对未知异常的判别, 仅限于地质、地球化学条件大致相同者。否则应重新建立适应新情况的判别函数。

3. 如果只考虑平均值进行判别而不考虑方差和协方差, 可建立很简单的判别式, 但判别能力较差。例如以两个不同母体平均值的中点作为分界线进行判别。当只有两个变量时可直接在散点图上选择分界的直线, 这样可以省却许多计算工作。