

地质工作中多变量统计分析

下 册

云南大学数学系编写
武汉地质学院计算站翻印

1981. 9

第六章 聚类分析

§1 什么是聚类分析

聚类分析又称群分析，它是研究（对样品或指标）分类问题的一种多元统计方法，首先认为我们所研究的样品或指标（变量）之间是存在着程度不同的相似性（亲疏关系）。于是，根据一批样品的多个观测指标具体找出一些能够度量样品或指标之间相似程度的统计量。以这些统计量为划分类型的依据，把一些相似程度较大的样品（或指标）聚合为一类，把另一些彼此之间相似程度较大的样品（或指标）又聚合为另外一类，……。关系密切的聚合到一个小的分类单位，关系疏远的聚合到一个大的分类单位，直到把所有的样品（或指标）都聚合完毕，把不同的类型一一划分出来，形成一个由小到大的分类系统。最后再把整个分类系统画成一张分群图（又称谱系图）用它把所有样品（或指标）间的亲疏关系表示出来。

运用聚类分析方法进行分类，不像判别分析那样，必须事先知道各种判别的类型和数目，并且要有一批来自各类的样品，才能建立判别函数对未知属性的样品判别和归类，对于聚类分析来说，一批给定样品划分的类型和数目事先都不知道，正好是需要通过聚类分析以后结合地质解释来确定的问题。

一般，根据分类对象的不同又有Q型群分析（对样品分类）和R型群分析（对指标分类）两种类型。

在地质工作中，分类问题很多，如岩石分类、矿物分类、地层分类、化石分类等等。应用聚类分析的方法可以突破传统地质学所建立的一些定性分类系统，形成一些定量的分类关系。譬如，利用聚类分析可以：

(1) 了解成矿元素究竟和那些地质因素有关，如Cr究竟和那些元素的形成有关。

(2) 找出和稀有元素伴生及相关的元素，以利用和稀有元素关系密切的普通元素作为稀有元素的找矿标志，从而帮助解决一些稀有元素的找矿问题。

(3) 研究在次生分散晕中异常元素究竟和那些元素共生组合在一起。

(4) 在评价测区化探异常时，通过对样品进行分类区分不同类型的异常，如与某类工业矿床有关的异常，无矿异常等等。

(5) 通过对样品（或观测点）进行分类，作为划分地层、岩相、各种地质构造带及侵入体界限的基础。

§ 2 聚类分析的内容和方法

2.1 分类统计量

设有 N 个样品，每个样品测得 p 项指标（变量），我们把每个样品看成 p 维空间中的一个向量； $X_i = [x_{i1} \cdots x_{ip}]'$ $i = 1 \cdots N$ 。这样， N 个样品可以排成一个矩阵，即

$$(1) X = \begin{matrix} & X_1 & X_2 & \cdots & X_N \\ \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{matrix} & \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p1} & x_{p2} & \cdots & x_{pN} \end{bmatrix} \end{matrix}$$

其中 x_{ij} ($i = 1, \cdots, p$; $j = 1, \cdots, N$) 为第 j 个样品的第 i 个指标的观测数据， X 即为原始资料矩阵，第 j 个样品 X_j 为矩阵 X 的第 j 列所描述，第 i 个变量 x_i 为矩阵 X 的第 i 行所描述，所以，任两个样品 X_j 与 X_k 之间的相似性可以通过矩阵 X 中的第 j 列与第 k 列的相似程度来刻画；任两个变量 x_j 与 x_k 之间的相似性可以通过矩阵 X 中的第 j 行与第 k 行的相似程度来刻画。

我们既然要定异地进行分类，首先必须确定一些划分类型的度量指标，即引进一些能够表示样品（或指标）间相似程度的度量，这些

数易叫做分类统计量。

1. Q型聚类分析(对样品进行分类)常用的统计量:

(1) 距离系数

如果把 N 个样品(X 中 N 个列)看成 p 维空间中 N 个点,则两个样品之间相似程度可用 p 维空间中两点的距离来度量,这种距离实际是Mahalanobis距离,但是如果标准化变量 $x_1, \dots, x_p$ 互不相关,那末马氏距离也就是通常的欧氏距离,这时样品 X_j 与 X_k 之间的距离为:

$$d_{jk} = \sqrt{\sum_{\alpha=1}^p (x_{\alpha j} - x_{\alpha k})^2}$$

有时,为使 d_{jk} 确定在某范围内变化,通常取定一个常数 C ,采用公式

$$(2) d_{jk} = \sqrt{\frac{1}{C} \sum_{\alpha=1}^p (x_{\alpha j} - x_{\alpha k})^2}$$

计算任意二个样品 X_j 与 X_k 之间的距离, d 值越小表示两个样品相似程度越大, d 值越大表示两个样品相似程度越小,如果把任何两个样品的距离都算出后,可排成距离系数矩阵

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ \dots & \dots & \dots & \dots \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{bmatrix} \quad \text{其中 } d_{11} = d_{22} = \dots = d_{NN} = 0$$

它是一个对称矩阵,所以只须计算出上三角形部分或下三角形部分即可,根据 D 可对 N 个点进行分类,距离近的点归为一类,距离远的点属于不同的类。

为了减少计算结果所产生的偏倚,当变量 $x_1, \dots, x_p$ 彼此相关时最好采用主成分分析的方法(第七章)找出几个正交因子,用它们代

替原始变量计标距阵D

(ii) 相似系数(夹角余弦)

把任何二个样品 Σ_j, Σ_k 看成 p 维空间的两个向量, 这两个向量的夹角余弦(即相似系数)用 $\cos Q_{jk}$ 来表示, 即:

$$\Sigma_j = [x_{1j} \cdots x_{pj}]'$$

$$\Sigma_k = [x_{1k} \cdots x_{pk}]'$$

$$(3) \cos Q_{jk} = \frac{\Sigma_j \cdot \Sigma_k}{|\Sigma_j| |\Sigma_k|} =$$

$$\frac{\sum_{\alpha=1}^p x_{\alpha j} x_{\alpha k}}{\sqrt{\sum_{\alpha=1}^p x_{\alpha j}^2 \cdot \sum_{\alpha=1}^p x_{\alpha k}^2}}$$

$-1 \leq \cos Q_{jk} \leq 1$, $\cos Q_{jk} = 1$ 说明 Σ_j 与 Σ_k 完全相似;
 $\cos Q_{jk}$ 接近 1, 说明两个样品 Σ_j 与 Σ_k 相似密切。 $\cos Q_{jk} = 0$
 说明两个样品 Σ_j 与 Σ_k 完全不一样; $\cos Q_{jk}$ 接近 0, 说明两个样品 Σ_j 与 Σ_k 差别大。

把所有两两样品的相似系数 $\cos Q_{jk}$ ($j, k = 1, \cdots, N$) 都标出后, 将它们排成一个相似系数矩阵。

$$Q = [q_{jk}] = \begin{bmatrix} \cos Q_{11} & \cos Q_{12} & \cdots & \cos Q_{1N} \\ \cos Q_{21} & \cos Q_{22} & \cdots & \cos Q_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \cos Q_{N1} & \cos Q_{N2} & \cdots & \cos Q_{NN} \end{bmatrix}$$

其中 $\cos Q_{11} = \cos Q_{22} = \cdots = \cos Q_{NN} = 1$ 这是一个实的, 对称的矩阵, 所以只须计标上三角形部份或下三角形部分, 根据 Q 可对 N 个样品进行分类, 把比较相似的样品归为一类, 不怎么相似的样品属于不同的类。

(iii) 相关系数

第 j 个样品与第 K 个样品之间的相关系数可规定为:

$$(4) \quad \tilde{r}_{jk} = \frac{\sum_{\alpha=1}^p (x_{\alpha j} - \tilde{x}_j)(x_{\alpha k} - \tilde{x}_k)}{\sqrt{\sum_{\alpha=1}^p (x_{\alpha j} - \tilde{x}_j)^2 \cdot \sum_{\alpha=1}^p (x_{\alpha k} - \tilde{x}_k)^2}}$$

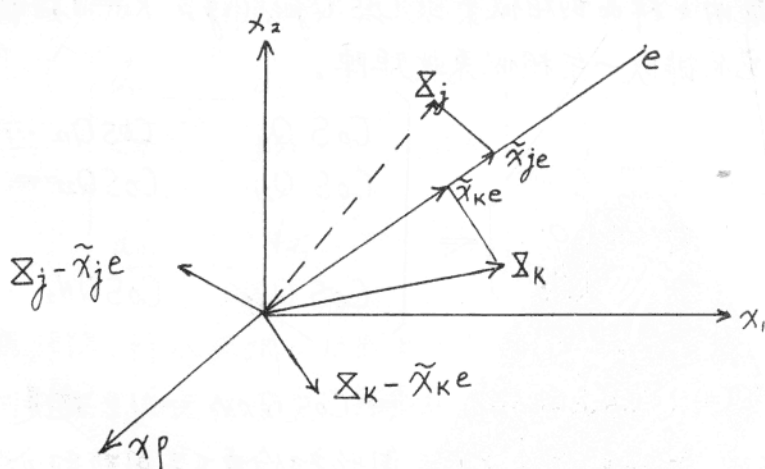
式中,

$$\tilde{x}_j = \frac{1}{p} \sum_{\alpha=1}^p x_{\alpha j}, \quad \tilde{x}_k = \frac{1}{p} \sum_{\alpha=1}^p x_{\alpha k}$$

由(4)可以看出, $\tilde{r}_{jk}$ 实际上是两个向量 $\tilde{\Sigma}_j - \tilde{\Sigma}_j$ ($\tilde{\Sigma}_j = [\tilde{x}_j, \dots, \tilde{x}_j]$) 与 $\tilde{\Sigma}_k - \tilde{\Sigma}_k$ ($\tilde{\Sigma}_k = [\tilde{x}_k, \dots, \tilde{x}_k]$) 的夹角余弦, 当 $\tilde{\Sigma}_j = \tilde{\Sigma}_k = 0$ 时 $\tilde{r}_{jk} = \cos \theta_{jk}$, $-1 \leq \tilde{r}_{jk} \leq 1$.

为了进一步解释 $\tilde{\Sigma}_j - \tilde{\Sigma}_j$ 及 $\tilde{\Sigma}_k - \tilde{\Sigma}_k$ 的几何意义, 我们引进等角线, 这就是通过原点与 $e = [1, \dots, 1]$ 这点的直线 (图 6-1) 此时, 向量 $\tilde{\Sigma}_j$ 在向量 $e = [1, \dots, 1]$ 上的投影向量为:

$$\left(\frac{\tilde{\Sigma}_j' e}{e' e} \right) e = \left(\frac{\sum_{\alpha=1}^p x_{\alpha j}}{\sum_{\alpha=1}^p 1} \right) e = \tilde{x}_j e = [\tilde{x}_j, \tilde{x}_j, \dots, \tilde{x}_j]' = \tilde{\Sigma}_j$$



(图 6-1)

如果将 $\bar{x}_j$ 分解为 $\tilde{x}_j$ 和 $\bar{x}_j - \tilde{x}_j$, 那么前者是 $\bar{x}_j$ 在 e 上的投影向量, 后者就是 $\bar{x}_j$ 在垂直于等角线的平面上的投影向量, 它的长度为 $[\bar{x}_j - \tilde{x}_j][\bar{x}_j - \tilde{x}_j] = \sum_{\alpha=1}^p (\alpha_{\alpha j} - \tilde{\alpha}_j)^2$; 完全类似, $\bar{x}_k - \tilde{x}_k$ 是 $\bar{x}_k$ 在垂直于等角线平面上的投影向量, 它的长度为 $\sum_{\alpha=1}^p (\alpha_{\alpha k} - \tilde{\alpha}_k)^2$ 将 $\bar{x}_j - \tilde{x}_j$ 及 $\bar{x}_k - \tilde{x}_k$ 平移使其一端都落到坐标原点, 则可看出由 (4) 所定义的样品相关系数 $\tilde{r}_{jk}$ 就是投影向量 $\bar{x}_j - \tilde{x}_j$ 与 $\bar{x}_k - \tilde{x}_k$ 的夹角余弦。

如果把两々样品的相关系数都标出后, 可排队样品相关系数矩阵

$$\tilde{R} = \begin{bmatrix} \tilde{r}_{11} & \tilde{r}_{12} & \cdots & \tilde{r}_{1N} \\ \tilde{r}_{21} & \tilde{r}_{22} & \cdots & \tilde{r}_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \tilde{r}_{N1} & \tilde{r}_{N2} & \cdots & \tilde{r}_{NN} \end{bmatrix}$$

其中 $\tilde{r}_{11} = \tilde{r}_{22} = \cdots = \tilde{r}_{NN} = 1$, 根据 $\tilde{R}$ 可对 N 个样品进行分类。

2. R型聚类分析(对指标进行分类)常用的统计量,

对于 p 个指标(变量)之间相似性的定义, 与样品相似性相仿, 此时是在 N 维空间中来研究, N 个坐标轴分别是 N 次观测, 变量之间相似性可以通过原始资料矩阵 X 中 p 行间相似关系来研究。这时的真或向量是 $(x_{j1}, \cdots, x_{jN})$ 与 $(x_{k1}, \cdots, x_{kN})$ 。衡量它们相似关系的统计量仍然可定义:

(i) 距离系数

$$(5) \quad d_{jk}^* = \sqrt{\frac{1}{C} \sum_{\beta=1}^N (x_{j\beta} - x_{k\beta})^2}$$

其中 C 为一常数, 把所有两々行之间距离系数标出后可排队距离系数矩阵

$$D^* = \begin{bmatrix} d_{11}^* & d_{12}^* & \dots & d_{1p}^* \\ d_{21}^* & d_{22}^* & & d_{2p}^* \\ \dots & \dots & \dots & \dots \\ d_{p1}^* & d_{p2}^* & \dots & d_{pp}^* \end{bmatrix}$$

根据 D^* 可对 p 个变量进行分类,

(ii) 相似系数 (夹角余弦)

$$(b) \cos Q_{jk}^* = \frac{\sum_{\beta=1}^N x_{j\beta} \cdot x_{k\beta}}{\sqrt{\sum_{\beta=1}^N x_{j\beta}^2 \cdot \sum_{\beta=1}^N x_{k\beta}^2}} \quad -1 \leq \cos Q_{jk}^* \leq 1$$

把 D 中两行间相似系数求出后可排列相似系数 (夹角余弦) 矩阵:

$$Q^* = [q_{jk}^*] = \begin{bmatrix} \cos Q_{11}^* & \dots & \cos Q_{1p}^* \\ \cos Q_{21}^* & \dots & \cos Q_{2p}^* \\ \dots & \dots & \dots \\ \cos Q_{p1}^* & \dots & \cos Q_{pp}^* \end{bmatrix}$$

根据它可对 p 个变量进行分类.

(iii) 相关系数

通常我们所说相关系数, 一般即指变量间相关系数, 作为刻画变量间相似关系, 它是最常用的.

$$(c) r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} \cdot S_{kk}}} = \frac{\sum_{\beta=1}^N (x_{j\beta} - \bar{x}_j)(x_{k\beta} - \bar{x}_k)}{\sqrt{\sum_{\beta=1}^N (x_{j\beta} - \bar{x}_j)^2 \sum_{\beta=1}^N (x_{k\beta} - \bar{x}_k)^2}}$$

其中 $\bar{x}_j = \frac{1}{N} \sum_{\beta=1}^N x_{j\beta}$, $\bar{x}_k = \frac{1}{N} \sum_{\beta=1}^N x_{k\beta}$ 分别是第 j 和第 k 个变量的样本平均值, $-1 \leq r_{jk} \leq 1$. 当把两变量间的相关系数都求

出后，可排成相关系数矩阵。

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{bmatrix}$$

其中 $r_{11} = r_{22} = \dots = r_{pp} = 1$ 。根据 R 可对 p 个变量进行分类。

2.2 数据变换

由于所研究的各个变量的单位和等级可能是不一样的，即便有些变量的度量一样，但各变量的绝对值大小也不一样，因此，直接用原始数据进行计算就会突出那些绝对值大的变量而压低了那些绝对值小的变量的作用，因此一般在计算前需对原始数据进行变换处理，常用的变换方法有下面几种：

(1) 标准差标准化

对于第 i 个变量进行标准化就是将 x_{ij} 变换为

$$x'_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i} \quad i = 1, \dots, p; \quad j = 1, \dots, N$$

$$\bar{x}_i = \frac{1}{N} \sum_{j=1}^N x_{ij} \quad s_i = \sqrt{\frac{1}{N-1} \sum_{j=1}^N (x_{ij} - \bar{x}_i)^2}$$

通过变换以后数据 $\{x_{ij}\}$ 中每个变量的样本标准差都化为 1，与变量的等级就没有关系了。

(2) 极差标准化

将 x_{ij} 变换为

$$x'_{ij} = \frac{x_{ij} - \min\{x_{ij}\}}{\max\{x_{ij}\} - \min\{x_{ij}\}} \quad 1 \leq j \leq N$$

这时 $\{x'_{ij}\}$ 每个变量的样本极差都化为 1，也排除了另纲的干扰。

3) 极差正规化

将 x_{ij} 变换为

$$x'_{ij} = \frac{x_{ij} - \min_{1 \leq j \leq N} \{x_{ij}\}}{\max_{1 \leq j \leq N} \{x_{ij}\} - \min_{1 \leq j \leq N} \{x_{ij}\}}$$

这时 $\{x'_{ij}\}$ 的极差为 1，消除了另纲的干扰；且 $0 \leq x'_{ij} \leq 1$

($i = 1, \dots, P; j = 1, \dots, N$).

2.3 形成分群图

在已计算出相似性矩阵以后，可以它为依据，对样品（或指标）进行聚合归类，最后形成分群图，其聚合归类一般遵以下面四条原则，

(1) 若选出一对样品在已经分好的组中都未出现过，则把它们形成一个独立的新组。

(2) 若选出两个样品中，有一个是在已经分好的组中出现过，则把另一个样品也加入到该组中。

(3) 若选出一对样品，都分别出现在已经分好的两组中，则把这两个组联结在一起。

(4) 若选出的的一对样品都出现在同一组中，则这对样品就不用再分组了。

按上述四条原则反复进行，直到把所有样品都分类完毕为止。

1. 一次形成法

例 1. 对某地超基性岩的一批样品，经光谱分析得与矿化有关的某些元素（指标），其资料数据如下表，试用 Q 型聚类分析对六个样品进行分类，

样品 编号	样 品 组 队	岩 性 特 征	元 素 含 量 (指 标)					
			Ni	Co	Cu	Cr	S	As
1($\bar{x}_1$)	A孔 6个样品平均	蛇纹岩 A组矿化	1903	273	100	1178	8163	4
2($\bar{x}_2$)	B孔 11个样品平均	蛇纹岩 无矿化	2328	79	6	3175	586	14
3($\bar{x}_3$)	C孔 11个样品平均	蛇纹岩 无矿化	744	26	1	841	425	3
4($\bar{x}_4$)	D孔 12个样品平均	滑美岩 B组矿化	2782	273	150	2400	8234	37
5($\bar{x}_5$)	E孔 10个样品平均	滑美岩 无矿化	1775	94	13	3140	54	1
6($\bar{x}_6$)	F孔 12个样品平均	滑美岩 无矿化	104	44	6	2093	104	4

(1) 对称变换: (表 6-1) 中各指标数据数量级相差很大, 我们将原始数据取其对称值进行变换得:

(表 6-2)

样 品 \ 指 标	Ni	Co	Cu	Cr	S	As
1 ($\bar{x}_1$)	3.2794	2.4362	2.2041	3.0711	3.9119	0.6021
2 ($\bar{x}_2$)	3.3670	1.8976	0.7782	3.5018	2.7679	1.1461
3 ($\bar{x}_3$)	2.8716	1.4150	0	2.9248	2.6284	0.4771
4 ($\bar{x}_4$)	3.4443	2.4362	2.1761	3.3802	3.9156	1.5682
5 ($\bar{x}_5$)	3.2492	1.9731	1.1139	3.4969	1.7324	0
6 ($\bar{x}_6$)	3.0195	1.6435	0.7782	3.3207	2.0170	0.6021

(2) 相似系数矩阵: 由(表 6-2)数据按照(3)计算的两样品的相似系数(夹角余弦)得相似系数矩阵

$$Q = [\cos Q_{ij}] = [q_{ij}] \text{ 如下:}$$

$$Q = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} & \left[\begin{array}{cccccc} 1.00 & & & & & \\ 0.9624 & 1.00 & & & & \\ 0.9406 & 0.9848 & 1.00 & & & \\ 0.9907 & 0.9715 & 0.9352 & 1.00 & & \\ 0.9357 & 0.9664 & 0.9499 & 0.9256 & 1.00 & \\ 0.9472 & 0.9806 & 0.9623 & 0.9516 & 0.9698 & 1.00 \end{array} \right] \end{matrix}$$

(3) 用一次形成法形成分群图 (据矩阵Q对样品进行分群)

(i) 记下Q中的最大值 $g_{14} = 0.9907$. 划去矩阵的第4行第4列。

(ii) 记下Q中剩余元素的最大值 $g_{23} = 0.9848$ 划去矩阵中第3行第3列。

(iii) 记下Q中剩余元素的最大值 $g_{56} = 0.9698$. 划去矩阵中第6行第6列。

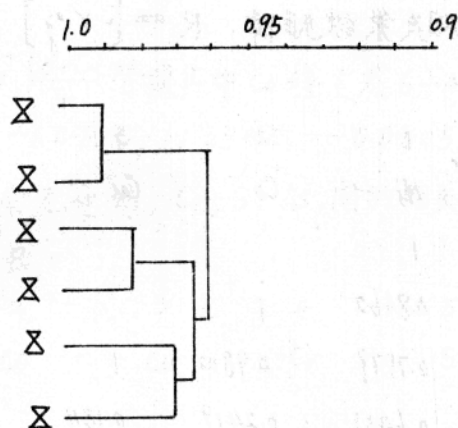
(iv) 记下Q中剩余元素的最大值 $g_{25} = 0.9664$. 划去矩阵中第5行第5列。

(v) 记下Q中剩余元素的最大值 $g_{21} = 0.9624$. 划去矩阵中第2行第2列。

总结以上步骤可列表如下 (表6-3)

连结顺序	连 结 样 品		相似系数
1	$\bar{x}_1$	$\bar{x}_4$	0.9907
2	$\bar{x}_2$	$\bar{x}_3$	0.9848
3	$\bar{x}_5$	$\bar{x}_6$	0.9698
4	$\bar{x}_2 \bar{x}_3$	$\bar{x}_5 \bar{x}_6$	0.9664
5	$\bar{x}_1 \bar{x}_4$	$\bar{x}_2 \bar{x}_3 \bar{x}_5 \bar{x}_6$	0.9624

据(表6-3)可作出对于六个样品的Q型一次划分分群图如下:



(图6-2)

(4) 成果解释: 由分群(图6-2)可以看出六个组合样品间关系都比较密切(相似系数在0.9以上), 而它们相互的聚合关系又能客观地反映岩体的性质, 把有矿化的 X_1, X_2 聚合为一群, 把无矿化的 X_3, X_4, X_5, X_6 聚合为另外一群。进一步, 又按岩性特征把蛭纹岩(X_2, X_3)和滑美岩(X_5, X_6)分开成两小群。从分群图上我们可以直观地看到各样品的亲疏聚合关系。

例2. 对(表6-1)中的各指标进行R型聚类分析以研究各指标间的相关关系。

(1) 运用(表6-2)对各指标进行标准化得:

(表6-4)

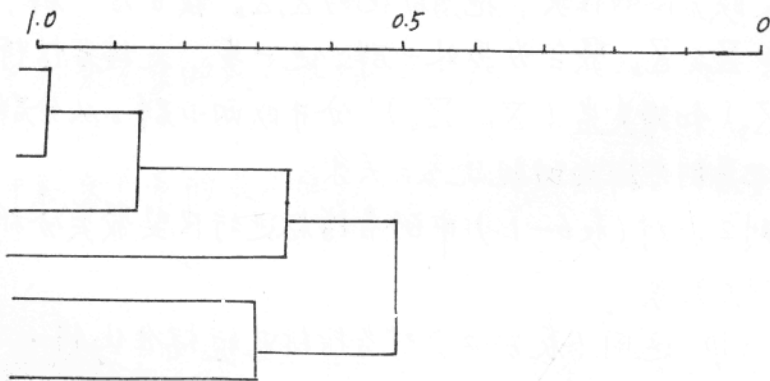
样 品 \ 指 标	Ni	Co	Cu	Cr	S	As
1 (X_1)	0.3738	1.2438	1.2999	0.9828	1.3007	-0.2605
2 (X_2)	0.8151	-0.1867	-0.5014	1.0185	-0.0733	0.8255
3 (X_3)	-1.6806	-1.4628	-1.4845	-1.6626	-0.2408	-0.5101
4 (X_4)	1.2045	1.2438	1.2645	-0.4535	1.3052	1.6682
5 (X_5)	0.2217	0.0164	-0.0773	0.9958	-1.3170	-1.4626
6 (X_6)	-0.9355	-0.8517	-0.5014	0.1770	-0.9751	-0.2605

(2) 相关系数矩阵 由(表 6-1)数据, 按照(7)计算两两变量的相关系数, 得相关系数矩阵 $R = [r_{ij}]$

如下:

$$R = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \text{ Ni} \\ 2 \text{ Co} \\ 3 \text{ Cu} \\ 4 \text{ Cr} \\ 5 \text{ S} \\ 6 \text{ As} \end{matrix} & \begin{bmatrix} \text{Ni} & \text{Co} & \text{Cu} & \text{Cr} & \text{S} & \text{As} \\ 1 & & & & & \\ 0.8462 & 1 & & & & \\ 0.7579 & 0.9802 & 1 & & & \\ 0.6431 & 0.2419 & 0.1811 & 1 & & \\ 0.5039 & 0.7370 & 0.7210 & -0.3075 & 1 & \\ 0.5603 & 0.4241 & 0.3930 & 0.1998 & 0.6802 & 1 \end{bmatrix} \end{matrix}$$

(3) 用一次划分法形成六个指标的分群图(如例 1.3)



(图 6-3)

一次划分法比较简便, 只要计算出相关矩阵以后, 一次即可形成分群图, 显示出变量的聚合情况, 但是, 一次形成可能把不相关的变量都放到同一个群理, 以后也就不能剔除。Cr 因为与 Ni 的关系而被聚集在 Co、Cu 这一群里, 但事实并不如此, 比较准确的划分常采用下面的逐步形成法。

2. 逐步形成法

下面, 我们对例 2 运用逐步形成法来作出其分群图。
在相关矩阵 R 中:

(i) 记下最大值 $r_{23} = 0.9802$ ($Co \sim Cu$ 的相关系数), 划掉 R 中第三行第三列。

(ii) 用 (表 6-4) 中 $\frac{Co+Cu}{2}$ 替换其中 Co 得 (表 6-4) 中 Co' 即:
 $Co' = (1.2719, -0.3426, -1.4737, 1.2542, -0.0305, -0.6793)$

(iii) 重新计算 Co' 与其余元素 Ni, Cr, S, As 间的相关系数, 用来替换 R 中第二行第二列得到 R_1 :

$$R_1 = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \ Ni \\ 2 \ Co' \\ 3 \ Cu \\ 4 \ Cr \\ 5 \ S \\ 6 \ As \end{matrix} & \left[\begin{array}{cccccc} Ni & Co & Cu & Cr & S & As \\ 1 & & & & & \\ 0.8019 & 1 & & & & \\ \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ 0.6431 & 0.2116 & & & & \\ 0.5039 & 0.7290 & & -0.3075 & 1 & \\ 0.5603 & 0.4086 & & 0.1998 & 0.6802 & 1 \end{array} \right] \end{matrix}$$

(i) 记下 R_1 中最大值 $r_{12} = 0.8019$ ($Co' \sim Ni$ 的相关系数) 划去 R_1 中第二行第二列。

(ii) 用 $Ni' = \frac{2 \times Co' + Ni}{3}$ 替换 (表 6-4) 中 $Ni' = (0.9725, 0.0433, -1.5427, 1.2376, 0.0536, -0.7647)$

(iii) 计算 Ni' 与剩余元素 Cr, S, As 的相关系数以代替 R_1 中第一行第一列得 R_2 :

$$R_2 = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \ Ni' \\ 2 \ Co' \\ 3 \ Cu \\ 4 \ Cr \\ 5 \ S \\ 6 \ As \end{matrix} & \left[\begin{array}{cccccc} Ni' & Co' & Cu & Cr & S & As \\ 1 & & & & & \\ \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ 0.3554 & & & 1 & & \\ 0.6539 & & & -0.3075 & 1 & \\ 0.4591 & & & 0.1998 & 0.6802 & 1 \end{array} \right] \end{matrix}$$

(i) 记下 R_2 中最大值 $r_{56} = 0.6802$ ($S \sim As$ 相关系数)。划去 R_2 中第六行第六列。

(ii) 用 $S' = \frac{S+As}{2}$ 替换(表6-4)中 S 即: $S' = (1.5201, 0.3761, -0.3755, 1.4867, -1.3898, -0.6178)$

(iii) 计算 S' 与 Ni', Cr 间的相关系数, 代替 R_2 中第五行第五列得 R_3 :

	1	2	3	4	5	6
	Ni'	Co'	Cu	Cr	S'	As
1 Ni'	1					
2 Co'						
$R_3 = 3 Cu$						
4 Cr	0.3554			1		
5 S'	0.5565			-0.0538	1	
6 As						

(i) 记下 R_3 中最大值 $r_{15} = 0.5565$ (Ni' 与 S' 的相关系数) 划去 R_3 中第5行第5列。

(ii) 用 $Ni'' = \frac{3 \times Ni' + 2 \times S'}{5}$ 替换(表6-4)中 Ni' 即: $Ni'' = (0.7915, 0.1764, -1.0758, 1.3372, -0.5234, -0.7039)$

(iii) 最后计算 Ni'' 与 Cr 的相关系数为 0.1918, 用它替换 R_3 中第一列而得 R_4 :

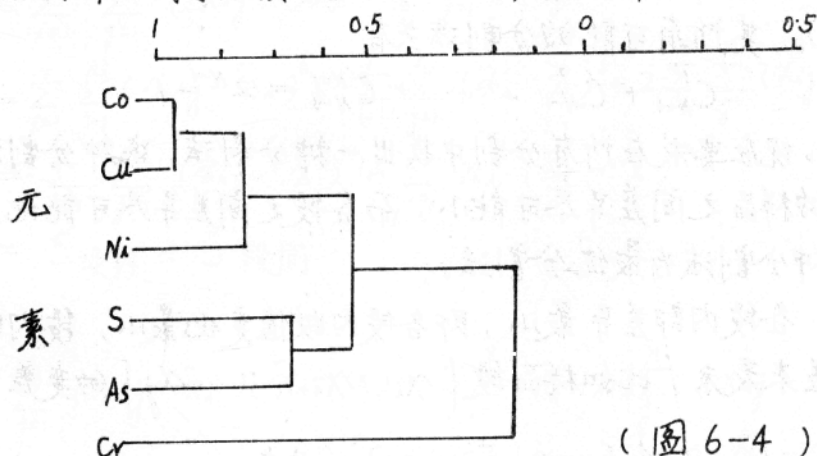
	1	2	3	4	5	6
	Ni''	Co'	Cu	Cr	S'	As
Ni''						
Co'						
$R_4 = Cu$						
Cr	0.1918			1		
S'						
As						

4. 总结列表作图

(表6-5)

连结顺序	连 结 元 素		相关系数
1	Co	Cu	0.9802
2	Ni	Co Cu	0.8019
3	S	As	0.6802
4	S As	Ni Co Cu	0.5565
5	Cr	Ni Co Cu S As	0.1918

根据列表可作出对于元素(指标)的R型聚类分析的分群图如下:



由分群图可以看出,在相关系数0.2~0.5的水平上,可将元素分为二群,一群是亲S的Co, Cu, Ni, S, As,另一是单独的一个元素——亲氧元素Cr;在相关系数为0.6左右的水平上又可将五元素分为二组,一组是亲S阳离子Co, Cu, Ni,另一组是阴离子S, As,同时还可看出,在Co, Cu, Ni三元素中Co~Cu之间关系更为密切,上述对元素的R型聚类分析很好地揭示了这些元素在超基性岩中的相互关系。

§3. 有序样品的聚类——最优分割法

3.1 什么是最优分割

这是对一批有序样本进行划分分段的一种统计分析方法。这类