第13卷第3期　　石　油　实　验　地　质　　Vol.13, No.3
1991年9月　　EXPERIMENTAL PETROLEUM GEOLOGY　　Sept., 1991

储层评价中的大样本聚类方法

杨德永　　李　超

（长庆石油勘探开发研究院，甘肃庆阳）

本文介绍了普通聚类分析和大样本聚类分析方法，並对谱系图绘制问题、克服重心反转问题、变量选择及结果分析问题进行了探讨.

在陕甘宁盆地中生界储层评价中，为了深入研究不同沉积类型不同含油层系的孔隙结构特征，对全区储层进行分类评价，以寻找较好的储层发育区，在全盆地范围内选取了819个样品，每个样品若干项参数（包括渗透率、孔隙度、排驱压力、喉道均值、孔容等）作聚类分析。从而使我们有机会对大样本聚类方法进行了一些探索，并在实践中加深了认识。

一、普通聚类分析回顾

聚类分析是研究样品或变量之间分类问题的一种多元统计分析方法。一批样品或一组变量之间存在着程度不同的相似性。我们可以根据一批样品的多个观测指标找出一些能够度量样品或变量之间相似程度的统计量，以这些统计量为分类依据，把一些相似程度较大的样品或变量聚合为一类，把另一些彼此之间相似程度较大的样品或变量聚合为另外一类。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变量都聚合完毕，从而形成一个由小到大由亲到疏的分类系统。

限于篇幅，本文只谈样品分类，即Q型聚类分析问题，变量分类即R型聚类分析问题方法与此完全类似。

设有N个样品，每个样品测得M个变量的值，原始观测数据矩阵为

$$X=(x_{ij})=\begin{pmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1m} \\ x_{21} & x_{22} & \cdots\cdots & x_{2m} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ x_{n1} & x_{n2} & \cdots\cdots & x_{nm} \end{pmatrix} \qquad (1)$$

根据需要可对原始数据矩阵X的数据进行预处理，即作标准化或正规化处理。

为了把样品按照彼此相似程度进行分类，首先要求出表示样品间相似程度的统计量。表示样品之间相似程度的统计量可分为两大类：

一是把每个样品看成多维空间中的一个点，用点与点之间的距离（如欧氏距离）表

示样品间的亲疏关系，距离系数越小表示关系越密切。欧氏距离系数的计算公式为

$$d_{ij}=\sqrt{\frac{1}{M}\sum_{k=1}^{M}(x_{ik}-x_{jk})^2} \qquad (2)$$

$$i,\ j=1,\ 2,\ \cdots\cdots N$$

二是用样品间的相似系数（如向量之间的夹角余弦）表示，相似系数越大关系越密切。相似系数（夹角余弦）计算公式为

$$r_{ij}=\frac{\sum_{k=1}^{M}x_{ik}\cdot x_{jk}}{\sqrt{\sum_{k=1}^{M}x_{ik}^2\cdot\sum_{k=1}^{M}x_{jk}^2}} \qquad (3)$$

$$i,j=1,2\cdots\cdots N$$

在聚合归类过程中，都是把最相似的（如距离系数最小的）样品或样品组聚合为一类，并且一般遵从下述原则：

1.若选出的一对样品在已经分好的类中都未出现过，则形成一个独立的新类。

2.若选出的一对样品中，有一个在已经分出的类中出现过，则把另一个样品也加入到该类中。

3.若选出的一对样品，都分别出现在已经分好的两类中，则把这两类归并为一类。

普通聚类分析的具体作法如下：

首先用公式(2)计算出距离系数矩阵D（或用公式(3)计算出相似系数矩阵R，方法类似），然后进行逐步聚类连结。连结方法有多种，通常采用重心法计算连结。其步骤是：

1.开始时将每个样品看成一类，全部样品共有N类，每类中样品个数为1。设置数组ID，指示每类中的样品个数。此时置ID（i）=1，类号i=1，2，……N。

2.在矩阵D的上三角阵中，找出最小值d_{pq}，在上三角阵中必有p<q。

3.把p、q二类（第一次是p、q二样品）合并，这时要作三件事：

1）新类号记为p，把类号q划掉，记下距离系数d_{pq}，且类数减1。

2）把挑出的成对样品或样品组的相应变量加权平均，形成新类的代表性样品

$$X_p'=\frac{n_pX_p+n_qX_q}{n_p+n_q} \qquad (4)$$

用以更新原始数据X矩阵(1)的第P行数据，其中n_p、n_q分别为相连结的样品组中的样品个数，$n_p=ID(p)$，$n_q=ID(q)$，x_p、x_q为相应变量的数据。新类p的样品数更新为$ID(p)=n_p+n_q$。

3）重新计算新类p与其它类间的距离系数，用以取代矩阵D中的第P行和第p列元素，并划去第q行和第q列元素，设置标志ID（q）=0。矩阵D的第p行元素为

$$d_{pj}=\sqrt{\frac{1}{M}\sum_{k=1}^{M}(x_{pk}-x_{jk})^2} \qquad (5)$$

$$j=p+1,p+2,\cdots\cdots N;\ ID(j)\neq 0$$

第p列元素为

$$d_{ip}=\sqrt{\frac{1}{M}\sum_{k=1}^{M}(x_{ik}-x_{pk})^2} \qquad (6)$$
$$i=1,2,\cdots\cdots p-1;\ ID(i)\neq 0$$

4.重复2、3步，并进行N－1次，最后就把全部样品聚为一类。

然后对逐次计算连结的结果进行排队，并打印出聚类谱系图。

二、大样本聚类方法

当样品数较大时，一般的计算机特别是微型机、小型机均因内存太小而无法进行普通聚类分析计算。以801个样品为例，作普通Q型聚类分析时需计算距离系数矩阵D。矩阵D是实对称矩阵，其上三角阵为

$$D_{上}=(d_{ij})=\begin{pmatrix} 0 & d_{12} & d_{13} & d_{14}\cdots\cdots & d_{1n} \\ & 0 & d_{23} & d_{24}\cdots\cdots & d_{2n} \\ & & 0 & d_{34}\cdots\cdots & d_{3n} \\ & & & \cdots\cdots\cdots\cdots & \\ & & & 0 & d_{n-1\,n} \\ & & & & 0 \end{pmatrix} \qquad (7)$$

为了存放这个上三角阵，需要在内存中建立一个大小为$\frac{1}{2}N(N-1)=\frac{1}{2}801\cdot 800=320400$的实型数组。一个实型数占用4个字节，仅此一个数组就需要1281600个字节，大大超过一般微机的全部内存大小。因此，普遍聚类法是行不通的，必须采用特殊的方法才有可能进行大样本聚类分析。

与普通聚类分析方法类似，开始时同样把每个样品看成一类，全部样品共N类，并设置数组ID(i)=1，i=1，2，……N。

用距离系数公式(2)计算样品间的距离系数矩阵D的行极小向量A及与行极小值对应的列号向量L

$$A=\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \end{pmatrix} \qquad L=\begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_{n-1} \end{pmatrix} \qquad (8)$$

其中　$a_i=d_{ik}=\min\{d_{ij}\}$

$l_i=k$

$i=1,2,\cdots\cdots n-1;\quad i<j\leqslant n$

在内存中保留向量A、L，不保留D矩阵，仅用2(n－1)个单元。

然后按照聚合准则，进行逐次计算连结，其步骤为：

1.在向量A中求极小值$a_p=\min\{a_i\}$及相应的列号$q=L_p$。

2.将p、q二类（第一次是p、q二样品）连结为一类，并以组合样x_p'（用公式(4)计算）作为新类p的代表性样品，以代替X矩阵(1)的第p行数据，新群p的样品数为ID（p）$=N_p+N_q$，划去X矩阵的第q行元素，置标志ID（q）$=0$。

3.重新计算向量A和向量L。这里需要重新计算的是a_1，a_2，……a_{q-1}中的某些值，即ID（i）≠0时对应的a_i，$i=1,2,\cdots\cdots q-1$。具体而言

1）用公式(5)计算D矩阵中上三角的第p行元素d_{pj}，并求出极小值a_p'。

2）用公式(6)计算D矩阵中上三角的第p列元素d_{ip}，并与对应的a_i比较求出极小值a'_i。

普通聚类法工作时，D矩阵在内存中，可以很容易在其中找出最小值d_{pq}。大样本聚类就不行了。组合样参加计算后，A向量和L向量中的有关的一些值就会发生变化。因而还需要计算。

3）当$L_j=p$或$L_i=q$时，重新计算D矩阵上三角的第i行元素

$$d_{ij}=\sqrt{\frac{1}{M}\sum_{k=1}^{M}(x_{ik}-x_{jk})^2} \qquad (9)$$

$L_i=p,q$

$j=i+1,i+2,\cdots\cdots N;\quad ID(j)\neq 0$

并求出极小值a'_i。仔细注意这一步的工作实际上包含了第1)步的工作，即当$L_p=q$时，要重新计算第p行元素并求出极小值a_p'的情况。注意到这点，可以简化程序设计。

4.重复1～3步共N－1次，即可得到N个样品从亲到疏的聚类结果。

为了节省计算时间，在计算距离系数d_{ij}时，可以不用公式(2)，而改用公式

$$d_{ij}=\sum_{k=1}^{M}(x_{ik}-x_{jk})^2 \qquad (10)$$

计算，这样省去了除法和开方运算，效果是一样的。

上述这种大样本聚类方法与普通聚类方法比较，不但大大节省内存，而且也不会增加很多计算时间。实践证明，这是一个比较好的、行之有效的方法。

三、几个问题探论

1.谱系图绘制问题

谱系图是聚类分析的结果图件，可以清晰直观地反映出分类系统，同时还可以定量地表达样品间的相似程度。

在普通聚类分析中，要对逐次聚类连结的结果进行排队，即按照作图的顺序重新排列，才能顺利地画出谱系图来。对于大样本聚类分析，仅仅这样作是不够的。由于样品数过多，聚类连结的次数过多，最后形成的谱系图过于庞大，一般的打印纸或绘图纸宽度又有限，因而绘出的谱系图中不少连结线挤在一起，分辨率很低，影响使用。

我们知道，谱系图一般是由叶聚为枝再聚为根，反映出样品间由亲到疏的关系。在逐步聚类过程中，倒数第k次连结时，就将全部样品分为k类，根是由全部样品最后一

次聚类连结的结果。但在实际上，只要是分类问题，人们总希望将一批样品分为若干类，全部样本聚为一类意义并不大。因此在画谱系图时，我们可以不要这个根，而是根据需要灵活地掌握所分类数。比方我们设想把一批样品分为10类，就可以把倒数第10次连结时的距离系数作为最大值来画图。对在其后连结的样品或样品组，在作图时距离系数也取此最大值。这样，画出的谱系图就清晰得多了，分辨率明显提高。

2.关于克服重心反转问题

在逐步聚类过程中，当相似性统计量为距离系数时，从理论上讲，各次计算出的距离系数的值应该单调增加，即第k次连结时算出的距离系数应比前k－1次连结时的距离系数要大，这才符合最亲聚类准则。但是实际上距离系数并非总是单调增加，而是有可能出现摆动，即第k次连结时的距离系数反而小于第k－1次连结时的距离系数。这种情况称为重心反转现象。解决的办法是，一旦出现这种情况，可以取第k次连结时的距离系数等于第k－1次的距离系数。不这样处理，在画谱系图时就会遇到麻烦，会出现类间交叉现象。但是，这种人为修改的因素多了，分类结果的可信度就降低了。产生这种现象的原因是什么？有些人认为是计算误差累积所致。据我们分析，主要原因恐怕是方法本身不够完善。要想从根本上解决这个问题，需要另找出路，例如可以采用误差平方和法。这种方法的主要思想是在聚类的每一阶段，找出这样两类，当二者合并时使对全部各类的组内误差平方和的增大达到极小。误差平方和函数是非降的，因而不会发生反转现象（详见文献1）。

3.变量选择及结果分析

参加聚类分析的变量并不是越多越好。要注意选择那些与所研究的问题有比较密切联系的具有较强分辨力的变量。我们在储层评价工作中，开始时尽量多取得一些分析参数，包括储层物性参数、压汞参数、铸体薄片参数等多达21项。通过回归分析，从中找出影响孔隙结构的若干主要因素，然后再作聚类分析，以便对全区储层进行分类评价。通过聚类分析，可以发现存在于原始数据中的规律性，可以突破传统地质学所建立的定性分类系统，可以对已知现象提出新的解释。

当然，聚类分析的结果不一定完全反映客观规律。聚类分析这种数学分类方法和在综合研究基础上提出的地质学分类方法二者有机结合相互补充，可能更有助于揭示客观存在的规律性，而勘探实践则是对这种规律性的最好检验。

大样本聚类分析程序用FORTRAN77语言写成，在IBM-PC-XT及兼容机上实现。程序提供的相似性统计量有欧氏距离、斜交距离、相似系数等。逐步聚类连结方法有重心法和误差平方和法可供选择。绘制谱系图时可以根据需要灵活掌握所分类数。该程序功能较全，使用方便，在陕甘宁盆地储层评价中发挥了应有作用。

参考文献

〔1〕于崇文等编著，数学地质的方法与应用，冶金工业出版社.

〔2〕王学仁编著，地质数据的多变量统计分析，科学出版社.

〔3〕胡远来等，大样本Q-型聚类分析，成都地质学院建院三十周年论文集.

〔4〕方开泰、潘恩沛著，聚类分析，地质出版社.

CLUSTERING ANALYSIS ON BULK SAMPLES FOR THE RESERVOIR EVALUATION

Yang Deyong　　Li Chao

(Institute of Petroleum Exploration and Development, Qingyang, Gansu)

Abstract

The procedures for conventional and bulk sample clustering analysis are introduced in this paper. Furthermore, the problems concerning lineage plotting, the reverse of gravity centre, selection of variables and result analysis are discussed in detail.