灰色系统聚类分析在工程地质分析中的应用

袁开先

（武汉电力学校）

内　容　提　要

本文应用灰色系统关联度分析的基本原理，借助模糊聚类的方法，提出一个灰色系统的聚类分析方法，用以对工程地质参数信息不完全的系统进行聚类。文中举出了实例，以说明应用效果。

一个系统，若信息（主要指参数信息、系统的内部结构信息、系统与环境之间的相互关系信息）不完全或不确知，这个系统就称为灰色系统。灰色系统理论是华中工学院邓聚龙教授于1982年创立的。该理论包括信息处理、系统分析、系统建模、系统的预测决策和区划等。这一理论在经济、社会、工农业、工程建设等方面有许多重要研究成果，有着广阔的应用前景。

一、问题的提出

在我们对地质体进行研究时，常要借助某些参数指标（包括物理、力学、化学等指标）来说明研究对象的水文地质、工程地质性质，这些参数或者是测试过程中随机因素的干扰，或是地质体本身非均质性影响，或是来自不同精度的报告资料等等，总之由于灰因素的影响而使得这些参数信息不确定或不完全，以致有时不得不用一个“区间数”或“概略值”来表示。在这种情况下若要借助众多的参数指标来比较地质体若干部分（单元）的性质，仅用列出资料表让人们去直观判断的方法显然是不够的。聚类分析给我们提供了一个恰当的分析方法。它用于多因素具有多指标情况下的归纳聚类。例如表1是一个说明我们某些地区红粘土的工程地质性质的资料表，从这张表上是不容易判断出一般粘土与其它红土在工程地质性质上有什么显著差异的，也不易看出各类红粘土在性质上的差异，即回答不了哪些土（因素）性质相近，哪些相差较大这样的问题，这些问题在工程地质分析中常会遇到。

为此，我们可以把研究对象看成是一个参数信息（还可能包含其它信息）不完全确知的灰色系统。把指标值视为具有白化值的灰数；若为区间数则把它视为区间灰数，并认为区间灰数是在某一基本值（本文采取区间中值为基本值）附近变动的，取基本值作为灰数的白化值（如表1所载）。这样，我们就可以此白化值建立数列，进行下面的具体分析。

表1 某些红土的工程地质性质对比表

代号	数列 / 土类	ε		W(%)		γ(g/cm³)		W_L 小于		α_{1-2}(cm²/kg)		C(kg/cm²)		φ(度)	
		区间数	白化值	区间数	白化值	区间数	白化值	区间数	白化值	区间数	白化值	区间数	白化值	区间数	白化值
1	一般粘土	0.75～1.00	0.875	25～40	32.5	1.85～2.05	1.95	40～55	47.5	中～高压缩 0.01～0.1	0.05	0.1～0.6	0.35	8～20	14
2	碳酸岩风化红土	0.80～2.20	1.500	20～60	40.0	1.6～2.0	1.80	50～100	75.0	0.01～0.035	0.0225	0.3～1.50	0.90	12～21	16.5
3	玄武岩风化红土	1.01～1.96	1.485	31～65	48.0	1.6～1.8	1.70	52～70	61.0	0.01～0.035	0.0225	0.4～0.57	0.485	12～21	16.5
4	花岗岩风化红土	0.60～1.60	1.100	25～52	38.5	1.7～1.9	1.80	35～65	50.0	0.02～0.04	0.0300	0.24～0.60	0.42	14～32	23.0
5	碎屑岩风化红土	0.50～1.07	0.785	9～38	23.5	1.6～2.08	1.84	23～61	42.0	0.004～0.024	0.0140	0.02～1.80	0.91	10～55	32.5
6	网纹红土	0.5～0.8	0.650	18～28	23.0	1.9～2.0	1.95	27～54	40.5	0.007～0.028	0.0175	0.23～1.35	0.79	19～33	26.0
7	大连红土	1.01～1.48	1.245	30～50	40.0	1.3～2.2	1.75	46～66	56.0	0.02～0.045	0.0325	0.36～0.38	0.37	17～28	22.5

注：资料摘自《水文地质工程地质》杂志1986年第6期“综合评述红土问题”。其中一般粘土的压缩系数 α_{1-2} 原表中为中～高压缩土；表头代号系为后面聚类分析用的土样代号。

二、方法原理

1.应用灰色系统理论的关联度分析，以建立等价关系矩阵

设有 m个因素，有m个数列：$\{X_K(i)\}$，$(K=1,2,3,\cdots\cdots m;\ i=1,2,3,\cdots\cdots n)$，式中 n 为数列中数据个数。关联度分析就是根据这些因素数列的几何形状、发展态势的接近程度来衡量因素关联程度的方法。几何形状愈相似，发展态势愈接近，则因素的关联度就愈大，否则就愈小。

今令 $\{X_K(i)\}$ 中某一数列为母数列（参考数列）X_0，其余为子数列（被比较数列）X_j，研究 X_j与X_0的关联性

由　　$X_0=(X_0(1),\ X_0(2),\ \cdots\cdots,\ X_0(n))$

$X_j=(X_j(1),\ X_j(2))\ \cdots\cdots,\ X_j(n))$，$(j=1,\ 2,\ 3,\ \cdots\cdots,\ m-1)$

进行标么化处理，以便比较，于是

$$\begin{cases}\overline{X}_0=\sum\limits_{i=1}^{n}X_0(i)/n\\ \overline{X}_j=\sum\limits_{i=1}^{n}X_j(i)/n\end{cases}\qquad(1)$$

得：

$$\begin{cases}x_0(i)=X_0(i)/\overline{X}_0\\ x_j(i)=X_j(i)/\overline{X}_j\end{cases}\quad(i=1,2,\cdots\cdots n,j=1,2\cdots\cdots,m-1)\qquad(2)$$

以 $x_0(1)$为参考点，将被比较数列 $\{x_j(i)\}$ 中的第一点$x_j(1)$进行坐标平移，使之与参考点重合，其余各点也作相应的平移。经过这样平移处理后，子线与母线的几何形状、发展态势的接近程度就可以用线上各点的距离大小来衡量了。

第 j 条子线对于母线在第 i 点的距离为

$$D_{0j}(i)=|x_0(i)-x_j(i)|\qquad(3)$$

定义X_j对于X_0在第i点的关联系数为

$$\xi_{0j}(i)=\frac{\sigma\max\limits_{j}\max\limits_{i}|x_0(i)-x_j(i)|}{|x_0(i)-x_j(i)|+\sigma\max\limits_{j}\max\limits_{i}|x_0(i)-x_j(i)|}\qquad(4)$$

式中$\sigma\in[0,\ 1]$，是取定的数，本文取$\sigma=0.5$。称γ_{0j}为X_j对于X_0的关联度

$$\gamma_{0j}=\frac{1}{n}\sum_{i=1}^{n}\xi_{0j}(i),\qquad\gamma_{0j}\in[0,\ 1]\qquad(5)$$

如果存在$\gamma_{0K}>\gamma_{0L}>\gamma_{0P}$，则为$X_K$对于$X_0$的关联程度最大，$X_L$次之，$X_P$更次。以上这些就是求关联度的基本方法。

为了分析各因素之间的相互关联性，必须分别以m个因素作母因素，以便得到γ_{Kj}，用以评价X_j对于X_K的关联性，并以γ_{Kj}为基础建立等价关系矩阵（$K=1,2,\cdots\cdots m$，$j=1,2,3,\cdots\cdots m$）。

若关系矩阵$R(\gamma_{Kj})$中γ_{Kj}有(1)$\gamma_{KK}=1$（自反性），(2)$\gamma_{Kj}=\gamma_{jK}$（对称性），且(3)$R^{2P}=R^P$（$P=1,2,3\cdots\cdots$）（传递性），则称R为等价关系。若仅满足(1)、(2)两条，则称R为相似关系矩阵。

为了形成相似关系矩阵，在计算关联系数$\xi(i)$时，必须使(4)式中的$\max\limits_{j}\max\limits_{i}|x_0(i)-x_j(i)|$值定为常值，并令为$D(\max)$。(4)式改写为

$$\xi_{Kj}(i)=\frac{\sigma D(\max)}{|x_K(i)-x_j(i)|+\sigma D(\max)}\quad(K=1,2,\cdots\cdots,m;j=1,2,\cdots\cdots m)\tag{6}$$

式中$D(\max)=\max\limits_{j}\max\limits_{i}|x_1(i)-x_j(i)|$，其余符号同前。$D(\max)$的意义是当开始把第一个因素作为母素列计算关联系数时，其值就定了，以后再计算$\xi_{Kj}(i)$，其值不变。这样可得到相似关系$R=(\gamma_{Kj})_{m\times n}$。

将相似关系改造成等价关系的方法是用模糊数学中的关系合成法则完成的。即：$R\cdot R=R^2,\cdots\cdots R^P\cdot R^P=R^{2P}$，当$R^{2P}=R^P$时，$R^P$即为等价关系。其中，$R=(\gamma_{Kj})$，有$R\cdot R=R^2=(\gamma^*_{Kj})$，而$\gamma^*_{Kj}=\bigvee(\gamma_{K1}\wedge\gamma_{1j})$（符号$\vee$，$\wedge$分别为取大、取小值运算）。以下$\gamma^*_{Kj}$与$\gamma_{Kj}$不加区别。

2.在等价关系的基础上进行聚类

等价关系矩阵$R=(\gamma_{Kj})$，可视为一模糊等价m阶方阵，$\gamma_{Kj}\in[0,1]$。若γ_{Kj}蜕化为：$\gamma_{Kj}\in\{0,1\}$，则R蜕化为普通集合等价关系。

对于普通集合，$\gamma_{Kj}=1$时，x_K与x_j有关系，$\gamma_{Kj}=0$，x_K与x_j无关。

若对模糊等价关系$R(\gamma_{Kj})$取截集λ，$\lambda\in[0,1]$。$\lambda\leqslant\gamma_{Kj}$时，使$\gamma_{Kj}=1$；$\lambda>\gamma_{Kj}$时，$\gamma_{Kj}=0$，从而使$R$蜕化成了普通集合的等价关系。这样，根据$\gamma_{Kj}$就可判定在$\lambda$水平上$x_K$与$x_j$因素的关系了。最后还可用聚类树图表示各因素在不同$\lambda$水平上的关系。

三、实例应用

现以表1所列白化值作原始资料建立数列，按上述方法进行聚类分析，结果是用BASIC语言在PC—1500袖珍机上实现的，成果如下：（源程序略）

原　始　数　据

CLUSTERING ANALYSIS

M=7　N=7

——DATA——

(1) 0.875 32.5 1.95 47.5 0.05 0.35 14
(2) 1.5 40 1.8 75 0.0225 0.9 16.5
(3) 1.485 48 1.7 61 0.0225 0.4 85 16.5
(4) 1.1 38.5 1.8 50 0.03 0.42 23
(5) 0.785 23.5 1.84 42 0.014 0.91 32.5
(6) 0.65 23 1.95 40.5 0.0175 0.79 26
(7) 1.245 40 1.75 56 0.0325 0.37 22.5

相　似　关　系　矩　阵

SIMILAR RELATION MATRIX

(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0.85	0.89	0.88	0.75	0.77	0.89
0.85	1	0.85	0.79	0.73	0.75	0.82
0.89	0.85	1	0.84	0.73	0.75	0.87
0.88	0.79	0.84	1	0.79	0.83	0.93
0.75	0.73	0.73	0.79	1	0.9	0.77
0.77	0.75	0.75	0.83	0.9	1	0.8
0.89	0.82	0.87	0.93	0.77	0.8	1

等价关系矩阵
EQUAL IN VALUE
RELATION MATRIX

(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0.85	0.89	0.89	0.83	0.83	0.89
0.85	1	0.85	0.85	0.83	0.83	0.85
0.89	0.85	1	0.89	0.83	0.83	0.89
0.89	0.85	0.89	1	0.83	0.83	0.93
0.83	0.83	0.83	0.83	1	0.9	0.83
0.83	0.83	0.83	0.83	0.9	1	0.83
0.89	0.85	0.89	0.93	0.83	0.83	1

聚类结果
CRITICAL VALUE = 0.93

(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	1
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	1	0	0	1

聚类结果
CRITICAL VALUE = 0.89

(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0	0	1	0	0	1
0	1	0	0	0	0	0
0	0	1	0	0	0	0
1	0	0	1	0	0	1
0	0	0	0	1	1	0
0	0	0	0	1	1	0
1	0	0	1	0	0	1

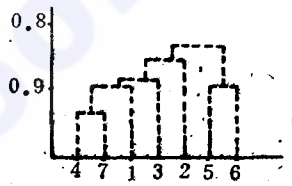
聚类结果
CRITICAL VALUE = 0.88

(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0	1	1	0	0	1
0	1	0	0	0	0	0
1	0	1	1	0	0	1
1	0	1	1	0	0	1
0	0	0	0	1	1	0
0	0	0	0	1	1	0
1	0	1	1	0	0	1

聚类结果
CRITICAL VALUE = 0.85

(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1	1	1	0	0	1
1	1	1	1	0	0	1
1	1	1	1	0	0	1
1	1	1	1	0	0	1
0	0	0	0	1	1	0
0	0	0	0	1	1	0
1	1	1	1	0	0	1

聚类图
CLUSTERING GRAPH



经以上聚类结果如下：

$\lambda > 0.93$，　七种土分为七类：(1)，(2)，(3)，(4)，(5)，(6)，(7)

$0.89 < \lambda \leqslant 0.93$，可分为六类：(4，7)，(1)，(2)，(3)，(5)，(6)

$0.88 < \lambda \leqslant 0.89$，可分为四类：(1，4，7)，(3)，(2)，(5，6)

$0.85 < \lambda \leqslant 0.88$，可分为三类：(1，4，7，3)，(2)，(5，6)

$0.82 < \lambda \leqslant 0.85$，可分为二类：(1，4，7，3，2)，(5，6)

$\lambda \leqslant 0.82$，　七种土合为一类。

由以上聚类可知，表1所列土的工程地质性质，花岗岩风化红土与大连红土性质最为相近，且与一般粘土无太大差别；碎屑岩风化红土与网纹红土性质相近，且与一般粘土差别较大。碳酸岩风化红土及玄武岩红土与网纹红土和一般粘土都有一定差别。

若取表1中的指标上界值进行聚类，其分类结果与上述结论一致。

四、结　论

1.水文地质工程地质问题分析中，经常用归类的方法来说明研究对象的性质，聚类分析是常用的一种分析方法。本文应用灰色系统理论的关联度分析于聚类分析，得到了研究对象性质差异程度的明确信息，这实质上是一种信息的开发。

2.当我们的研究对象关系比较复杂，因素众多数据复杂时，用本文介绍的方法，可以得到因素间相关性的明确概念。计算工作量不大（只要有一般计算器就可以计算），只要各因素具有相对应的一组数列，即可进行运算，效果是良好的。

3.应该要注意的是，任何一种数学统计分析方法，只能是一种分析问题的工具。计

（下转第7页）

税率5%的税金后，年收入约10亿元，用以进行雅砻江梯级开发。估计在二滩电站投产后20年左右，公司有能力建成下游段全部梯级电站。这样，国家只拿出该河段总投资的30%（相当于全河梯级总设资的11%），就可完成全部梯级开发。后期，公司可还清国家贷款全部本息，并积累有雄厚资金，进行雅砻江中上游或其他河流水电开发。

4.四川是占全国人口十分之一的大省，如经济发展受阻，必将拖全国四化建设的后腿。同时，四川江河水电开发，对长江三峡工程和长江流域经济发展，以及全国能源与水资源综合平衡都将有着重要的效益和影响。国家应对四川水电开发给予高度的重视。

5.建议研究采取一些特殊政策和措施，包括利用外资等多种方式，以加快四川水电开发。可否考虑划出某个河段或工程地区，向国外招标，由外国公司按照我们的规划要求，进行设计和投资建设，保证其获得一定年限的发电收益（用四川的工农业产品等支付），然后全部收归国有。

6.水能资源是四川突出的、持久的优势。开发水电是推动四川经济持续发展的制约因素。希望国家和四川省组建一个专门的领导机构，加强对四川水电开发的领导和研究。

（上接19页）

算结果一定要回到具体问题中来，结合实际条件进行判定，才能得出合理的结论。

主要参考资料

〔1〕 邓聚龙，灰色控制系统，华中工学院出版社

〔2〕 邓聚龙，本征灰色系统的主要方法，系统工程理论与实践，1986年第1期

〔3〕 贺仲雄编《模糊数学及应用》，天津科学技术出版社，1983年元月