袁连新,余勇.聚类分析方法及其环境监测(水质分析)中的应用[J]. 环境科学与技术, 2011, 34(12H):267- 270. Yuan Lian- xin, Yu- Yong. Application of clustering analysis methods in environmentai monitor(water analysis)[J]. Environmental Science & Technology, 2011, 34(12H):267- 270.

聚类分析方法及其环境监测(水质分析)中的应用

袁连新[1], 余勇[2]

(1. 湖北省环境监测中心站,湖北 武汉 430072; 2. 武汉大学资环学院,湖北 武汉 430072)

摘 要:文章对比分析了"系统聚类法"、"模糊聚类法"、"灰色聚类法"三种聚类分析方法的原理、原则、聚类依据、聚类分析步骤,以及其各自的优缺点,并举例说明了其在环境监测领域(水质分析)的应用。

关键词:统聚类法; 模糊聚类法; 灰色聚类法; 环境监测; 水质分析

中图分类号:X82 文献标志码:A doi:10.3969/j.issn.1003-6504.2011.12H.064 文章编号:1003-6504(2011)12H-0267-04

Application of Clustering Analysis Methods in Environmentai Monitor(Water Analysis)

YUAN Lian-xin, YU-Yong

(1. Central Station Of Environmental Monitor Hubei Province;
2. Resource and Environment Institute of Wuhan University)

Abstract: In this paper, the theories, principles, basis, analysis procedures, advantages and disadvantages of three clustering analysis methods, including system clustering method, fuzzy clustering method and grey clustering method, were analysed and contrasted. And their applications in environmental monitoring (water quality analysis) were presented in the paper.

Key words: system clustering method; fuzzy clustering method; grey clustering method; environmental monitoring; water quality analysis.

聚类分析是将研究的对象按其共性进行分类,以便系统地加以科学研究的一种有效方法, 其目的在于辨认在某些特征上相似的事物,并把事物就这些特征划分成若干类, 使在同一类的事物具有高度共质性,而不同类的事物具有高度相异性。

聚类分析方法主要针对的的问题是:对于样本空间中的元素含有多个属性,要求对其中的元素进行合理的分类。

常见的聚类分析方法有 "系统聚类法"、"模糊聚类法"、"灰色聚类法"等。本文将就这三种聚类分析方法对比分析。分析三种聚类方法的原理、原则、聚类依据、聚类分析步骤,以及其各自的优缺点,结合本人专业进行分析三种聚类方法在环境监测(水质分析)的适用领域,并列举应用实例。

1 聚类分析依据

聚类分析是在相似性或距离的基础上进行的,所以首先需要确定的就是确定相似性度量——聚类统计量,用以度量样品之间的联系(相似性),作为聚类分析的依据。

常用的聚类统计量主要有距离统计量、相似系数统计量、相关系数统计量等。当两个样品之间的距离越小,它们之间的相似性越高,反之亦然;当两个样品的相似系数越接近 1,它们之间的相似性越高;当两个样品的相关系数月接近于 1, 它们之间的相关程度越高,越密切[1]。

2 三种聚类分析

2.1 系统聚类分析

系统聚类分析(hierachical cluster analysis)在聚类分析中应用最为广泛,是多元统计中一种十分常见的手段。凡是具有数值特征的变量和样品都可以通过选择不同的距离和系统聚类方法而获得满意的数值分类效果。系统聚类法就是把个体逐个地合并成一些子

《环境科学与技术》编辑部:(网址) http://fjks.chinajournal.net.cn(电话) 027-87643502(电子信箱) hjkxyjs@126.com
收稿日期:2011-09-23;修回 2011-11-01
作者简介:袁连新(1976-),男,工程师,主要从事环境监测工作。

至整个总体都在一个集合之内为止。

系统聚类法是一种广为使用的聚类方法，内容极为丰富。分类的原则是：类与类之间的距离最近的两类合并。因此，系统聚类法的分类统计量一般采用距离系数统计量，其中最常见距离分类是最短距离法。

系统聚类分析的基本步骤为：

(1) 每个样品为一类，计算样品之间的距离系数，构造距离系数矩阵；

(2) 把距离最小的两类合并为一个新类（为距离最小的两类的平均值）；

(3) 计算新的类间距离；

(4) 重复 2、3 步一直到合并到一类为止。

2.2 模糊聚类分析

模糊聚类分析是考虑到事物的模糊性质而进行分类的一类数学方法，是在模糊数学理论的基础上，建立模糊等价关系，利用 λ 截关系进行分类的一种动态聚类分析法。

由于建立在论域 U 上的模糊等价关系 R 的每一个 λ 截关系 R_λ 都是普通等价关系，因此都可以得到论域 U 上的一个分划。对于 U 上模糊等价关系 R，当 λ 由 1 减至 0，根据 R_λ 可得论域 U 由细到粗的不同分类。其分类依据主要是利用不同的 λ 截关系 R_λ 进行的。

基于模糊等价关系的动态聚类分析法的步骤如下：

(1) 确立拟分类对象各元素之间的模糊相似系数 $r_{ij}=\mu_R(x_i,x_j)$，建立模糊相似方阵 $R=(r_{ij})_{n\times m}$；

(2) 由模糊相似方阵 $R=(r_{ij})_{n\times m}$ 建模糊等价矩阵 R^*；

(3) 选取恰当的 λ 进行截取，进行动态聚类分析。

2.3 灰色聚类分析

灰色聚类是在聚类分析方法中引进灰色理论的白化权函数而形成的是将聚类对象对不同聚类指标拥有的白化数按几个灰类进行归纳提出的以灰数的白化函数生成为基础的新的聚类方法。

灰色聚类分析方法的常见步骤为：

(1) 确定灰类的白化函数（白化函数反映了聚类指标对灰类的亲疏关系）；

(2) 确定聚类标准权（聚类权是各指标对某一灰类的权重）；

(3) 确定聚类系数（聚类系数反映了聚类样本对灰类的亲疏程度）；

(4) 构造聚类行向量，进行聚类分析。

3 三种聚类分析比较

“系统聚类法”、“模糊聚类法”和“灰色聚类法”作为三种最为常见的聚类分析方法。是在多元统计[illegible]础上，结合其他数学理论（如模糊数学理论、灰色数学理论）发展起来的。现将三种聚类分析方法的区别简要分析如下：

表 1 “系统聚类法”、“模糊聚类法”和“灰色聚类法”的区别

	系统聚类分析	模糊聚类分析	灰色聚类分析
理论基础	多元统计	模糊数学	灰色理论
分类原则	类与类之间的距离最近的两类合并一类	隶属度的类为 1 的各类合并为一类	聚类系数最大的类合并为一类
分类依据	距离系数	λ 截关系	聚类系数

4 三种聚类分析在水资源与环境科学中的应用

聚类分析作为一种有效的科学手段，在环境监测领域有着十分广泛的应用，如水质评价，水化学分类、地下水成因分类、水功能区分类等。其中在水质评价中的应用最为广泛和成熟。下面将以“系统聚类”为例介绍聚类分析方法在水质评价方面的应用(其中模糊聚类和灰色聚类的文献资料较多[2-7]，本处就不做赘述)。

4.1 实例介绍

例：对于某地地下水水质监测的数据如表 2，对使用聚类分析方法对其进行分类。

表 2 某地地下水水质分析结果 (mg/L)

监测点	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
1	157.5	62.4	15.6	320	1.08	6.4
2	306	109.2	80.6	650	1.05	27.4
3	1 100	546.9	226.9	2 111	1.48	30.95

已知国家地下水质量分类标准（GB/T14848-93）中对于地下水质量分类的标准中各指标的允许范围如表 3。

表 3 地下水质量分类标准 (mg/L)

类别	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
Ⅰ	≤150	≤50	≤50	≤300	≤1.0	≤2.0
Ⅱ	≤300	≤150	≤150	≤500	≤2.0	≤5.0
Ⅲ	≤450	≤250	≤250	≤1 000	≤3.0	≤20
Ⅳ	≤550	≤350	≤350	≤2 000	≤10	≤30
Ⅴ	>550	>350	>350	>2 000	>10	>30

4.2 聚类分析

首先将数据进行数学处理。设每个监测点的 6 个监测指标为：

$$x_i=(x_{i1},x_{i2},x_{i3},x_{i4},x_{i5},x_{i6});$$

为了方便分类，将分类标准的中间值作为样本一起加入样本空间。见表 4。

其中 i=1，2，3，Ⅰ，Ⅱ，Ⅲ，Ⅳ，Ⅴ；Ⅰ—Ⅴ分别代表

表 4　地下水质量分类标准值处理　(mg/L)

类别	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
Ⅰ	75	25	25	150	0.5	1.0
Ⅱ	225	100	100	400	1.5	3.5
Ⅲ	375	200	200	750	2.5	12.5
Ⅳ	500	300	300	1 500	6.5	25
Ⅴ	550	350	350	2 000	10	30

Ⅰ—Ⅴ类地下水。

4.2.1　数据预处理

在进行聚类分析之前，必须首先对数据进行预处理，预处理的方法有标准化和正规化。在本例中进行正规化处理[8]。

4.2.2　系统聚类分析方法在水质评价中的应用

表 5　第一次数据处理(规范化处理结果)

监测点	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
1	0.080 5	0.071 7	0.000 0	0.086 7	0.061 1	0.180 3
2	0.225 4	0.161 3	0.194 4	0.255 0	0.057 9	0.881 5
3	1.000 0	1.000 0	0.631 9	1.000 0	0.103 2	1.000 0
Ⅰ	0.000 0	0.000 0	0.028 1	0.000 0	0.000 0	0.000 0
Ⅱ	0.146 3	0.143 7	0.252 4	0.127 5	0.105 3	0.083 5
Ⅲ	0.292 7	0.335 3	0.551 4	0.306 0	0.210 5	0.384 0
Ⅳ	0.414 6	0.526 9	0.850 5	0.688 4	0.631 6	0.801 3
Ⅴ	0.463 4	0.622 7	1.000 0	0.943 4	1.000 0	0.968 3

利用欧拉距离公式 $d_{ij}=\left[\frac{1}{n}\sum_{k=1}^{m}(x_{ik}-x_{jk})^2\right]^{\frac{1}{2}}$，计算得到距离系数见表 5，选出其中最小的距离为 $d_{3\text{Ⅴ}}=0.011\ 6$，将测点 3 和Ⅴ分为一类，去它们各检测指标的平均值，并成一类(表 6)。

表 6　第一次计算距离系数

监测点	1	2	3	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
1	0	0.247 9	0.289 8	0.063 7	0.034 2	0.072 0	0.219 6	0.278 6
2		0	0.041 9	0.311 6	0.282 1	0.175 9	0.028 3	0.030 7
3			0	0.353 6	0.324 0	0.217 8	0.070 2	0.011 2
Ⅰ				0	0.029 5	0.135 8	0.283 3	0.342 3
Ⅱ					0	0.106 2	0.253 8	0.312 8
Ⅲ						0	0.147 6	0.206 6
Ⅳ							0	0.059 0
Ⅴ								0

表 7　第二次数据处理

监测点	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
1	0.080 5	0.071 7	0.000 0	0.086 7	0.061 1	0.180 3
2	0.225 4	0.161 3	0.194 4	0.255 0	0.057 9	0.881 5
3，Ⅴ	0.731 7	0.811 4	0.816 0	0.971 7	0.551 6	0.984 2
Ⅰ	0.000 0	0.000 0	0.028 1	0.000 0	0.000 0	0.000 0
Ⅱ	0.146 3	0.143 7	0.252 4	0.127 5	0.105 3	0.083 5
Ⅲ	0.292 7	0.335 3	0.551 4	0.306 0	0.210 5	0.384 0
Ⅳ	0.414 6	0.526 9	0.850 5	0.688 4	0.631 6	0.801 3

表 8　第二次计算距离系数

监测点	1	2	3，Ⅴ	Ⅰ	Ⅱ	Ⅲ	Ⅳ
1	0	0.256 0	0.687 3	0.065 8	0.035 4	0.074 4	0.226 8
2		0	0.511 7	0.321 8	0.291 4	0.181 6	0.029 2
3，Ⅴ			0	0.758 9	0.635 7	0.448 7	0.208 0
Ⅰ				0	0.030 5	0.140 2	0.292 6
Ⅱ					0	0.109 7	0.262 1
Ⅲ						0	0.152 4
Ⅳ							0

根据表 6，再次计算欧拉距离，得到表 7。并重复以上步骤，直至分成一类。见表 8~表 18。

由此，可以看出 1 号点位的水质应属于Ⅲ类水质，2 号点位的水质应属于Ⅳ水质，3 号点位的水质应属于Ⅴ水质。

5　结语

聚类分析是将研究的对象按其共性进行分类，以便系统地加以科学研究的一种有效方法，其目的在于辨认在某些特征上相似的事物，并把事物就这些特征

表 9 第三次数据处理

监测点	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
1	0.080 5	0.071 7	0.000 0	0.086 7	0.061 1	0.180 3
2, Ⅳ	0.475 0	0.592 6	0.735 6	0.816 3	0.467 9	0.884 9
3, Ⅴ	0.731 7	0.811 4	0.816 0	0.971 7	0.551 6	0.984 2
Ⅰ	0.000 0	0.000 0	0.028 1	0.000 0	0.000 0	0.000 0
Ⅱ	0.146 3	0.143 7	0.252 4	0.127 5	0.105 3	0.083 5
Ⅲ	0.292 7	0.335 3	0.551 4	0.306 0	0.210 5	0.384 0

表 10 第三次计算距离系数

监测点	1	2, Ⅳ	3, Ⅴ	Ⅰ	Ⅱ	Ⅲ
1	0	0.600 3	0.246 1	0.065 8	0.035 4	0.074 4
2, Ⅳ		0	0.163 9	0.676 5	0.546 4	0.344 2
3, Ⅴ			0	0.276 6	0.221 3	0.165 9
Ⅰ				0	0.0305	0.140 2
Ⅱ					0	0.109 7
Ⅲ						0

表 11 第四次数据处理

监测点	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
1	0.080 5	0.071 7	0.000 0	0.086 7	0.061 1	0.180 3
2, Ⅳ	0.475 0	0.592 6	0.735 6	0.816 3	0.467 9	0.884 9
3, Ⅴ	0.731 7	0.811 4	0.816 0	0.971 7	0.551 6	0.984 2
Ⅰ, Ⅱ	0.0706	0.133 9	0.133 7	0.367 4	0.098 9	0.446 6
Ⅲ	0.292 7	0.335 3	0.551 4	0.306 0	0.210 5	0.384 0

表 12 第四次计算距离系数

监测点	1	2, Ⅳ	3, Ⅴ	Ⅰ, Ⅱ	Ⅲ
1	0	0.600 3	0.246 1	0.186 0	0.074 4
2, Ⅳ		0	0.163 9	0.138 7	0.344 2
3, Ⅴ			0	0.666 4	0.165 9
Ⅰ, Ⅱ				0	0.238 5
Ⅲ					0

表 13 第五次数据处理

监测点	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
1, Ⅲ	0.186 6	0.203 5	0.275 7	0.196 4	0.135 8	0.282 2
2, Ⅳ	0.475 0	0.592 6	0.735 6	0.816 3	0.467 9	0.884 9
3, Ⅴ	0.731 7	0.811 4	0.816 0	0.971 7	0.551 6	0.984 2
Ⅰ, Ⅱ	0.070 6	0.133 9	0.133 7	0.367 4	0.098 9	0.446 6

表 14 第五次计算距离系数

监测点	1, Ⅲ	2, Ⅳ	3, Ⅴ	Ⅰ, Ⅱ
1, Ⅲ	0	0.126 1	0.745 9	0.180 9
2, Ⅳ		0	0.163 9	0.138 7
3, Ⅴ			0	0.666 4
Ⅰ, Ⅱ				0

划分成若干类，使在同一类的事物具有高度共质性，而不同类的事物具有高度相异性。其在环境监测(水质监测)将具有十分广泛应用。

表 15 第六次数据处理

监测点	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
1, 2, Ⅲ, Ⅳ	0.280 9	0.340 1	0.392 4	0.476 9	0.232 1	0.693 4
3, Ⅴ	0.731 7	0.811 4	0.816 0	0.971 7	0.551 6	0.984 2
Ⅰ, Ⅱ	0.070 6	0.133 9	0.133 7	0.367 4	0.098 9	0.446 6

表 16 第六次计算距离系数

监测点	1, 2, Ⅲ, Ⅳ	3, Ⅴ	Ⅰ, Ⅱ
1, 2, Ⅲ, Ⅳ	0	0.587 7	0.518 4
3, Ⅴ		0	0.666 4
Ⅰ, Ⅱ			0

表 17 第七次数据处理

监测点	总硬度	硫酸根离子	氯离子	溶解性总固体	高锰酸盐指数	氨氮
1, 2, Ⅰ, Ⅱ, Ⅲ, Ⅳ	0.177 0	0.206 0	0.266 3	0.270 3	0.142 4	0.367 6
3, Ⅴ	0.731 7	0.811 4	0.816 0	0.971 7	0.551 6	0.984 2

表 18 第七次计算距离系数

监测点	1, 2, Ⅰ, Ⅱ, Ⅲ, Ⅳ	3, Ⅴ
1, 2, Ⅰ, Ⅱ, Ⅲ, Ⅳ	0	1.003 9
3, Ⅴ		0

表 19 系统聚类分析过程表

连接顺序	连接样品		距离函数
1	3	Ⅴ	0.011 2
2	2	Ⅳ	0.029 2
3	Ⅰ	Ⅱ	0.030 5
4	1	Ⅲ	0.074 4
5	2, Ⅳ	1, Ⅲ	0.126 1
6	Ⅰ, Ⅱ	1, 2, Ⅲ, Ⅳ	0.518 4
7	1, 2, Ⅰ, Ⅱ, Ⅲ, Ⅳ	3, Ⅴ	1.003 9

[参考文献]

[1] 向东进,李宏伟,刘小雅. 实用多元统计分析[M]. 武汉: 中国地质大学出版社, 2005.

[2] 于皓, 刘志斌, 王昭君. 基于灰色聚类分析法的矿井水质评价[J]. 辽宁工程技术大学学报, 2003(22): 74–76.

[3] 史瑞卿, 孙维芬. 灰色聚类分析在地下水功能区划中的应用.科技纵横. 1999(1): 27–30.

[4] 王艳芬. 地下水灌溉水质评价的灰色聚类分析 [J]. 宁夏农学院学报, 1997, 18(4): 81–85.

[5] 朱艳红, 周志芳. 某水电站坝址处地下水水质的模糊聚类分析及成因研究[J]. 勘察科学技术, 2001(6): 15–19.

[6] 孙建华, 汪志林, 许春莲. 徐州市区浅层地下水水质的模糊聚类分析[J]. 能源技术与管理, 2006(1): 74–76.

[7] 郝瑞森.模糊数学在地下水水质评价中的应用[J]. 科技情报开发与经济, 2003, 13(11): 115–116.

[8] 彭放, 杨瑞琰, 罗文强, 等. 数学建模方法[M]. 北京:科学出版社, 2007.