

区域地貌中的聚类分析及贝叶斯判别方法^{*}

田絮资

(宝鸡文理学院 计算机科学系, 陕西 宝鸡 721007)

[摘 要] 为了对区域地貌的各项数据进行科学解读, 以寻求建立区域地貌数字系统的方法与模式, 利用系统聚类法对区域地貌进行了分类, 并用贝叶斯判别法检验了分类的合理性, 结果与用系统聚类法得到的分类一致, 说明系统聚类法对区域地貌数据分析的结果是科学的。

[关键词] 区域地貌; 聚类分析; 分类; 贝叶斯判别

[中图分类号] O 213

[文献标识码] A

[文章编号] 1671-9387(2003)05-0178-05

中国数字地球是本世纪我国地球科学领域的重要战略目标, 随着数字城市^[1,2]在全国的广泛开展及其相关系统的建立, 中国数字地球的进程已逐步进入到数字区域^[3]阶段。数字区域建立的基础性工作首先是地貌的数字化。在RS、GIS等现代技术手段与方法的支撑下, 现代地理学的信息获取在空间范围、时间效率、信息内涵等方面产生了质的飞跃, 将这些利用现代技术与方法获取的信息与数据, 同传统地理学的理论、观点和方法相融合, 是实现数字区域的关键环节。

笔者拟应用聚类分析法及贝叶斯判别法^[4,5], 对由RS中所获取的有关区域地貌数据进行分析, 以寻求建立区域地貌数字系统的方法与模式。通过这一数学方法的移入, 力图解决地貌研究中只重感性、经验性的定性研究, 而轻数据性的定量研究的不足, 同时, 也使利用现代技术手段与方法获取的信息与数据科学化、理性化, 构建数字区域的地貌数字系统。

在区域地貌研究中, 不同区域的地貌情况是复杂多样的。为了通过对各区域的地貌进行分析, 从而找出其普遍规律, 往往需要将众多研究对象按要求进行分类, 这样问题的研究和处理就大为方便了。然而, 仅凭经验和专业知识是难以进行确切分类的, 必须借助数学方法。聚类分析是一种根据多项指标分类的多元统计分析方法^[6,7], 利用该方法可以按要求把相似事物进行归类, 它广泛应用于自然科学、社会科学及国民经济等各个领域。为了保证使用聚类分析得到的分类结果更具科学性, 可以利用判别分析

进行检验。本文基于区域地貌多方面的统计数据, 用系统聚类法对地貌类型进行了分类, 并且用贝叶斯判别法检验了分类结果的合理性, 以期为区域地貌的研究和地貌数字系统的建立提供参考。

1 系统聚类分析法

聚类分析的模型是, 针对 n 个研究对象选取 n 种变量, 构成论域 X^m , 每个对象可看成 X^m 中的一个样本, 根据各对象的数字特征, 给出样本之间的接近程度, 可以用距离或相似系数来描述样本间的相似程度, 再依照样本间的相似程度数据进行分类。

设 $X = \{X_1, X_2, \dots, X_n\}$ 是均含有 m 种变量的 n 个样本的论域, 样本 X_i 和 X_j 间的距离记为 d_{ij} ($i, j = 1, 2, \dots, n$)。最常用的绝对值距离为:

$$d_{ij} = \sum_{k=1}^m |x_{ik} - x_{jk}| \quad (1)$$

系统聚类法的具体步骤如下:

(1) 利用公式

$$y_{ik} = \frac{x_{ik} - \bar{x}_k}{S_k} \quad i = 1, 2, \dots, n; k = 1, 2, \dots, m \quad (2)$$

将样本的原始数据标准化, 其中 $\bar{x}_k$ 和 S_k 分别为第 k 个分量样本的均值和标准差。

(2) 计算 n 个样本中两两间的距离 d_{ij} , 得到距离矩阵 $D^{(0)}$ 。

(3) 构造 n 个类, 每个类恰好包含 1 个样本。

(4) 根据类与类间的距离 $D(p, q)$, 合并距离最

^{*} [收稿日期] 2003-07-02

[基金项目] 陕西省自然科学基金资助项目(2001G11)

[作者简介] 田絮资(1954-), 女, 湖南株洲人, 副教授, 主要从事应用数学研究。

近的 2 类为 1 个新类, 计算新类与当前各类间的距离, 得到新的距离矩阵 $D^{(u)}$, $u = 1, 2, \cdots, n-1$ 。

- (5) 若类的个数为 1, 转下一步, 否则转回到步骤(4)。
- (6) 根据需要选取阈值 λ , 得到相应的分类, 并画出聚类图。
- (7) 决定类的个数和各类包含的样本。

2 区域地貌的系统聚类分析

为了应用系统聚类分析法对区域地貌进行分类, 选取了某省 15 个区域地貌的 5 项指标: 绝对高度(x_1), 相对高度(x_2), 山麓线高度(x_3), 平均坡度(x_4)和切割坡度(x_5), 其数据如表 1, 即论域 $X = \{X_1, X_2, \cdots, X_{15}\}$ 中含 15 个样本, 5 个变量。

表 1 地貌指标的原始数据
Table 1 Firsthand data of the landforms

样本 Sample	x_1	x_2	x_3	x_4	x_5
1	1 032	700	300	25	2 00
2	1 108	600	360	30	1 45
3	1 545	700	260	25	2 40
4	875	350	250	17	1 60
5	635	450	100	12	1 50
6	480	300	100	12	1 80
7	250	200	60	12	1 40
8	531	150	200	10	1 90
9	923	450	240	17	2 50
10	736	350	200	20	1 60
11	269	130	100	10	2 20
12	350	150	50	10	1 90
13	278	150	90	15	1 90
14	1 257	650	320	25	2 30
15	286	160	110	13	1 70

现根据需要要把这些区域的地貌分为 3 类, 即中山、低山和丘陵。
利用公式(2)将表 1 中的数据标准化, 再利用绝

对值距离公式(1) 求出样本最初的距离矩阵 $D^{(0)}$, 显然是一对角矩阵, 可简化为表 2。

表 2 绝对值距离矩阵
Table 2 Matrix of absolute value distance

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0														
2	2 49	0													
3	2 81	5 90	0												
4	4 83	5 18	6 85	0											
5	7 45	7 25	9 46	3 58	0										
6	7 75	9 19	9 77	4 01	3 97	0									
7	10 31	9 84	12 33	5 48	2 87	2 56	0								
8	7 37	9 36	9 39	4 17	4 08	2 40	4 00	0							
9	4 60	7 12	4 44	3 02	5 57	5 88	8 43	5 50	0						
10	5 25	5 60	7 27	1 29	2 16	3 59	5 06	3 76	3 76	0					
11	9 41	11 93	10 33	6 75	4 68	2 74	3 25	2 58	4 90	6 33	0				
12	9 32	11 30	11 34	6 12	4 05	2 11	2 25	1 95	6 00	5 70	1 62	0			
13	8 41	10 39	10 43	5 21	4 01	2 07	2 38	2 48	5 09	4 79	1 74	1 32	0		
14	1 83	4 05	1 85	6 18	8 80	9 11	11 66	8 73	3 35	6 60	9 67	10 68	9 76	0	
15	8 96	9 85	10 97	4 67	3 09	1 69	1 74	2 55	5 68	4 25	2 08	1 78	1 12	10 31	0

系统聚类法中求类与类之间的距离有多种方法, 这里采用其解为最优的最短距离法, 即用公式:
 $D_s(p, q) = \min\{d_{ij} | x_i \in G_p, x_j \in G_q\}$ (3)
求类 G_p 和 G_q 中相距最近的 2 个样本间的距离。

先将每个产品各自作为 1 类, 共有 15 类。在距离矩阵的非对角线元素中, 找出最短距离 $D_s = (15,$

$13) = 1 12$, 将类 G_{13} 和 G_{15} 并成 1 个新类 $G_{16} = G_{13} \cup G_{15}$, 利用(3)式和下式
 $D_s(16, p) = \min\{D(15, p), D(13, p)\} \quad p \neq 13, 15$
计算 G_{16} 与其余各类间的距离, 得到新的距离矩阵 $D^{(1)}$, 如表 3 所示。

表 3 距离矩阵 $D^{(1)}$ Table 3 Matrix of the distance $D^{(1)}$

	1	2	3	4	5	6	7	8	9	10	11	12	14	16
1	0													
2	2.49	0												
3	2.81	5.90	0											
4	4.83	5.18	6.85	0										
5	7.45	7.25	9.46	3.58	0									
6	7.75	9.19	9.77	4.01	3.97	0								
7	10.31	9.84	12.33	5.48	2.87	2.56	0							
8	7.37	9.36	9.39	4.17	4.08	2.40	4.00	0						
9	4.60	7.12	4.44	3.02	5.57	5.88	8.43	5.50	0					
10	5.25	5.60	7.27	1.29	2.16	3.59	5.06	3.76	3.76	0				
11	9.41	11.93	10.33	6.75	4.68	2.74	3.25	2.58	4.90	6.33	0			
12	9.32	11.30	11.34	6.12	4.05	2.11	2.25	1.95	6.00	5.70	1.62	0		
14	1.83	4.05	1.85	6.18	8.80	9.11	11.66	8.73	3.35	6.60	9.67	10.68	0	
16	8.41	9.85	10.43	4.67	3.09	1.69	1.74	2.48	5.09	4.25	1.74	1.32	9.76	0

在 $D^{(1)}$ 中 $D_s(10, 4) = 1.29$ 最小, 故将 G_4 和 G_{10} 并成一新类 $G_{17} = G_4 \cup G_{10}$, 再利用上述公式计算 G_{17} 与其余各类间的距离得 $D^{(2)}$, 依此逐步聚类, 得:

在 $D^{(2)}$ 中, $D_s(16, 12) = 1.32$, $G_{18} = G_{12} \cup G_{16}$, 求出 $D^{(3)}$;

.....

在 $D^{(12)}$ 中, $D_s = 1.83$, $G_{17} = G_7 \cup G_{16}$, 求出 $D^{(7)}$; 如此下去, 最后可将 15 个样本聚为一类, 但现在样本被聚为 3 类, 已满足需要, 从而得到聚类图, 如图 1 所示。

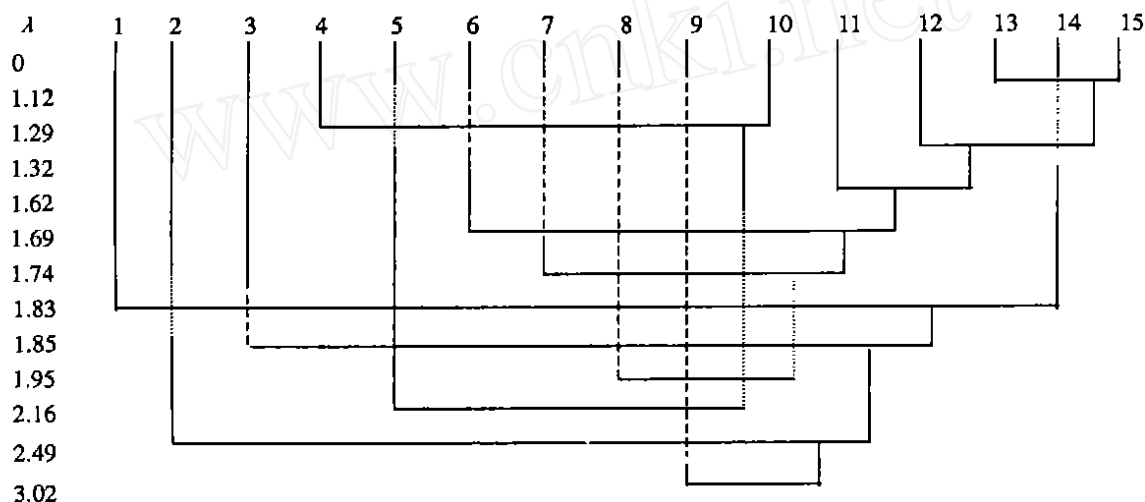


图 1 聚类示意图

Fig 1 Map showing cluster

从图 1 可以看出, 当 $2.49 \leq \lambda < 3.02$ 时, 可分为 3 类: $\{X_1, X_2, X_3, X_9, X_{14}\}$, $\{X_4, X_5, X_{10}\}$, $\{X_6, X_7, X_8, X_{11}, X_{12}, X_{13}, X_{15}\}$, 根据经验, 可分别视为对应中山、低山和丘陵的 3 类, 即为所得的分类结果。

3 分类结果的贝叶斯判别

聚类分析是基于观察的个体特性, 而对个体进行的分类, 其结果是否合理, 需要作进一步的检验。现用贝叶斯判别法来检验上述分类结果的合理性。

贝叶斯判别是在对研究对象已有一定认识的基础上来解决样本的归属问题。 $X = \{X_1, X_2, \dots, X_n\}$

是 n 个样本, 设有 k 个总体 $G_1, G_2, \dots, G_k$, 适当选取 X 的 1 个划分 $A_1, A_2, \dots, A_k$, 正好对应 k 个总体, 判别规则可表示为: X_j 归入 G_i 当且仅当 $X_j \in A_i$ ($j = 1, 2, \dots, n; i = 1, 2, \dots, k$)。贝叶斯判别准则就是分错率达到最小, 其方法是构造关于各总体的判别函数, 计算每个样本对应各判别函数的值, 使样本归入判别函数值最大的那个总体。

根据区域地貌的情况, 可设 3 个总体 G_1 (中山)、 G_2 (低山) 和 G_3 (丘陵), 分别为 3 种类型, 利用表 1 中的数据及上述分类结果, 计算出每种类型各要素平均值以及各要素的平均值, 见表 4。

表 4 要素平均值

Table 4 A verage value of key element

地貌类型 Regional landfom s type	x 1	x 2	x 3	x 4	x 5
中山M iddle mountain ($\bar{X}^{(1)}$)	1 152 10	612 50	290 00	24 25	2 11
低山Low mountain ($\bar{X}^{(2)}$)	748 67	383 33	183 33	16 33	1 56
丘陵 Hill ($\bar{X}^{(3)}$)	347 67	180 00	100 00	11 52	1 85
平均A verage ($\bar{X}$)	687 72	360 00	177 96	16 54	1 87

令

$$M_i = \sum_{j=1}^{U_i} (X_i^{(j)} - \bar{X}^{(i)}) (X_j^{(j)} - \bar{X}^{(i)})^T \quad t = 1, 2, 3$$

$$Y_2 = (8\ 437E - 4) x_1 + 0\ 185x_2 - (1\ 141E - 2) x_3 + 2\ 856x_4 + 93\ 478x_5 - 131\ 281$$

式中, U_i 为上述分类结果中第 t 类包含的样本数, 分别是 5, 3, 7。

丘陵的判别函数为:

$$Y_3 = (-\ 4\ 216E - 3) x_1 + 0\ 152x_2 - (3\ 678E - 2) x_3 + 2\ 909x_4 + 90\ 821x_5 - 111\ 861$$

$$R = \frac{1}{15 - 3} (M_1 + M_2 + M_3)$$

$$B_t = R^{-1} \bar{X}^{(t)}, C_t = \frac{1}{2} (\bar{X}^{(t)})^T R^{-1} \bar{X}^{(t)} \quad t = 1, 2, 3$$

由此可得 3 个判别函数

$$Y_t(X) = B^T X + C_t, t = 1, 2, 3$$

经计算得中山的判别函数为:

利用以上 3 个判别函数, 计算 15 个区域地貌的判别函数值, 将每个区域划归到函数值最大的一类地貌。例如, 对于第 2 个区域, 把 5 个要素数据分别代入 3 个判别函数, 得到 $Y_1 = 269\ 73, Y_2 = 244\ 52, Y_3 = 225\ 91$, 因 $Y_1 > Y_2 > Y_3$, 故第 2 个区域的地貌划归第 1 类(中山)。一般取

$$Y_1 = (6\ 624E - 4) x_1 + 0\ 276x_2 - (1\ 129E - 3) x_3 + 3\ 328x_4 + 132\ 539x_5 - 272\ 228$$

$$Y^*(X) = \max_{1 \leq t \leq 3} \{Y_t(X)\}$$

作为类型的判别, 15 个区域地貌的判别函数值如表 5 所示。

低山的判别函数为:

表 5 判别函数值

Table 5 Value of distinguish function

样本 Sample	Y_1	Y_2	Y_3	Y^*	类别 D lassification
1	284 28	254 08	233 64	Y_1	1
2	269 73	244 52	225 91	Y_1	1
3	284 67	254 93	232 95	Y_1	1
4	103 35	129 48	123 29	Y_2	2
5	98 02	125 86	121 40	Y_2	2
6	96 30	125 96	126 78	Y_3	3
7	15 67	70 39	77 37	Y_3	3
8	60 34	100 87	103 03	Y_3	3
9	250 24	232 27	220 41	Y_1	1
10	115 10	138 50	134 44	Y_2	2
11	94 53	126 07	132 01	Y_3	3
12	60 39	102 37	109 30	Y_3	3
13	79 93	116 13	122 68	Y_3	3
14	296 64	272 78	251 48	Y_1	1
15	41 29	93 35	99 42	Y_3	3

显然, 对 15 个区域地貌的系统聚类分析结果, 与贝叶斯判别的结果完全一致。

4 结 论

上述研究表明, RS 作为现代化的数据采集手段, 是区域地形地貌数字系统的信息源。数学方法是全息解读 RS 数据库的语言与方法, 聚类分析法正

是一种行之有效的数学方法。聚类分析法能对区域地貌的各项要素和指标的数据进行科学解读, 并在此基础上完成区域地貌类型的划分。系统聚类分析法对区域地貌数据分析的结果是科学的, 也与传统地理学的野外调查研究结果相一致。

[参考文献]

- [1] 承继成, 林 晖, 周成虎 数字城市——理论、方法与应用[M]. 北京: 科学出版社, 2003 304- 311.
- [2] 梁 军, 何建邦 数字城市建设的核心问题[J]. 地球信息科学, 2002, 4(1): 21- 26
- [3] 李 硕, 曾志远, 张运生 数字地形分析技术在分布式水文建模中的应用[J]. 地球科学进展, 2002, 17(5): 769- 775
- [4] 杨伦标, 高英仪 模糊数学原理及其应用[M]. 广州: 华南理工大学出版社, 1998
- [5] 曾文艺, 张 颜, 宋文彦 研究生招生中的模糊聚类分析方法[J]. 北京师范大学学报, 2001, 37(4): 436- 439
- [6] 方开泰 实用多元统计分析[M]. 上海: 华东师范大学出版社, 1989
- [7] 王秀红 多元统计分析在分区研究中的应用[J]. 地理科学, 2003, 23(1): 66- 70

Method of cluster analysis and Bayes distinguish on regional landforms

TIAN Xu-zi

(Department of Computer Science, Baoji University of Arts and Science, Baoji, Shaanxi 721007, China)

Abstract: In order to explain scientifically the datum of regional landforms, and to seek the model of establishing figures system on regional landforms, the regional landforms are classified by the method of system cluster analysis, and rationality of the classifying is tested by the method of Bayes distinguish. The classifying results are identical with results of Bayes distinguish. It showed that classifying result by method of system cluster analysis to regional landforms is scientific.

Key words: regional landforms; cluster analysis; classify; Bayes distinguish

(上接第 177 页)

The design and application of the precision agriculture information system to the planning of agricultural land consolidation

AIL ang-hui¹, WU Ci-fang¹, ZHAO Xiao-m in²

(1 Southeast and Management College of Zhejiang University, Hangzhou, Zhejiang 310029, China;

2 Land Resource and Environment College of Jiangxi Agriculture University, Nanchang, Jiangxi 330045, China)

Abstract: This paper briefly introduce the concept and features of the precision agriculture information system, brought forward the design targets and structure of the precision agriculture information system, elaborated the function and structure of seven subsystem of the precision agriculture information system. The paper applied the precision agriculture information system developed on 3S technology to the planning of agricultural land consolidation in Jiangxi province, specified the range of area of consolidation in quantity and spatial location, pointed out the measures of consolidation in different planning area.

Key words: precision agriculture information system 3S technology agricultural land consolidation