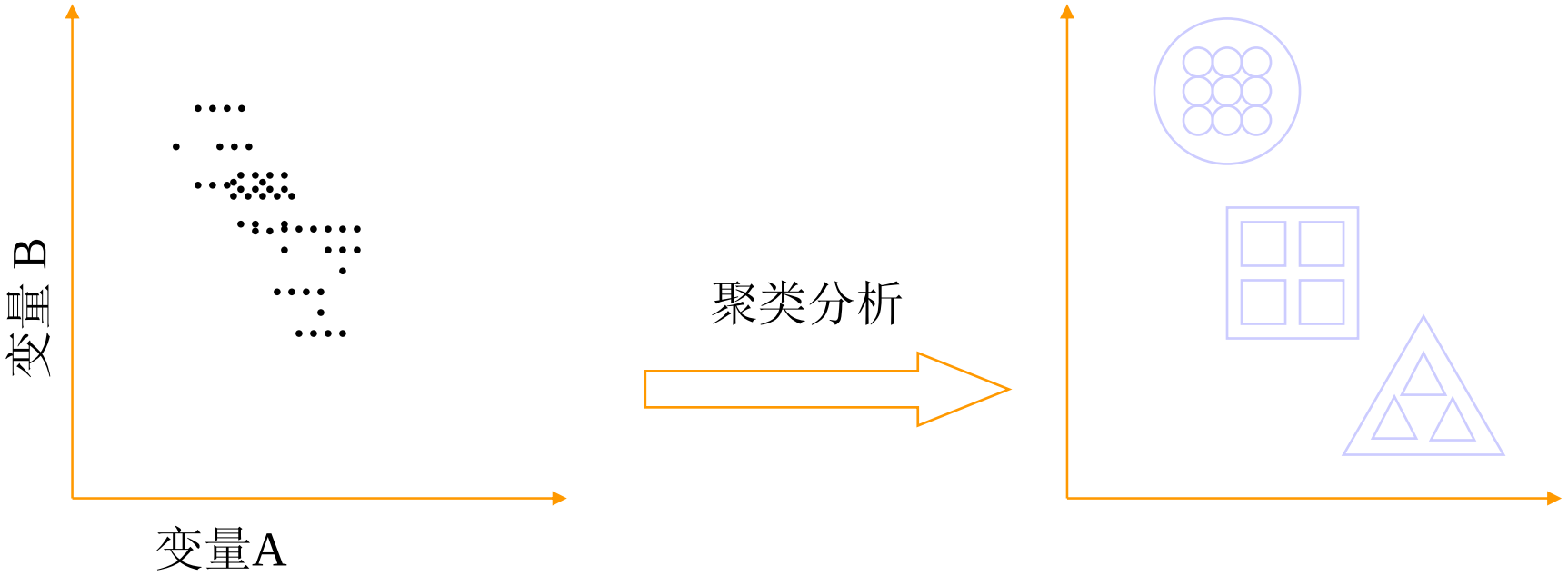
第十二讲（第**16**章）聚类分析

Cluster Analysis

聚类的概念

把研究对象分割成为具有相同属性的小的群体



聚类分析做什么?

- 把研究对象（人，城市，品牌等）分割成为更加同质的细分群体
- 描述对象的整体结构或者各个类之间的关系
- 根据每个类的描述资料进行该类特征的定位
- 评价一种判别类与类之间定性区别的方法 (例如：根据背景、品牌使用、心理因素)

聚类分析怎样使用?

- 去识别细分市场 —了解购买行为
- 为市场测试确定相匹配的城市
- 在市场结构分析中去识别竞争者
- 缩减数据以便进一步的分析

2个市场方面的概念

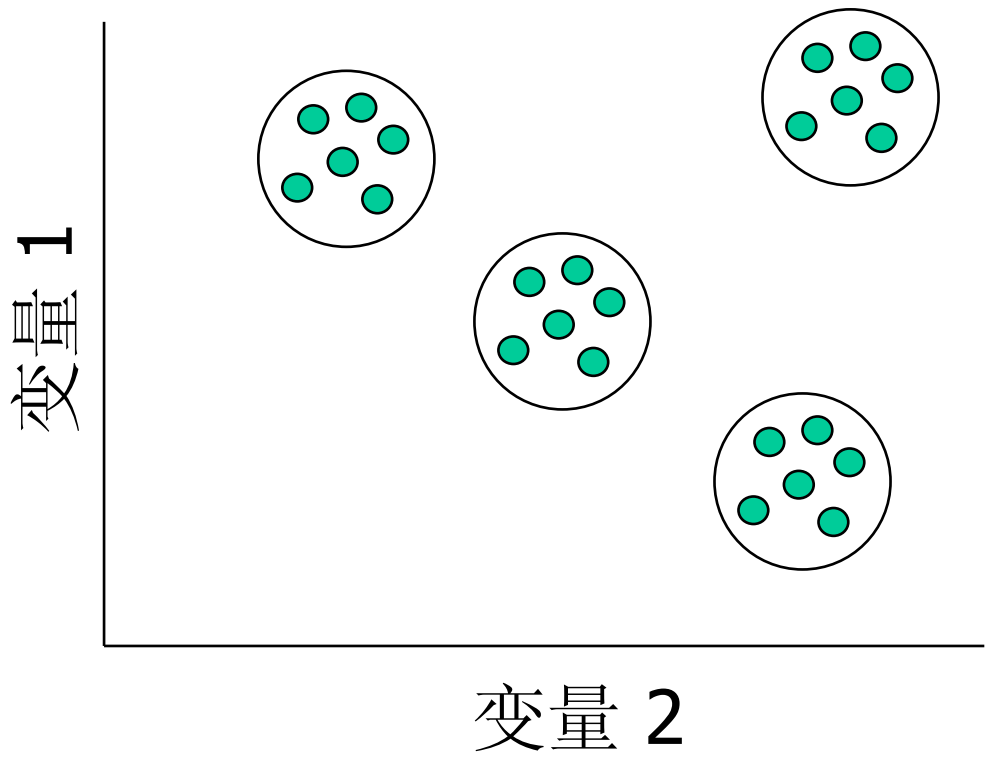
大众营销（Mass marketing）：

一种产品 ==>所有的消费者

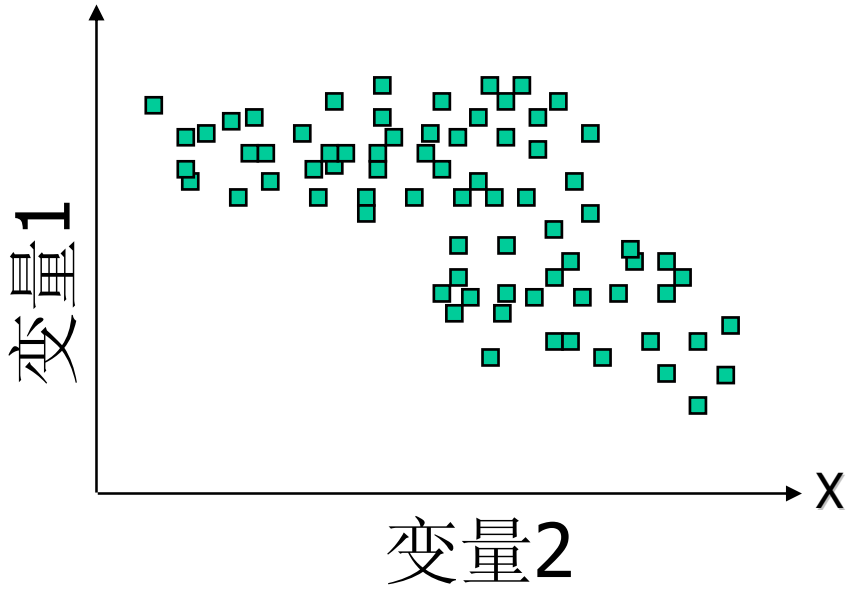
目标营销（Target marketing）：

产品和营销的组合 ==>不同的细分

理想的聚类情形



实际的聚类情形



有关统计量

- 聚合过程表(**Agglomeration schedule**)：提供系统聚类过程中每一步聚合的对象的信息。
- 群重心(**Cluster centroid**)：某一群中全部成员的变量均值。
- 群中心(**Cluster centers**)：非系统聚类时的起始值，群组围绕这些中心形成。
- 群别(**Cluster membership**)：表示每一对象所属的群。

有关统计量

- 树状图**(Dendrogram)**：显示聚类结果的树状图。垂直线代表聚在一起的组，水平尺度的位置表示群组聚合时的距离。
- 群间距离**(Distances between cluster centers)**：表示某一对群之间的距离。

变量的标准化方法

1.标准正态变换(Z值)

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad (i = 1,2,\cdots,m;\ j = 1,2,\cdots,n)$$

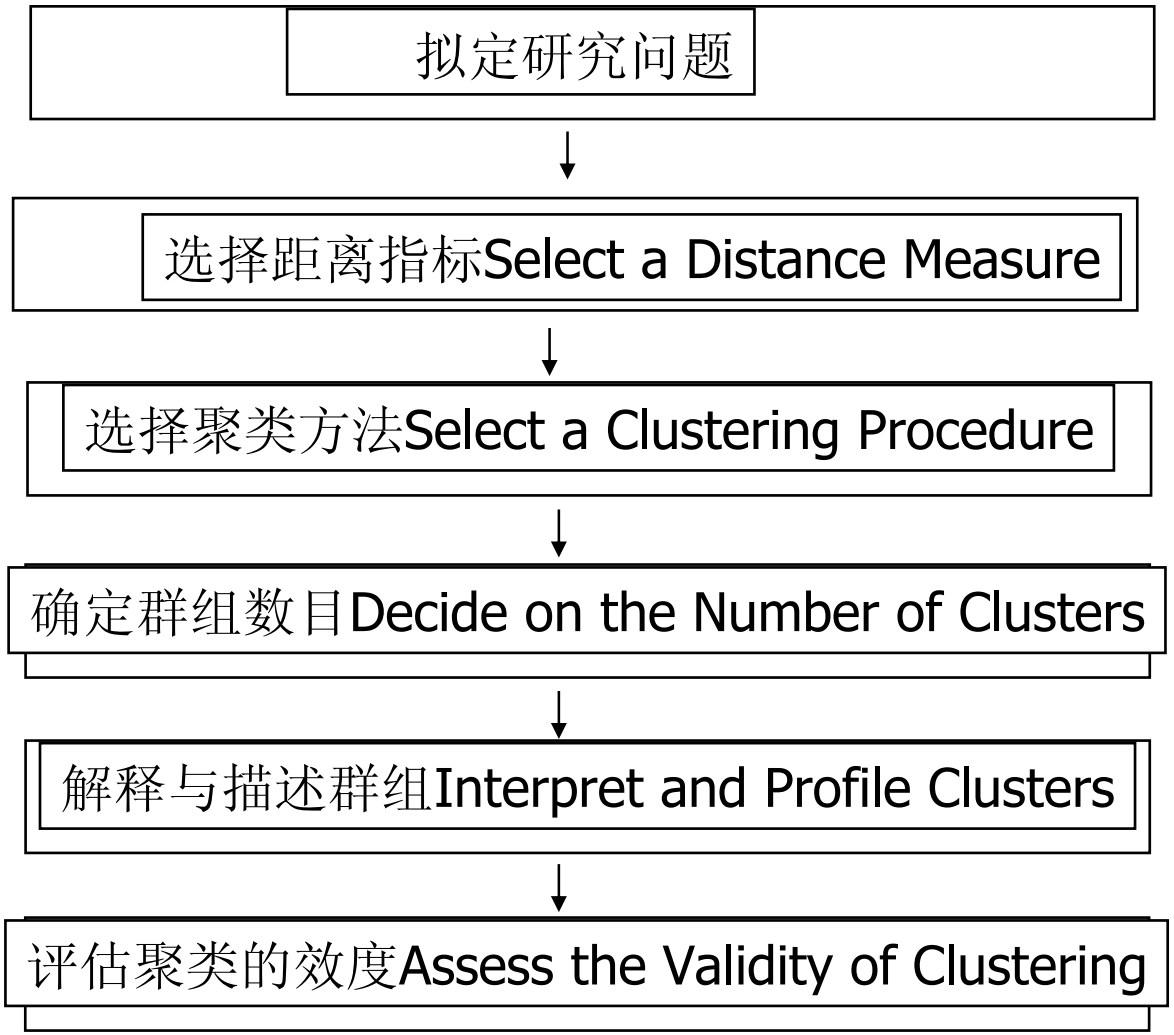
2.极差标准化(范围从0到1)

$$x_{ij} = \frac{x_{ij} - \min\limits_i \{x_{ij}\}}{\max\limits_i \{x_{ij}\} - \min\limits_i \{x_{ij}\}} \qquad (i = 1,2,\cdots,m;\ j = 1,2,\cdots,n)$$

3.极大值标准化(极大值为1)

$$x'_{ij} = \frac{x_{ij}}{\max\limits_i \{x_{ij}\}} \qquad (i = 1,2,\cdots,m;\ j = 1,2,\cdots,n)$$

聚类分析步骤


拟定研究问题
选择距离指标Select a Distance Measure
选择聚类方法Select a Clustering Procedure
确定群组数目Decide on the Number of Clusters
解释与描述群组Interpret and Profile Clusters
评估聚类的效度Assess the Validity of Clustering

选择距离指标

- 什么是距离？
- 设dij表示第i个样品与第j个样品的距离。则dij需满足如下四个条件。

$$X=\begin{bmatrix} x_{11}x_{12}\cdots x_{1p} \\ x_{21}x_{22}\cdots x_{2p} \\ \vdots \\ x_{n1}x_{n2}\cdots x_{np} \end{bmatrix}$$

- $d_{ij}\geq 0$ 对一切的i和j;
- $d_{ij}=0$ 等价于i等于j;
- $d_{ij}=d_{ji}$ 对一切的i和j;
- $d_{ij}\leq d_{ik}+d_{kj}$ 对一切的i,j,k。

常用距离测量方法

- 明氏（Minkowski）距离：

$$d_{ij}(q)=\left(\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right|^{q}\right)^{1/q}$$

- 当q=1时，称为绝对（city-block)距离：

$$d_{ij}(1)=\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right|$$

- 当q=2时，称为欧氏(euclidean)距离：

$$d_{ij}(2)=\left(\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right|^{2}\right)^{1/2}$$

- 当q= ∞ 时，称为切比雪夫(chebychev)距离：

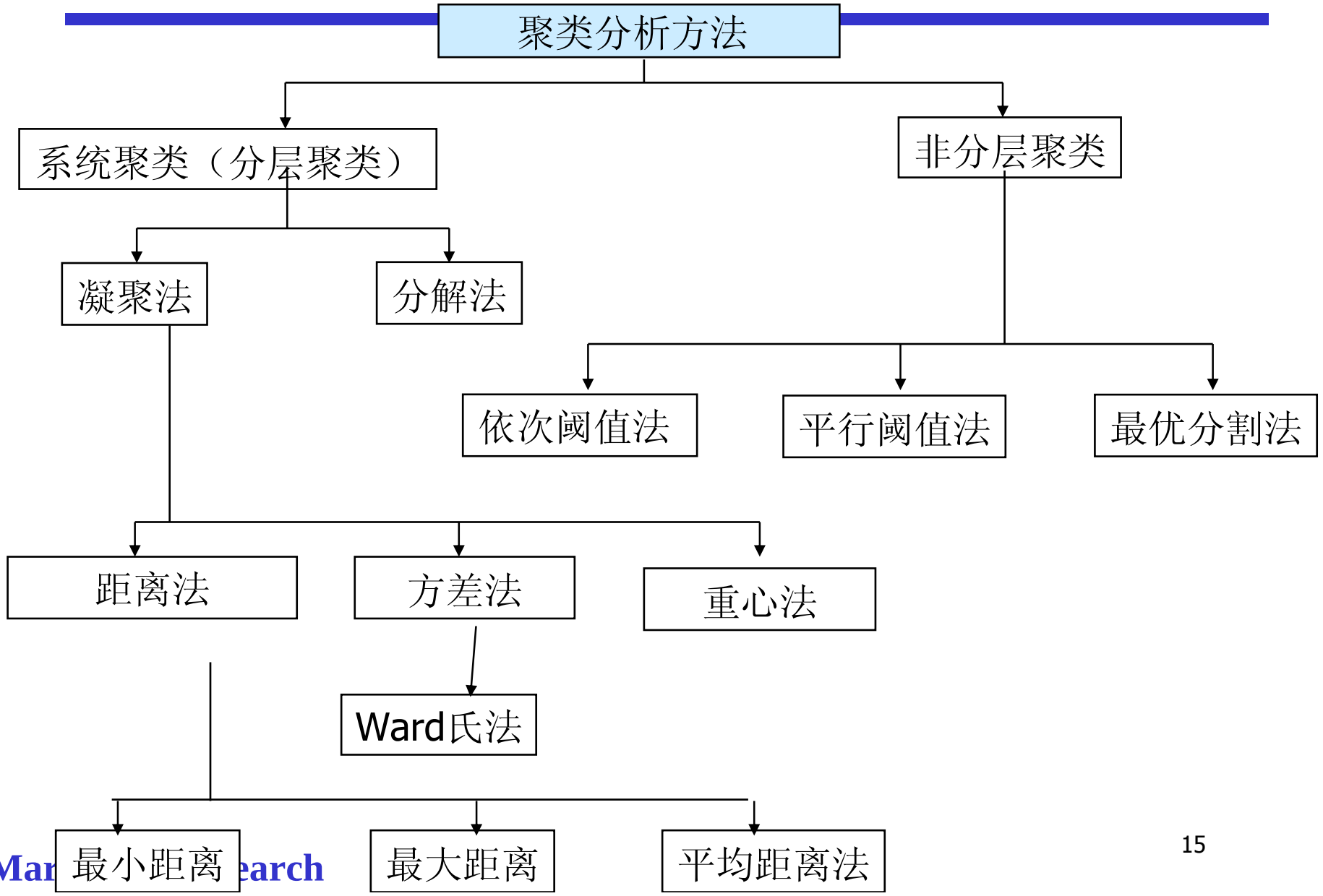
$$d_{ij}(\infty)=\max_{1\le k\le p}\left|X_{ik}-X_{jk}\right|$$

距离的测量方法

- 自定义距离(customized)

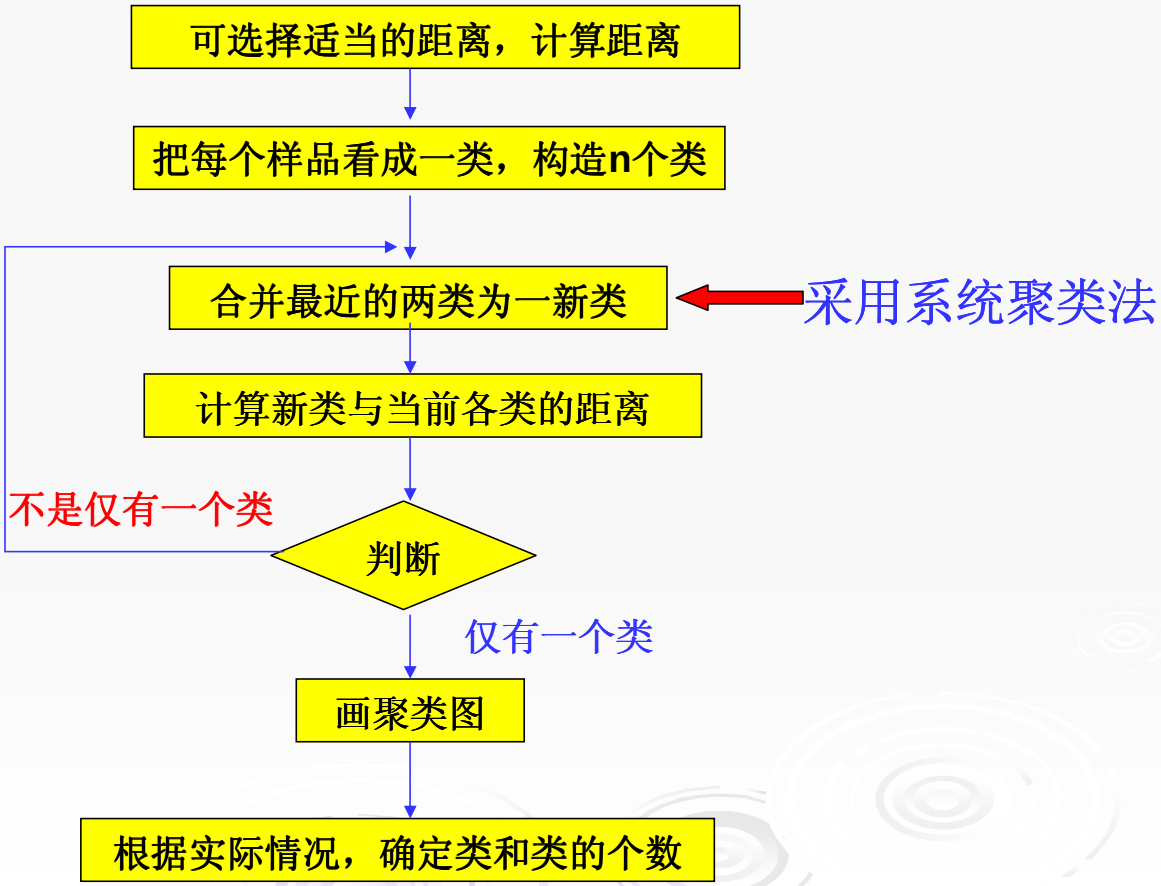
$$d_{ij}(q,r) = \left(\sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|^{q}\right)^{1/r}$$

聚类分析方法分类


聚类分析方法
系统聚类（分层聚类）
非分层聚类
凝聚法
分解法
依次阈值法
平行阈值法
最优分割法
距离法
方差法
重心法
Ward氏法
最小距离
最大距离
平均距离法

系统聚类____系统聚类法步聚

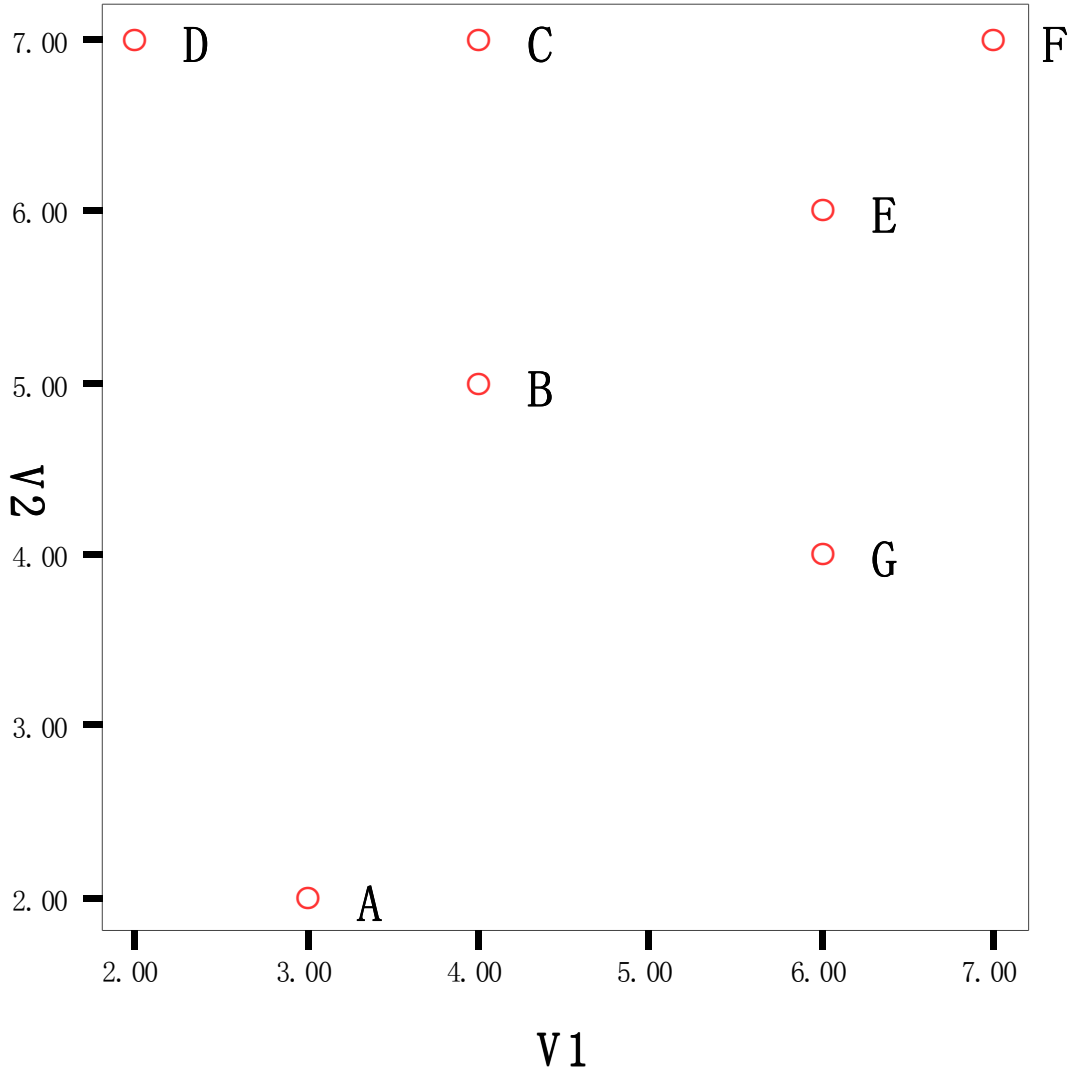
系统聚类法的步骤



引例

	A	B	C	D	E	F	G
V1	3	4	4	2	6	7	6
V2	2	5	7	7	6	7	4

引例



引例

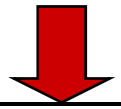
Proximity Matrix

	Euclidean Distance						
	1:A	2:B	3:C	4:D	5:E	6:F	7:G
1:A	.000	3.162	5.099	5.099	5.000	6.403	3.606
2:B	3.162	.000	2.000	2.828	2.236	3.606	2.236
3:C	5.099	2.000	.000	2.000	2.236	3.000	3.606
4:D	5.099	2.828	2.000	.000	4.123	5.000	5.000
5:E	5.000	2.236	2.236	4.123	.000	1.414	2.000
6:F	6.403	3.606	3.000	5.000	1.414	.000	3.162
7:G	3.606	2.236	3.606	5.000	2.000	3.162	.000

This is a dissimilarity matrix

最短距离法(Single Linkage)

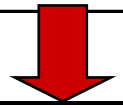
	1:A	2:B	3:C	4:D	5:E	6:F	7:G
1:A	0						
2:B	3.16	0.00					
3:C	5.10	2.00	0.00				
4:D	5.10	2.83	2.00	0.00			
5:E	5.00	2.24	2.24	4.12	0.00		
6:F	6.40	3.61	3.00	5.00	1.41	0.00	
7:G	3.61	2.24	3.61	5.00	2.00	3.16	0



	1:A	2:B	3:C	4:D	(E, F)	7:G
1:A	0					
2:B	3.16	0.00				
3:C	5.10	2.00	0.00			
4:D	5.10	2.83	2.00	0.00		
(E, F)	5.00	2.24	2.24	4.12	0.00	
7:G	3.61	2.24	3.61	5.00	2.00	0

最长距离法(Complete Linkage)

	1:A	2:B	3:C	4:D	5:E	6:F	7:G
1:A	0						
2:B	3.16	0.00					
3:C	5.10	2.00	0.00				
4:D	5.10	2.83	2.00	0.00			
5:E	5.00	2.24	2.24	4.12	0.00		
6:F	6.40	3.61	3.00	5.00	1.41	0.00	
7:G	3.61	2.24	3.61	5.00	2.00	3.16	0



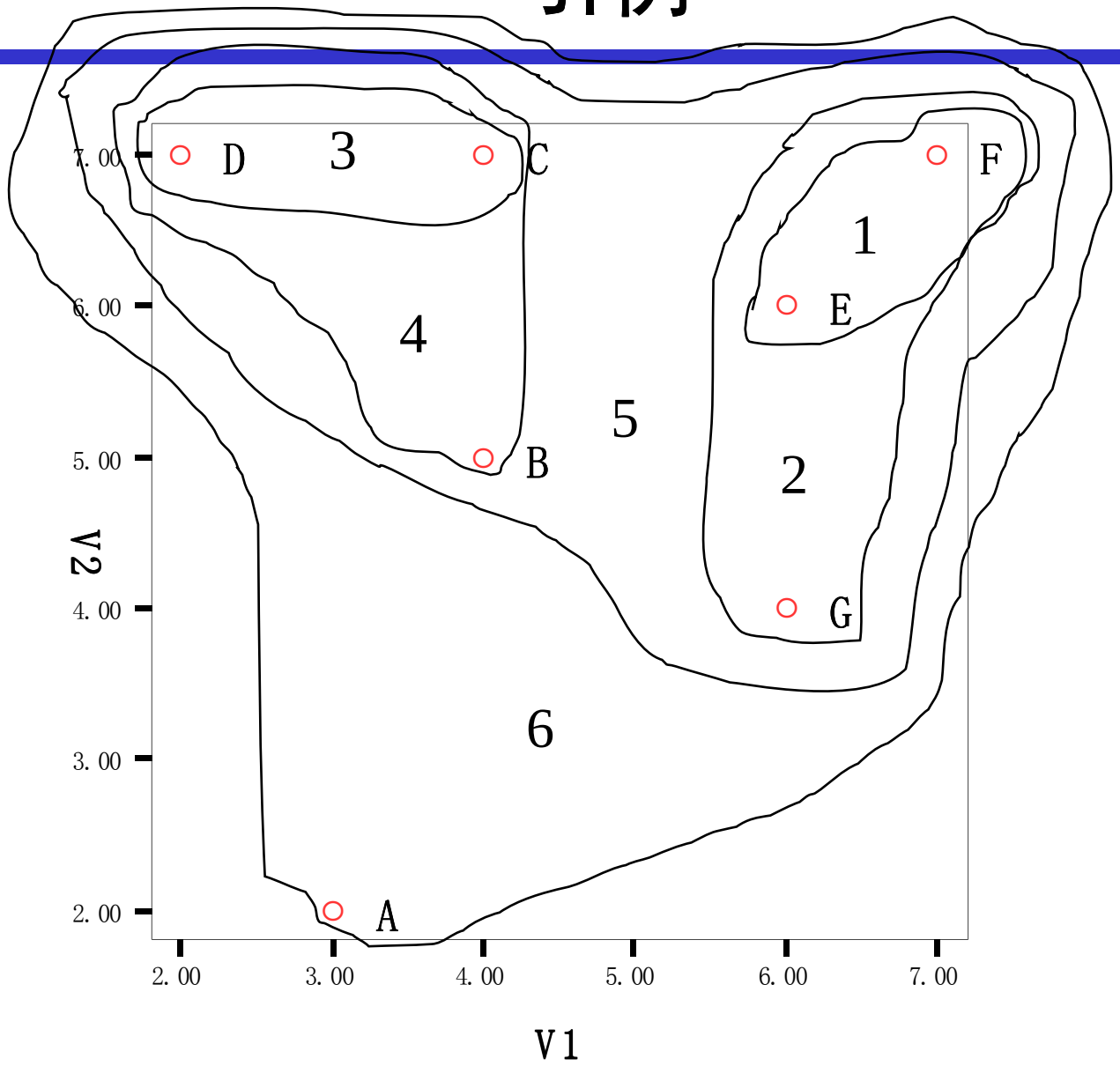
	1:A	2:B	3:C	4:D	(E, F)	7:G
1:A	0					
2:B	3.16	0.00				
3:C	5.10	2.00	0.00			
4:D	5.10	2.83	2.00	0.00		
(E, F)	6.40	3.61	3.00	5.00	0.00	
7:G	3.61	2.24	3.61	5.00	3.16	0

聚合过程表

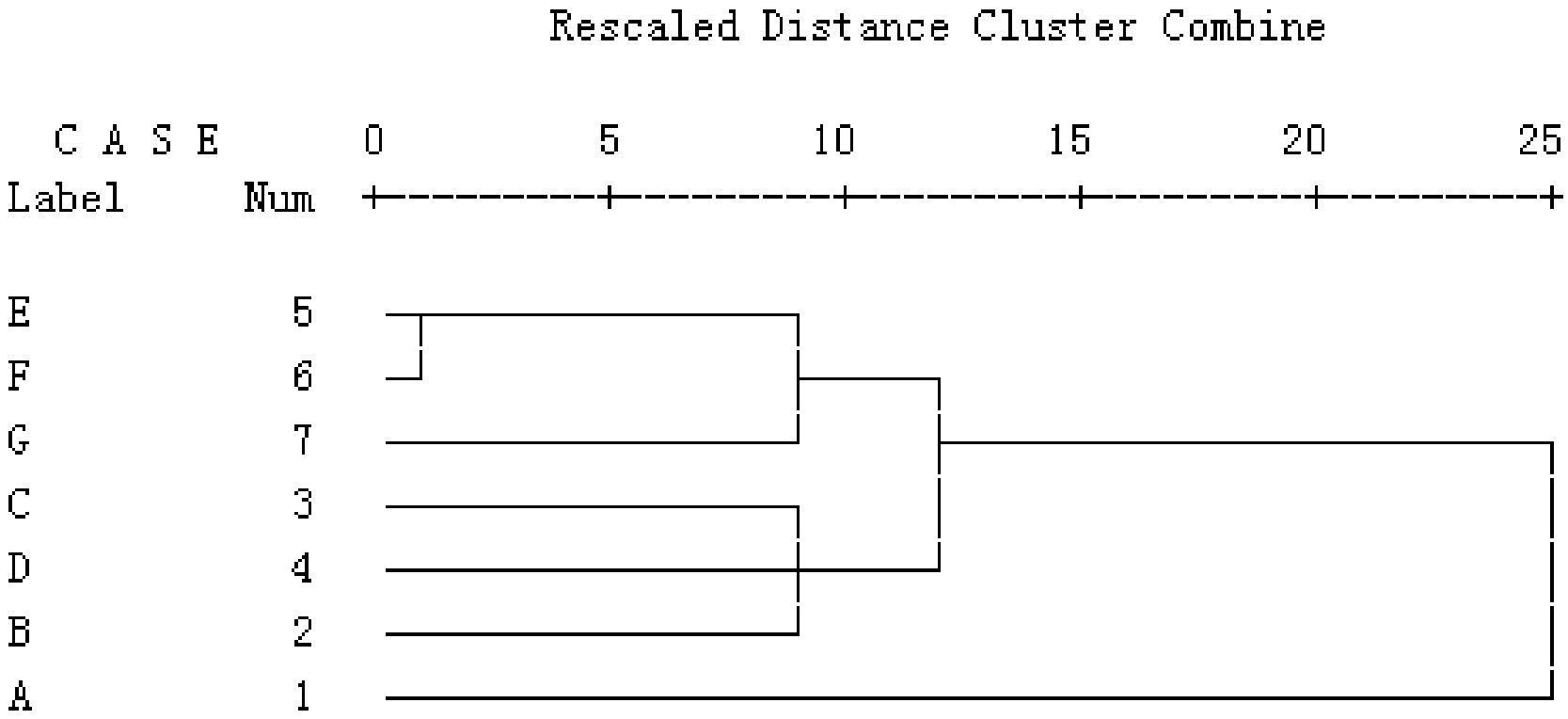
Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	6	1.414	0	0	2
2	5	7	2.000	1	0	5
3	3	4	2.000	0	0	4
4	2	3	2.000	0	3	5
5	2	5	2.236	4	2	6
6	1	2	3.162	0	5	0

引例


D
3
C
F
1
E
4
5
B
2
G
6
A
7. 00
6. 00
5. 00
4. 00
3. 00
2. 00
V2
2. 00
3. 00
4. 00
5. 00
6. 00
7. 00
V1

树状图(Dendrogram)



系统聚类法

- 重心法——**Centroid Clustering**

$$D_{pq} = \min d(\overline{x}_p, \overline{x}_q)$$

- 类平均法——**Between-groups Linkage**

$$D_{pq} = \frac{1}{n_1 n_2} \sum_{x_i \in G_p} \sum_{x_j \in G_q} d(x_i, x_j)$$

系统聚类法

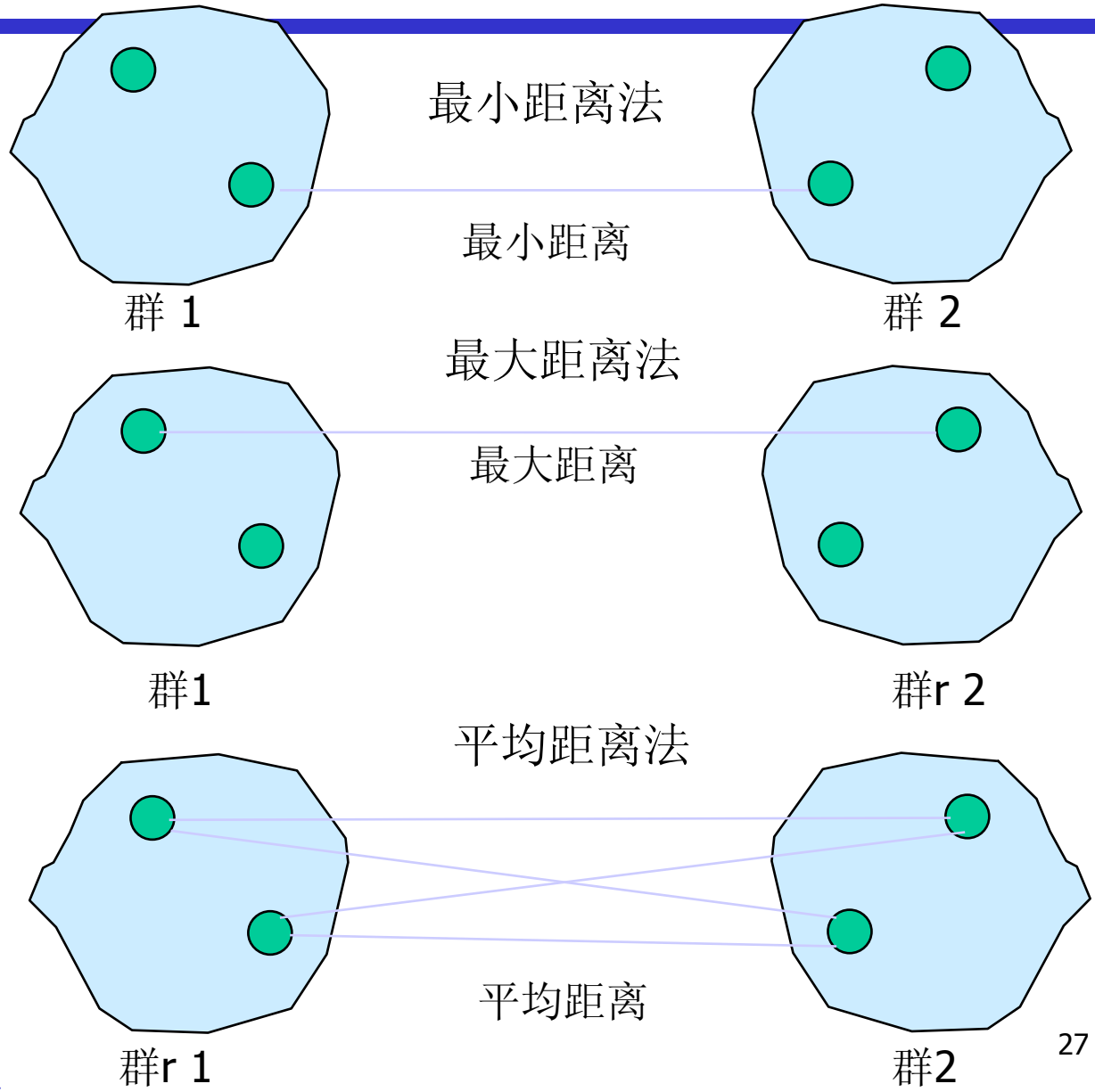
- 离差平方和法——**Word's Method**

$$D_1 = \sum_{x_i \in G_p} (x_i - \overline{x}_p)'(x_i - \overline{x}_p), D_2 = \sum_{x_j \in G_q} (x_j - \overline{x}_q)'(x_j - \overline{x}_q),$$

$$D_{1+2} = \sum_{x_k \in G_p \cup G_q} (x_k - \overline{x})'(x_i - \overline{x}) \Rightarrow D_{pq} = D_{1+2} - D_1 - D_2$$

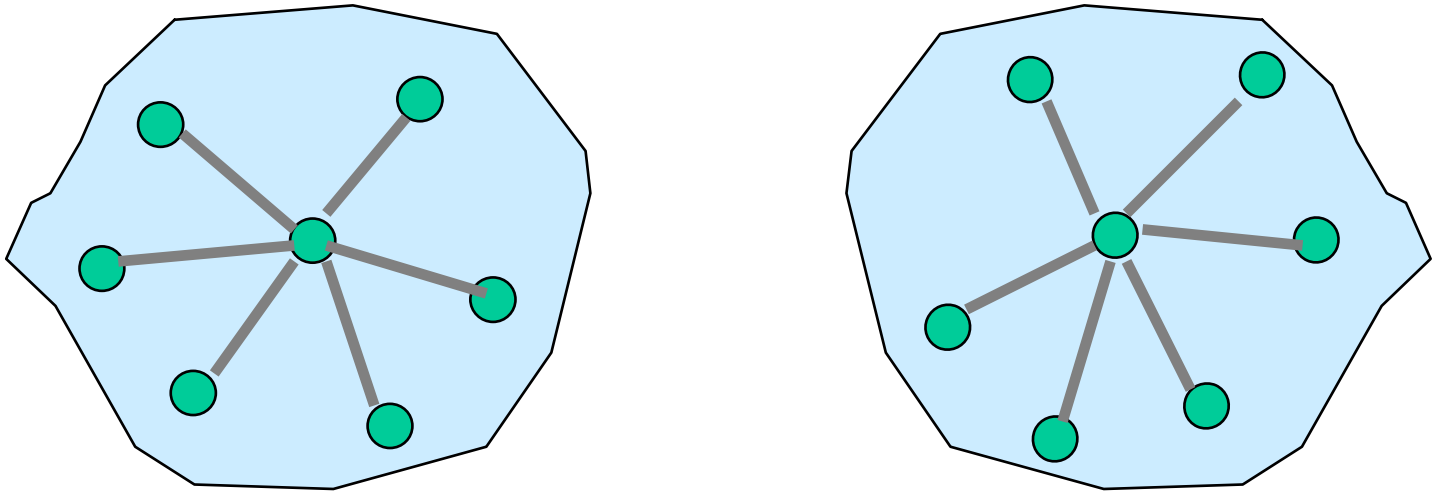
它的思想来源于方差分析

距离聚类法

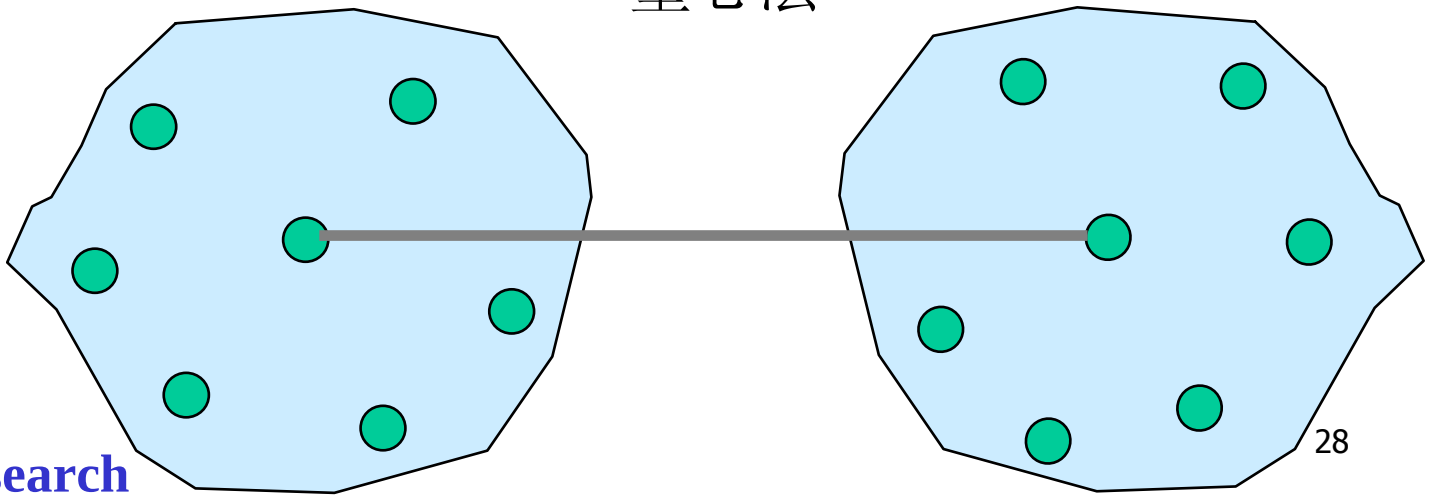


其他聚合聚类方法

Ward氏法



重心法



快速聚类法

- 快速聚类法（ *k*-means clustering）包括三种方法：依次阈值法 （sequential threshold）, 平行阈值法（parallel threshold）, 最优分割法（ optimizing partitioning）.
- 快速聚类（K-means)的优点：

  第一，无需计算所有个体间距离；

  第二，可允许研究者自己决定丛数；

  第三，一个个体进入某丛后可以退出来加入另外一丛。换句话说，为了使聚类结果尽可能完善，已被聚类的对象能被重新聚类。

快速聚类法与系统聚类法应用区别

- 系统聚类法的聚类过程是单方向的，一旦某个样品（case）进入某一类，就不可能从该类出来，再归入其他的类。
- 而快速聚类法受奇异值、相似测度和不合使得聚类变量的影响较小，对于不合适的初始分类可以进行反复调整。
- 在聚类分析发展的早期，系统聚类法应用普遍，其中尤以组间类平均法和离差平方和法应用最广。
- 后来快速聚类方法逐步被人们接受，应用日益增多。现在是两者相结合，取长补短。
- 首先使用系统聚类法确定分类数，检查是否有奇异值，去除奇异值后，对剩下的案例重新进行分类，把用系统聚类法得到的各个类的重心，作为迭代法的初始分类中心，对样本进行重新调整。

快速聚类法与系统聚类法应用区别

- 聚类结果中，如果孤类点太多，则说明该种聚类方法不好。如果从减少孤类来看，一般情况下用Ward氏方法最好。

系统聚类法上机结果

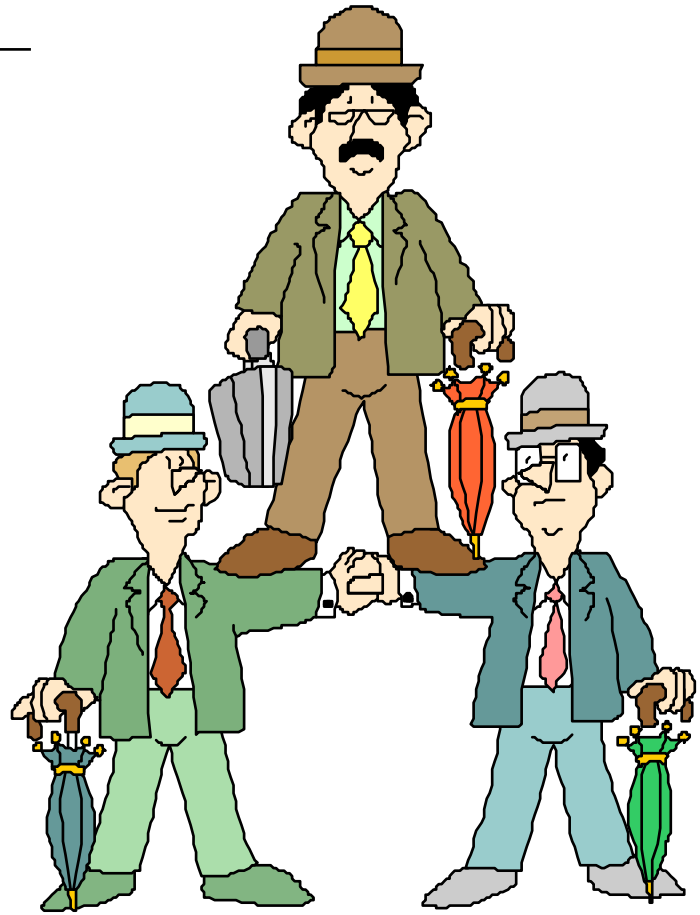
Ward氏法聚合过程表

	合并的群			群首次出现的步骤		
步骤	群 1	群 2	系数	群 1	群 2	下一步
1	14	16	1.000000	0	0	6
2	6	7	2.000000	0	0	7
3	2	13	3.500000	0	0	15
4	5	11	5.000000	0	0	11
5	3	8	6.500000	0	0	16
6	10	14	8.160000	0	1	9
7	6	12	10.166667	2	0	10
8	9	20	13.000000	0	0	11
9	4	10	15.583000	0	6	12
10	1	6	18.500000	6	7	13
11	5	9	23.000000	4	8	15
12	4	19	27.750000	9	0	17
13	1	17	33.100000	10	0	14
14	1	15	41.333000	13	0	16
15	2	5	51.833000	3	11	18
16	1	3	64.500000	14	5	19
17	4	18	79.667000	12	0	18
18	2	4	172.662000	15	17	19
19	1	2	328.600000	16	18	0

系统聚类法上机结果

Ward氏法的各群成员构成

样本编号	群数 4	3	2
1	1	1	1
2	2	2	2
3	1	1	1
4	3	3	2
5	2	2	2
6	1	1	1
7	1	1	1
8	1	1	1
9	2	2	2
10	3	3	2
11	2	2	2
12	1	1	1
13	2	2	2
14	3	3	2
15	1	1	1
16	3	3	2
17	1	1	1
18	4	3	2
19	3	3	2
20	2	2	2

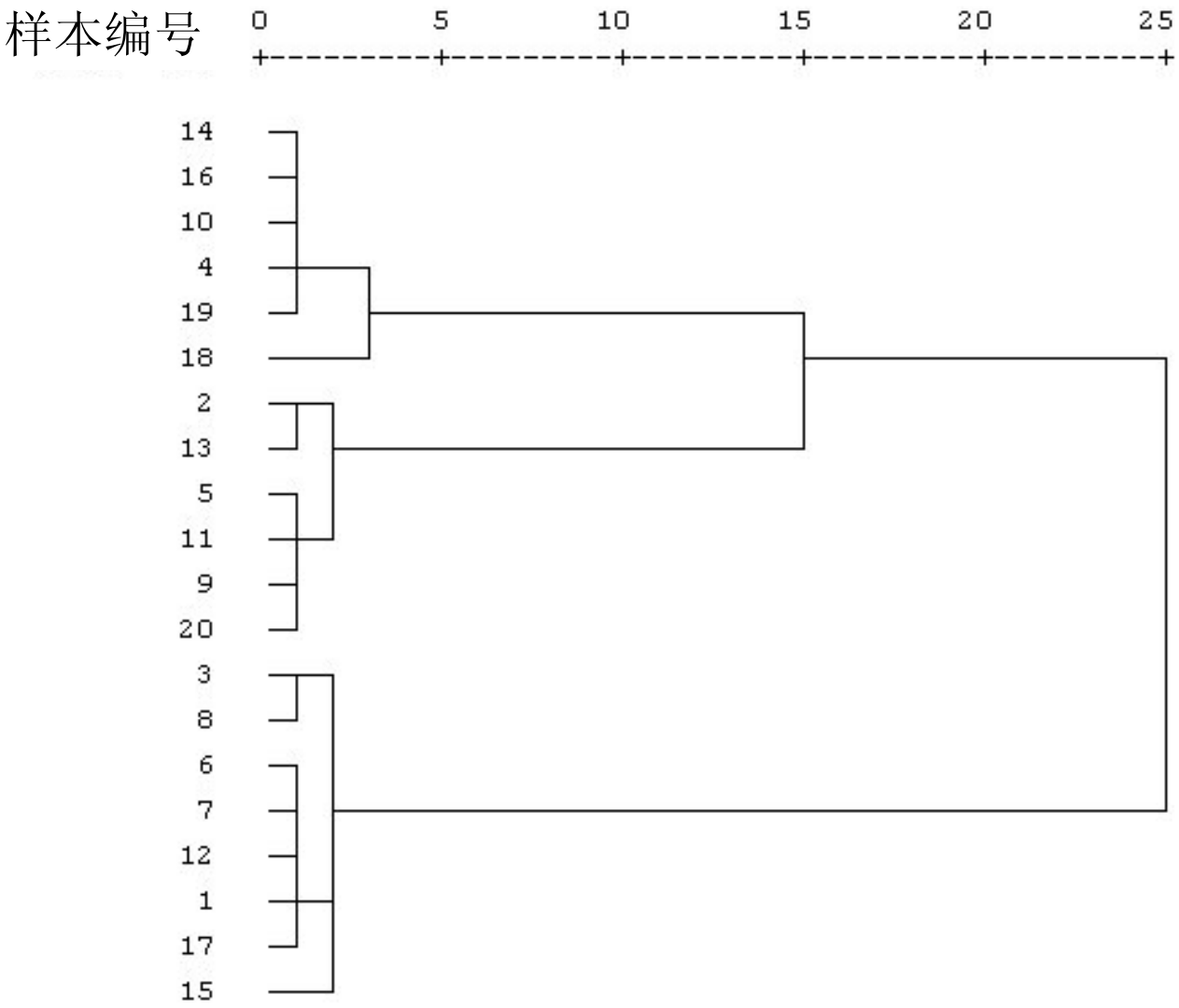


冰柱图 Vertical Icicle Plot

群数 Number of clusters	18		19		16		14		10		4		20		9		11		5
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
4	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
5	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
6	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
7	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
8	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
9	X		X		X	X	X	X	X	X	X		X	X	X	X	X	X	X
10	X		X		X	X	X	X	X	X	X		X	X	X		X	X	X
11	X		X		X	X	X	X	X	X	X		X	X	X		X	X	X
12	X		X		X	X	X	X	X		X		X	X	X		X	X	X
13	X		X		X	X	X	X	X		X		X		X		X	X	X
14	X		X		X	X	X	X	X		X		X		X		X	X	X
15	X		X		X	X	X		X		X		X		X		X	X	X
16	X		X		X	X	X		X		X		X		X		X	X	X
17	X		X		X	X	X		X		X		X		X		X		X
18	X		X		X	X	X		X		X		X		X		X		X
19	X		X		X	X	X		X		X		X		X		X		X

树状图 Dendrogram

标准化后的群间距离



聚类分析——确定群组数目

- 理论、概念和实际的考虑可能提供确切群数的依据
- 分层聚类时，各群合并时的距离可以作为确切群数的标准。可以从聚合表或树状图获得这一信息
- 非分层聚类时，可以将群内总方差与群间方差的比值和相应的群数制图。折点处就是合适的群数，群数超过这一点后通常是不值得的
- 群的相对大小应当合适

聚类分析——解释与描述群组

- 对群的解释与描述涉及对群重心的考察，可以根据群重心给每一群起一个合适的名字来描述该群的特征
- 通常还需要用聚类时没有用到的变量来描述各群，这些变量可以是人口统计变量、心理测试变量、产品使用、媒体使用或其他变量

群重心

变量均值

群编号	V_1	V_2	V_3	V_4	V_5	V_6
1	5.750	3.625	6.000	3.125	1.750	3.875
2	1.667	3.000	1.833	3.500	5.500	3.333
3	3.500	5.833	3.333	6.000	3.500	6.000

聚类分析——评估信度与效度

1. 用不同的距离指标对同一数据进行聚类分析，然后对结果加以比较，以确定其稳定性
2. 采用不同的聚类方法并比较其结果
3. 将数据随机的分成两半，分别进行聚类分析，然后比较两个子样本的群重心
4. 随机删除一些变量，用剩下的变量进行聚类分析，然后比较两个子样本的群重心
5. 非分层聚类的结果可能取决于数据中样本的排列顺序，用不同的顺序反复进行聚类分析直至结果稳定

快速聚类法步骤

- 1.先用分层聚类法判断需分几层。若有必要，需对数据标准化。
- 2.在用K－MEANS法前先对数据标准化。
- 3.样本容量小时，用每次加入一个个体后即更新平均数的办法。
- 4.可给出聚类后的方差分析表。

非分层聚类结果

起始群中心

	群		
	1	2	3
V1	4	2	7
V2	6	3	2
V3	3	2	6
V4	7	4	4
V5	2	7	1
V6	7	2	3

迭代过程

	群中心变化		
迭代	1	2	3
1	2.154	2.102	2.550
2	0.000	0.000	0.000

非分层聚类结果

对象所属群别

编号	群	距离
1	3	1.414
2	2	1.323
3	3	2.550
4	1	1.404
5	2	1.848
6	3	1.225
7	3	1.500
8	3	2.121
9	2	1.756
10	1	1.143
11	2	1.041
12	3	1.581
13	2	2.598
14	1	1.404
15	3	2.828
16	1	1.624
17	3	2.598
18	1	3.555
19	1	2.154
20	2	2.102

非分层聚类结果

最终群中心

	群		
	1	2	3
V1	4	2	6
V2	6	3	4
V3	3	2	6
V4	6	4	3
V5	4	6	2
V6	6	3	4

群中心之间的距离

群	1	2	3
1		5.568	5.698
2	5.568		6.928
3	5.698	6.928	

非分层聚类结果

方差分析

	群		误差			
	Mean Square	df	Mean Square	df	F	Sig.
V1	29.108	2	0.608	17	47.888	0.000
V2	13.546	2	0.630	17	21.505	0.000
V3	31.392	2	0.833	17	37.670	0.000
V4	15.713	2	0.728	17	21.585	0.000
V5	22.537	2	0.816	17	27.614	0.000
V6	12.171	2	1.071	17	11.363	0.001

F检验只起描述性作用，因为以群间差异最大化原则将对象系统地分配到各群，因此F检验的概率不能作为检验群间无差异的零假设的依据。

每群例数

群	1	6.000
	2	6.000
	3	8.000
有效例数		20.000
缺失		0.000

两步法聚类

优点：

- 分析时既可以用分类变量，也可以用连续变量
- 可以自动选择组群数
- 特别适宜于大样本数据

两步法聚类原理

- 第一步先建立一个聚类谱系树(Cluster Features Tree)。首先把样本中的第一个体放入一个分支点(leaf node). 后续个体根据和第一个体的相似性或距离(即不相似性)，或者放入同一个分支点，或者开辟另一个分支点。每个分支点都对所含的个体按照聚类所用的变量总结出其特征。把所有分支点信息加总即为整个样本数据的总结。
- 第二步在谱系树基础上进行整合聚类过程(agglomerative clustering). 该过程可以给出不同的组群数。要知道那个组群数是最佳的，就要把含有不同组群数的聚类结果和斯瓦氏贝叶斯聚类标准(Schwarz's Bayesian Criterion (BIC))或赤池信息标准(Akaike Information Criterion (AIC))进行比较。

- 如果是小样本并且想选择不同聚类方法，或者有变量需要转换，或需要测量不同从间距离就应该用分层聚类法(Hierarchical Cluster Analysis)。后者还可以对变量而不是个体聚类。
- 如果有大样本，且变量都是连续变量则应该用 K 平均数法(K-Means Cluster Analysis)。后者还可以保存每个个体到从重心的距离。

SPSS 窗口

要选择SPSS for Windows的该程序，点击：

Analyze>Classify>Hierarchical Cluster …

Analyze>Classify>K-Means Cluster …

Analyze>Classify>TwoStep Cluster …