

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



第三章 聚类分析

主讲教师 周仕君

- 聚类分析是研究如何将研究对象按照多个方面的特征进行综合分类的一种统计方法。
- 以往的分类学中，主要靠经验和专业知识做定性分类，往往带有主观性和任意性，不能揭示客观事物内在的本质差别和联系；或者只根据事物单方面的特征进行分类，往往难以反映各类事物之间的综合差异。
- 聚类分析有效解决了科学研究中多因素、多指标的分类问题。

分类与聚类的区别：

- **分类**：是对当前所研究的问题已知它的类别数目及各类的特征（例如分布规律或来自各类的训练样本），目的是要将另一些未知类别的个体正确地归属于其中某一类。判别分析就是一种分类技术。
- **聚类**：是事先不知道研究的问题应分为几类，更不知道观测到的个体的具体分类情况。目的是通过对观测数据所进行的分析处理，选定一种度量个体接近程度的统计量，确定分类数目，建立一种分类方法，并按接近程度对观测对象给出合理的分类。

-
- 聚类分析就是分析如何对样品（或变量）按照它们在性质上的亲疏程度进行量化分类的问题。
 - 通常聚类分析分为Q型聚类和R型聚类。
 - Q型聚类是对样品进行分类处理，
 - R型聚类是对变量进行分类处理。

聚类方法的分类：

 系统聚类法——（分层聚类Hierarchical Cluster）系统聚类法是应用最广泛的一种聚类方法。

 聚类原则：都是相近的聚为一类，即距离最近或最相似的聚为一类。

 非系统聚类法——（快速聚类法—K-均值聚类法K-means Cluster）;

第三章 聚类分析

- 第一节 相似性的量度
- 第二节 系统聚类分析法
- 第三节 K均值聚类

第一节 相似性的量度

□ 样品相似性的度量

□ 变量相似性的度量

一、样品相似性的度量

- Q型聚类分析，常用距离来测度样品之间的相似程度。两个样品间相似程度就可用 p 维空间中的两点距离公式来度量。样本阵如下：

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

- 因此，任何两个样品 X_i 与 X_j 之间的相似性可以通过矩阵中的第 i 行与第 j 行的相似程度来刻画。

1. 明考夫斯基距离

□ 令 d_{ij} 表示样品 X_i 与 X_j 的距离，则明考夫斯基的距离公式为：

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

其中，当 q 取不同的值时，又会产生不同的距离。

明考夫斯基距离简称明氏距离，按的取值不同
又可分为：

□ 绝对距离($q=1$)

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

□ 欧氏距离 ($q=2$)

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

□ 切比雪夫距离($q=\infty$)

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

注意：

明氏距离的不足之处及解决办法：

- ① 明氏距离没有考虑指标的数量级水平及量纲。
当各变量数量级相差悬殊时且量纲不同时，
采用明氏距离并不合理。
- ◆ 解决办法：先对数据标准化，然后用标准化后的数据计算距离。

② 使用欧氏距离要求各坐标对欧氏距离的贡献是同等且**方差**大小也是相同的。这样的欧氏距离才合适，效果也较好，否则就不能如实的反映情况。

◆ 解决办法是对坐标加权，产生“统计距离”。

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

2.马氏距离

- 设 X_i 与 X_j 是来自均值向量为 μ ，协方差为 Σ 的总体 G 中的 p 维样品，则两个样品间的马氏距离为：

$$d_{ij}^2(M) = (\mathbf{X}_i - \mathbf{X}_j)' \Sigma^{-1} (\mathbf{X}_i - \mathbf{X}_j)$$

- 实际应用中，总体协差阵未知可以用样本协差阵来代替。

注意：

马氏距离的优点在于：

- ① 马氏距离又称为广义欧氏距离。马氏距离考虑了观测变量之间的相关性。
- ② 马氏距离还考虑了观测变量之间的变异性，不再受各指标量纲的影响。
- ③ 将原始数据作线性变换后，马氏距离不变。

3. 兰氏距离

□ 兰氏距离的计算公式为：

$$d_{ij}(L) = \frac{1}{p} \sum_{k=1}^p \frac{|X_{ik} - X_{jk}|}{X_{ik} + X_{jk}}$$

□ 它仅适用于一切 $X_{ij} > 0$ 的情况。

兰氏距离的特点：

- ① 这个距离也可以克服各个指标之间量纲的影响，这是一个自身标准化的量。
- ② 由于它对大的奇异值不敏感，它特别适合于高度偏倚的数据。
- ③ 虽然这个距离有助于克服明氏距离的第一个缺点，但它也没有考虑指标之间的相关性。

例：某地区9个农业区的7项指标。

区代 号	人均 耕地 X_1	劳均 耕地 X_2	水田 比重 X_3	复种 指数 X_4	粮食 单产 X_5	人均粮 食 X_6	稻谷占 粮食比 重 X_7
G_1	0.294	1.093	5.63	113.6	4 510.5	1 036.4	12.2
G_2	0.315	0.971	0.39	95.1	2 773.5	683.7	0.85
G_3	0.123	0.316	5.28	148.5	6 934.5	611.1	6.49
G_4	0.179	0.527	0.39	111	4 458	632.6	0.92
G_5	0.081	0.212	72.04	217.8	12 249	791.1	80.38
G_6	0.082	0.211	43.78	179.6	8 973	636.5	48.17
G_7	0.075	0.181	65.15	194.7	10 689	634.3	80.17
G_8	0.293	0.666	5.35	94.9	3	771.7	7.8

极差标准化处理后的数据

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
G_1	0.91	1.00	0.07	0.15	0.18	1.00	0.14
G_2	1.00	0.87	0.00	0.00	0.00	0.24	0.00
G_3	0.20	0.15	0.07	0.44	0.44	0.08	0.07
G_4	0.44	0.38	0.00	0.13	0.18	0.13	0.00
G_5	0.03	0.03	1.00	1.00	1.00	0.45	1.00
G_6	0.03	0.03	0.61	0.69	0.65	0.13	0.59
G_7	0.00	0.00	0.90	0.81	0.84	0.13	1.00
G_8	0.91	0.53	0.07	0.00	0.10	0.43	0.09
G_9	0.38	0.26	0.04	0.00	0.15	0.00	0.00

- 把任何两个样品的距离算出后可形成距离阵，聚类即是在距离阵的基础上进行的，其中的距离值越小表明两样品越相似。
- 上例中9个农业区之间的绝对值距离矩阵如下

$$D = (d_{ij})_{9 \times 9} = \begin{bmatrix} 0 & & & & & & & & \\ 1.52 & 0 & & & & & & & \\ 3.10 & 2.70 & 0 & & & & & & \\ 2.19 & 1.47 & 1.23 & 0 & & & & & \\ 5.86 & 6.02 & 3.64 & 4.77 & 0 & & & & \\ 4.72 & 4.46 & 1.86 & 2.99 & 1.78 & 0 & & & \\ 5.79 & 5.53 & 2.93 & 4.06 & 0.83 & 1.07 & 0 & & \\ 1.32 & 0.88 & 2.24 & 1.29 & 5.14 & 3.96 & 5.03 & 0 & \\ 2.62 & 1.66 & 1.20 & 0.51 & 4.84 & 3.06 & 3.32 & 1.40 & 0 \end{bmatrix}$$

4. 距离选择的原则

通常选择距离公式应注意遵循以下的基本原则：

- ① 要考虑所选择的距离公式在实际应用中的意义。
- ② 要综合考虑对样本观测数据的预处理和将要采用的聚类分析方法。
- ③ 要考虑研究对象的特点和计算量的大小。

5.数据的标准化

1. 中心化变换

称变换

$$x_{ij}^* = x_{ij} - \bar{x}_j \quad (i = 1, 2, \dots, n; j = 1, \dots, m)$$

为中心化变换. 变换后数据的均值为 0, 而协方差阵不变, 即协方差阵为

$$S^* = S = (s_{ij})_{m \times m},$$

$$\text{其中 } s_{ij} = \frac{1}{n-1} \sum_{i=1}^n (x_{i\cdot} - \bar{x}_i)(x_{ij} - \bar{x}_j) = \frac{1}{n-1} \sum_{i=1}^n x_{i\cdot}^* x_{ij}^*.$$

中心化变换是一种方便地计算样本协方差阵的变换.

2. 标准化变换 称变换

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} \quad \begin{cases} i = 1, 2, \dots, n \\ j = 1, 2, \dots, m \end{cases}$$

为标准化变换. 变换后的数据, 每个变量的样本均值为 0, 标准差为 1, 而且标准化变换后的数据 $\{x_{ij}^*\}$ 与变量的量纲无关.

3. 极差标准化变换 称变换

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{R_j} \quad \begin{cases} i = 1, 2, \dots, n \\ j = 1, 2, \dots, m \end{cases}$$

为极差标准化变换. 变换后的数据, 每个变量的样本均值为 0, 极差为 1, 且 $|x_{ij}^*| < 1$, 在以后的分析计算中可以减少误差的产生; 同时变换后的数据也是无量纲的量.

4. 极差正规化变换(规格化变换) 称变换

$$x_{ij}^* = \frac{x_{ij} - \min_{1 \leq t \leq n} x_{tj}}{R_j} \quad \begin{cases} i = 1, 2, \dots, n \\ j = 1, 2, \dots, m \end{cases}$$

为极差正规化变换. 变换后的数据 $0 \leq x_{ij}^* \leq 1$; 极差为 1, 也是无量纲的量.

5. 对数变换 称变换

$$x_{ij}^* = \ln(x_{ij}) \quad (\text{要求 } x_{ij} > 0, i = 1, 2, \dots, n; j = 1, 2, \dots, m)$$

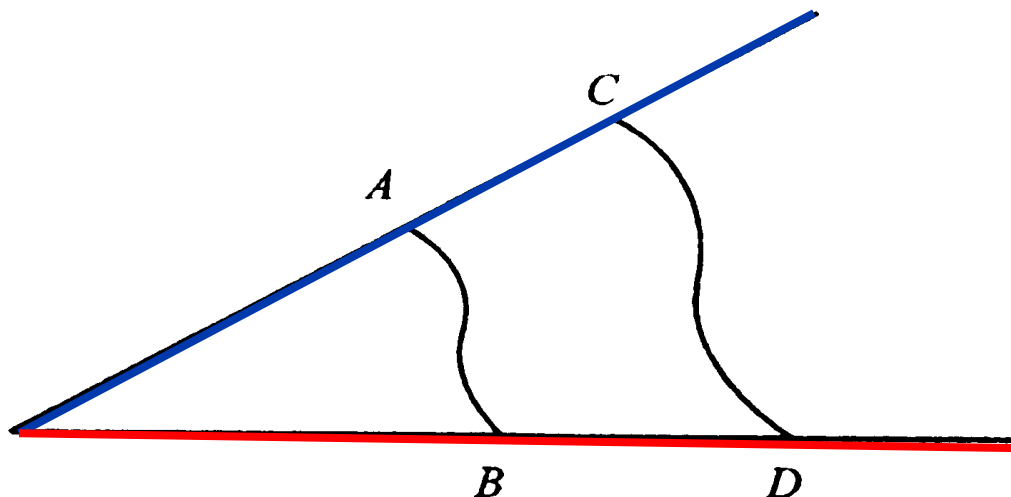
为对数变换. 它可将具有指数特征的数据结构变换为线性数据结构.

二、变量相似性的度量

- 变量间的相似性要从它们的方向趋同性或“相关性”进行考察，度量方法有：“夹角余弦法”和“相关系数”。

1. 夹角余弦

□ 两变量 X_i 与 X_j 看作 n 维空间(n 个样品所张成的空间)的两个向量, 则:



□ 夹角余弦是两向量在 origin 处的夹角余弦。如果二者的取值关系密切, 那么二者的夹角就接近于零, 其夹角余弦就接近于1。

2. 相关系数

- 相关系数经常用来度量变量间的相似性。变量 X_i 与 X_j 的相关系数定义为：

$$r_{ij} = \frac{\sum_{k=1}^n (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{\sqrt{\sum_{k=1}^n (X_{ki} - \bar{X}_i)^2 \sum_{k=1}^n (X_{kj} - \bar{X}_j)^2}}$$

- 显然 $|r_{ij}| < 1$ 。事实上，相关系数是将样本观测数据中心化或标准化后的夹角余弦。

注意：

□ 无论是夹角余弦还是相关系数，把它们统称为相似系数，记为 c_{ij} 。

- ① 当 $|c_{ij}| = 1$ 时，说明变量 X_i 与 X_j 完全相似；
- ② 当 $|c_{ij}|$ 近似于1时，说明变量 X_i 与 X_j 非常密切；
- ③ 当 $|c_{ij}| = 0$ 时，说明变量 X_i 与 X_j 完全不一样；
- ④ 当 $|c_{ij}|$ 近似于0时，说明变量 X_i 与 X_j 差别很大。

□ 如果需要用距离来测定变量间的亲疏程度，
则变量之间常借助于相似系数来定义距离：

$$d_{ij} = 1 - |c_{ij}|$$

或者

$$d_{ij}^2 = 1 - c_{ij}^2$$

第二节 系统聚类分析法

- 系统聚类的基本思想
- 类间距离与系统聚类法
- 类间距离的统一性

一、系统聚类的基本思想

- 系统聚类的基本思想是：**距离**相近的样品（或变量）先聚成类，**距离**相远的后聚成类，过程一直进行下去，每个样品（或变量）总能聚到合适的类中。
- 有时为了直观地反映系统聚类过程，可以把整个分类系统画成一张谱系图(**dendrogram**)。因此，系统聚类也称为谱系分析。

系统聚类过程是：

- ① 假设总共有 n 个样品（或变量），首先将每个样品（或变量）独自聚成一类，共有 n 类；然后根据所确定的样品（或变量）“距离”公式，形成初始距离阵。之后，将其中距离较近的两个样品（或变量）聚合为一类，其它的样品（或变量）仍各自聚为一类。
- ② 第二步再根据新合并类与其他类的“距离”计算公式，在形成的新的距离阵中，将“距离”最近的两个类进一步再聚成一类；
- ③ 以上步骤一直进行下去，最后将所有的样品（或变量）全聚成一类。

二、类间距离与系统聚类法

- 系统聚类法有8种，分别为最短距离法、最长距离法、中间距离法、重心法、类平均法、可变类平均法、可变法和离差平方和法。
- 它们的归类步骤基本上是一致的，主要差异是：①两类间距离；②新合并类与其他类间距离的计算方法不同。

1. 最短距离法

 用 d_{ij} 表示样品 X_i 与 X_j 之间距离，用 D_{ij} 表示类 G_i 与 G_j 之间的距离。定义类 G_i 与 G_j 之间的距离为两类最近样品的距离，即为：

$$D_{ij} = \min_{X_i \in G_i, X_j \in G_j} d_{ij}$$

 最短距离法进行聚类分析的步骤如下：

① 定义样品之间距离，计算样品的两两距离，得一距离阵记为 $D_{(0)}$ ，因为开始每个样品自成一类，所以 $D_{ij} = d_{ij}$ 。

② 找出 $D_{(0)}$ 的非对角线上最小元素，设为 D_{pq} ，则将 G_p 和 G_q 合并成一个新的类，记为 G_r ，则 $G_r = \{G_p, G_q\}$

③ 计算新类与其它类的距离，然后重复第二步。

$$\begin{aligned} D_{kr} &= \min_{X_i \in G_k, X_j \in G_r} d_{ij} \\ &= \min \left\{ \min_{X_i \in G_k, X_j \in G_p} d_{ij}, \min_{X_i \in G_k, X_j \in G_q} d_{ij} \right\} = \min \{D_{kp}, D_{kq}\} \end{aligned}$$

④ 不断合并直到所有元素并成一类为止。如果某一步距离最小的元素不止一个，则对应这些最小元素的类可以同时合并。

例1：设有五位推销员，其教育水平和推销能力的评分如下，试用最短距离法将它们分类。

推销员	推销能力	教育程度
1	1.2	1.2
2	1.2	2.4
3	7.1	3.6
4	9.6	2.4
5	9.6	0

□ 在聚类前先标准化，用标准化后的数据进行聚类：

推销员	1	2	3	4	5
推销能力	0	0	0.7	1	1
教育程度	0.33	0.67	1	0.67	0

◆ 样品间采用绝对值距离: $d_{ij} = |X_i - X_j| + |Y_i - Y_j|$

$$D_0 = \begin{matrix} & G_1 & G_2 & G_3 & G_4 & G_5 \\ \begin{matrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{matrix} & \begin{bmatrix} 0 & 0.34 & 1.37 & 1.34 & 1.33 \\ & 0 & 1.03 & 1 & 1.67 \\ & & 0 & 0.63 & 1.30 \\ & & & 0 & 0.67 \\ & & & & 0 \end{bmatrix} \end{matrix}$$

结果: G1与G2合并成新类G6。

□ 计算G6与其他类的距离：

$$D_{63} = \min(d_{13}, d_{23}) = \min(1.37, 1.03) = 1.03$$

$$D_{64} = \min(d_{14}, d_{24}) = \min(1.34, 1) = 1$$

$$D_{65} = \min(d_{15}, d_{25}) = \min(1.33, 1.67) = 1.33$$

□ 由此得，

$$D_1 = \begin{matrix} & G_6 & G_3 & G_4 & G_5 \\ \begin{matrix} G_6 \\ G_3 \\ G_4 \\ G_5 \end{matrix} & \begin{bmatrix} 0 & 1.03 & 1 & 1.33 \\ & 0 & 0.63 & 1.30 \\ & & 0 & 0.67 \\ & & & 0 \end{bmatrix} \end{matrix}$$

□ G4与G3合并成新类G7，它与其它各类的距离如下：

$$D_{67} = \min(d_{63}, d_{64}) = \min(1.03, 1) = 1$$

$$D_{75} = \min(d_{35}, d_{45}) = \min(1.30, 01.67) = 0.67$$

□ 由此得，

$$D_2 = \begin{matrix} & G_6 & G_7 & G_5 \\ G_6 & \begin{bmatrix} 0 & 1 & 1.33 \\ & 0 & 0.67 \\ & & 0 \end{bmatrix} \\ G_7 & \\ G_5 & \end{matrix}$$

□ G7与G5合并成新类G8，它与其它各类的距离如下：

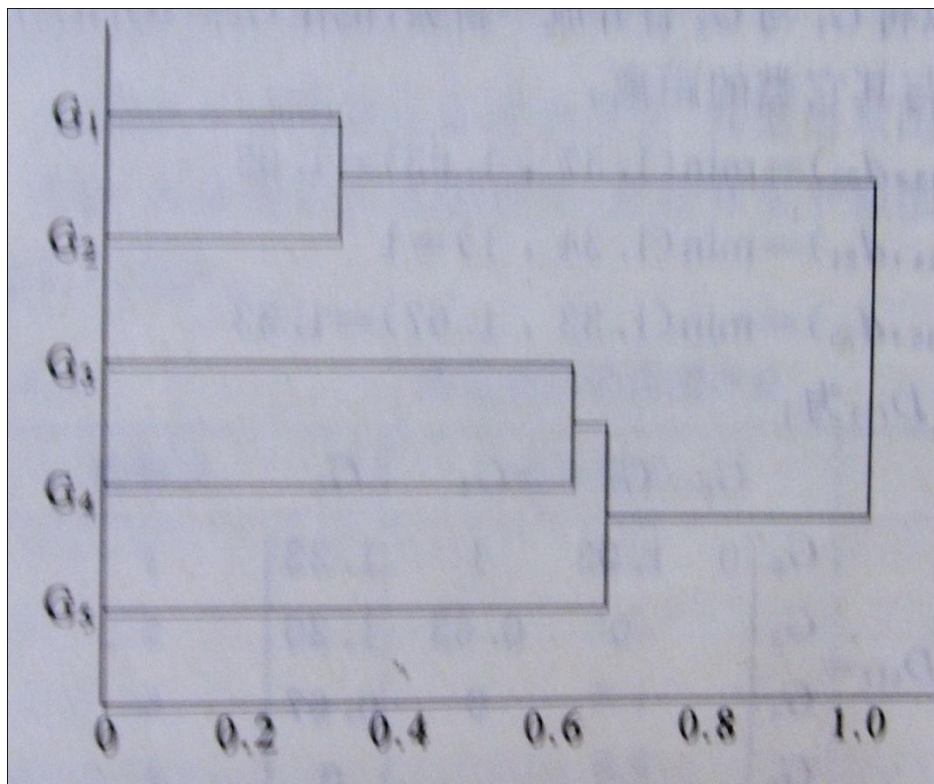
$$D_{68} = \min(d_{67}, d_{65}) = \min(1, 1.33) = 1$$

□ 由此得，

$$D_3 = \begin{matrix} & G_6 & G_8 \\ \begin{matrix} G_6 \\ G_8 \end{matrix} & \begin{bmatrix} 0 & 1 \\ & 0 \end{bmatrix} \end{matrix}$$

□ 最后，将G6与G8合并为一类，由此结束聚类。

□ 将聚类过程通过谱系图反映出来。



- ◆ 谱系聚类图的作用在于：根据实际问题希望分为几类，都可以从谱系聚类图中得到分类果。
- ◆ 但到底分为几类最合适？并没有绝对正确的原则。一般可根据实际问题的不同，可以从谱系聚类图直观看出或通过分界值（阈值）给出分类。

2. 最长距离法

- 定义类 G_i 与 G_j 之间的距离为两类最远样品的距离，即为：

$$D_{pq} = \max_{X_i \in G_p, X_j \in G_q} d_{ij}$$

- 然后将距离最小的两类合并。将类 G_p 与 G_q 合并为 G_r ，则任一类 G_k 与 G_r 的类间距离公式为：

$$\begin{aligned} D_{kr} &= \max_{X_i \in G_k, X_j \in G_r} d_{ij} \\ &= \max \left\{ \max_{X_i \in G_k, X_j \in G_p} d_{ij}, \max_{X_i \in G_k, X_j \in G_q} d_{ij} \right\} = \max \{ D_{kp}, D_{kq} \} \end{aligned}$$

◆ 仍是上例，采用绝对距离的初始距离阵如下：

$$D_0 = \begin{matrix} & \begin{matrix} G_1 & G_2 & G_3 & G_4 & G_5 \end{matrix} \\ \begin{matrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{matrix} & \begin{bmatrix} 0 & 0.34 & 1.37 & 1.34 & 1.33 \\ & 0 & 1.03 & 1 & 1.67 \\ & & 0 & 0.63 & 1.30 \\ & & & 0 & 0.67 \\ & & & & 0 \end{bmatrix} \end{matrix}$$

⑩ 因此，G1与G2合并成新类G6。

□ 计算G6与其他类的距离：

$$D_{63} = \max(d_{13}, d_{23}) = \max(1.37, 1.03) = 1.37$$

$$D_{64} = \max(d_{14}, d_{24}) = \max(1.34, 1) = 1.34$$

$$D_{65} = \max(d_{15}, d_{25}) = \max(1.33, 1.67) = 1.67$$

□ 由此得，

$$D_0 = \begin{matrix} & G_1 & G_2 & G_3 & G_4 & G_5 \\ \begin{matrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{matrix} & \begin{bmatrix} 0 & 0.34 & 1.37 & 1.34 & 1.33 \\ & 0 & 1.03 & 1 & 1.67 \\ & & 0 & 0.63 & 1.30 \\ & & & 0 & 0.67 \\ & & & & 0 \end{bmatrix} \end{matrix}$$

□ G4与G3合并成新类G7，它与其它各类的距离如下：

$$D_{67} = \max(d_{63}, d_{64}) = \max(1.37, 1.34) = 1.37$$

$$D_{75} = \max(d_{35}, d_{45}) = \max(1.30, 0.67) = 1.30$$

□ 由此得，

$$D_1 = \begin{matrix} & G_6 & G_3 & G_7 & G_4 & G_5 & G_5 \\ \begin{matrix} G_6 \\ G_7 \\ G_3 \\ G_4 \\ G_5 \end{matrix} & \begin{bmatrix} 0 & 1.37 & 1.37 & 1.67 & 1.67 \\ 1.37 & 0 & 1.34 & 1.67 & 1.67 \\ 0 & 0 & 0.63 & 1.30 & 1.30 \\ 0 & 0 & 0.63 & 0 & 0.67 \\ 0 & 0 & 1.30 & 0.67 & 0 \end{bmatrix} \end{matrix}$$

- G7与G5合并成新类G8，它与其它各类的距离如下：

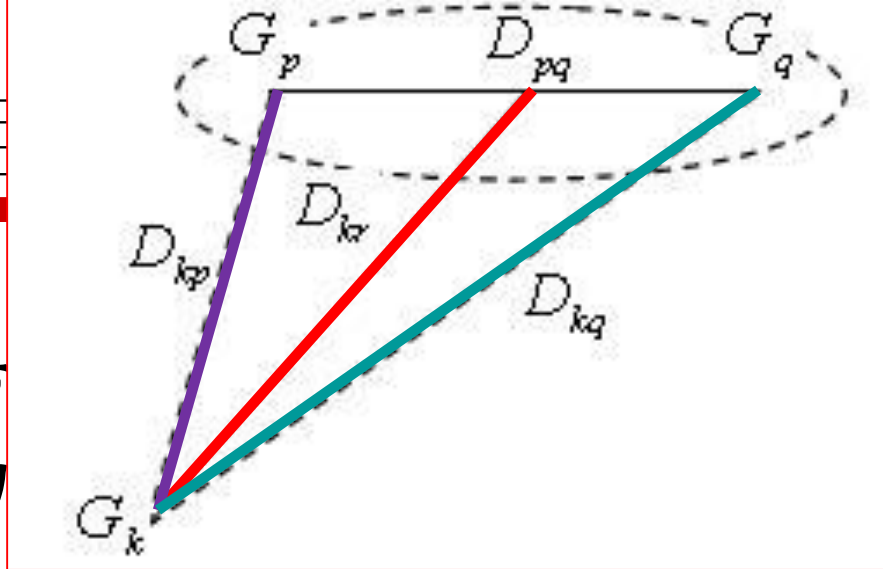
$$D_{68} = \max(d_{67}, d_{65}) = \max(1.37, 1.67) = 1.67$$

- 由此得，

$$D_2 = \begin{array}{c} \begin{array}{c} D_3 \\ \equiv G_6 \end{array} \begin{array}{c} G_6 \\ G_7 \end{array} \begin{array}{c} G_6 \quad G_8 \quad G_5 \\ \left[\begin{array}{ccc} 0 & 1.37 & 1.67 \\ & 0 & 0 \\ & & 1.30 \end{array} \right] \\ \begin{array}{c} G_5 \\ 0 \end{array} \end{array}$$

- 最后，将G6与G8合并，聚类结束。

3. 中间距



两者之间

在定义类的距离的

设将类 G_p 与 G_q 类合并为类 G_r , 则任意的类 G_k 和 G_r 的距离公式:

- 若 $D_{kq} > D_{kp}$, 如果采用最短距离法, 则 $D_{kr} = D_{kp}$, 如果采用最长距离法, 则 $D_{kr} = D_{kq}$ 。
- 而中间距离法是取它们 (最长距离与最短距离) 的中间一点作为计算 D_{kr} 的根据。

-
- 中间距离法取最短距离与最长距离的中线，则由初等几何可知这个中线的平方为：

$$D_{kr}^2 = \frac{1}{2} D_{kp}^2 + \frac{1}{2} D_{kq}^2 - \frac{1}{4} D_{pq}^2$$

- 由于此公式中出现的全是距离的平方，所以为了计算的方便，距离矩阵的元素也可以都为平方。

◆ 仍是上例，采用欧氏距离，则初始距离矩阵为：

$$D_0^2 = \begin{matrix} & \begin{matrix} G_1 & G_2 & G_3 & G_4 & G_5 \end{matrix} \\ \begin{matrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{matrix} & \begin{bmatrix} 0 & 0.12 & 0.94 & 1.12 & 1.11 \\ & 0 & 0.60 & 1.00 & 1.45 \\ & & 0 & 0.20 & 1.09 \\ & & & 0 & 0.45 \\ & & & & 0 \end{bmatrix} \end{matrix}$$

$D_{12}^2 = (0-0)^2 + (0.33-0.67)^2 \approx 0.12$

⑩ 因此，G1与G2合并成新类G6。

□ 计算G6与其他类的距离：

□ 由此得，

$$D_1^2 = \begin{matrix} & G_6 & G_3 & G_4 & G_5 \\ \begin{matrix} G_6 \\ G_3 \\ G_4 \\ G_5 \end{matrix} & \begin{bmatrix} 0 & 0.74 & 1.03 & 1.25 \\ & 0 & 0.20 & 1.09 \\ & & 0 & 0.45 \\ & & & 0 \end{bmatrix} \end{matrix}$$

□ G4与G3合并成新类G7，它与其它各类的距离如下：

□ 由此得，

$$D_2^2 = \begin{matrix} & G_6 & G_7 & G_5 \\ \begin{matrix} G_6 \\ G_7 \\ G_5 \end{matrix} & \begin{bmatrix} 0 & 0.84 & 1.25 \\ & 0 & 0.72 \\ & & 0 \end{bmatrix} \end{matrix}$$

□ G7与G5合并成新类G8，它与其它各类的距离如下：

$$D_{68}^2 = \frac{1}{2} D_{56}^2 + \frac{1}{2} D_{76}^2 - \frac{1}{4} D_{57}^2 = 1.49$$

□ 由此得，

$$D_3^2 = \begin{matrix} & G_6 & G_8 \\ \begin{matrix} G_6 \\ G_8 \end{matrix} & \begin{bmatrix} 0 & 1.49 \\ & 0 \end{bmatrix} \end{matrix}$$

□ 最后，将G6与G8合并，聚类结束。

4. 重心法

- 重心法是定义类间距离为两类样品重心（各类样品的均值）的距离的系统聚类方法。重心指标对类有很好的代表性，但利用各样本的信息不充分。
- 设 G_p 与 G_q 分别有样品 n_p, n_q 个，其重心分别为 $\bar{X}_p$ 和 $\bar{X}_q$ ，则 G_p 与 G_q 之间的距离定义为 $\bar{X}_p$ 和 $\bar{X}_q$ 之间的距离，用欧氏距离来表示，即

$$D_{pq}^2 = (\bar{X}_p - \bar{X}_q)'(\bar{X}_p - \bar{X}_q)$$

□ 设将 G_p 和 G_q 合并为 G_r , 则 G_r 内样品个数为 $n_r = n_p + n_q$, 它的重心是 $\bar{X}_r = \frac{1}{n_r}(n_p \bar{X}_p + n_q \bar{X}_q)$, 类 G_k 的重心是 $\bar{X}_k$, 那么依据上式它与新类 G_r 的距离为:

$$\begin{aligned} D_{kr}^2 &= (\bar{X}_k - \bar{X}_r)' (\bar{X}_k - \bar{X}_r) \\ &= \left[\bar{X}_k - \frac{1}{n_r} (n_p \bar{X}_p + n_q \bar{X}_q) \right]' \left[\bar{X}_k - \frac{1}{n_r} (n_p \bar{X}_p + n_q \bar{X}_q) \right] \\ &= \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 - \frac{n_p n_q}{n_r^2} D_{pq}^2 \end{aligned}$$

□ 例：设有六个样品，每个只测量一个指标，分别是1，2，5，7，9，10，试用重心法将它们聚类。

(1) 样品采用欧氏距离，计算样品间的平方距离阵 $D^2_{(0)}$ ，见下表。

	G_1	G_2	G_3	G_4	G_5	G_6
G_1	0					
G_2	1	0				
G_3	16	9	0			
G_4	36	25	4	0		
G_5	64	49	16	4	0	
G_6	81	64	25	9	1	0

(2) $D^2_{(0)}$ 中最小的元素是 $D^2_{12} = D^2_{56} = 1$ ，于是将 G_1 和 G_2 合并成 G_7 ， G_5 和 G_6 合并成 G_8 ，并利用重心距离公式计算新类与其它类的距离得到距离阵 $D^2_{(1)}$ ，见表：

	G_7	G_3	G_4	G_8
G_7	0			
G_3	12.25	0		
G_4	30.25	4	0	
G_8	64	20.25	6.25	0

其中，

$$\begin{aligned}
 D^2_{37} &= \frac{1}{2} D^2_{31} + \frac{1}{2} D^2_{32} - \frac{1}{2} \cdot \frac{1}{2} D^2_{12} \\
 &= \frac{1}{2} \times 16 + \frac{1}{2} \times 9 - \frac{1}{2} \cdot \frac{1}{2} \times 1 = 12.25
 \end{aligned}$$

(3) 在 $D^2_{(1)}$ 中最小值是 $D^2_{34} = 4$, 那么 G_3 与 G_4 合并一个新类 G_9 , 其与与其它类的距离 $D^2_{(2)}$, 见下表:

	G_7	G_9	G_8
G_7	0		
G_9	20.25	0	
G_8	64	12.5	0

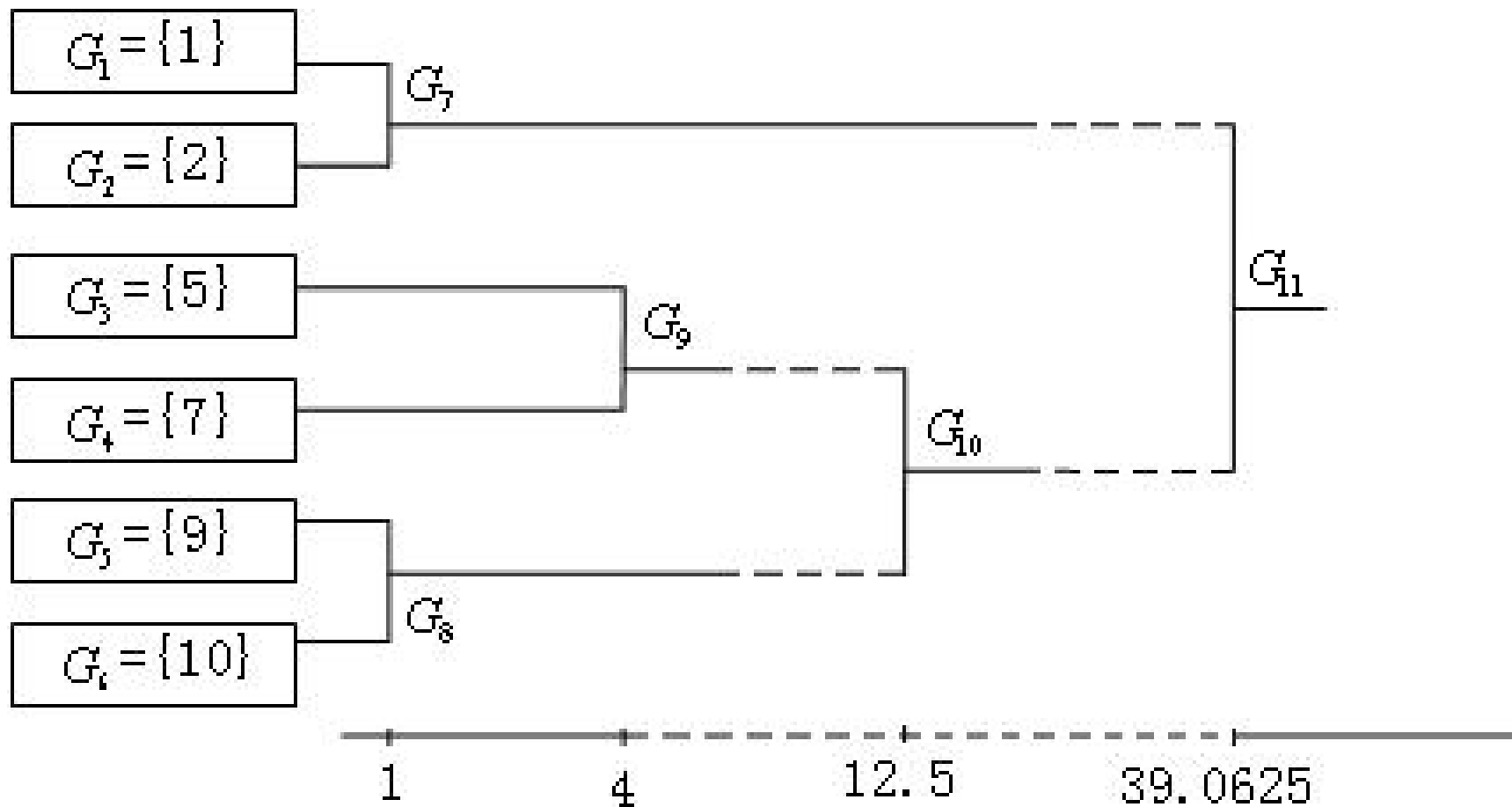
其中,

$$D^2_{79} = \frac{1}{2} D^2_{73} + \frac{1}{2} D^2_{74} - \frac{1}{2} \cdot \frac{1}{2} D^2_{34}$$
$$= \frac{1}{2} \times 12.25 + \frac{1}{2} \times 30.25 - \frac{1}{2} \cdot \frac{1}{2} \times 4 = 20.25$$

(4) 在 $D^2_{(2)}$ 中最小值是 $= 12.5$, 那么 G_8 和 G_9 合

	G_7	G_{10}
G_7	0	
G_{10}	39.0625	0

(5) 最后将 G_7 和 G_{10} 合并成 G_{11} , 这时所有的六个样品聚为一类, 其过程终止。



5. 类平均法

重心法虽然有较好的代表性，但并未充分利用样本的信息。而类平均法定义两类之间的距离平方为两类两两样品之间距离平方的平均数，即为：

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{X_i \in G_p} \sum_{X_j \in G_q} d_{ij}^2$$

➤ 设聚类的某一步将 G_p 和 G_q 合并为新类 G_r , 则类 G_k 与新并类 G_r 的距离为:

$$\begin{aligned} D_{kr}^2 &= \frac{1}{n_k n_r} \sum_{X_i \in G_k} \sum_{X_j \in G_r} d_{ij}^2 \\ &= \frac{1}{n_k n_r} \left(\sum_{X_i \in G_k} \sum_{X_j \in G_p} d_{ij}^2 + \sum_{X_i \in G_k} \sum_{X_j \in G_q} d_{ij}^2 \right) \\ &= \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 \end{aligned}$$

- 例：设有六个样品，每个只测量一个指标，分别是1, 2, 5, 7, 9, 10，试用类平均法将它们聚类。（1）样品采用欧氏距离，计算样品间的平方距离阵 $D^2_{(0)}$ ，见下表。

	G_1	G_2	G_3	G_4	G_5	G_6
G_1	0					
G_2	1	0				
G_3	16	9	0			
G_4	36	25	4	0		
G_5	64	49	16	4	0	
G_6	81	64	25	9	1	0

(2) $D^2_{(0)}$ 中最小的元素是 $D^2_{12} = D^2_{56} = 1$ ，于是将 G_1 和 G_2 合并成 G_7 ， G_5 和 G_6 合并成 G_8 ，并利用重心距离公式计算新类与其它类的距离得到距离阵 $D^2_{(1)}$ ，见表：

	G_7	G_3	G_4	G_8
G_7	0			
G_3	12.5	0		
G_4	30.5	4	0	
G_8	64.5	20.5	6.5	0

其中，

$$\begin{aligned}
 D^2_{37} &= \frac{1}{2} D^2_{31} + \frac{1}{2} D^2_{32} \\
 &= \frac{1}{2} \times 16 + \frac{1}{2} \times 9 = 12.5
 \end{aligned}$$

$$\begin{aligned}
 D^2_{78} &= \frac{1}{4} (d^2_{15} + d^2_{16} + d^2_{25} + d^2_{26}) \\
 &= \frac{1}{4} (64 + 81 + 49 + 64) = 64.5
 \end{aligned}$$

(3) 在 $D^2_{(1)}$ 中最小值是 $D^2_{34} = 4$, 那么 G_3 与 G_4 合并一个新类 G_9 , 其与与其它类的距离 $D^2_{(2)}$, 见下表:

	G_7	G_9	G_8
G_7	0		
G_9	21.5	0	
G_8	64.5	13.5	0

其中,

$$D^2_{79} = \frac{1}{2} D^2_{37} + \frac{1}{2} D^2_{47}$$

$$= \frac{1}{2} \times 12.5 + \frac{1}{2} \times 30.5 = 21.5$$

$$D^2_{89} = \frac{1}{2} D^2_{38} + \frac{1}{2} D^2_{48}$$

$$= \frac{1}{2} \times 20.5 + \frac{1}{2} \times 6.5 = 13.5$$

(4) 在 $D^2_{(2)}$ 中最小值是 $= 13.5$, 那么 G_8 和 G_9 合

	G_7	G_{10}
G_7	0	
G_{10}	43	0

(5) 最后将 G_7 和 G_{10} 合并成 G_{11} , 这时所有的六个样品聚为一类, 其过程终止。

6. 可变类平均法

- 由于类平均法中没有反映出 G_p 和 G_q 之间的距离 D_{pq} 的影响，信息利用不充分。为了充分利用各类距离的信息，可将类平均法和中间距离法进行组合，得到一个组合模型，此组合模型称为可变类平均法。
- 如果将 G_p 和 G_q 合并为新类 G_r ，类 G_k 与新并类 G_r 的距离公式为：

$$D_{kr}^2 = (1 - \beta) \left(\frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 \right) + \beta D_{pq}^2$$

$$D_{kr}^2 = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2$$

$$D_{kr}^2 = \frac{1}{2} D_{kp}^2 + \frac{1}{2} D_{kq}^2 - \frac{1}{4} D_{pq}^2$$

-
- 之所以称该方法为可变类平均法是因为 β 是可变的。
 - 系数 β 称为聚类强度系数， β 的取值不同，聚类的结果就会不同。通常 β 取负值时可以使聚类的分辨能力提高；一般情况下，取 $\beta = -1/4$ 。

- 例：设有六个样品，每个只测量一个指标，分别是1, 2, 5, 7, 9, 10，试用可变类平均法（取 $\beta=-1/4$ ）将它们聚类。（1）样品采用欧氏距离，计算样品间的平方距离阵 $D^2_{(0)}$ ，见下表。

	G_1	G_2	G_3	G_4	G_5	G_6
G_1	0					
G_2	1	0				
G_3	16	9	0			
G_4	36	25	4	0		
G_5	64	49	16	4	0	
G_6	81	64	25	9	1	0

(2) $D^2_{(0)}$ 中最小的元素是 $D^2_{12} = D^2_{56} = 1$ ，于是将 G_1 和 G_2 合并成 G_7 ， G_5 和 G_6 合并成 G_8 ，并利用可变类平均公式计算新类与其它类的距离得到距离阵 $D^2_{(1)}$ ，见表：

	G_7	G_3	G_4	G_8
G_7	0			
G_3	15.375	0		
G_4	37.875	4	0	
G_8	100.21875	25.375	7.9	0

其中，

$$D^2_{37} = \frac{(1 - (-1/4))}{2} D^2_{31} + \frac{(1 - (-1/4))}{2} D^2_{32} - \frac{1}{4} D^2_{12}$$

$$= \frac{5}{8} \times (16 + 9) - \frac{1}{4} \times 1 = 15.375$$

(3) 在 $D^2_{(1)}$ 中最小值是 $D^2_{34} = 4$, 那么 G_3 与 G_4 合并一个新类 G_9 , 其与与其它类的距离 $D^2_{(2)}$, 见下表:

	G_7	G_9	G_8
G_7	0		
G_9	32.281	0	
G_8	64.5	19.797	0

其中,

$$D^2_{79} = \frac{(1 - (-1/4))}{2} D^2_{73} + \frac{(1 - (-1/4))}{2} D^2_{74} - \frac{1}{4} D^2_{34}$$

$$= \frac{5}{8} \times (15.375 + 37.875) - \frac{1}{4} \times 4 \approx 32.281$$

(4) 在 $D^2_{(2)}$ 中最小值是 $D^2_{89} = 19.797$, 那么 G_8 和 G_9 合并一个新类 G_{10} , 它与其它类的距离, 见表:

	G_7	G_{10}
G_7	0	
G_{10}	55.54	0

(5) 最后将 G_7 和 G_{10} 合并成 G_{11} , 这时所有的六个样品聚为一类, 其过程终止。

$$\begin{aligned} D^2_{7.10} &= \frac{5}{8} (D^2_{78} + D^2_{79}) - \frac{1}{4} D^2_{89} \\ &= \frac{5}{8} (25.375 + 7.9) - \frac{1}{4} D^2_{89} 19.797 = 55.54 \end{aligned}$$

7. 可变法

针对于中间法而言，如果将中间法的前两项的系数也依赖于 β ，或者在可变类平均法中，不考虑 G_p 和 G_q 两类中各自样品的个数，而是类同等看待，则得到了可变法。

如果将 G_p 和 G_q 合并为新类 G_r ，类 G_k 与新并类 G_r 的距离公式为：

$$D_{kr}^2 = \frac{1-\beta}{2} (D_{kp}^2 + D_{kq}^2) + \beta D_{pq}^2$$

$$\text{可变类平均法: } D_{kr}^2 = (1-\beta) \left(\frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 \right) + \beta D_{pq}^2$$

8. 离差平方和法

- ❄ 该方法是Ward提出来的，所以又称为Ward法。
- ❄ 该方法的基本思想来自于方差分析，如果分类正确，同类样品的离差平方和应当较小，类与类的离差平方和较大。
- ❄ 具体做法是先将 n 个样品各自成一类，然后每次缩小一类，每缩小一类，离差平方和就要增大，**选择使方差增加最小的两类合并**，直到所有的样品归为一类为止。

□ 设将 n 个样品分成 k 类 $G_1, G_2, \dots, G_k$, 用 X_{it} 表示 G_t 中的第 i 个样品, n_t 表示 G_t 中样品的个数, $\bar{X}_t$ 是 G_t 的重心, 则 G_t 的样品离差平方和为:

$$S_t = \sum_{i=1}^{n_t} (\mathbf{X}_{it} - \bar{\mathbf{X}}_t)' (\mathbf{X}_{it} - \bar{\mathbf{X}}_t)$$

□ 如果 G_p 和 G_q 合并为新类 G_r ，则类内离差平方和分别为：

$$S_p = \sum_{i=1}^{n_p} (\mathbf{X}_{ip} - \bar{\mathbf{X}}_p)' (\mathbf{X}_{ip} - \bar{\mathbf{X}}_p)$$

$$S_q = \sum_{i=1}^{n_q} (\mathbf{X}_{iq} - \bar{\mathbf{X}}_q)' (\mathbf{X}_{iq} - \bar{\mathbf{X}}_q)$$

$$S_r = \sum_{i=1}^{n_r} (\mathbf{X}_{ir} - \bar{\mathbf{X}}_r)' (\mathbf{X}_{ir} - \bar{\mathbf{X}}_r)$$

□ 它们反映了各自类内样品的分散程度，如果 G_p 和 G_q 这两类相距较近，则合并后所增加的离散平方和 $S_r - S_p - S_q$ 应较小；否则，应较大。于是定义 G_p 和 G_q 之间的平方距离为： $D_{pq}^2 = S_r - S_p - S_q$

□ 如果将 G_p 和 G_q 合并为新类 G_r ，类 G_k 与新并类 G_r 的距离公式为：

$$D_{kr}^2 = \frac{n_k + n_p}{n_r + n_k} D_{kp}^2 + \frac{n_k + n_q}{n_r + n_k} D_{kq}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2$$

 例：设有五个样品，每个只测量一个指标，分别是1，2，3.5，7，9，试用离差平方和法将它们聚类。

- ① 五个样品各成一类，所以类内平方和为0；
- ② 计算两两样品合并后增加的离差平方和，形成初始平方距离阵 $D^2_{(0)}$ ，见下表。

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	0.5	0			
G_3	3.125	1.125	0		
G_4	18	12.50	6.125	0	
G_5	32	24.5	15.125	2	0

$$D_{12}^2 = (1-1.5)^2 + (2-1.5)^2 = 0.5$$

(2) $D^2_{(0)}$ 中最小的元素是 $D^2_{12} = 0.5$ ，于是将 G_1 和 G_2 合并成 G_6 ，计算新类与其它类的距离得到距离阵 $D^2_{(1)}$ ，见表：

	G_6	G_3	G_4	G_5
G_6	0			
G_3	2.667	0		
G_4	20.167	6.125	0	
G_5	37.5	15.125	2	0

其中，

$$D^2_{k6} = \frac{n_1 + n_k}{n_6 + n_k} D^2_{k1} + \frac{n_2 + n_k}{n_6 + n_k} D^2_{k2} - \frac{n_k}{n_6 + n_k} D^2_{12}$$

$$D^2_{36} = \frac{2}{3} D^2_{31} + \frac{2}{3} D^2_{32} - \frac{1}{3} D^2_{12} = \frac{2}{3} 3.125 + \frac{2}{3} 1.125 - \frac{1}{3} 0.5 = 2.667$$

(3) 在 $D^2_{(1)}$ 中最小值是 $D^2_{45} = 2$, 那么 G_5 与 G_4 合并一个新类 G_7 , 其与与其它类的距离 $D^2_{(2)}$, 见下表:

	G_6	G_3	G_7
G_6	0		
G_3	2.667	0	
G_7	42.25	13.5	0

其中,

$$D^2_{k7} = \frac{n_4 + n_k}{n_7 + n_k} D^2_{k4} + \frac{n_5 + n_k}{n_7 + n_k} D^2_{k5} - \frac{n_k}{n_7 + n_k} D^2_{45}$$

$$D^2_{76} = \frac{2+1}{2+2} D^2_{64} + \frac{3}{4} D^2_{65} - \frac{2}{4} D^2_{45} = \frac{3}{4} 20.167 + \frac{3}{4} 37.5 - \frac{2}{4} 2 \approx 42.25$$

(4) 在 $D^2_{(2)}$ 中最小值是 $= 2.667$, 那么 G_3 和 G_6 合并一个新类 G_8 , 它与其它类的距离, 见表:

	G_7	G_{10}
G_7	0	
G_{10}	40.83	0

(5) 最后将 G_7 和 G_{10} 合并为 G_{87} , 这时所有样本

$$\begin{aligned}
 \text{聚为 } D_{87}^2 &= \frac{n_7 + n_3}{n_7 + n_8} D_{73}^2 + \frac{n_7 + n_6}{n_7 + n_8} D_{76}^2 - \frac{n_7}{n_7 + n_8} D_{36}^2 \\
 &= \frac{2+1}{2+3} 13.5 + \frac{2+2}{5} 42.25 - \frac{2}{5} 2.667 = 40.83
 \end{aligned}$$

三、类间距离的统一性

- 上述八种系统聚类法的步骤完全一样，只是距离的递推公式不同。兰斯（Lance）和威廉姆斯（Williams）于1967年给出了一个统一的公式，即将 G_p 和 G_q 合并为新类 G_r ，类 G_k 与新并类 G_r 的距离公式为：

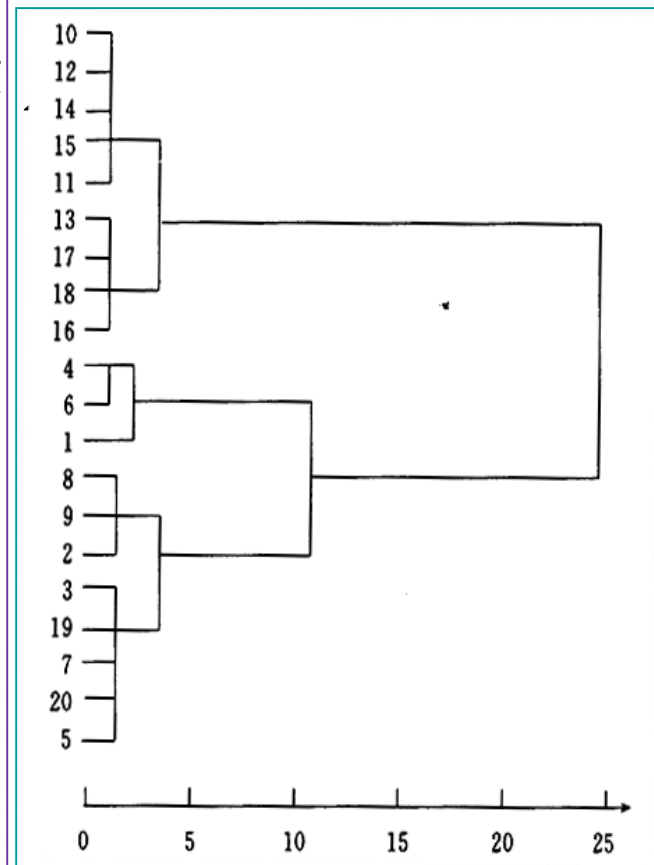
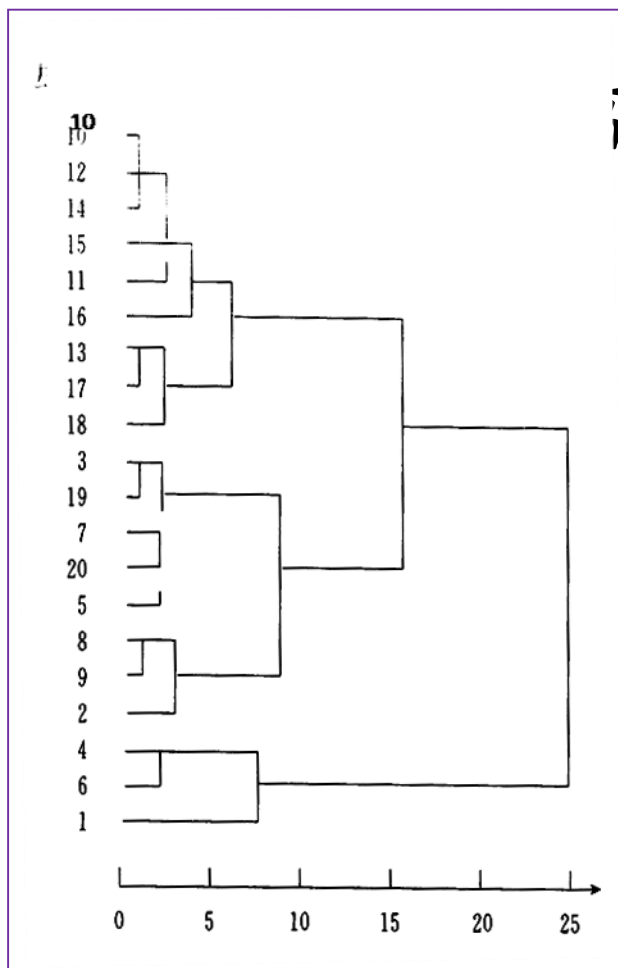
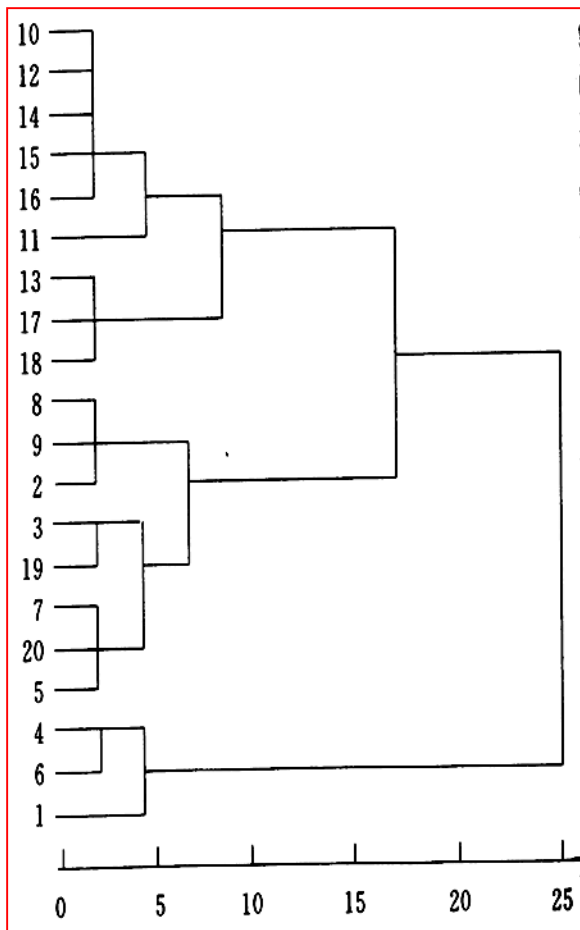
$$D_{kr}^2 = \alpha_p D_{kp}^2 + \alpha_q D_{kq}^2 + \beta D_{pq}^2 + \gamma |D_{kp}^2 - D_{kq}^2|$$

其中 α_p 、 α_q 、 β 、 γ 是参数，不同的系统聚类法，它们取不同的数。

系统聚类法参数表

方 法	α_p	α_q	β	γ
最短距离法	1/2	1/2	0	-1/2
最长距离法	1/2	1/2	0	1/2
中间距离法	1/2	1/2	-1/4	0
重心法	n_p/n_r	n_q/n_r	$-\alpha_p\alpha_q$	0
类平均法	n_p/n_r	n_q/n_r	0	0
可变类平均法	$(1-\beta)n_p/n_r$	$(1-\beta)n_q/n_r$	$\beta(<1)$	0
可变法	$(1-\beta)/2$	$(1-\beta)/2$	$\beta(<1)$	0
离差平方和法	$\frac{(n_p + n_k)}{(n_r + n_k)}$	$\frac{(n_q + n_k)}{(n_r + n_k)}$	$-n_k/(n_k + n_r)$	0

	country	call	movecall	fee	computer	mips	net
1	美 国	631.60	161.90	0.36	403.00	26073.00	35.34
2	日 本	498.40	143.20	3.57	176.00	10223.00	6.26
3	德 国	557.60	70.60	2.18	199.00	11571.00	9.48
4	瑞 典	684.10	281.80	1.40	286.00	16660.00	29.39
5	瑞 士	644.00	93.50	1.98	234.00	13621.00	22.68
6	丹 麦	620.30	248.60	2.56	296.00	17210.00	21.84
7	新加坡	498.40	147.50	2.50	284.00	13578.00	13.49
8	中国台湾	469.40	56.10	3.68	119.00	6911.00	1.72
9	韩 国	434.50	73.00	3.36	99.00	5795.00	1.68
10	巴 西	81.90	16.30	3.02	19.00	876.00	0.52
11	智 利	138.60	8.20	1.40	31.00	1411.00	1.28
12	墨西哥	92.20	9.80	2.61	31.00	1751.00	0.35
13	俄罗斯	174.90	5.00	5.12	24.00	1101.00	0.48
14	波 兰	169.00	6.50	3.68	40.00	1796.00	1.45
15	匈牙利	262.20	49.40	2.66	68.00	3067.00	3.09
16	马来西亚	195.50	88.40	4.19	53.00	2734.00	1.25
17	泰 国	78.60	27.80	4.95	22.00	1662.00	0.11
18	印 度	13.60	0.30	6.28	2.00	101.00	0.01
19	法 国	559.10	42.90	1.27	201.00	11702.00	4.76
20	英 国	521.10	122.50	0.98	248.00	14461.00	11.91



从聚类图看，本例用三种方法聚类结果基本一致，而最长距离法和重心法所得结果更接近一致。

四、类个数的确定

□ 由适当的阈值确定

□ 根据数据点的散布图直观地确定类的个数。

如果考察的指标只有两个($m=2$),则可通过数据点的散布图直观地确定类的个数.如果有三个变量,可以绘制三维散布图并通过旋转三维坐标轴由数据点的分布来确定应分几个类(使用 SAS 软件).当考察的指标在三个以上时,可以由这些指标综合出两个或三个综合变量,然后再绘制数据点在综合变量上的散布图,从而直观地确定分类个数.

(1) R^2 统计量：假定已将 n 个样品分为 k 类，记为 $G_1, G_2, \dots, G_k$ ， n_t 表示 G_t 类的样品个数 ($n_1 + \dots + n_k = n$)， $\bar{X}^{(t)}$ 表示 G_t 的重心。 $X_{(i)}^{(t)}$ 表示 G_t 中第 i 个样品 ($i=1, \dots, n_t$)， $\bar{X}$ 表示所有样品的重心，则 G_t 类中 n_t 个样品的离差平方和为

$$W_t = \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \bar{X}^{(t)})' (X_{(i)}^{(t)} - \bar{X}^{(t)}),$$

其中 $X_{(i)}^{(t)}$, $\bar{X}^{(t)}$ 和 $\bar{X}$ 均为 m 维向量， W_t 为一数值；所有样品的总离差平方和为

$$T = \sum_{t=1}^k \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \bar{X})' (X_{(i)}^{(t)} - \bar{X}),$$

类内离差平方和

T 又可以分解为

$$\begin{aligned} T &= \sum_{t=1}^k \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \bar{X}^{(t)} + \bar{X}^{(t)} - \bar{X})' (\dots) \\ &= \sum_{t=1}^k W_t + \sum_{t=1}^k n_t (\bar{X}^{(t)} - \bar{X})' (\bar{X}^{(t)} - \bar{X}) = P_k + B_k. \end{aligned}$$

类间离差平方和

$$\text{令 } R_k^2 = \frac{B_k}{T} = 1 - \frac{P_k}{T},$$

则 R_k^2 值越大, 也就是 B_k/T 越大, 表示 k 个类的类间偏差平方和的总和 B_k 在总离差平方和 T 中占的比例越大, 这说明 k 个类越能够区分开. 因此 R_k^2 统计量可用于评价合并为 k 个类时的聚类效果. R_k^2 越大, 聚类效果越好.

例：16个地区农民生活水平的调查

地区	食品 (X_1)	衣着 (X_2)	燃料 (X_3)	住房 (X_4)	生活用品及 其他(X_5)	文化生活服务 支出(X_6)
北京	190.33	43.77	9.73	60.54	49.01	9.04
天津	135.20	36.40	10.47	44.16	36.49	3.94
河北	95.21	22.83	9.30	22.44	22.81	2.80
山西	104.78	25.11	6.40	9.89	18.17	3.25
内蒙古	128.41	27.63	8.94	12.58	23.99	3.27
辽宁	145.68	32.83	17.79	27.29	39.09	3.47
吉林	159.37	33.38	18.37	11.81	25.29	5.22
黑龙江	116.22	29.57	13.24	13.76	21.75	6.04
上海	221.11	38.64	12.53	115.65	50.82	5.89
江苏	144.98	29.12	11.67	42.60	27.30	5.74
浙江	169.92	32.75	12.72	47.12	34.35	5.00
安徽	153.11	23.09	15.62	23.54	18.18	6.39
福建	144.92	21.26	16.96	19.52	21.75	6.73
江西	140.54	21.50	17.64	19.19	15.97	4.94
山东	115.84	30.26	12.20	33.61	33.77	3.85
河南	101.18	23.26	8.46	20.20	20.50	4.30

16 个地区的分类及平均消费水平

类 别	第一类	第二类	第三类	第四类	第五类
该类所包含的地区	河北,河南 山西,内蒙古	黑龙江,安徽 福建,江西	天津,山东,江苏 浙江,辽宁,吉林	北京	上海
食品平均消费	107.395	138.698	145.165	190.33	221.11
衣着平均消费	24.708	23.855	32.457	43.77	38.64
燃料平均消费	8.275	15.865	13.870	9.73	12.53
住房平均消费	16.278	19.000	34.431	60.54	115.65
生活用品以及其他平均消费	21.368	19.413	32.715	49.01	50.82
文化生活服务支出平均消费	3.405	6.025	4.537	9.04	5.89

分为几类最合适,则应该看 R_k^2 值的变化. 比如,在分为 4 个类之前的并类过程中 R_k^2 的值减少是逐渐的,改变不大;假定分为 4 类时的 $R_4^2=0.797$,而下一次合并后分为 3 类时 R_3^2 的值下降较多,比如 $R_3^2=0.402$,这时通过分析 R_k^2 统计量的变化可得出,分为 4 个类是较合适的.

4. 根据谱系图确定分类个数的准则

Bemirmen(1972)提出了应根据研究的目的来确定适当的分类方法,并提出了一些根据谱系图来分析的准则:

准则 A 各类重心之间的距离必须很大;

准则 B 确定的类中,各类所包含的元素都不要太多;

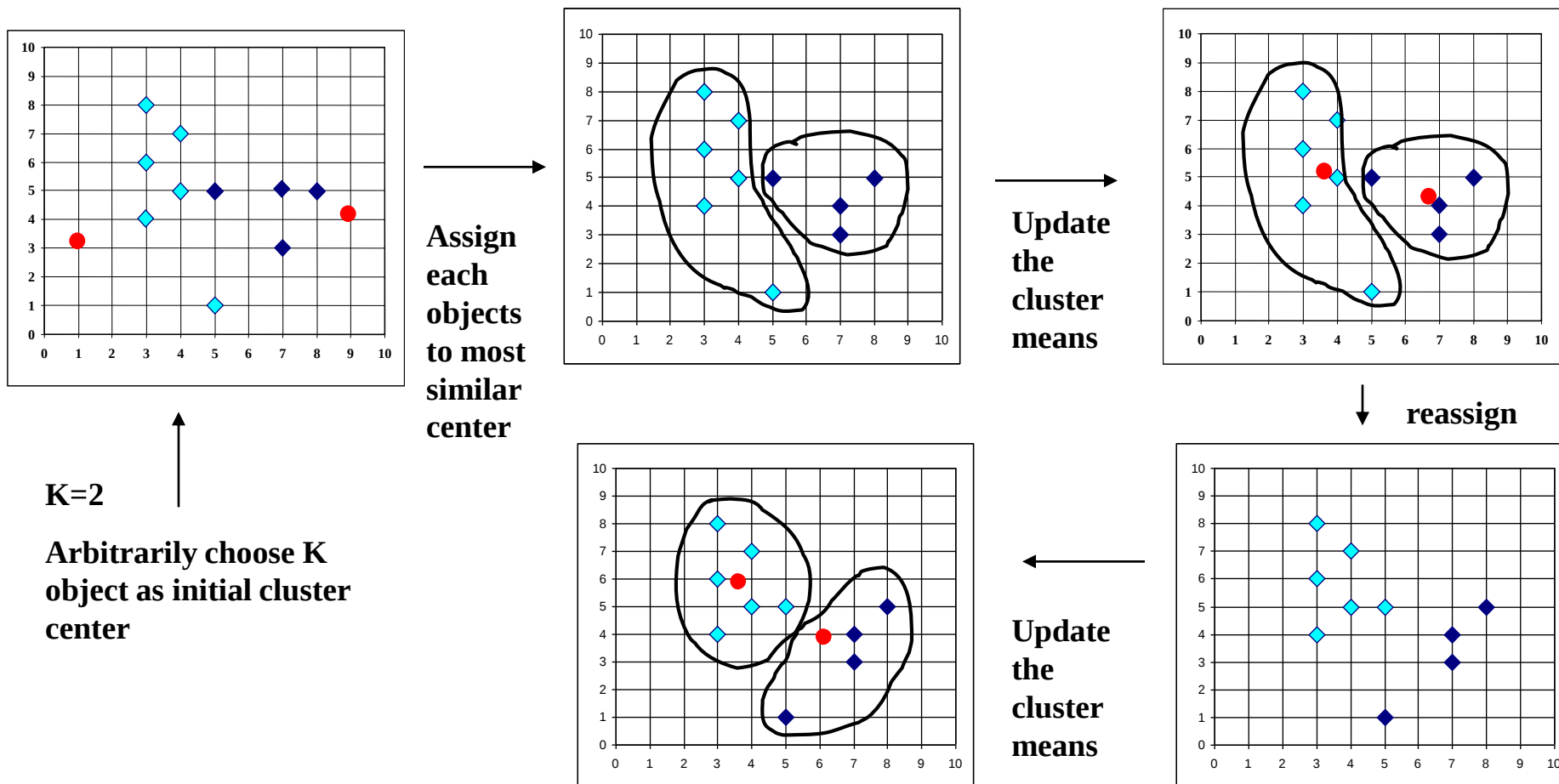
准则 C 类的个数必须符合实用目的;

准则 D 若采用几种不同的聚类方法处理,则在各自的聚类图中应发现相同的类.

第四节 K均值聚类分析

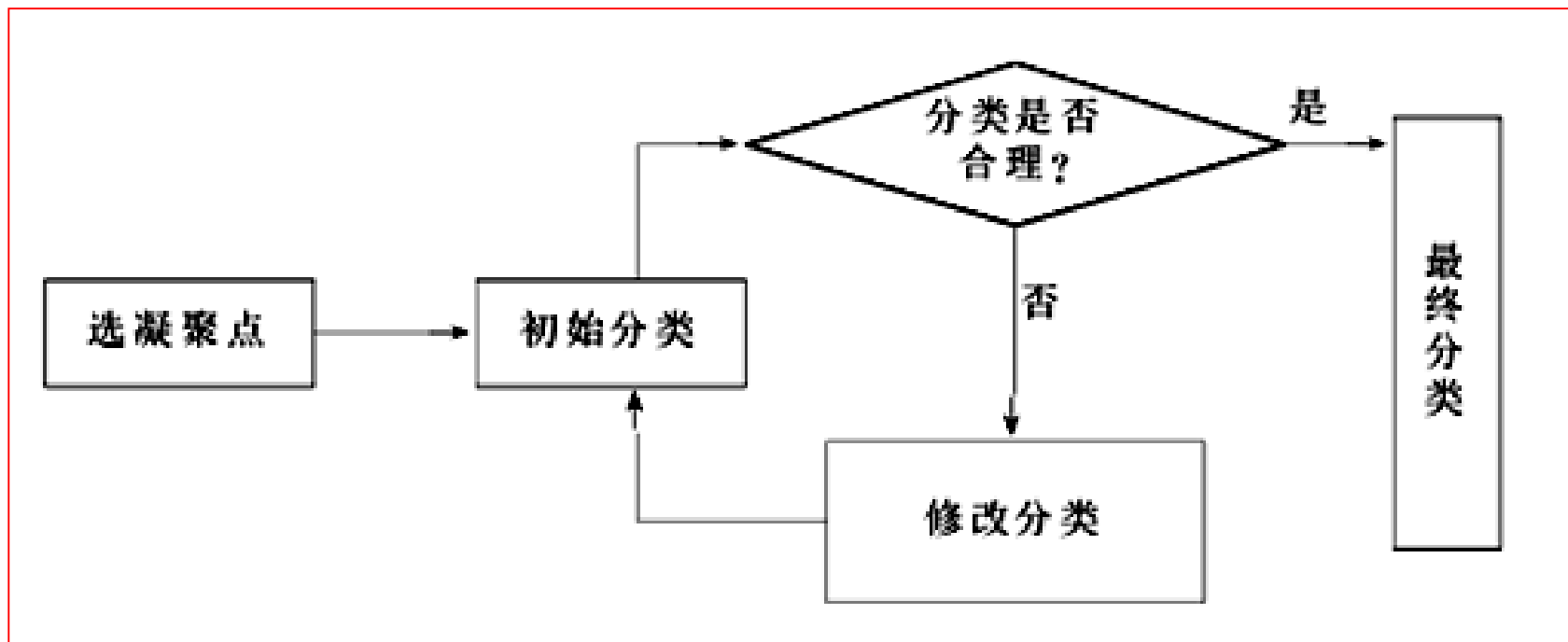
- 系统聚类法一次形成类以后就不能改变了，这就要求一次分类分得比较准确，对分类的方法就提出较高的要求，相应的计算量自然也比较大。
- 如Q型系统聚类法，聚类过程是在样品间的距离矩阵基础上进行的，而且在并类过程中，需要将每类样品和其他类样品间的距离逐一加以比较，以决定应合并的类别，因此当样本容量很大时，需要占据足够大的计算机内存空间和计算时间。
- 基于这种情况，产生了动态聚类。

K-Means Clustering *K*-均值聚类方法



-
- 动态聚类法又称为逐步聚类法，其基本思想是：开始先粗略地分一下类，然后按照某种最优的原则修改不合理的分类，直至类分得比较合理为止，这样就形成一个最终的分类结果。
 - 该方法具有计算量较小，占用计算机内存空间较少，方法简单，**适用于大样本的Q型聚类分析。**

-
- K均值法和系统聚类法一样，都是以距离的远近亲疏为标准进行聚类的。
 - 但是两者的不同之处也是明显的：系统聚类对不同的类数产生一系列的聚类结果，而K—均值法只能产生指定类数的聚类结果。

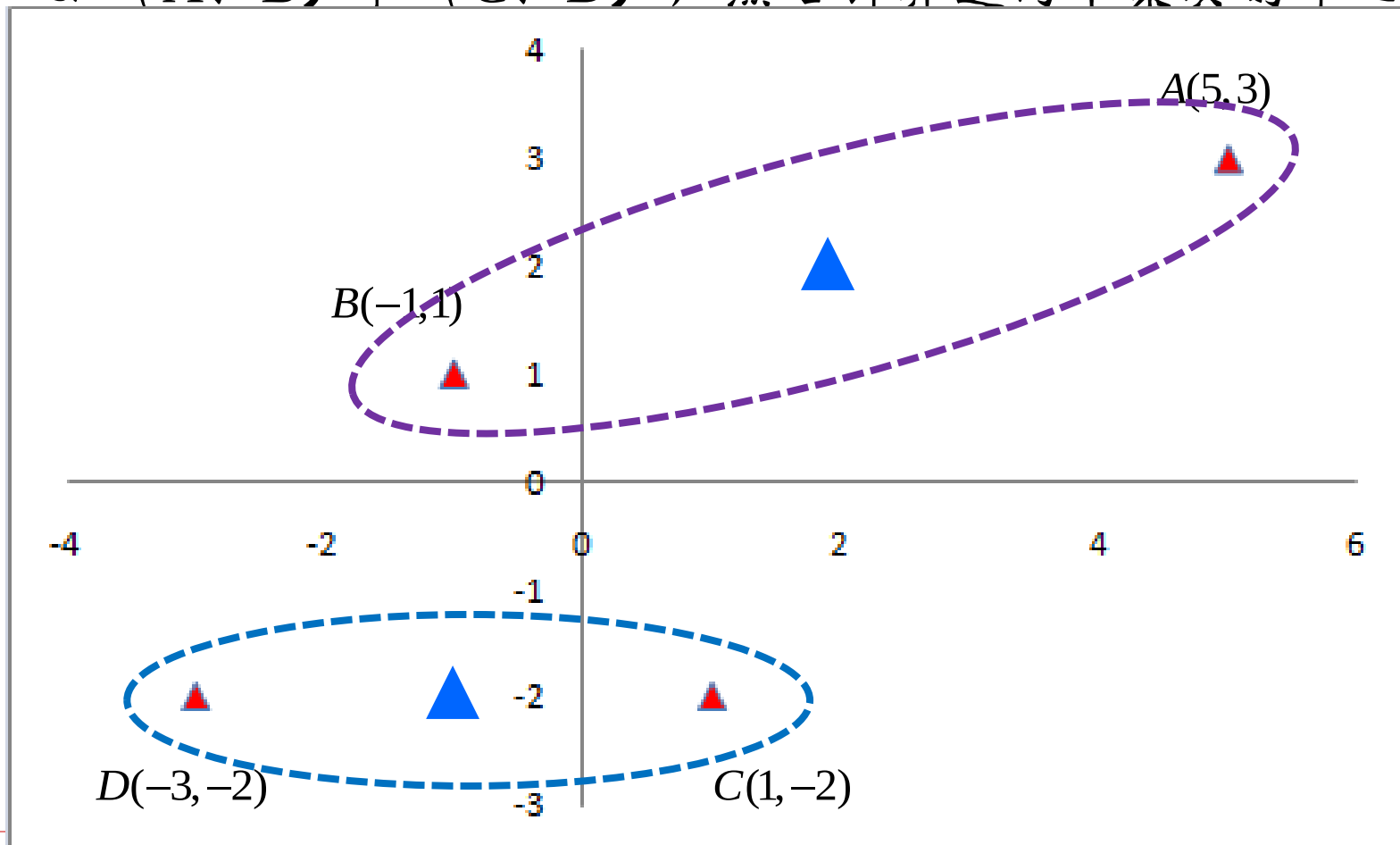


❄️ 【例5.3】 假定我们对A、B、C、D四个样品分别测量两个变量和得到结果见下表。

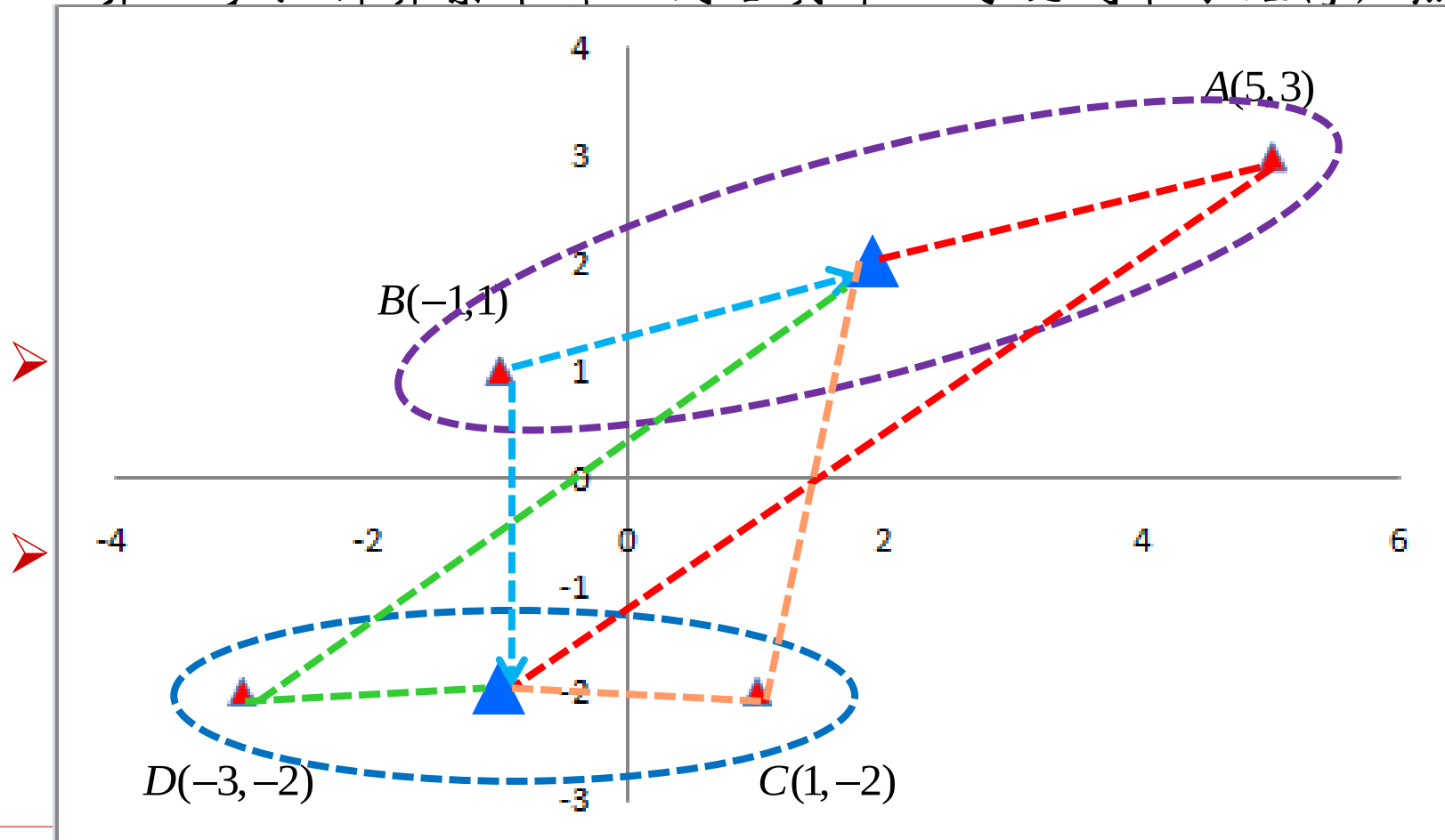
样品	变量	
A	5	3
B	-1	1
C	1	-2
D	-3	-2

试将以上的样品聚成两类。

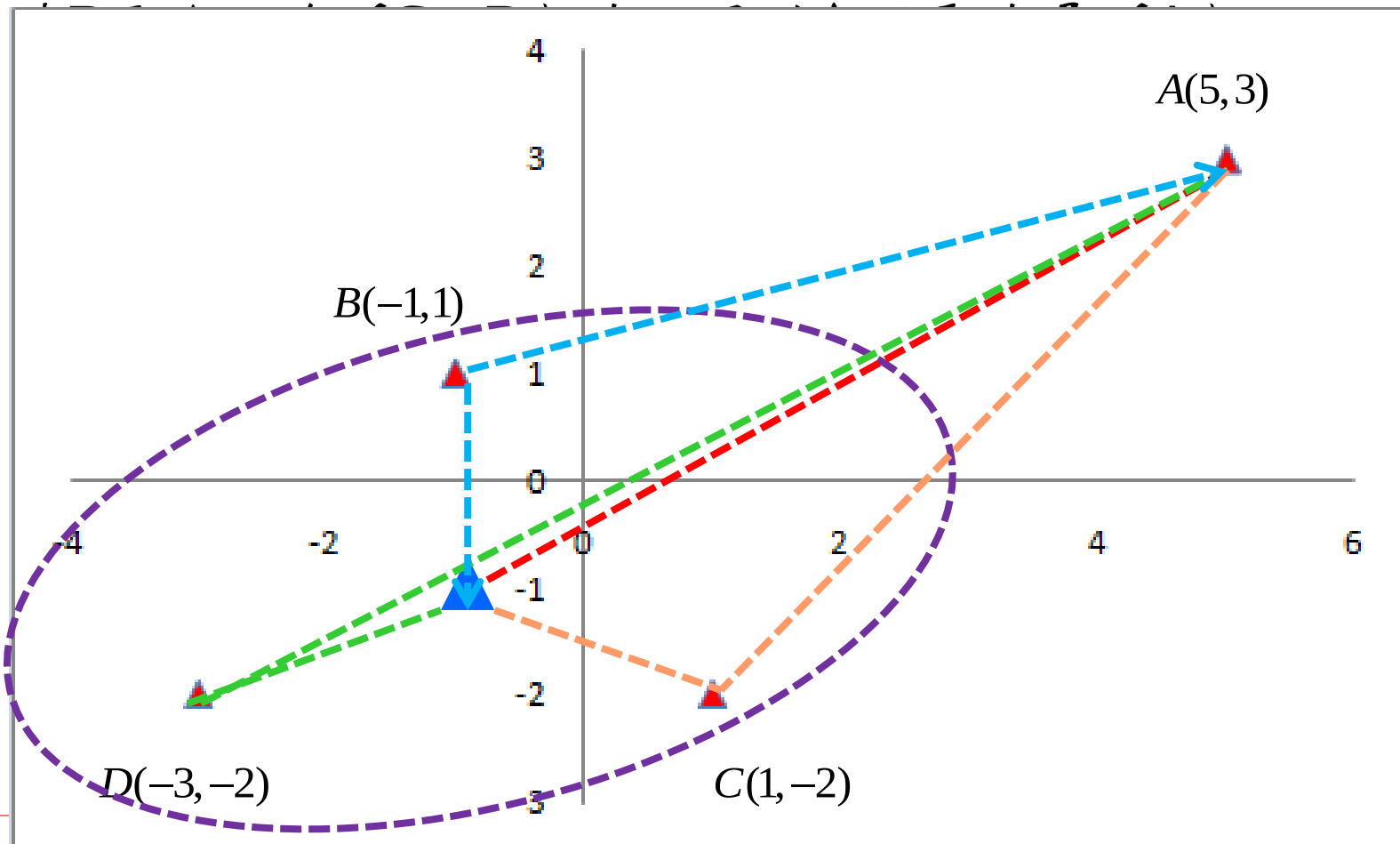
□ 第一步：按要求取 $K=2$ ，将这些样品随意分成两类，比如 (A、B) 和 (C、D)，然后计算这两个聚类的中心



□ 第二步：计算某个样品到各类中心的欧氏平方距离，然



✿ 由于B到 (A、B) 的距离大于到 (C、D) 的距离，因





第三步：再次检查每个样品，以决定是否需要重新分类。计算各样品到各中心的距离平方，得：

聚类	样品到中心的距离平方			
	A	B	C	D
(A)	0	40	41	89
(B、C、D)	52	4	5	5

- 到现在为止，每个样品都已经分配给距离中心最近的类，因此聚类过程到此结束。最终得到K=2的聚类结果是A独自成一类，B、C、D聚成一类。

疑问

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他人回复问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

