应用统计学II 第4讲 聚类分析

Instructor: 郝壮

haozhuang@buaa.edu.cn
School of Economics and Management
Beihang University

May 22, 2022

聚类分析的意义

现代的数据信息特点:

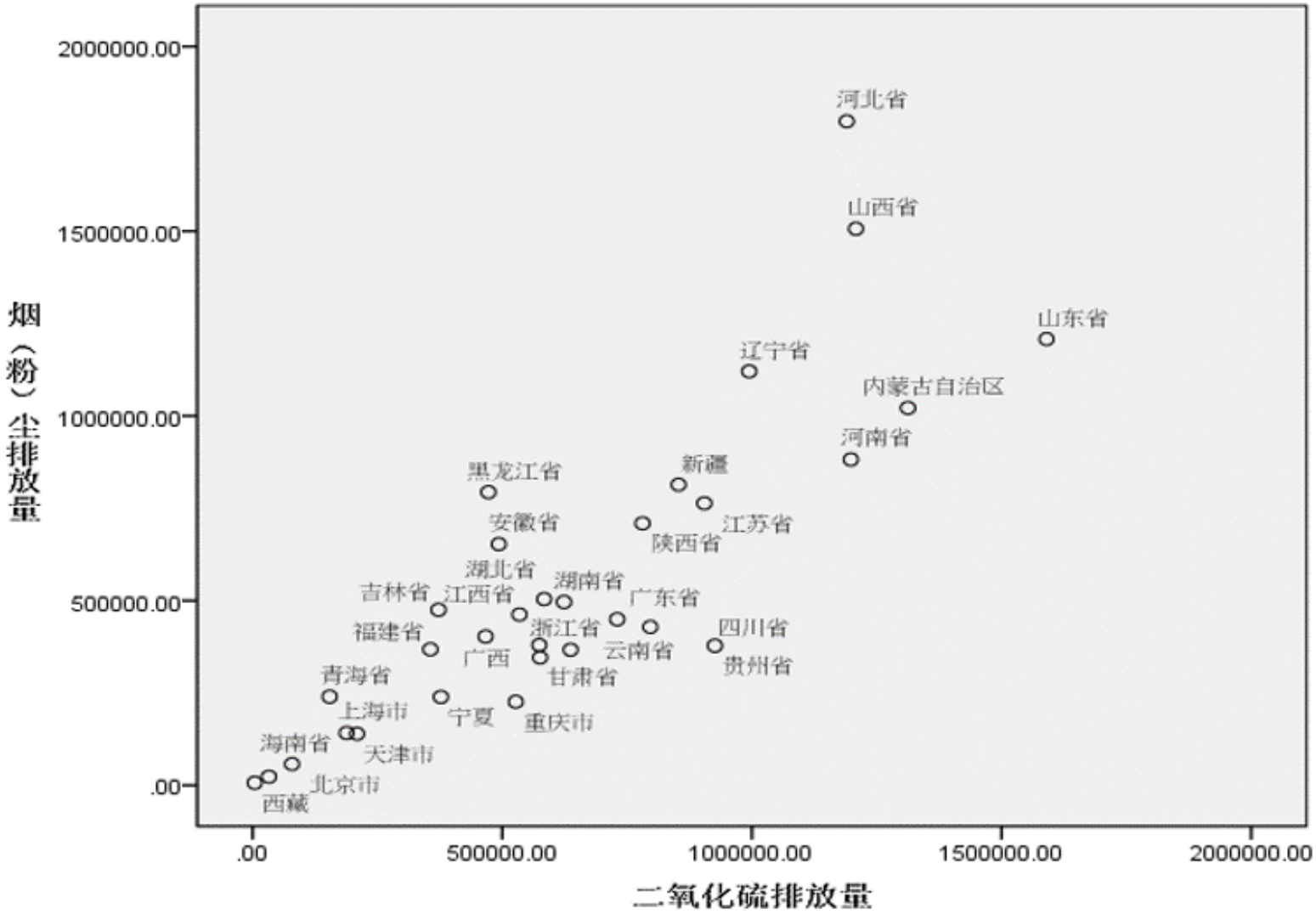
- 样本点数量巨大
- 指标变量众多

如何消除规模与复杂程度之间的关系?

- 对数据分类-聚类分析
- 对指标分类-提炼主要因素 (PCA)

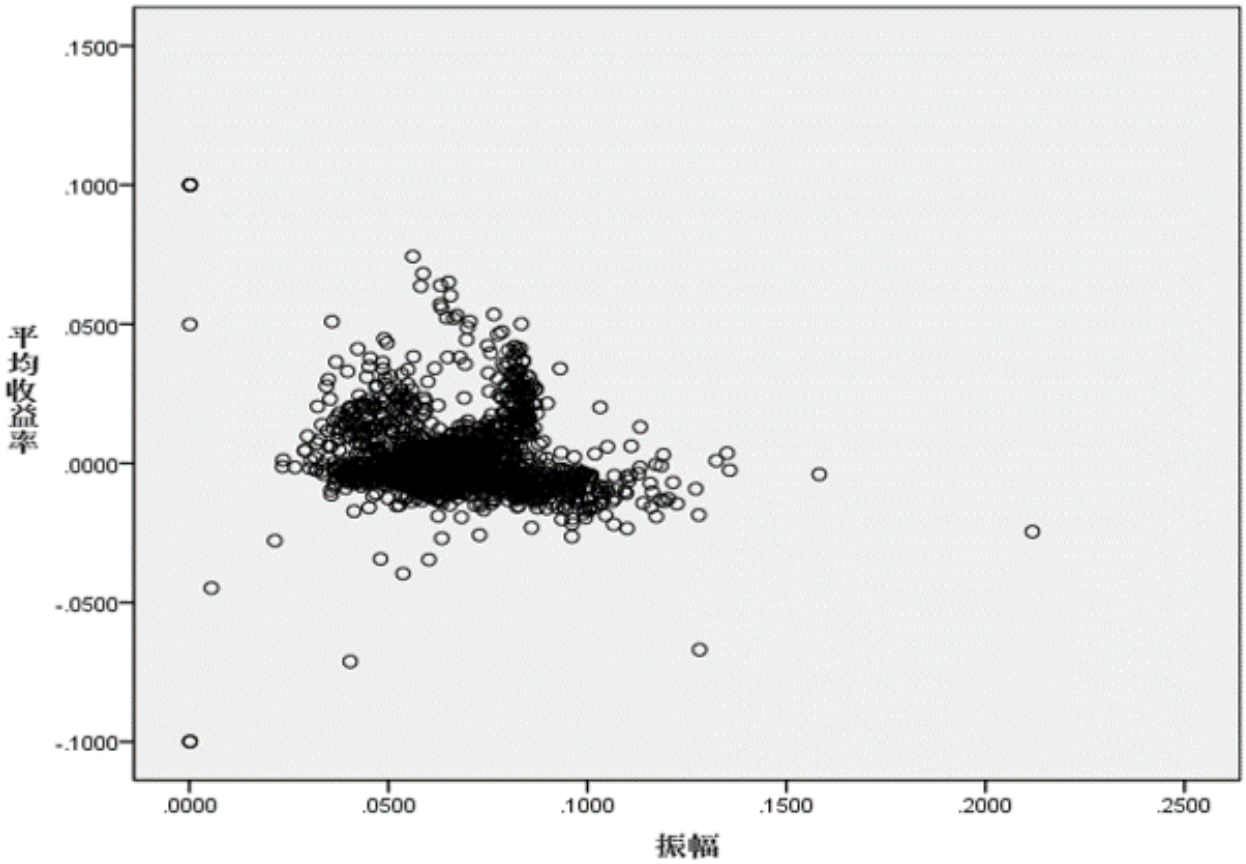
例1. 各地区二氧化硫与粉尘排放数据(2014)

散点图的应用: 两个变量, 多个样本点. 散点图是一个非常有用的可视化工具.



例2. 2015.5.4- 2015.7.31. 2617支股票

变量: 振幅(风险), 收益率.



直接使用原始数据, 很难观察股市特征

聚类分类示例

为了更清晰地对数据进行描述, 我们可以对股票按照风格进行分类

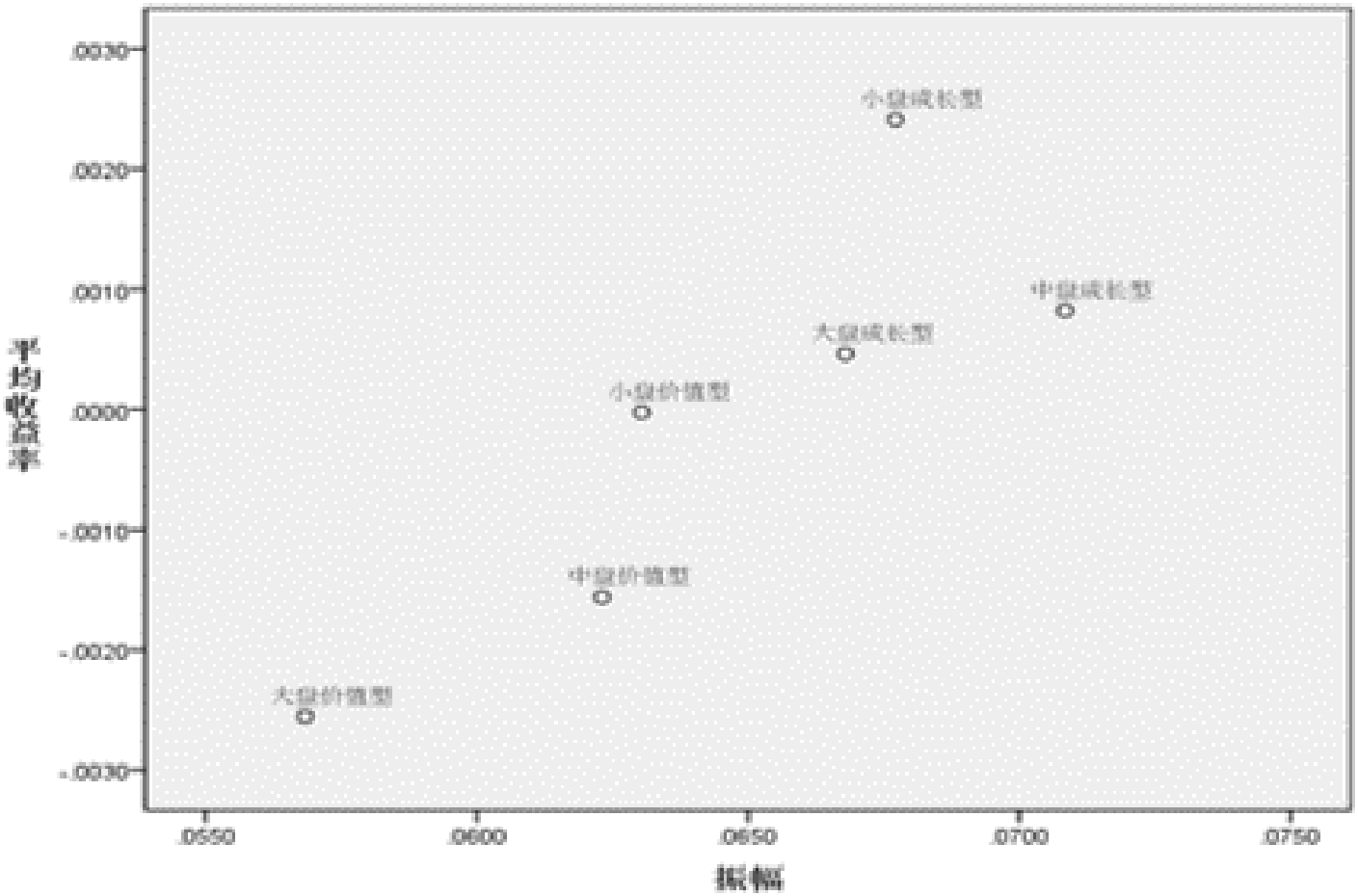
中信风格指数划分的股票风格分类

- 规模因素: 以流通市值表示
- 净市值比(简称B/P值):净资产除以总市值

表 4 – 1　中信风格股票分类

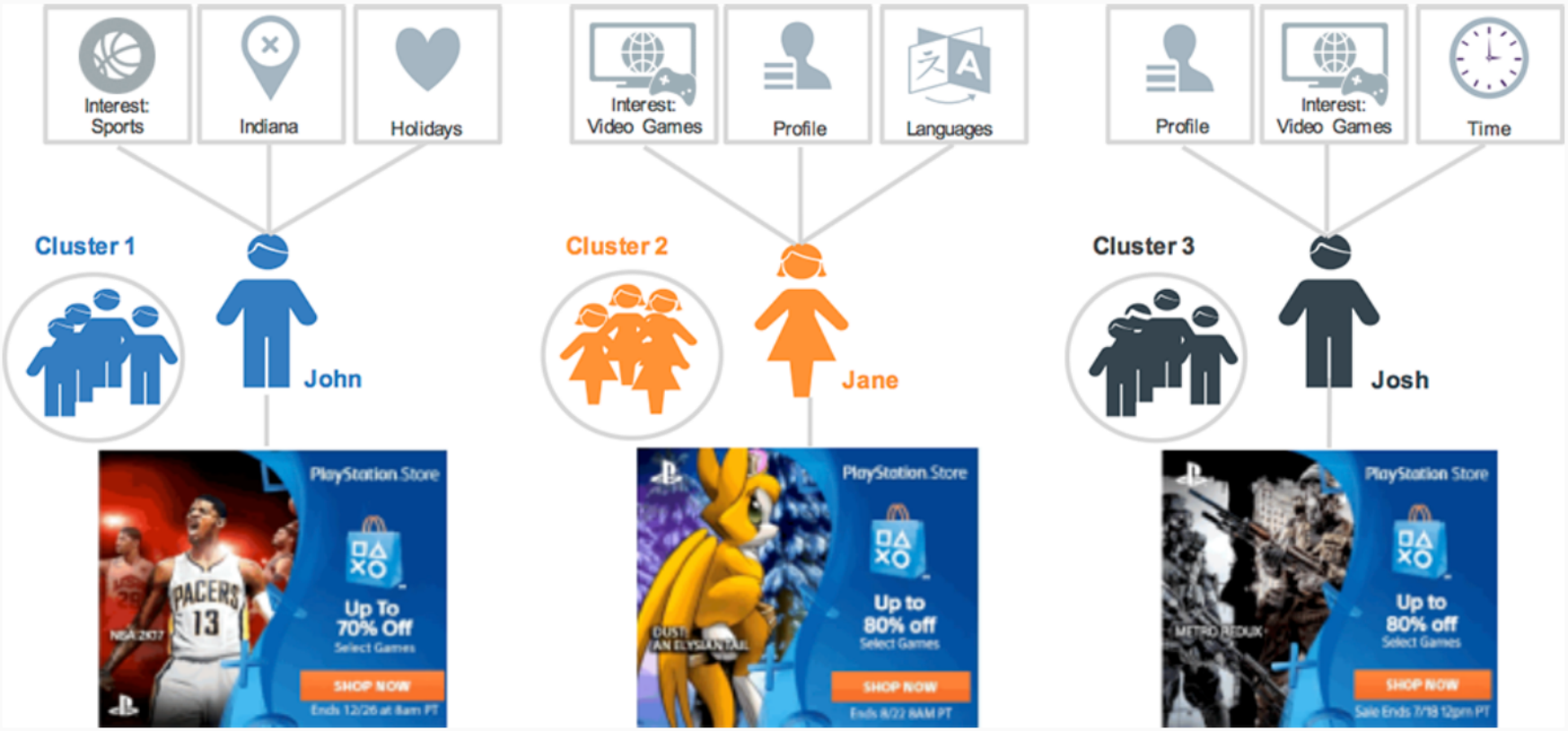
	低 B/P 值	高 B/P 值
高市值	大盘成长	大盘价值
中等市值	中盘成长	中盘价值
低市值	小盘成长	小盘价值

- 直接使用海量数据, 很难观察股市特征
- 分类之后, 更容易发现数据中隐含的模式

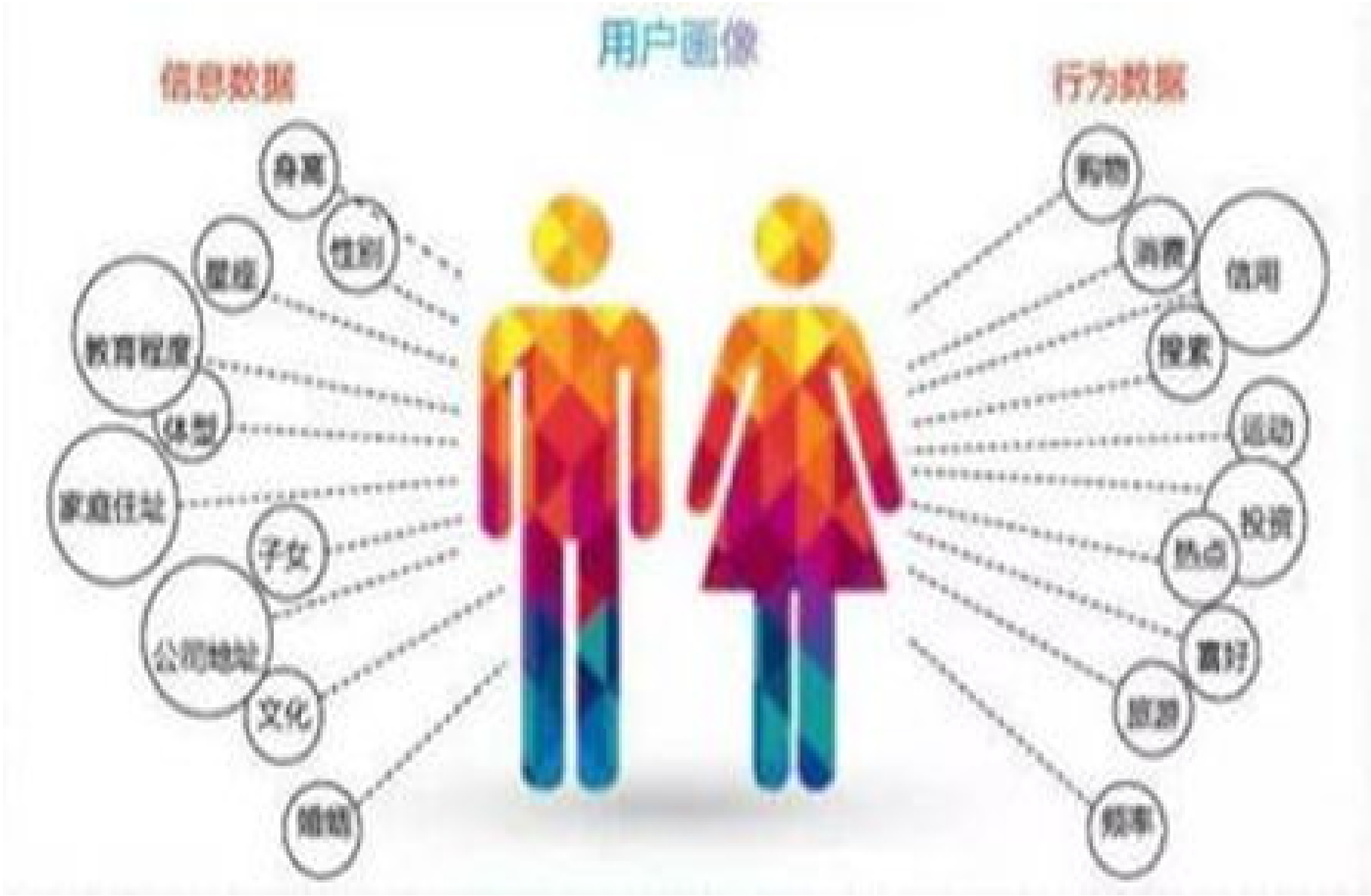


分类-用户画像-精准营销

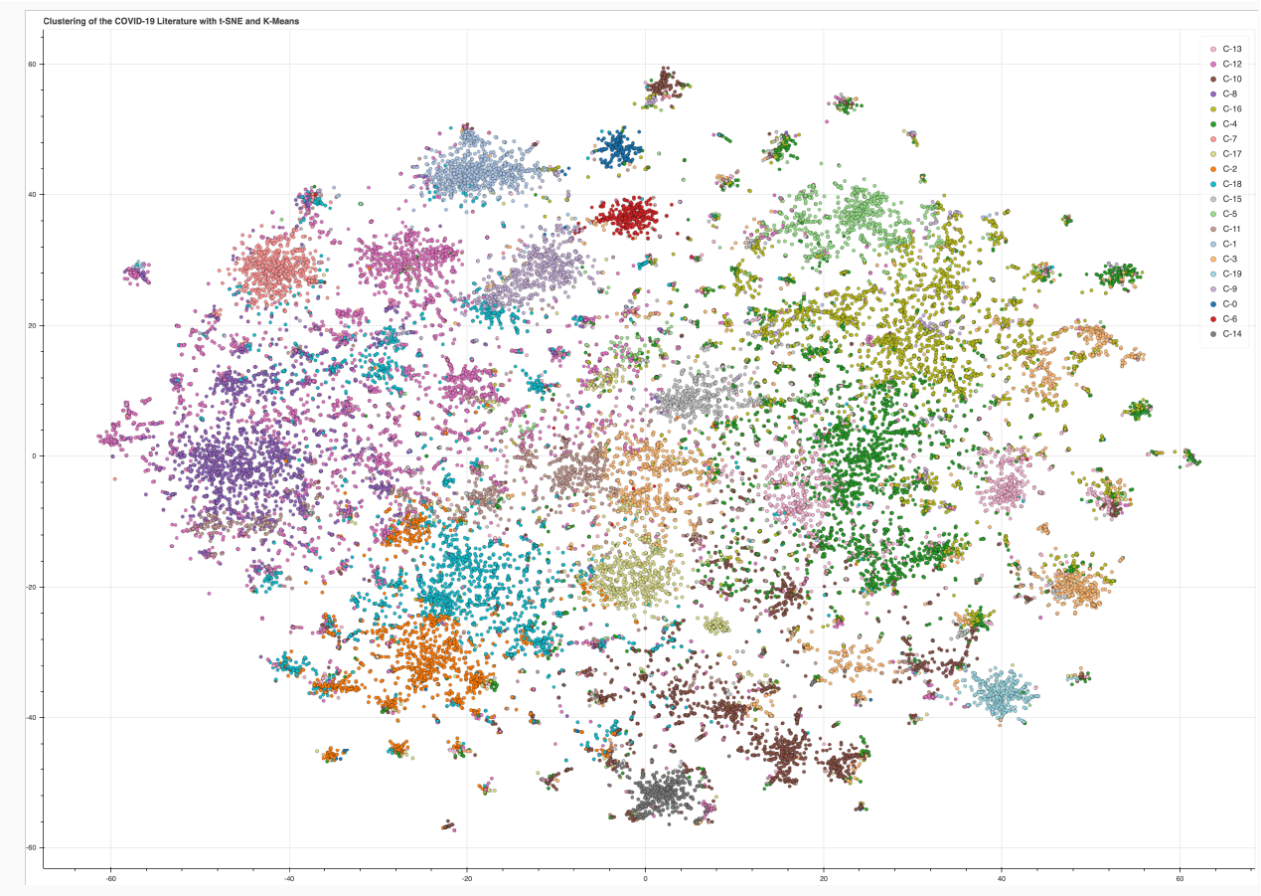
分类后，可以根据用户画像预测用户类型，并根据预测的类型进行精准营销



分类-用户画像-精准营销


用户画像
信息数据
行为数据
身高
性别
星座
教育程度
体型
家庭住址
子女
公司地址
文化
婚姻
购物
消费
信用
搜索
运动
投资
热点
置好
旅游
股票

文本聚类



Source: https://maksimekin.github.io/COVID19-Literature-Clustering/plots/t-sne_covid-19_interactive.html

用什么指标测度样本点相似性?

一. 样本点之间的相似性测度-距离

$\boldsymbol{e}_i \in \boldsymbol{R}^p$　e_i 与 e_k 的相似程度: $\mathrm{d}\left(e_i, e_k\right)$

定义: 距离 $\mathrm{d}(\cdot,\cdot) : \Omega \times \Omega \to \mathbf{R}^+$

距离应该满足如下性质

- (1) $\mathrm{d}(x,y) \geq 0, \quad \forall x, y \in \Omega$ (正定性)
  $\mathrm{d}(x,y) = 0$　当且仅当 $x = y$
- (2) $\mathrm{d}(x,y) = \mathrm{d}(y,x)$ (对称性)
- (3) $\mathrm{d}(x,y) \leq \mathrm{d}(x,z) + \mathrm{d}(z,y)$ (三角不等式)

相似性度量 - 距离

- 在聚类分析中, 如果样本点为有限维定量指标, 常用明考夫斯基距离 (Minkowski distance).
- 余弦距离(衡量文本之间的相似度常用).

明考夫斯基距离

明考夫斯基距离:

$$d_q(x,y) = \left[\sum_{j=1}^{p} |x_j - y_j|^q\right]^{1/q}$$

p为指标的个数. x和y表示任意两个样本点. 如根据家庭成员个数, 平均年龄, 收入, 对家庭进行聚类, 则指标个数有3个.

- $q = 1$ 时, 即绝对值距离: $d_1(x,y) = \sum_{j=1}^{p} |x_j - y_j|$
- $q = 2$ 时, 即欧式距离: $d_2(x,y) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$ (最常用)
- $q = \infty$ 时, 即切比雪夫距离: $d_\infty(x,y) = \max_{1\le j\le p} |x_j - y_j|$

- 最常用的是**欧氏距离.** 它的优点是: 坐标经旋转变换后, 点和点之间距离保持不变.
- 采用明氏距离时, 应采用相同量纲的变量. (如果量纲不同, 首先做数据标准化)
- 尽可能避免数据的多重相关性.

文本数据的相似性测度-余弦距离

例: N个文本中关键词出现与否的记录. 要识别N个文本的相似性(以2个记录为例).

	记录1	记录2
安检	1	0
安排	0	0
安全	0	0
安心	1	1
安装	0	1
奥蒂斯	0	0
奥林匹克	1	0
奥运	0	0
八里	0	1
白河	0	0
百吨	0	0
办法	0	0

文本数据的相似性测度-余弦距离

夹角余弦 C_{jk}

$$C_{jk} = \frac{\sum_i x_{ij} \cdot x_{ik}}{\sqrt{\sum_i x_{ij}^2}\sqrt{\sum_i x_{ik}^2}}$$

其中, jk表示第j列数据和第k列数据(本例中为第j和第k个文本), i表示第i个指标, 变量 (本例中为第i个关键词).

$0 \leq c_{jk} \leq 1$, C_{jk}越接近1, 两个文本越相似.
其中

$$x_{ij} \cdot x_{ik} = \begin{cases} 1 & x_{ij} = 1, \quad x_{ik} = 1 \\ 0 & x_{ij} = 1, \quad x_{ik} = 0 \\ 0 & x_{ij} = 0, \quad x_{ik} = 1 \\ 0 & x_{ij} = 0, \quad x_{ik} = 0 \end{cases}$$

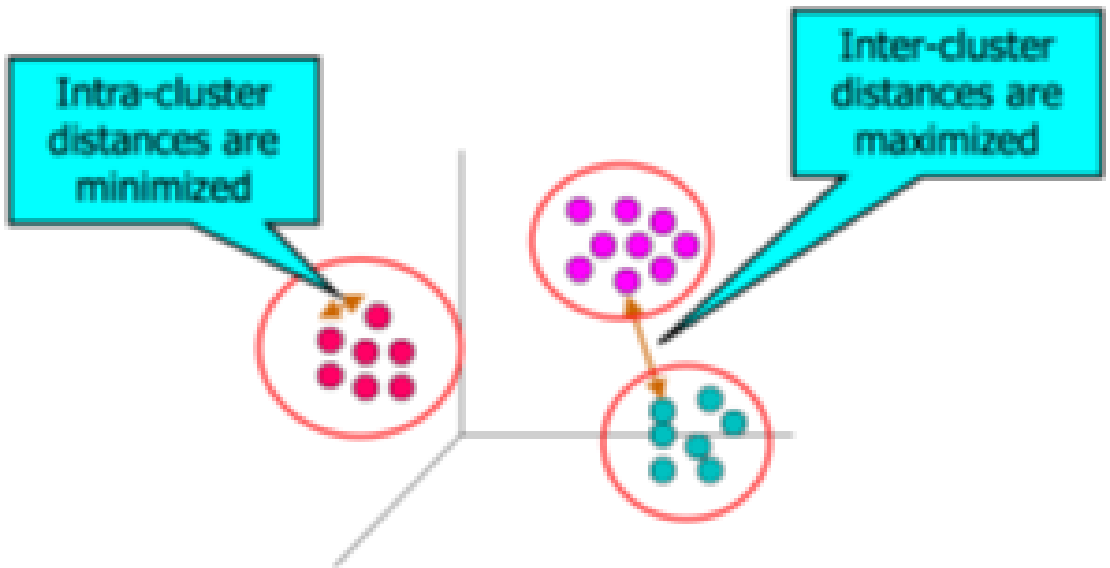
聚类方法

- K-均值聚类, K-means clustering (动态聚类)
- 分层聚类, Hierarchical clustering (系统聚类)

K-均值聚类 (K-means cluster, quick cluster)

K-均值聚类

- ①有n个样本点, 要分成K类;
- ②每一类中的元素能充分聚合;
- ③类与类之间的要充分区分.



组内的相似性越大, 组间差别越大, 聚类就越好

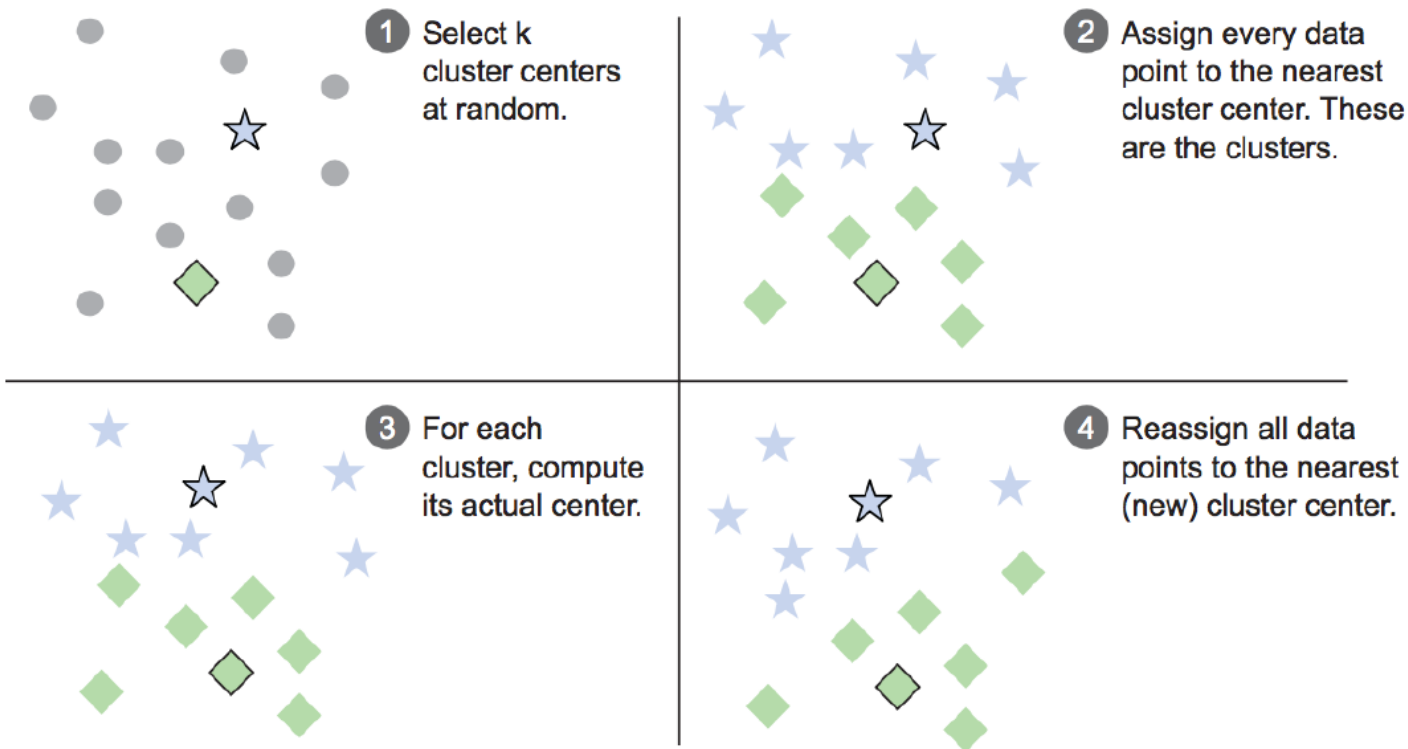
k -均值聚类 (K-means clustering)

k-均值聚类基本步骤

1. 确定类的个数 k
2. 随机抽取 k 个样本点作为 k 个类中心 (也可根据经验选取k个类中心)
3. 将每个样本点分配到距离其最近的类中心
4. 重新计算类中心
5. 重复第 3, 4 步直到样本点的类别不再变化或者达到了最大迭代次数

基本思路

k-均值聚类是一个迭代聚类算法



特点

- ① K-means是聚类的经典方法; 适用于大型的据表; 计算速度很快;
- ② **要事先确定要划分的类数K.**

例: 市场调查时, 就4000个人对衣着偏好提问, 要求把他们的回答迅速分成K类.

如何选择K

- 如何衡量聚类质量(goodness of clustering)?
- 组内距越小越好, 组间距越大越好.
- 组内平方和:

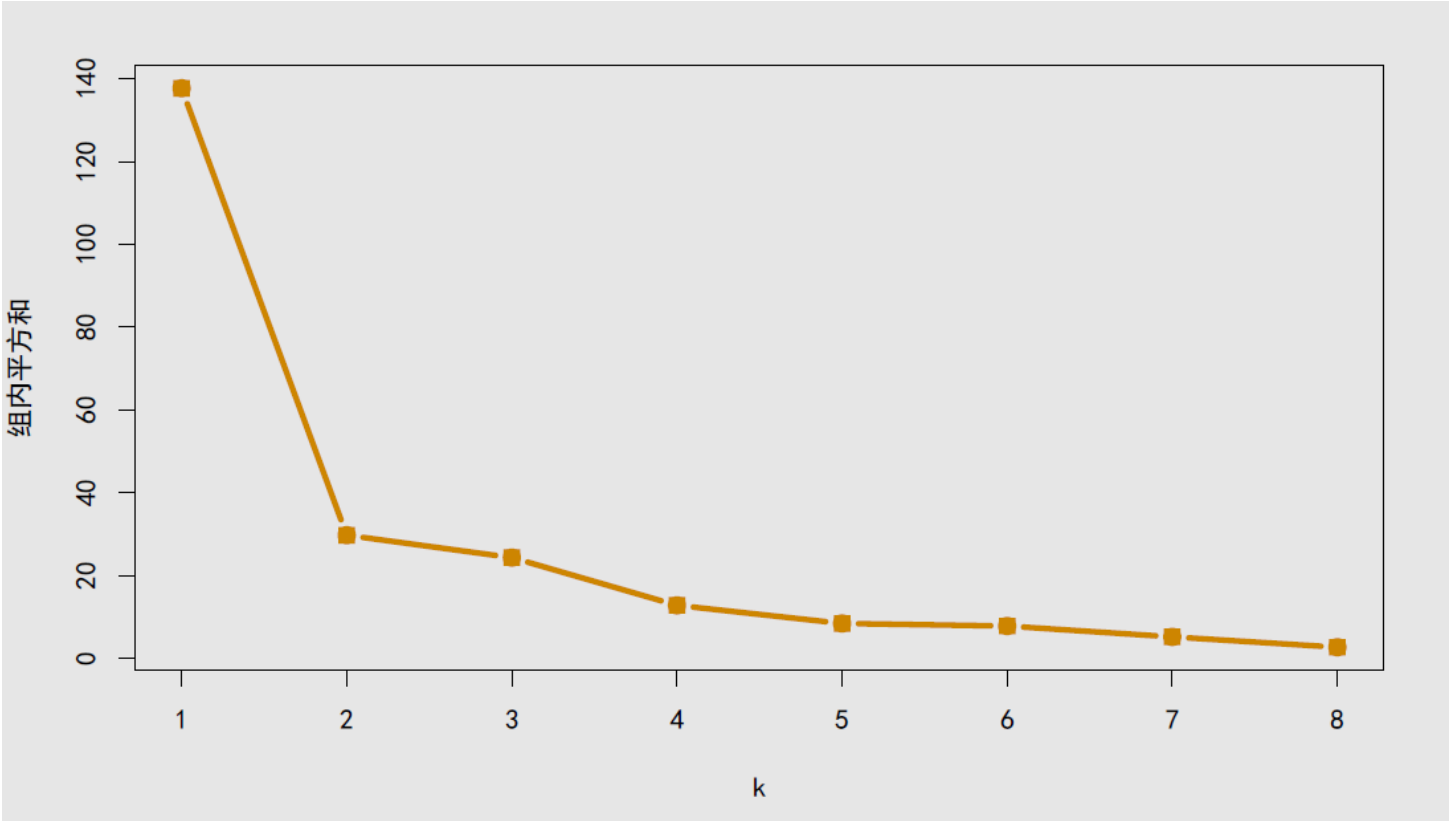
$$\text{tot.within} = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

其中 μ_k 为第 k 个类 C_k 的中心.

- **问题:** 组内平方和如何随 k 变化? (碎石图)

如何选择K

碎石图(Scree plot)



Stata中进行K-均值聚类(K-means clustering)

help cluster
help cluster kmeans

K-Means 案例: 身体指标分类

课堂演示-Stata案例: 体育课上对80个学生身体指标的测量, 包括柔韧度, 速度, 力量 flexibility, speed和strength 三个指标. 利用这三个指标, 对学生进行分组, 用于制定针对性的体育锻炼方案.

K-Means 案例: 身体指标分类

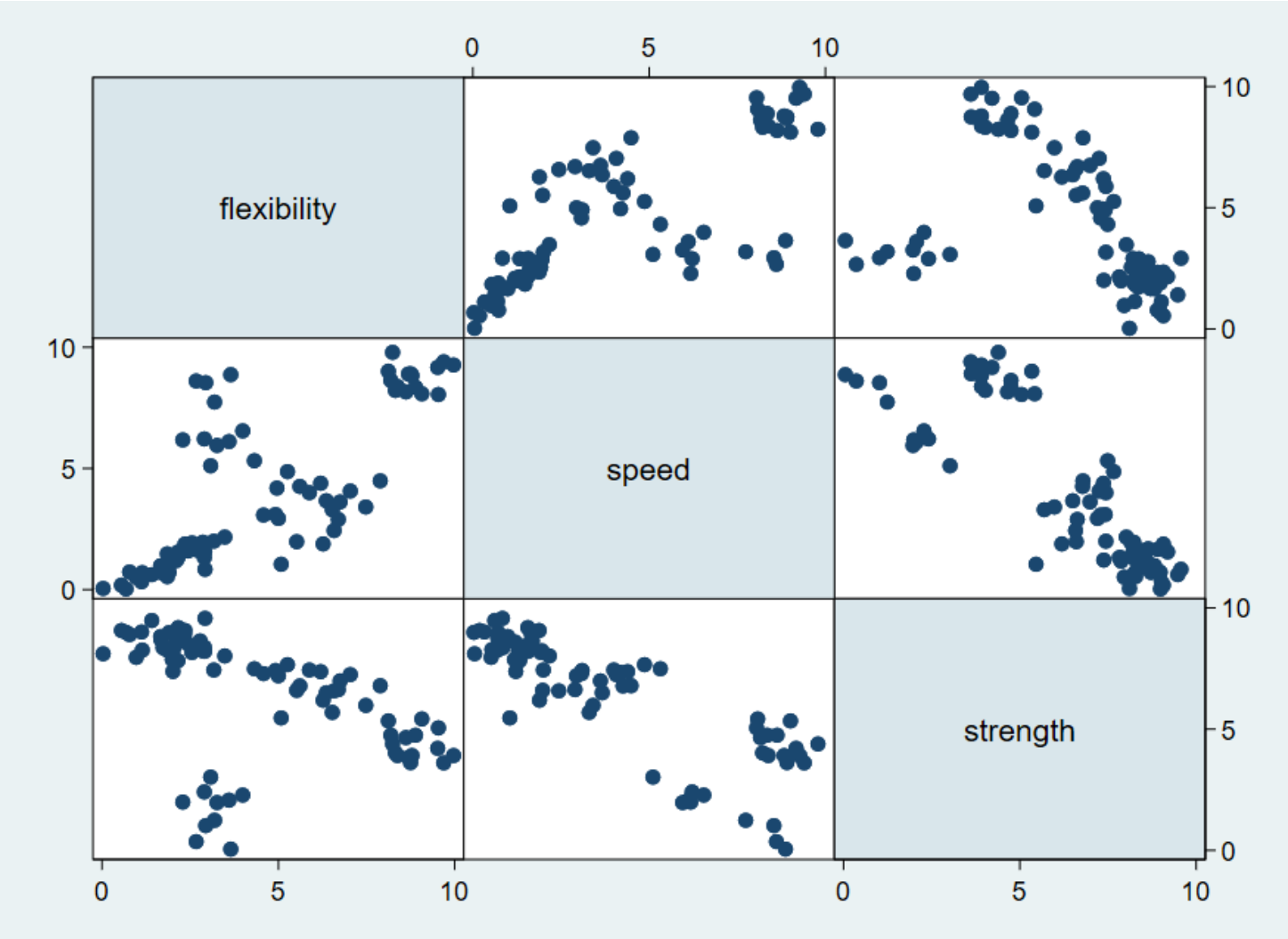
```
. use 身体指标.dta, clear

. **********描述数据**********
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
grp	80	2.8125	1.243655	1	4
flexibility	80	4.402625	2.788541	.03	9.97
speed	80	3.875875	3.121665	.03	9.79
strength	80	6.439875	2.449293	.05	9.57

K-Means 案例: 身体指标分类

```
graph matrix flex speed strength
```



K-Means 案例: 身体指标分类

由统计图可知, 学生身体指标存在差异化, 而且视觉上存在着明显分组, 所以可以进行聚类, 决定将数据分成4组, 以便于制定针对性训练方案.

K-Means 案例: 身体指标分类

如果想分成4组, 为便于复制结果, 用"seed"固定伪随机数生成起始点.

```
//1. 利用option krandom() 随机选择 k 个观测作为起始类中心

//2. 利用seed命令 s(kr(385617)) 赋予随机初始伪随机数以保证结果可复制性(reprodu

//3. 利用option keepcenters 生成4类的类均值并附在数据最后(第81-84观测)

. cluster k flex speed strength, k(4) name(g4abs) s(kr(385617)) mea(ab
```

K-Means 案例: 身体指标分类

结果: 1. 多出一列分组信息 2. 81-84观察为各组中心点

```
. list
     +-----------------------------------------------------+
     | grp   flexib~y      speed   strength    g4abs |
     |-----------------------------------------------------|
  1. |   2         3.6       6.11       2.07        3 |
  2. |   4        1.12        .33       9.01        4 |
  3. |   3        8.69        8.9       3.83        1 |
  4. |   2        2.67       8.61        .36        3 |
  5. |   4        2.78       1.69       8.64        4 |
     |-----------------------------------------------------|
...
     |-----------------------------------------------------|
 76. |   4        2.44        1.6       8.51        4 |
 77. |   4        2.34       1.66       8.91        4 |
 78. |   1        5.08       1.05       5.46        2 |
 79. |   4        1.67        .87        8.7        4 |
 80. |   3        9.52       9.17       4.21        1 |
     |-----------------------------------------------------|
 81. |   .       8.852   8.743333      4.358        . |
 82. |   .      5.9465     3.4485     6.8325        . |
 83. |   .       3.157      6.988      1.641        . |
 84. |   .    1.969429   1.144857   8.478857        . |
     +-----------------------------------------------------+
```

K-Means 案例: 身体指标分类

检查各组最大值, 最小值, 均值. 分析各组有什么特点.

```
drop in 81/L  // drop those with missing information

tabstat flex speed strength, by(g4abs) stat(min mean max)

Summary statistics: min, mean, max by categories of: g4abs
   g4abs |  flexib~y     speed  strength
---------+------------------------------
       1 |      8.12      8.05      3.61
         |     8.852  8.743333     4.358
         |      9.97      9.79      5.42
---------+------------------------------
       2 |      4.32      1.05      5.46
         |    5.9465    3.4485    6.8325
         |      7.89      5.32      7.66
---------+------------------------------
       3 |      2.29      5.11        .05
         |     3.157     6.988     1.641
         |      3.99      8.87      3.02
---------+------------------------------
       4 |        .03        .03      7.38
         |  1.969429  1.144857  8.478857
         |      3.48      2.17      9.57
---------+------------------------------
```

K-Means 案例: 身体指标分类

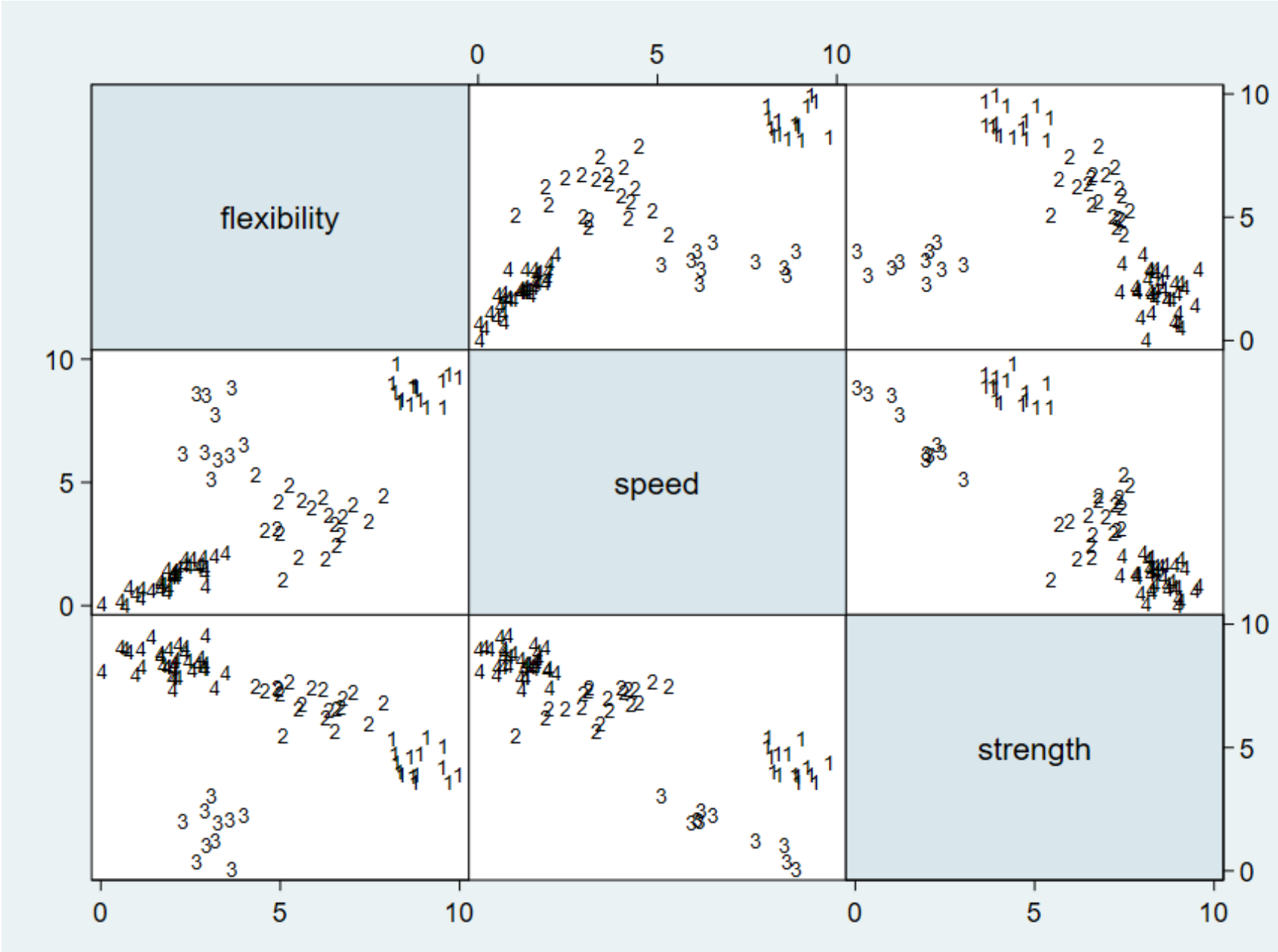
- 组1, 15人, 柔韧性和速度已经较好, 仅需对力量进行训练
- 组2, 20人, 速度训练要着重, 其他训练也要加强
- 组3, 10人, 加强柔韧性和力量训练
- 组4, 35人, 加强柔韧性和速度训练

K-Means 案例: 身体指标分类

数据可视化：利用画图命令graph可将分组结果展示在图上, 并将类别号用为画图符号

```
. graph matrix flex speed strength, m(i) mlabel(g4abs) mlabpos(0)
```

K-Means 案例: 身体指标分类

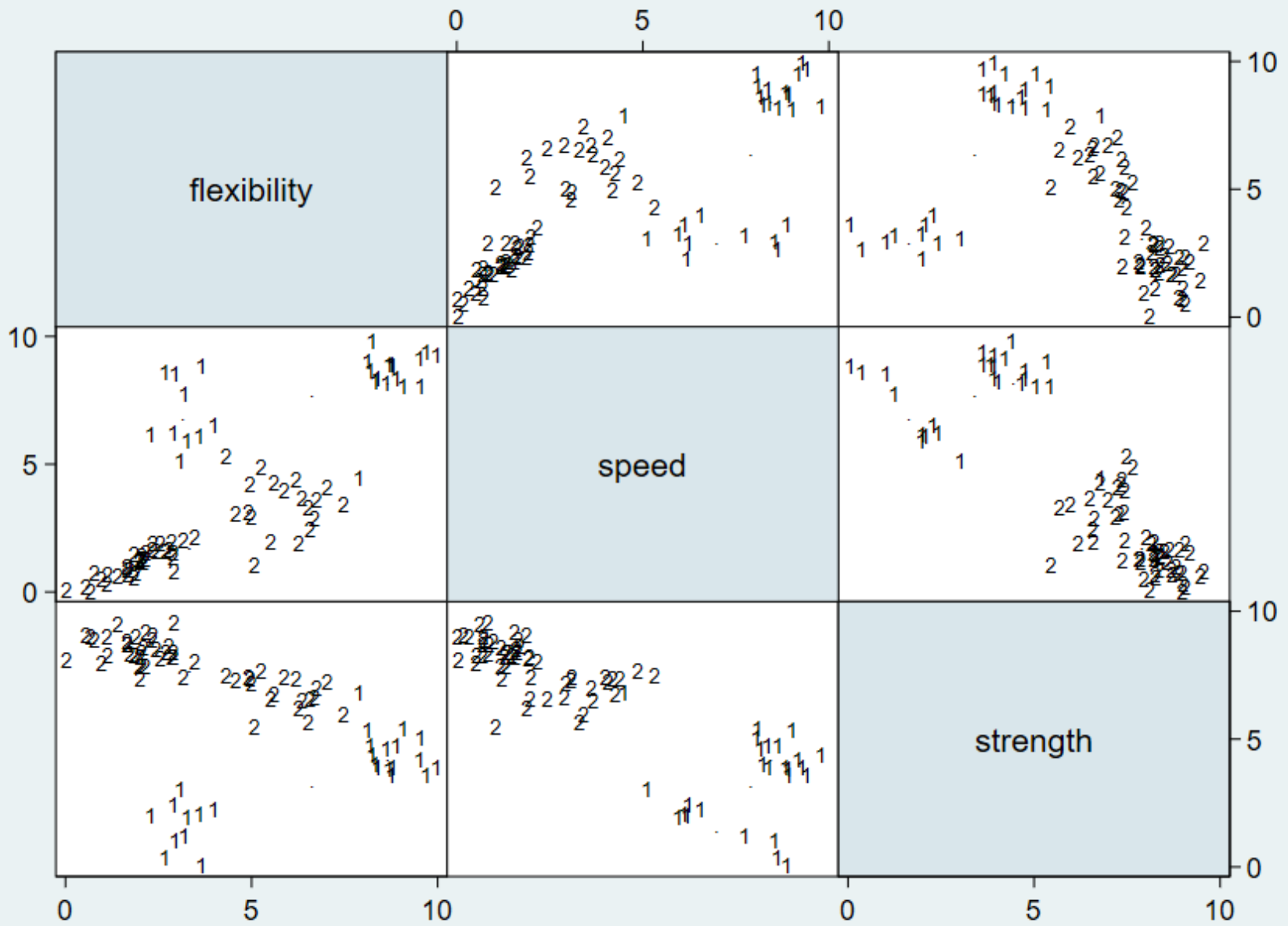


K-Means 案例: 身体指标分类

分成两组结果会怎样?

```
cluster k flex speed strength, k(3) name(g3abs) s(kr(385617)) mea(abs)
graph matrix flex speed strength, m(i) mlabel(g3abs) mlabpos(0)
```

K-Means 案例: 身体指标分类



系统聚类/分层聚类 (Hierachical clustering)

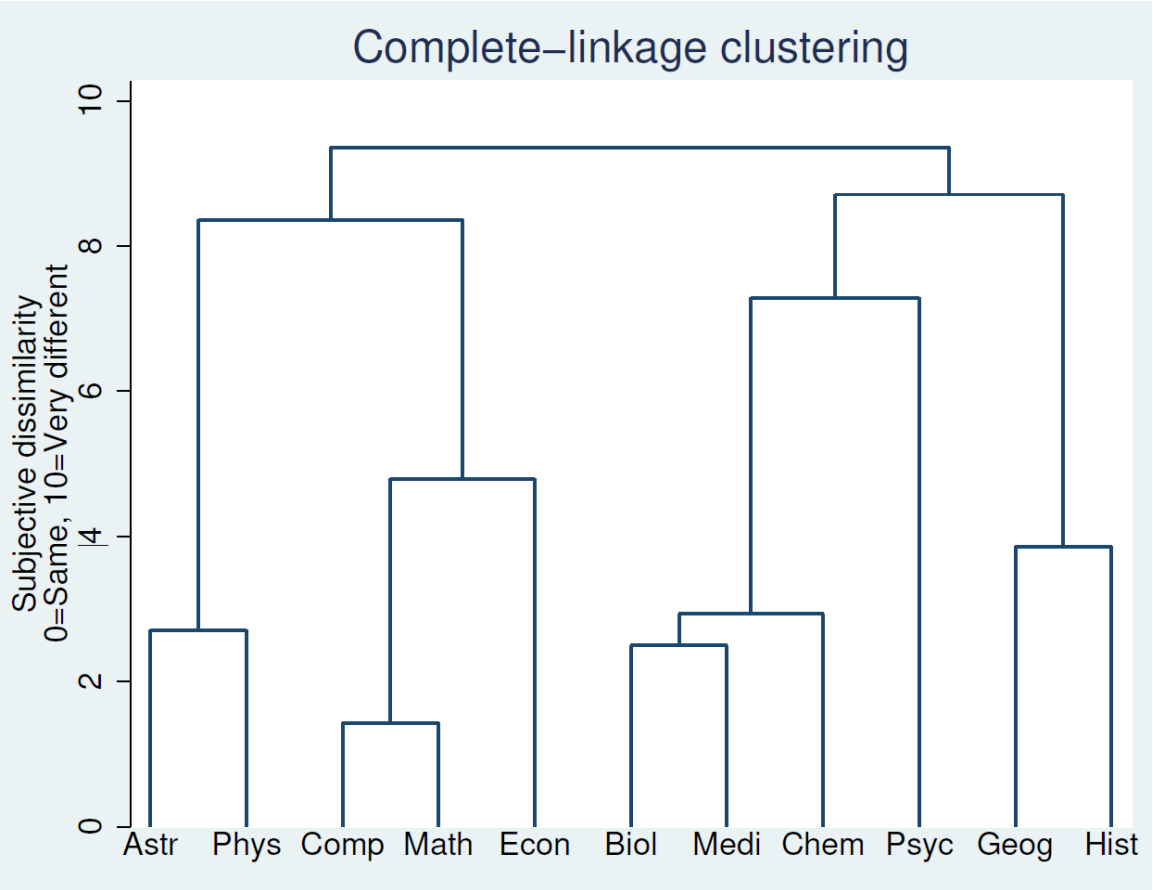
- 系统聚类/分层聚类算法也是一种迭代算法
- 它不需要提前指定分类个数
- 不适用大规模数据
- 分层聚类输出一个树状结构, 可以指出从粗到细的所有分类结果

系统聚类/分层聚类步骤

1. 每个样本点自成一类
2. 选择最近的两个类聚成一类
3. 计算新的类与类之间的距离
4. 重复第 2, 3 步直至所有的样本点聚为一类

系统树图/树状图/谱系图(dendrogram)

例: 学科相似程度



每层高度代表类间不同的程度(距离). 思考:

- 如果分成3类是哪3类
- 如果分成4类是哪4类

如何计算类 C_1 与类 C_2 之间的距离？聚合指数

聚合分析(agglomerative cluster analysis): 类(group) 与类(group)不同程度的度量. 主要的聚合方法：

1. 最短距离法(single linkage, default): $D(C_1, C_2) = \min_{\substack{x_i \in C_1 \\ y_j \in C_2}} \{d(x_i, y_j)\}$.
2. 最长距离法(complete linkage): $D(C_1, C_2) = \max_{\substack{x_i \in C_1 \\ y_j \in C_2}} \{d(x_i, y_j)\}$.
3. 重心法(centroid linkage): $D(C_1, C_2) = d(\bar{x}, \bar{y})$.
4. 类平均法(average linkage): $D(C_1, C_2) = \frac{1}{l \times m} \sum_{x_i \in C_1} \sum_{y_j \in C_2} d(x_i, y_j)$. (各类中所有样本点距离总和的平均值)
5. ...

其中 $\mathrm{d}(\cdot, \cdot) : \Omega \times \Omega \to \mathbf{R}^+$ 是样本点之间的距离.
(Stata: help cluster linkage -> Hierarchical cluster-analysis methods)

分层聚类案例-推销员问题

用系统聚类法对下面数据进行聚类.

要求:

- 使用绝对值距离计算样本点距离
  $d^2(w_i, w_k) = \sum_{j=1}^{2} |x_{ij} - x_{kj}|$;
- 使用最短距离法测度聚合指数
  $D(G_1, G_2) = \min_{\substack{x_i \in G_1 \\ y_j \in G_2}} \{d(x_i, y_j)\}$.

销售员	销售量 (百件)	回收款项 (万元)
w_1	1	0
w_2	1	1
w_3	3	2
w_4	4	3
w_5	2	5

(绝对值距离例: w_1和w_2间距离:
$d^2(w_1, w_2) = |1-1| + |0-1| = 1$)

分层聚类案例-推销员问题

I. 构造距离矩阵：

$$
\left(d_{ij}\right)_{5\times5}=\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array}\begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & w_4 & w_5 \end{array} \\ \begin{pmatrix} 0 & 1 & 4 & 6 & 6 \\ & 0 & 3 & 5 & 5 \\ & & 0 & 2 & 4 \\ & & & 0 & 4 \\ & & & & 0 \end{pmatrix} \end{array}
$$

II . $w_1 \ldots\ldots w_5$ 自成一类: $h_1 \ldots\ldots h_5$ 选择最接近的两元素聚成一类:

$$
\begin{aligned} h_6 &= w_1 \cup w_2 = h_1 \cup h_2 \\ D\left(h_1, h_2\right) &= d\left(w_1, w_2\right) = 1 \end{aligned}
$$

平台高度: $f\left(h_6\right) = 1$

分层聚类案例-推销员问题

III. 计算新类之间的关系： w_3, w_4, w_5, h_6

$$
\left(d_{ij}\right)_{5\times5}=\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array}\begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & w_4 & w_5 \end{array} \\ \begin{pmatrix} 0 & 1 & 4 & 6 & 6 \\ & 0 & 3 & 5 & 5 \\ & & 0 & 2 & 4 \\ & & & 0 & 4 \\ & & & & 0 \end{pmatrix} \end{array} \Rightarrow \begin{array}{c} \\ w_3 \\ w_4 \\ w_5 \\ h_6 \end{array}\begin{array}{c} \begin{array}{cccc} w_3 & w_4 & w_5 & h_6 \end{array} \\ \begin{pmatrix} 0 & 2 & 4 & 3 \\ & 0 & 4 & 5 \\ & & 0 & 5 \\ & & & 0 \end{pmatrix} \end{array}
$$

$$
\begin{aligned}
&D\left(w_3, h_6\right)=\min\left\{d\left(w_3, w_1\right), d\left(w_3, w_2\right)\right\}=\min\{4,3\}=3 \\
&D\left(w_4, h_6\right)=\min\left\{d\left(w_4, w_1\right), d\left(w_4, w_2\right)\right\}=\min\{6,5\}=5 \\
&D\left(w_5, h_6\right)=\min\{6,5\}=5
\end{aligned}
$$

取最相似的并成一类: $h_7=w_3 \cup w_4, f\left(h_7\right)=2$

分层聚类案例-推销员问题

IV. 计算新类之间的关系： w_5, h_6, h_7

$$\begin{array}{c} \\ w_3 \\ w_4 \\ w_5 \\ h_6 \end{array}\begin{array}{c} \begin{array}{cccc} w_3 & w_4 & w_5 & h_6 \end{array} \\ \begin{pmatrix} 0 & 2 & 4 & 3 \\ & 0 & 4 & 5 \\ & & 0 & 5 \\ & & & 0 \end{pmatrix} \end{array} \Rightarrow \begin{array}{c} \\ w_5 \\ h_6 \\ h_7 \end{array}\begin{array}{c} \begin{array}{ccc} w_5 & h_6 & h_7 \end{array} \\ \begin{pmatrix} 0 & 5 & 4 \\ & 0 & 3 \\ & & 0 \end{pmatrix} \end{array}$$

$$D(w_5, h_7) = \min\{d(w_5, w_3), d(w_5, w_4)\} = \min\{4, 4\} = 4$$

$$D(h_6, h_7) = \min\{D(h_6, w_3), D(h_6, w_4)\} = \min\{3, 5\} = 3$$

取最相似的并成一类: $\mathrm{h}_8 = \mathrm{h}_6 \cup \mathrm{h}_7$

平台高度: $\mathrm{f}(\mathrm{h}_8) = 3$

分层聚类案例-推销员问题

V. 计算新类之间的关系： w_5, h_8

$$\begin{array}{c} \\ w_5 \\ h_6 \\ h_7 \end{array}\begin{array}{c} \begin{array}{ccc} w_5 & h_6 & h_7 \end{array} \\ \begin{pmatrix} 0 & 5 & 4 \\ & 0 & 3 \\ & & 0 \end{pmatrix} \end{array} \Rightarrow \begin{array}{c} \\ w_5 \\ h_8 \end{array}\begin{array}{c} \begin{array}{cc} w_5 & h_8 \end{array} \\ \begin{pmatrix} 0 & 4 \\ & 0 \end{pmatrix} \end{array}$$

$$\mathrm{D}\left(\mathrm{w}_5,\ \mathrm{h}_8\right) = \min\left\{\mathrm{D}\left(\mathrm{w}_5,\ \mathrm{h}_6\right), \mathrm{D}\left(\mathrm{w}_5,\ \mathrm{h}_7\right)\right\} = \min\{4, 5\} = 4$$

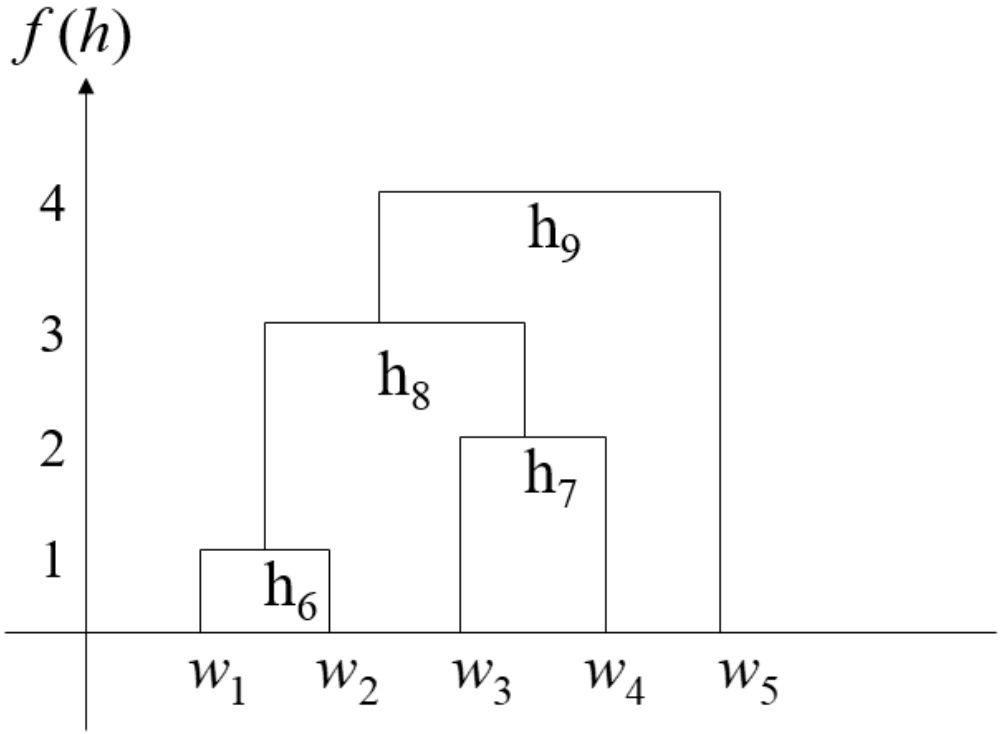
$$\mathrm{h}_9 = \mathrm{w}_5 \cup \mathrm{h}_8$$

$$\mathrm{f}\left(\mathrm{h}_9\right) = 4$$

由于所有点已聚为一类, 计算停止, 转入绘聚类图.

分层聚类案例-推销员问题

VI. 绘制聚类图：二分树法



分成二类: $(w_1, w_2, w_3, w_4)\,(w_5)$

分成三类: $(w_1, w_2)\,,(w_3, w_4)\,,(w_5)$

等等.

分层聚类案例-推销员问题

原始数据

销售员	销售量 (百件)	回收款项 (万元)
w_1	1	0
w_2	1	1
w_3	3	2
w_4	4	3
w_5	2	5

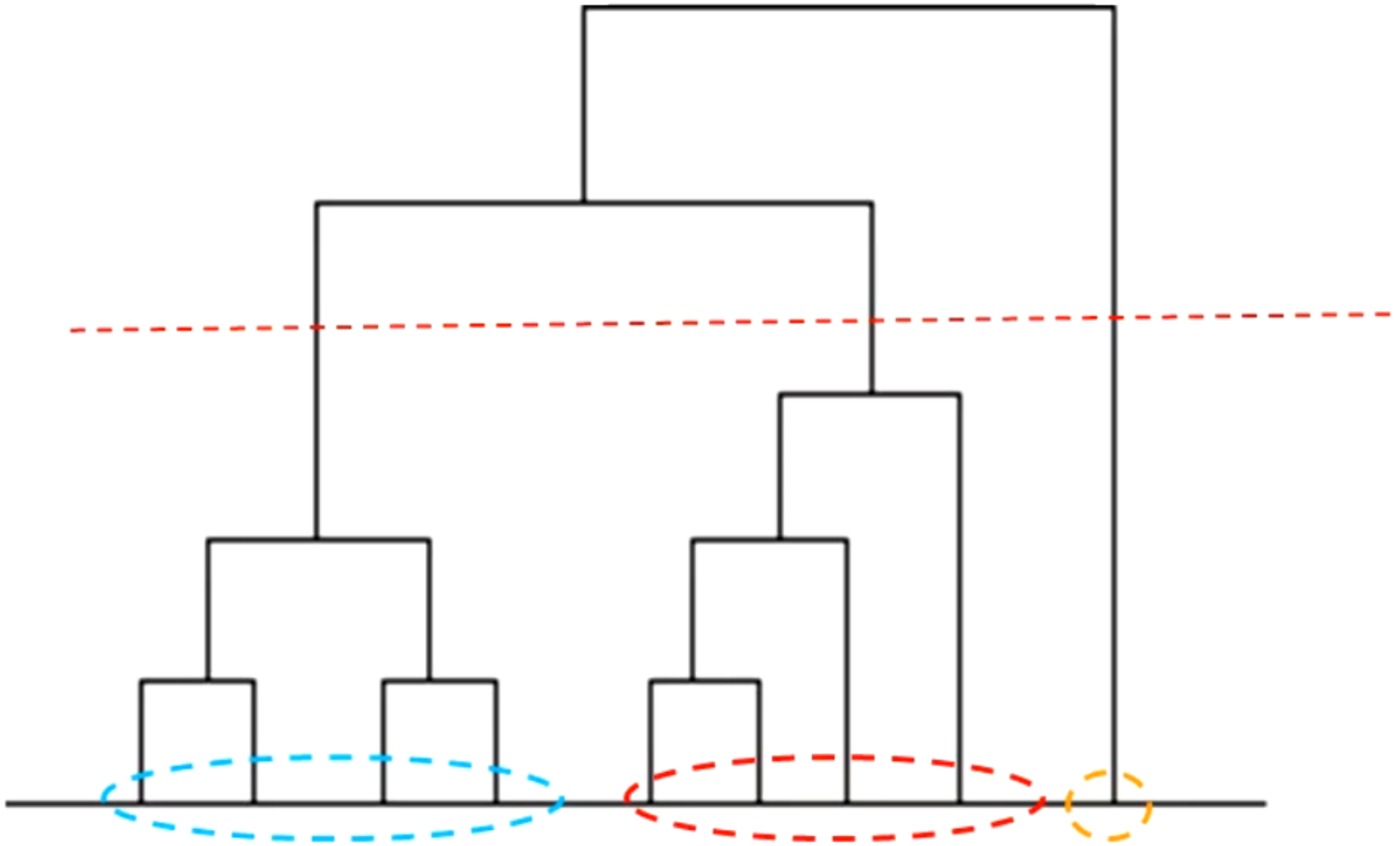
若分成3类的类重心:

销售员	销售量	回收款项
G1	**1**	**0.5**
G2	**3.5**	**2.5**
G3	**2**	**5**

(思考3类各有什么特点?)

怎样判断应分为几类更合适

标度突变法: 谱系图中截取到的高度变化均较大时的类数可以作为分类数量. Only a guideline.



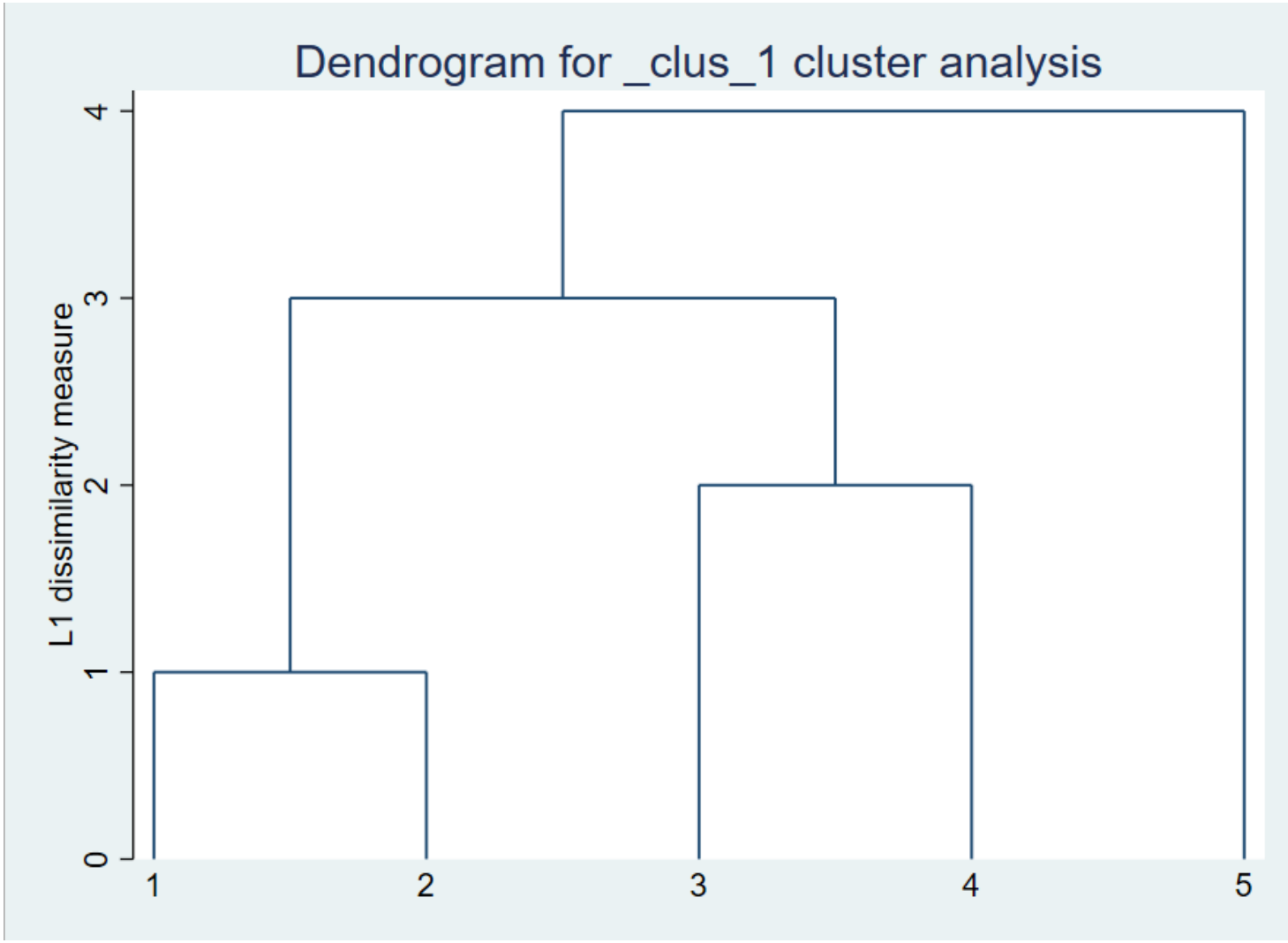
Stata中进行分层聚类

cluster linkage – Hierarchical cluster analysis
help cluster linkage

```
use 推销员问题.dta, clear
//所有聚类方法
help cluster
// 系统聚类方法
help cluster linkage

// 测度聚合指数: linkage  测度样本点距离: measure
// 最短距离测度聚合指数,  绝对值距离测度样本点距离
cluster s 销售量 回收款项, measure(L1) // or measure(absolute)
cluster tree
```

Stata中进行分层聚类



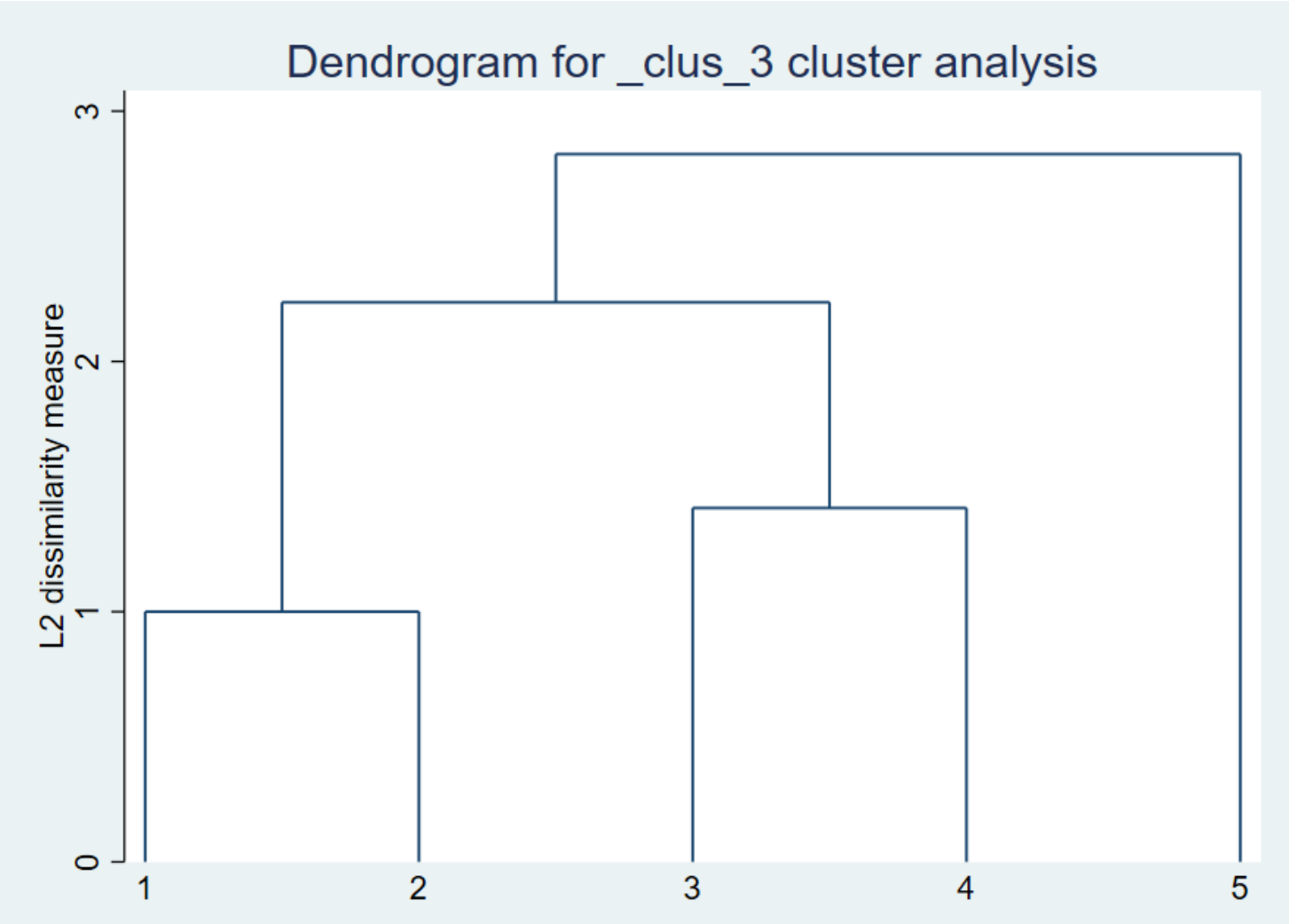
Stata中进行分层聚类

```
// 最短距离测度聚合指数， 欧式距离测度样本点距离

cluster s 销售量 回收款项 //or measure(Euclidean), or measure(L2)

cluster s 销售量 回收款项, measure(Euclidean)
cluster tree
```

Stata中进行分层聚类



* 距离高度变化(因为聚合指数测度方法变化), 其他不变

定性变量的聚类分析

假设有5个人, 他们的身体特征指标如下. 试对样本点进行分类(既有定量变量, 又有定性变量)

id	身高 (公分)	体重 (斤)	眼睛形状	鼻子形状	习惯用手	性别
1	166	120	单	高	右	女
2	175	145	双	低	右	男
3	168	135	单	高	右	男
4	167	100	双	低	右	女
5	174	150	双	低	左	男

若对该数据进行聚类, 由于身高体重有和dummy variable不同的量纲, 对距离的影响非常大!

如何消除量纲?

定性变量的聚类分析

可能处理方法: 把所有变量都化成哑变量处理.

$$x_1 = \begin{cases} 1 & \text{身高} \geq 170 \\ 0 & \text{身高} < 170 \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{体重} \geq 130 \\ 0 & \text{体重} < 130 \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{双} \\ 0 & \text{单} \end{cases}$$

等等

定性变量的聚类分析

可能处理方法: 把所有变量都化成哑变量处理.

id	身高 (x_1)	体重 (x_2)	眼睛 (x_3)	鼻子 (x_4)	习惯用手 (x_5)	性别 (x_6)
1^0	0	0	0	1	0	1
2^0	1	1	1	0	0	0
3^0	0	1	0	1	0	0
4^0	0	0	1	0	0	1
5^0	1	1	1	0	1	0

对样本点的分类主要是根据两点之间的共同特征的多少进行的：相似的样本点比不相似的样本点应具有更多的共同特征. (常用测度方法：匹配法 matching)

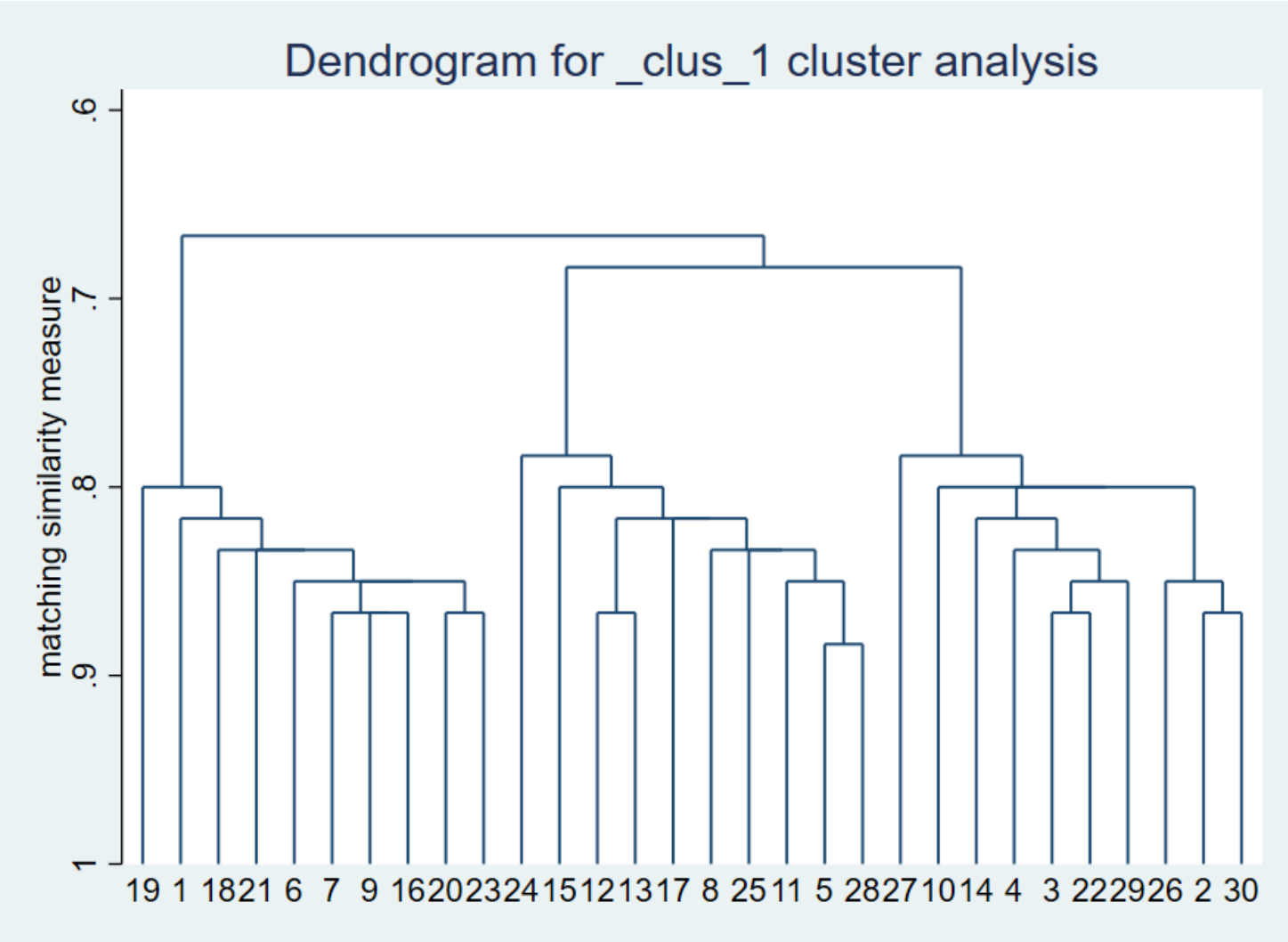
Stata中进行分层聚类案例2: 定性变量或哑变量聚类分析

```
use https://www.stata-press.com/data/r16/homework, clear

cluster s a1-a60, measure(matching)
cluster tree
```

	a1	a2	a3	a4	a5	a6
1	0	0	0	0	0	1
2	1	0	0	0	0	0
3	0	1	0	0	0	0
4	0	1	0	1	0	1
5	0	0	1	1	1	1
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	1	0	1	1
9	0	0	0	0	0	0
10	1	1	0	1	0	0
11	0	1	1	0	1	1

Stata中进行分层聚类案例2: 定性变量或哑变量聚类分析



拓展知识：其他常用的聚类分析模型

机器学习中的无监督方法

- 单指标分裂聚类法(DIV)
- 基于密度的聚类算法
- 自组织映射 (SOM)
- 混合模型聚类, EM(软聚类)
- ...

应用统计学II 作业4暨上机实验10

1, 下表给出6种精神治疗药物的3种临床测量指标数据, 请利用分层聚类方法做聚类分析(分别采用最短距离法和最长距离法), 给出树状图/谱系图, 并对聚类结果进行解释

变量 药物	吸入量	疗效	依赖性
速可眠	5	9	20
LSD	6	11	2
安定	4	5	20
吗啡	6	9	46
仙人球毒碱	5	7	1
酒精	3	1	12

2, 请采用聚类分析的方法, 对“污染数据(2014)”进行分类分析. 并解释各类的意义.