



数学建模

Mathematical Modeling

数学与统计学院数据科学系 郑斯斯

电子邮箱: zhengss@hzu.edu.cn

联系方式: 18814110413



聚类分析

- 聚类，就是将样本划分为由类似对象组成的多个类的过程。
- 聚类后，可以更准确在每个类中单独使用统计模型进行估计、分析或预测；也可以探究不同类之间的相关性和主要差异。
- 聚类和分类的区别：分类已知类别，聚类未知。



K-means 聚类算法

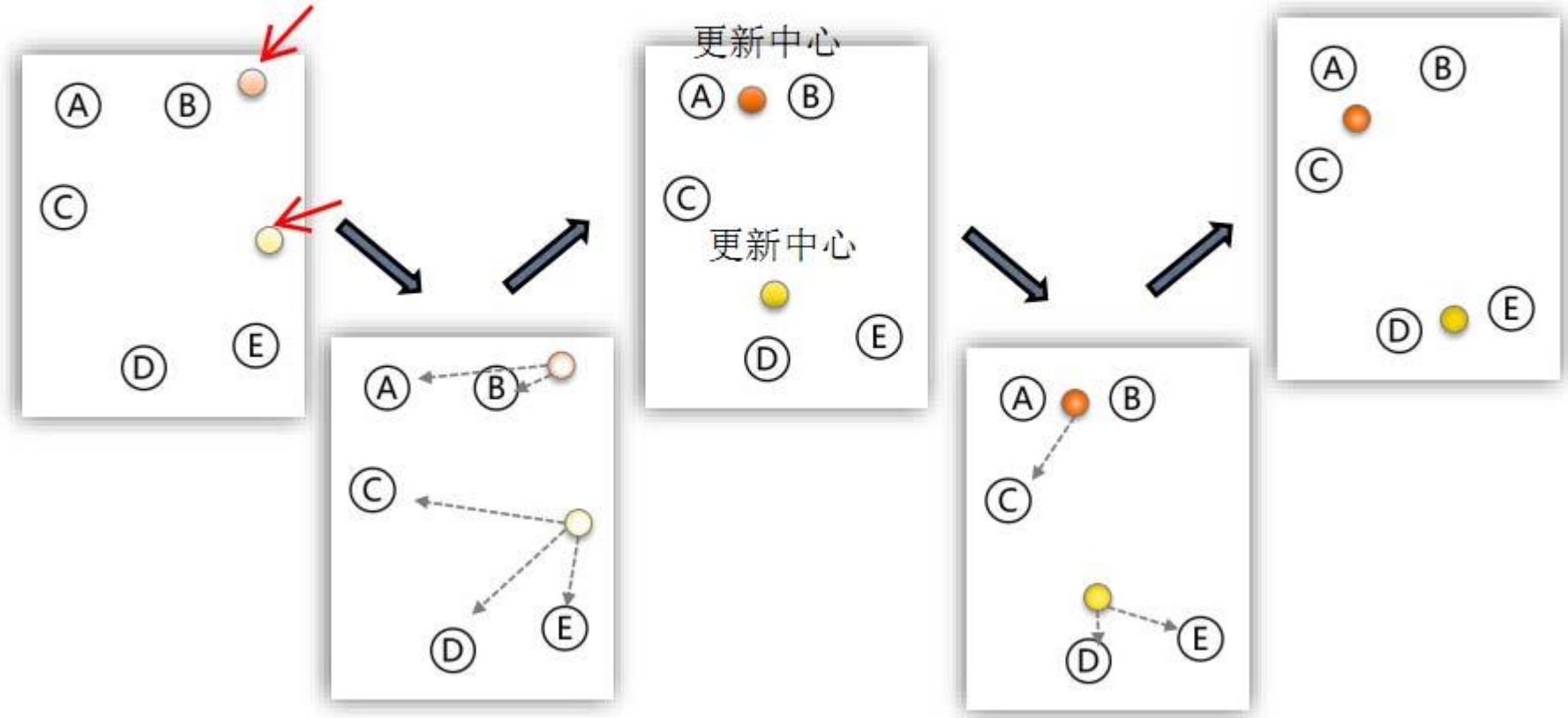
K-means 聚类的算法流程：

- 一、指定需要划分的簇的个数 K 值（类的个数）；
- 二、随机地选择 K 个数据对象作为初始的聚类中心（不一定是我们的样本点）；
- 三、计算其余的各个数据对象到这 K 个初始聚类中心的距离，把数据对象划归到距离它最近的那个中心所处所在的簇类中；
- 四、调整新类并且重新计算出新类的中心；
- 五、循环步骤三和四，看中心是否收敛（不变），如果收敛或达到迭代次数则停止循环；
- 六、结束。



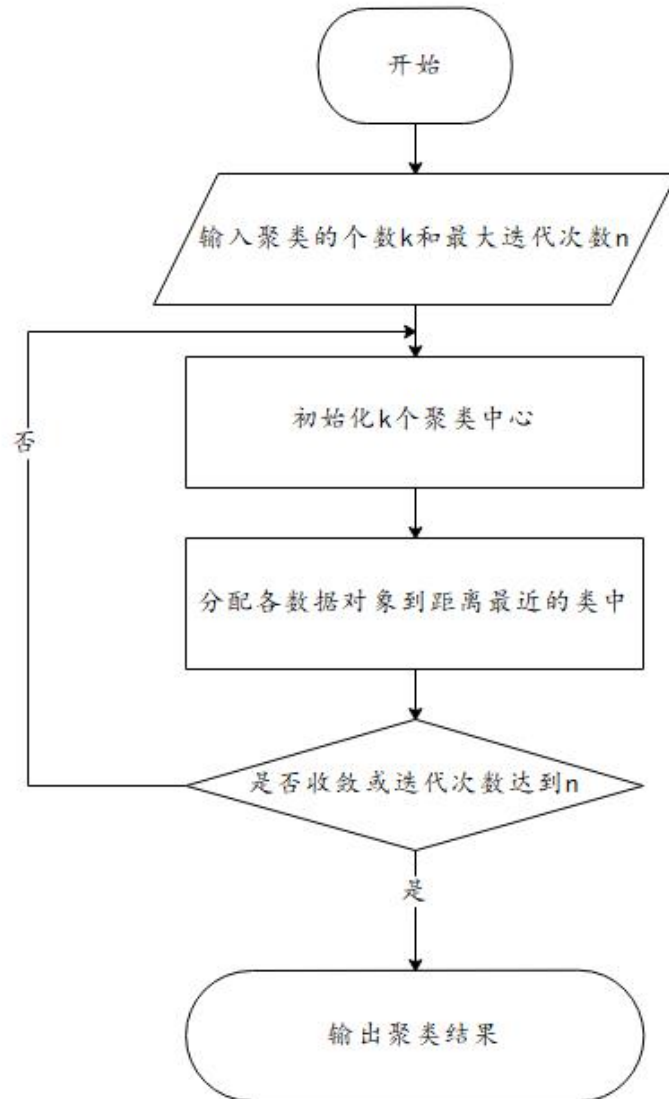
图解K-means算法

K-means聚类算法





算法流程图



亿图、
Visio、PPT
等软件都可
以画图。



K-means算法的评价

优点:

- (1) 算法简单、快速。
- (2) 对处理大数据集，该算法是相对高效率的。

缺点:

- (1) 要求用户必须事先给出要生成的簇的数目K。
- (2) 对初值敏感。
- (3) 对于孤立点数据敏感。

K - means++算法可解决2和3两个缺点。



K-means++ 算法

k-means++ 算法选择初始聚类中心的基本原则是：

初始的聚类中心之间的相互距离要尽可能的远。

只对K-means 算法“初始化K个聚类中心”这一步进行了优化。



K-means++ 算法

算法描述如下：

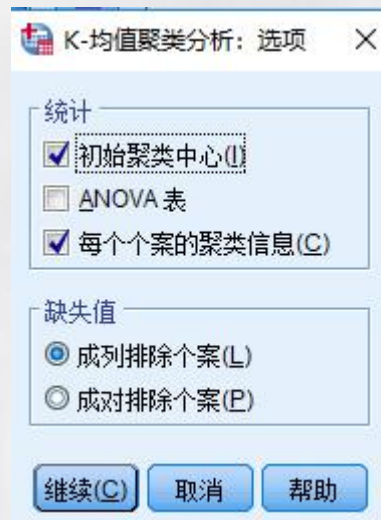
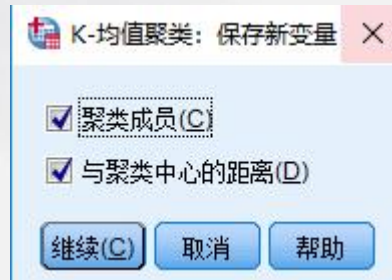
步骤一：随机选取一个样本作为第一个聚类中心；

步骤二：计算每个样本与当前已有聚类中心的最短距离（即与最近一个聚类中心的距离），这个值越大，表示被选取作为聚类中心的概率较大；最后，用轮盘法（依据概率大小来进行抽选）选出下一个聚类中心；

步骤三：重复步骤二，直到选出K个聚类中心。选出初始点后，就继续使用标准的K-means算法了。



SPSS软件操作



聚类成员			
个案号	城市	聚类	距离
1	北京	1	25.188
2	天津	2	6.856
3	河北	3	7.437
4	山西	3	16.033
5	内蒙古	3	3.107
6	辽宁	3	32.207
7	吉林	3	48.307
8	黑龙江	3	43.217
9	上海	1	29.578
10	江苏	2	45.684
11	浙江	2	25.136
12	安徽	3	44.677
13	福建	3	36.223
14	江西	3	35.947
15	山东	3	15.773
16	河南	3	30.747
17	湖南	3	30.383
18	湖北	3	13.547
19	广东	1	11.643
20	广西	3	22.203
21	海南	2	19.034
22	重庆	3	29.733
23	四川	3	27.393
24	贵州	3	4.587
25	云南	2	32.726
26	西藏	1	43.123
27	陕西	3	4.727
28	甘肃	3	32.123
29	青海	3	40.443
30	宁夏	3	.137
31	新疆	3	18.333



K-means 算法的一些讨论

(1) 聚类的个数K值怎么定？

答：分几类主要取决于个人的经验与感觉，通常的做法是多尝试几个K值，看分成几类的结果更好解释，更符合分析目的等。



K-means算法的一些讨论

(2) 数据的量纲不一致怎么办？

答：如果数据的量纲不一样，那么算距离时就没有意义。例如：如果X1单位是米，X2单位是吨，用距离公式计算就会出现“米的平方”加上“吨的平方”再开平方，最后算出的东西没有数学意义，这就是问题。

$$z_i = \frac{x_i - \bar{x}}{\sigma_x} \text{ (先减去均值再除以标准差)}$$



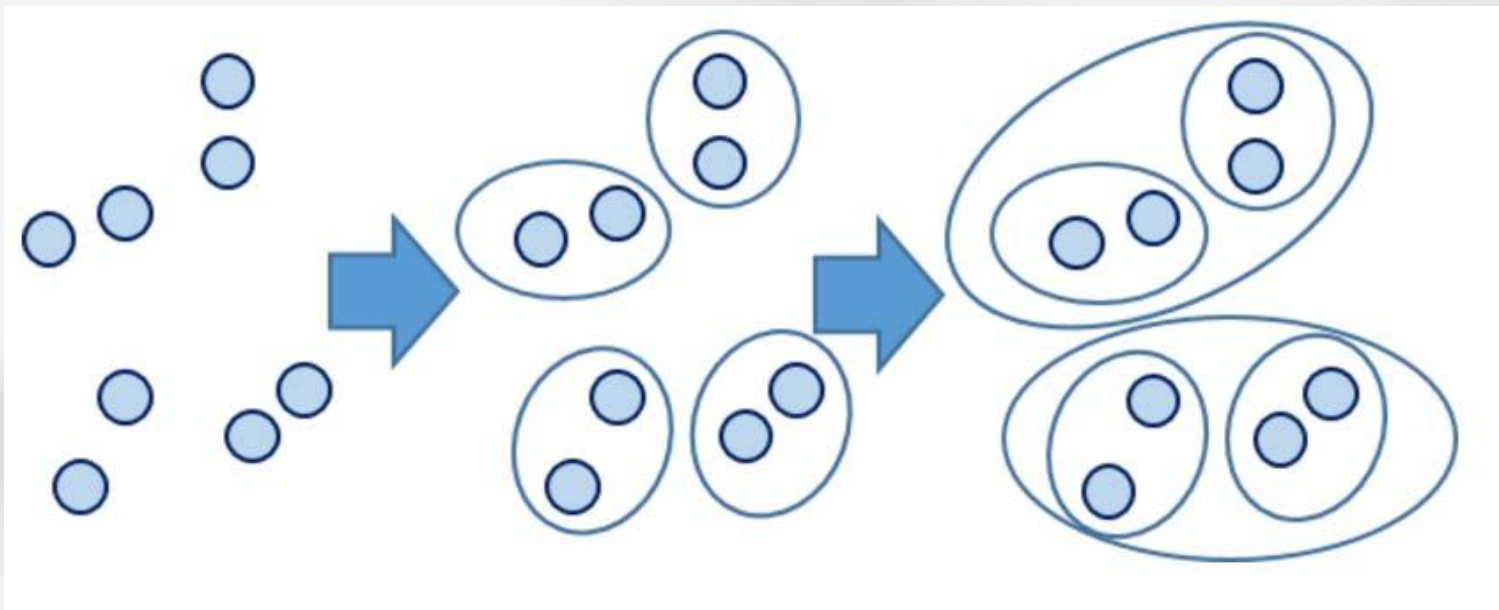
2021-9-3





系统（层次）聚类

系统聚类的合并算法通过计算两类数据点间的距离，对最为接近的两类数据点进行组合，并反复迭代这一过程，直到将所有数据点合成一类，并生成聚类谱系图。





系统（层次）聚类

参考：

于晶贤 - 辽宁石油化工大学 - 聚类分析之系统聚类法.pdf

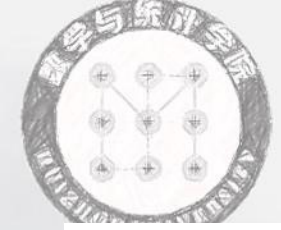


系统（层次）聚类算法流程

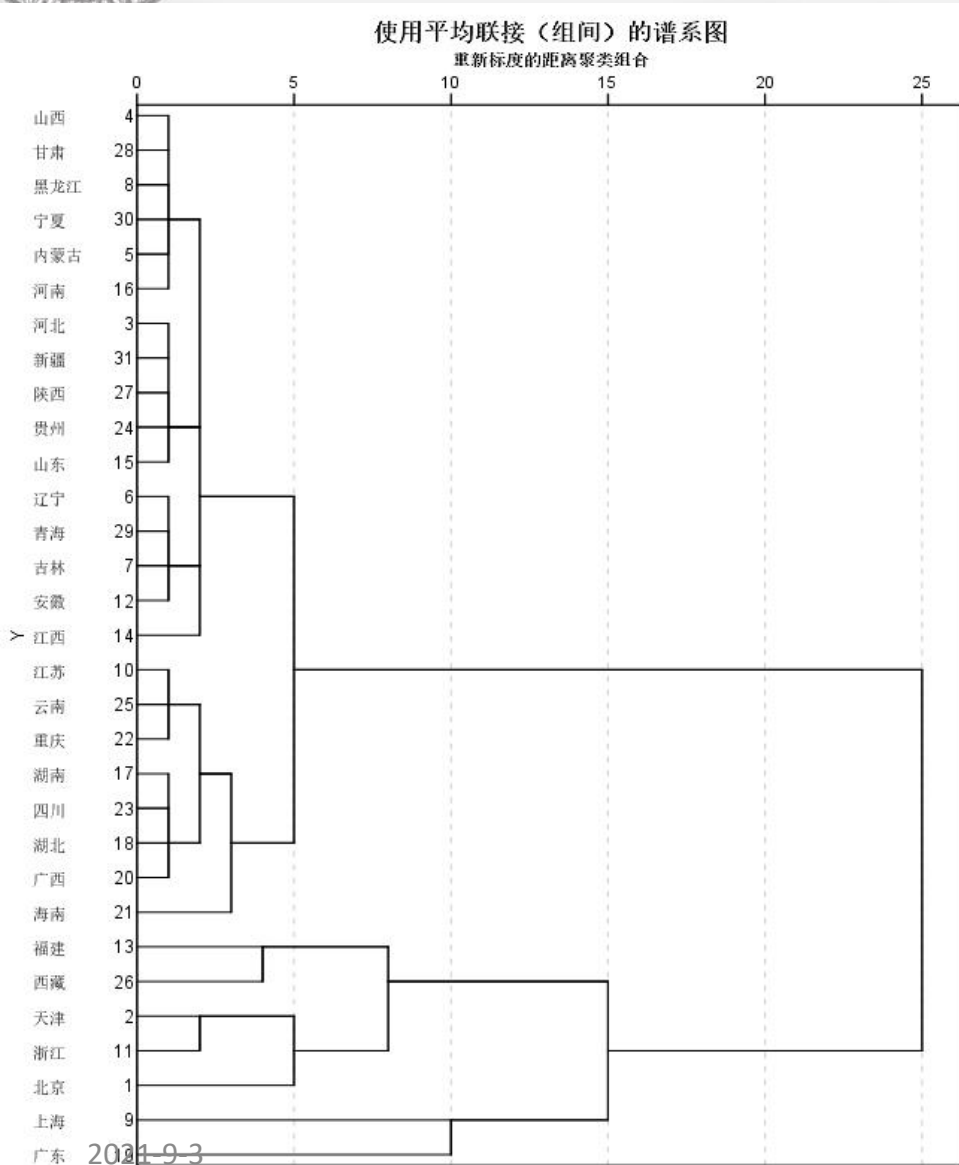
系统（层次）聚类的算法流程：

- 一、将每个对象看作一类，计算两两之间的最小距离；
- 二、将距离最小的两个类合并成一个新类；
- 三、重新计算新类与所有类之间的距离；
- 四、重复二三步，直到所有类最后合并成一类；
- 五、结束。

课后作业：将上述文字表述的流程绘制成一个流程图。



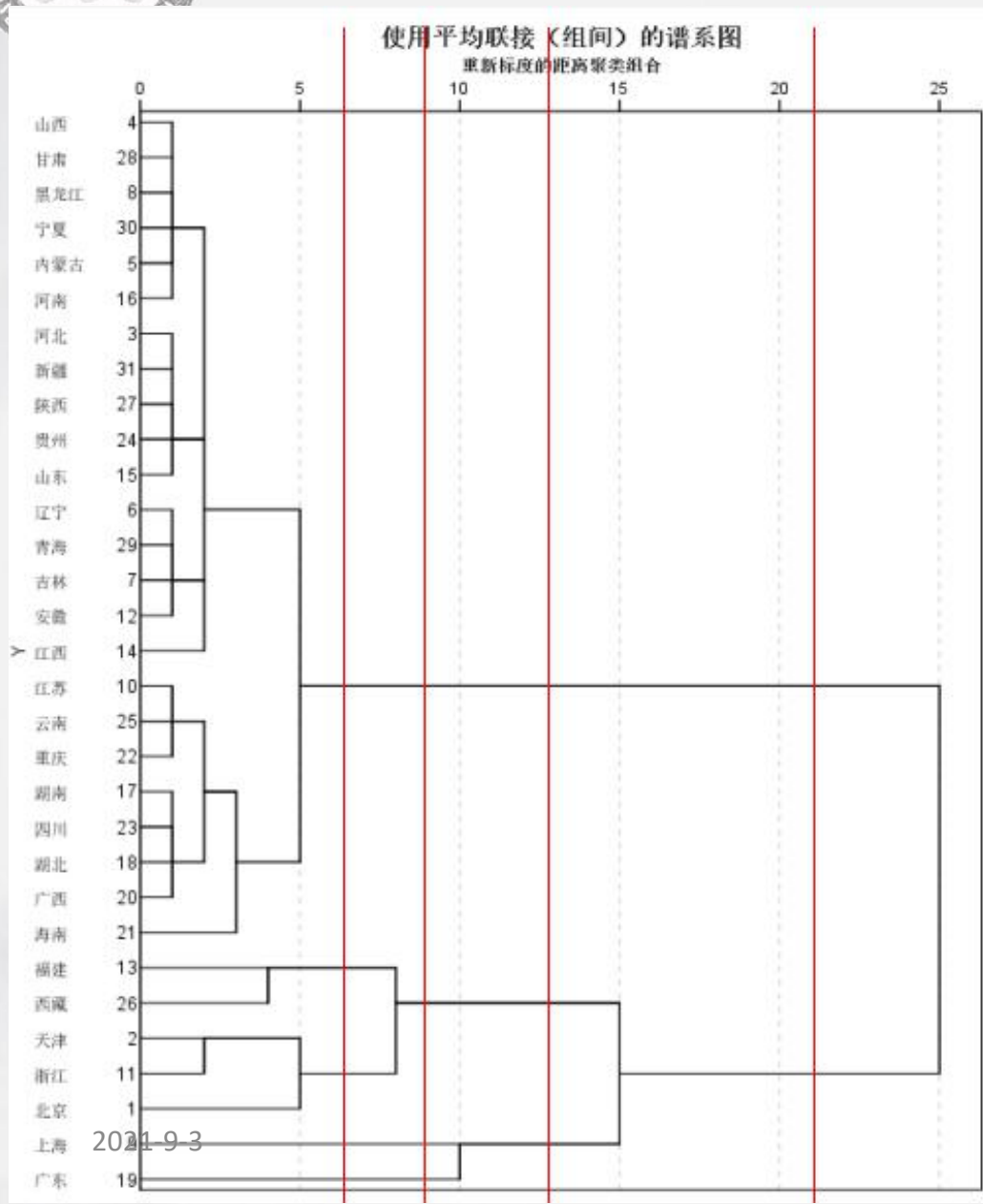
聚类谱系图（树状图）



- 谱系图是SPSS新版本才有的功能
- 横轴表示各类之间的距离（该距离经过重新标度）
- 聚类个数可以自己从图中决定



聚类谱系图（树状图）



- 谱系图是SPSS新版本才有的功能
- 横轴表示各类之间的距离（该距离经过重新标度）
- 聚类个数可以自己从图中决定



用图形估计聚类的数量

肘部法则：通过图形大致的估计出最优的聚类数量。

各个类畸变程度之和：各个类的畸变程度等于该类重心与其内部成员位置距离的平方和；

假设一共将 n 个样本划分到 K 个类中（ $K \leq n - 1$, 即至少有一类中有两个元素）

用 C_k 表示第 k 个类（ $k = 1, 2, \dots, K$ ），且该类重心的位置记为 u_k

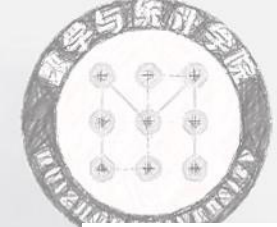
那么第 k 个类的畸变程度为：
$$\sum_{i \in C_k} |x_i - u_k|^2$$

（这里的绝对值符号的意义表示的是距离，可视为一种广义的记号）

定义所有类的总畸变程度：
$$J = \sum_{k=1}^K \sum_{i \in C_k} |x_i - u_k|^2$$

在部分多元统计教材中， J 又被称为聚合系数。

聚合系数折线图：横坐标为聚类的类别数 K ，纵坐标为聚合系数 J 。



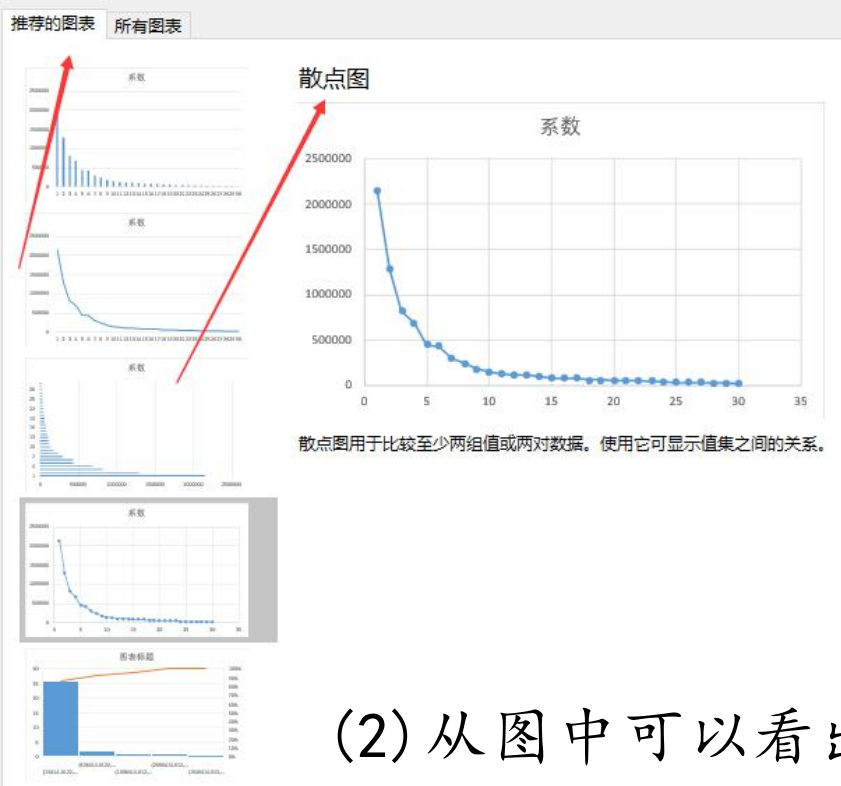
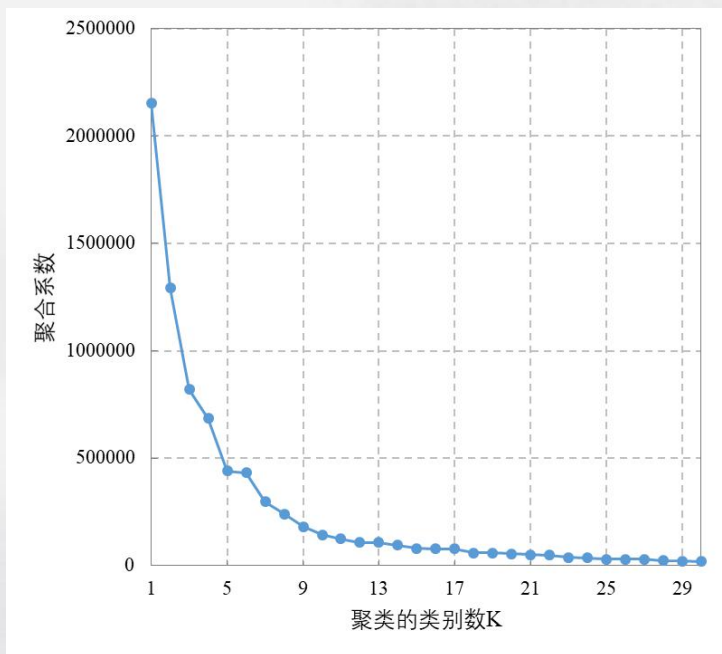
画图前先对数据进行处理

集中计划						
阶段	组合聚类		系数	首次出现聚类的阶段		下一个阶段
	聚类 1	聚类 2		聚类 1	聚类 2	
1	4	28	18414.012	0	0	7
2	8	30	20297.111	0	0	5
3	3	31	22161.800	0	0	10
4	6	29	28190.529	0	0	8
5	5	8	29537.326	0	2	7
6	17	23	29624.380	0	0	11
7	4	5	34367.611	1	5	9
8	6	7	38093.577	4	0	16
9	4	16	48752.992	7	0	20
10	3	27	49242.893	3	0	13
11	17	18	53983.761	6	0	17
12	10	25	57922.410	0	0	14
13	3	24	59288.905	10	0	15
14	10	22	77201.747	12	0	22
15	3	15	77370.143	13	0	19
16	6	12	79172.059	8	0	19
17	17	20	94389.183	11	0	22
18	2	11	107280.103	0	0	26
19	3	6	107957.125	15	16	20
20	3	4	124430.310	19	9	21
21	3	14	142727.072	20	0	25
22	10	17	180442.145	14	17	23
23	10	21	239702.036	22	0	25
24	13	26	296564.664	0	0	27
25	3	10	431624.250	21	23	30
26	1	2	440027.064	0	18	27
27	1	13	683977.679	26	24	29
28	9	19	819519.221	0	0	29
29	1	9	1291671.774	27	28	30
30	1	3	2152790.747	29	25	0

	A
1	系数
2	2152791
3	1291672
4	819519.2
5	683977.7
6	440027.1
7	431624.2
8	296564.7
9	239702
10	180442.1
11	142727.1
12	124430.3
13	107957.1
14	107280.1
15	94389.18
16	79172.06
17	77370.14
18	77201.75
19	59288.9
20	57922.41
21	53983.76
22	49242.89
23	48752.99
24	38093.58
25	34367.61
26	29624.38
27	29537.33
28	28190.53
29	22161.8
30	20297.11
31	18414.01

把数据粘贴到Excel
表格中，并按照降
序排好。

聚合系数折线图的画法

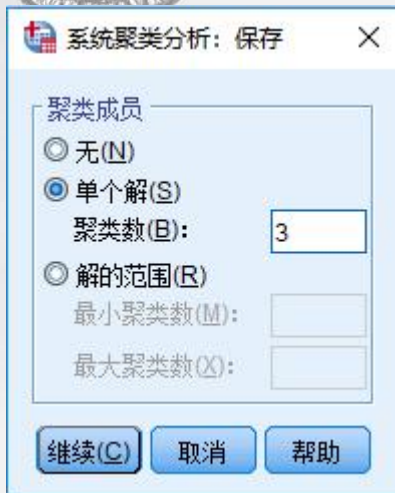


根据图来进行解释:

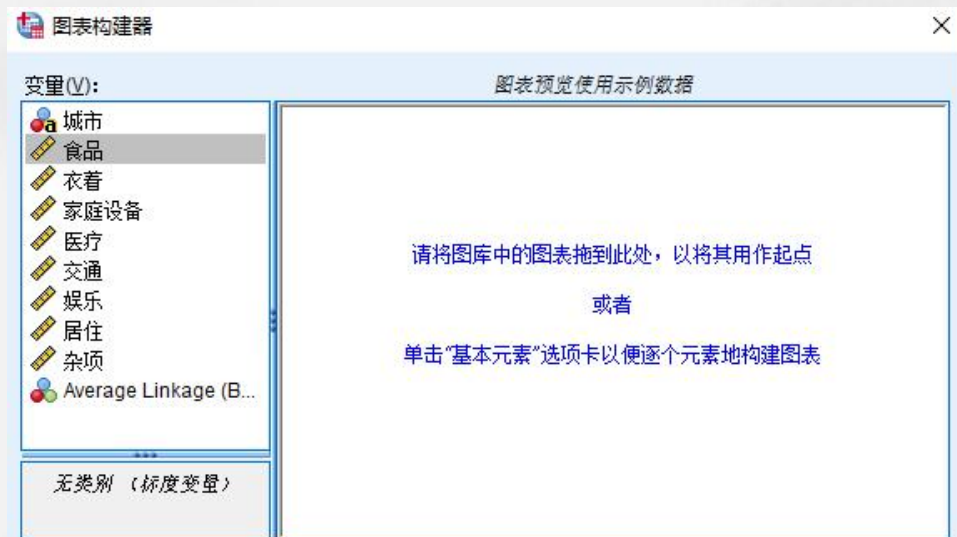
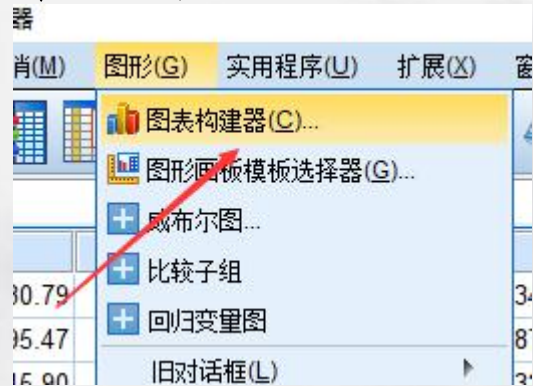
(1) 根据聚合系数折线图可知, 当类别数为5时, 折线的下降趋势趋缓, 故可将类别数设定为5。

(2) 从图中可以看出, K值从1到5时, 畸变程度变化最大。超过5以后, 畸变程度变化显著降低。因此肘部就是K=5, 故可将类别数设定为5。(当然, K=3也可以解释)

确定K后保存聚类结果并画图



重新聚类一次，确定
聚类个数

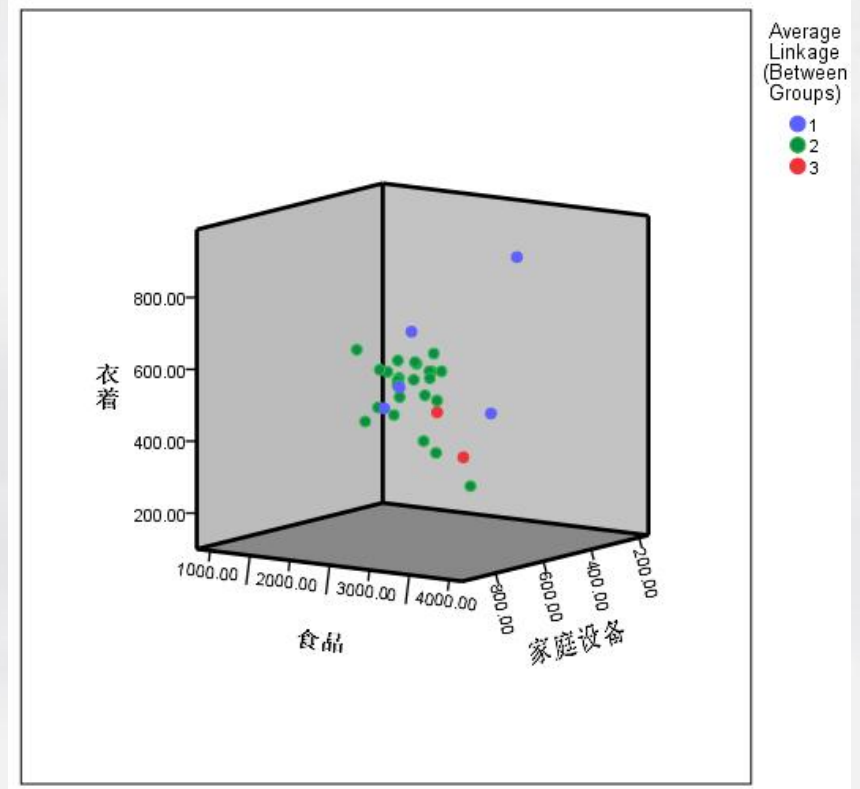
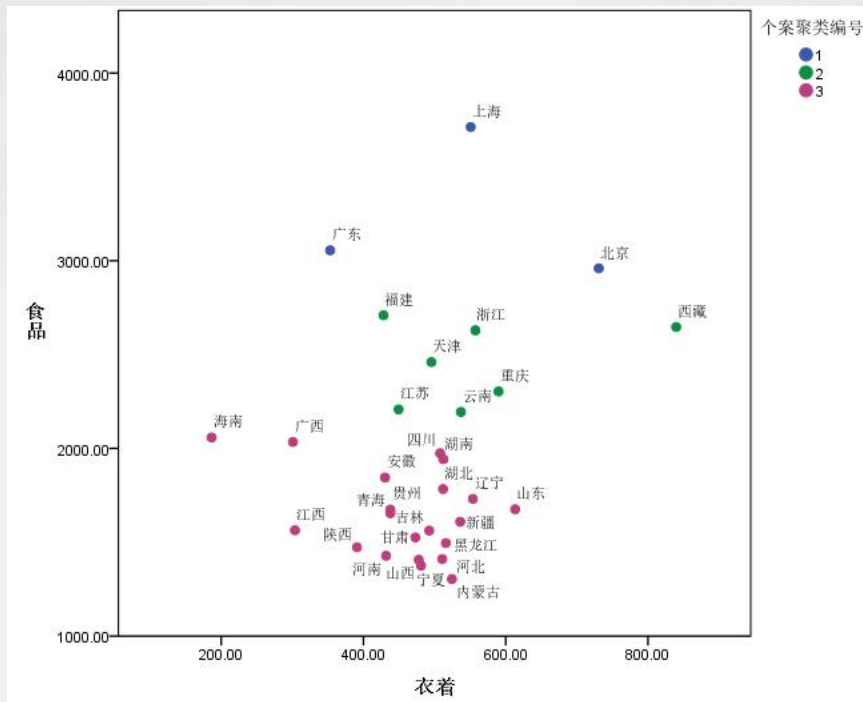


聚类结束后画图

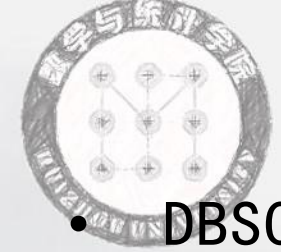


示意图

双击图中的任意元素，可对其进行调整。

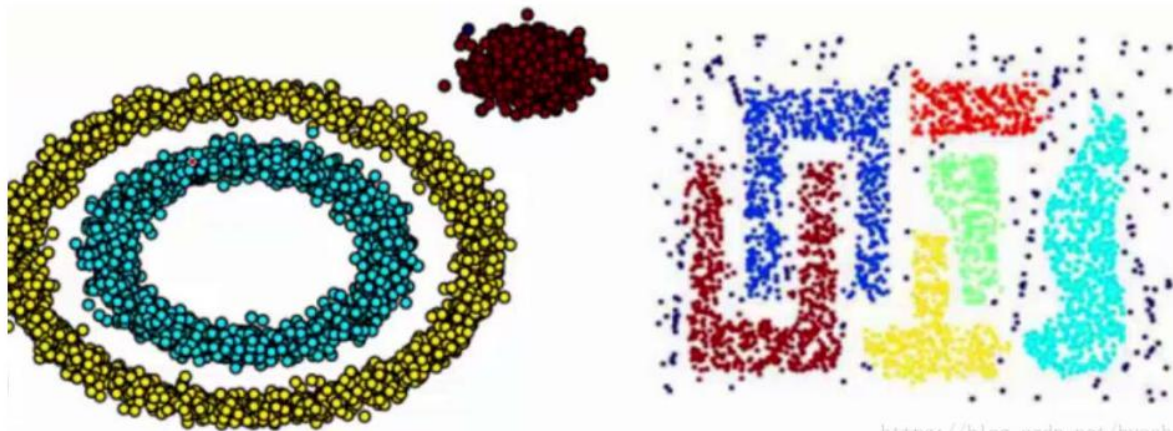


注意：只要当指标个数为2或者3的时候才能画图，上面两个图纯粹是为了演示作图过程，实际上本例中指标个数有8个，是不可能做出这样的图的。



DBSCAN算法

- DBSCAN是一种基于密度的聚类方法，聚类前不需要预先指定聚类的个数，生成的簇个数不定（和数据有关）。
- 算法利用基于密度的聚类概念，即要求聚类空间中的一定区域内所包含对象（点或其他空间对象）的数目不小于某一给定阈值。
- 该方法能在具有噪声的空间数据库中发现任意形状的簇，可将密度足够大的相邻区域连接，能有效处理异常数据。



<https://blog.csdn.net/huacha>

DBSCAN: 具有噪声的基于密度的聚类方法



基本概念

DBSCAN算法将数据点分为三类：

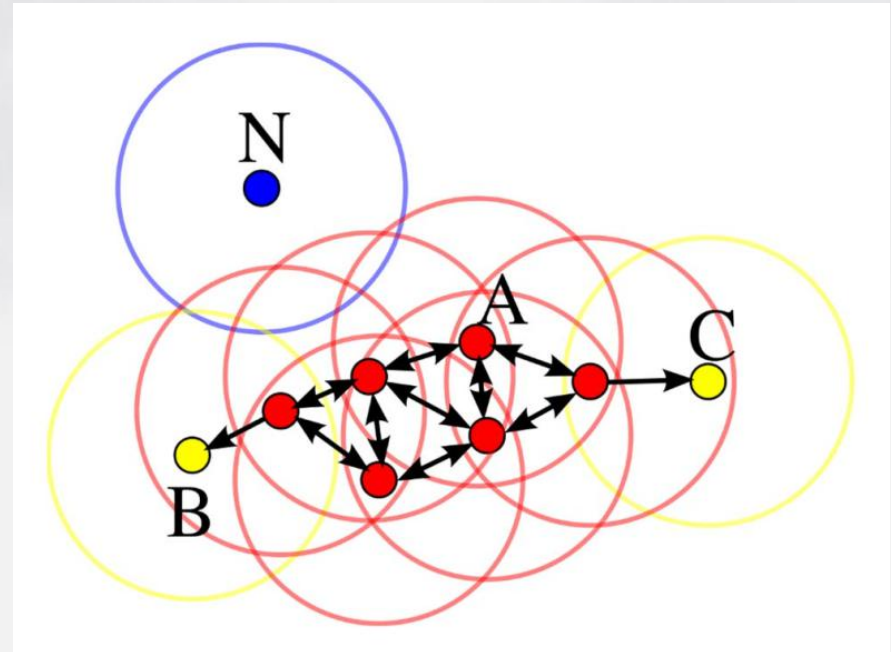
- 核心点：在半径 Eps 内含有不少于 $MinPts$ 数目的点；
- 边界点：在半径 Eps 内点的数量小于 $MinPts$ ，但是落在核心点的邻域内；
- 噪音点：既不是核心点也不是边界点的点。



基本概念

如图， $\text{MinPts} = 4$

- 点A和其它红色点是核心点，因为它们的 ε -邻域（图中红色圆圈）里包含最少4个点（包括自己），由于它们之间相互可达，它们形成了一个聚类。
- 点B和点C不是核心点，但可由A经其它核心点可达，所以和A属于同一个聚类。
- 点N是局外点，它既不是核心点，又不由其它点可达。





DBSCAN算法的评价

优点:

- 基于密度定义，能处理任意形状和大小的簇；
- 可在聚类的同时发现异常点；
- 与K-means比较起来，不需要输入要划分聚类个数。

缺点:

- 对输入参数 ϵ 和Minpts敏感，确定参数困难；
- DBSCAN算法中变量 ϵ 和Minpts是全局唯一的，当聚类的密度不均匀时，聚类距离相差很大时，聚类质量差；
- 当数据量大时，计算密度单元的计算复杂度大。



聚类算法的选取

建议：

- 只有两个指标，且做出散点图后发现数据表现得有密集的特点，这时候再用DBSCAN进行聚类；
- 其它情况下，全部使用系统聚类；
- K - means也可以用，不过用了的话论文上可写的东西比较少。



课后作业

(1) 用亿图、PPT或者Visio等软件中的一种画出系统聚类的流程图；

(2) 完成小作业：“各国森林、草原资源聚类.doc”。

提示：指标共三个，量纲不同，注意去除量纲的影响，论文中要交代清楚K的选择，当然也可以画出聚类结果的三维图。



THE END!

数学与统计学院数据科学系 郑斯斯

电子邮箱: zhengss@hzu.edu.cn

联系方式: 18814110413