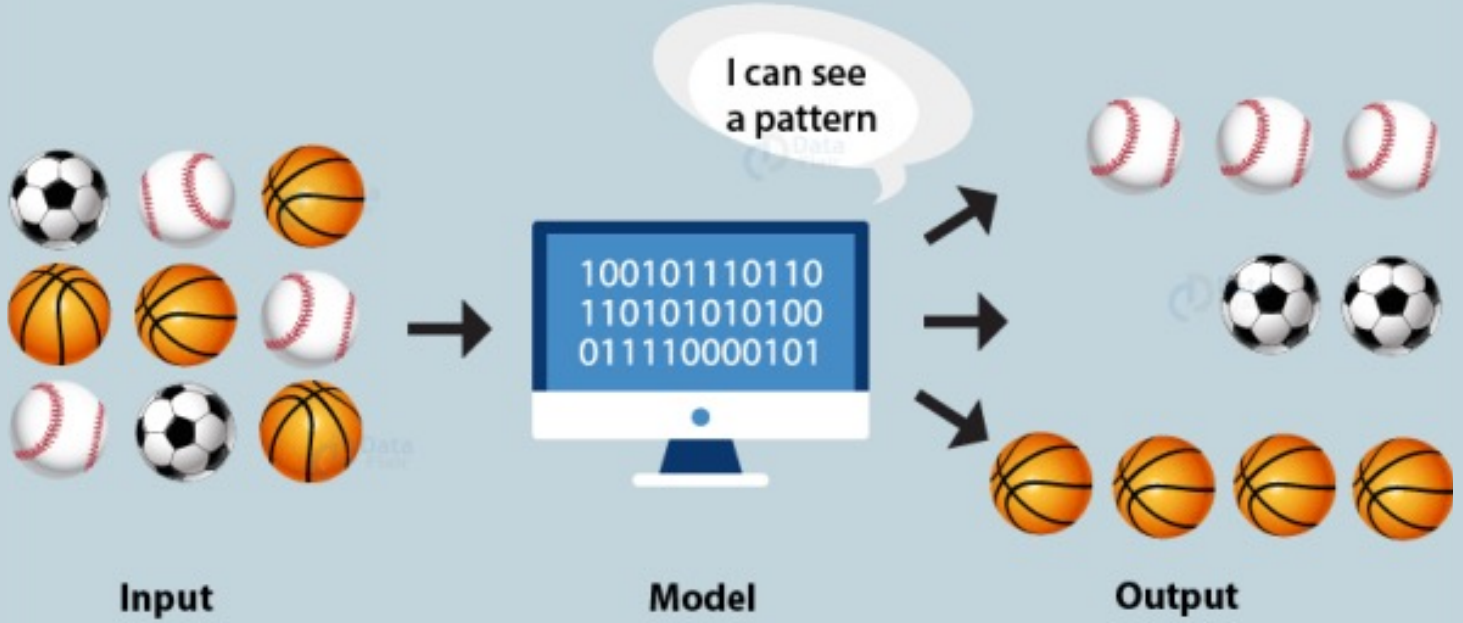


聚类分析

城市分析方法系列课程

苏州大学 王灿

大纲

- 聚类分析概述
- 层次聚类法
- K-means聚类法

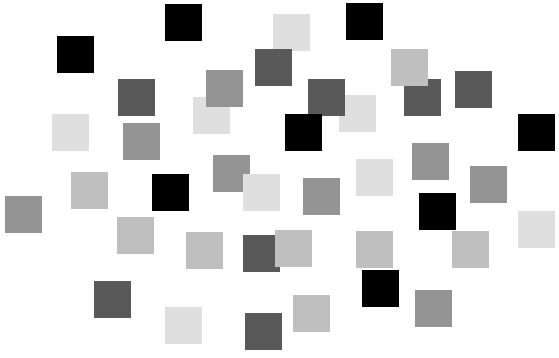
聚类分析概述

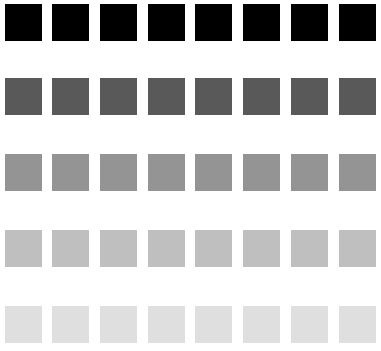
物以类聚，人以群分

birds of a feather flock together

聚类分析概述

庞大、复杂的数据 ⟶ ? ⟶ **有限**的类别/模式





聚类分析（clustering analysis），根据数据自身的**内在结构**特征，寻找相似元素的集合，将样本划分为不同的类别（或称子集、簇，cluster），使同一**类别内部**的相似性尽可能**高**，**类别之间**的相似性尽可能**低**。

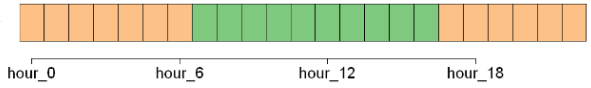
聚类分析概述

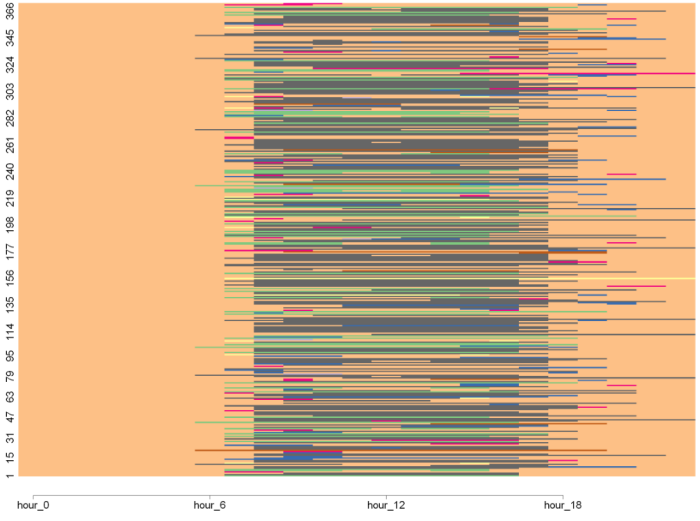
学校

家

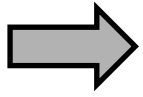
工作

居民的一日活动链

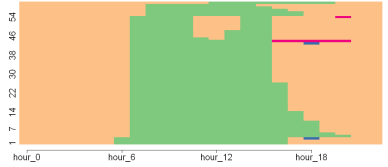




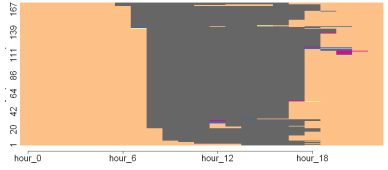
原始数据



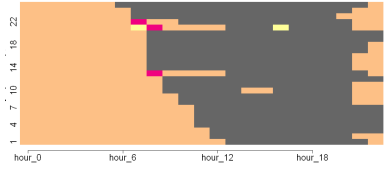
class1: 上学族



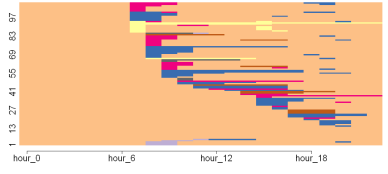
class2: 上班族



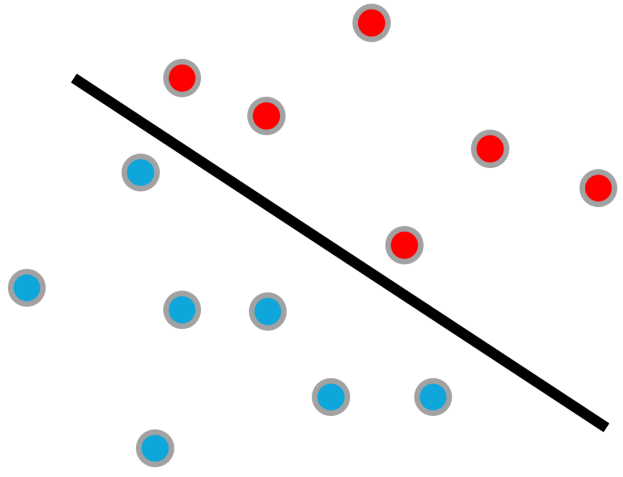
class3: 加班族



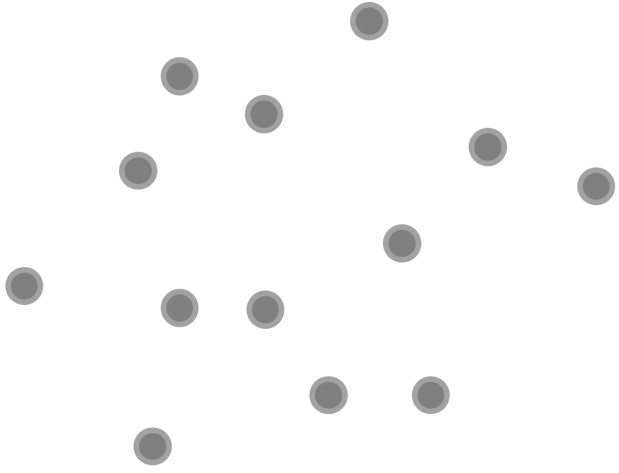
class4: 宅家族



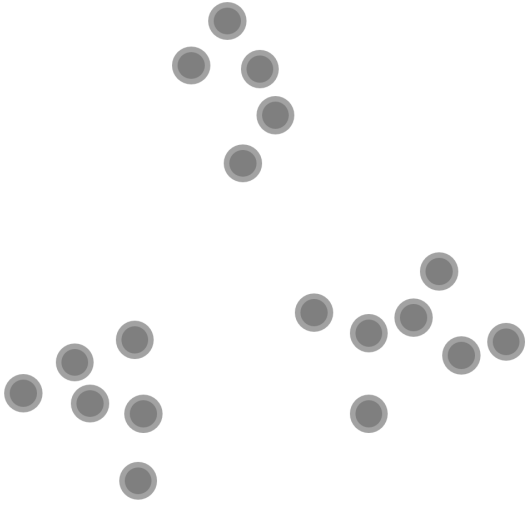
聚类分析概述



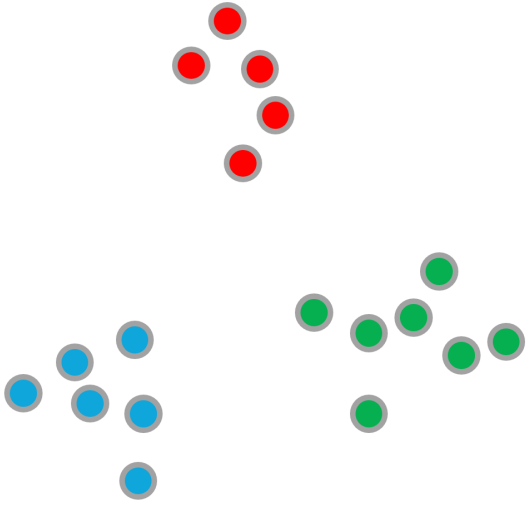
聚类分析概述



聚类分析概述

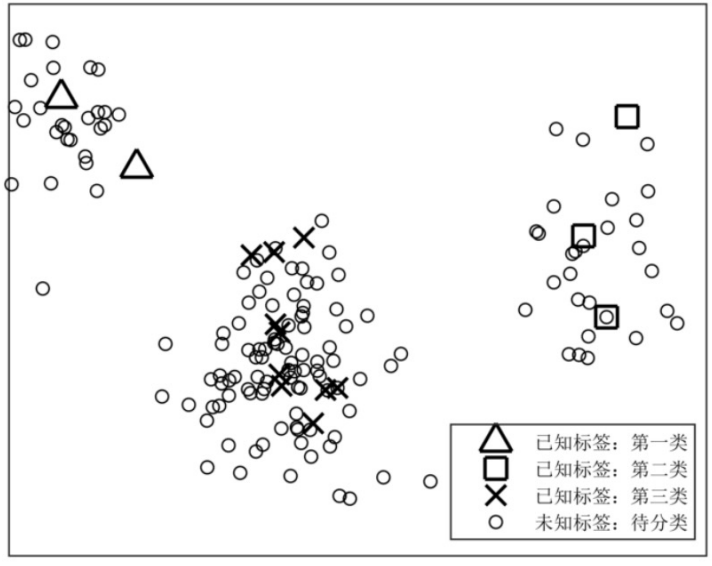


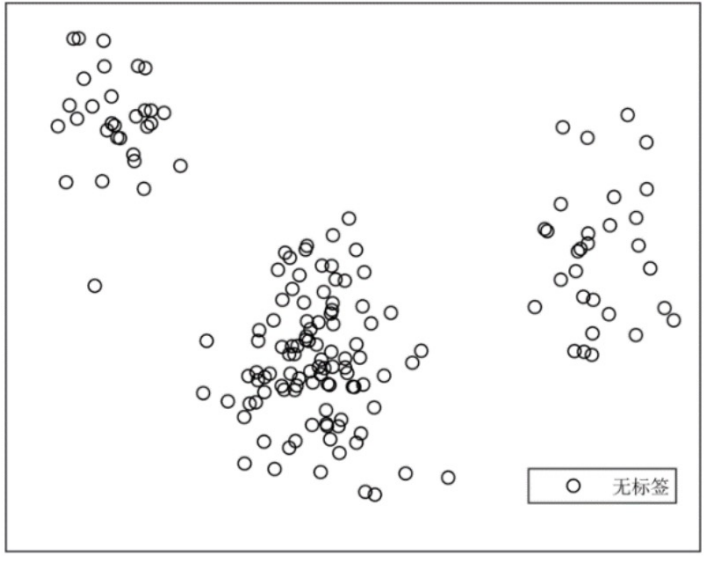
聚类分析概述



聚类分析概述

分类（classification） **vs** **聚类**（clustering）



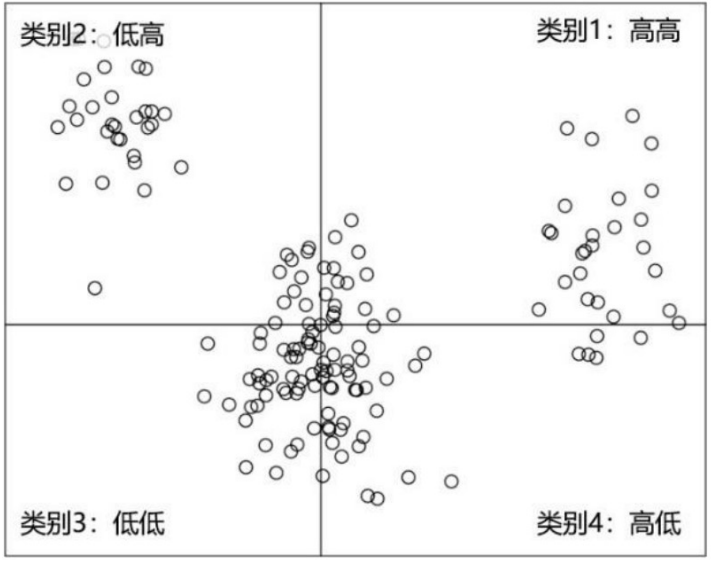


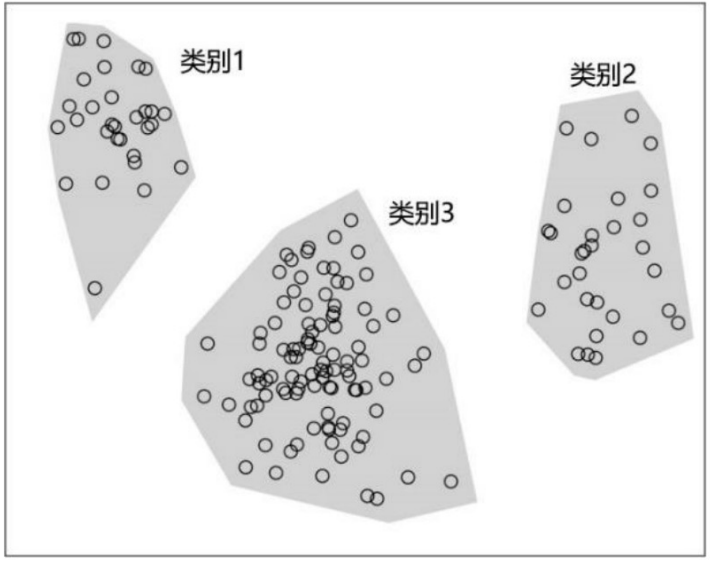
- 属监督学习（supervised learning）
- 预先知道所有可能类别，及部分样本的类别。
- 算法从已经打上标签的样本中，学习“数据$x \rightarrow$ 标签y”的关系，从而判断无标签的样本属于哪一类。

- 属无监督学习（unsupervised learning）
- 预先不知道有多少类别，也不知道任何一个样本属于哪一类别。
- 算法从无标签数据中，自动学习类别模式。

聚类分析概述

经验分类 vs 聚类分析





- 根据个人经验，进行人工分类。
- 当变量多时，经验分类难度大。

- 根据数据结构，进行自动分类。
- 当变量多时，聚类分析更容易。
- 类别的意义并非不言自明，需合理解读。

聚类分析概述

特点

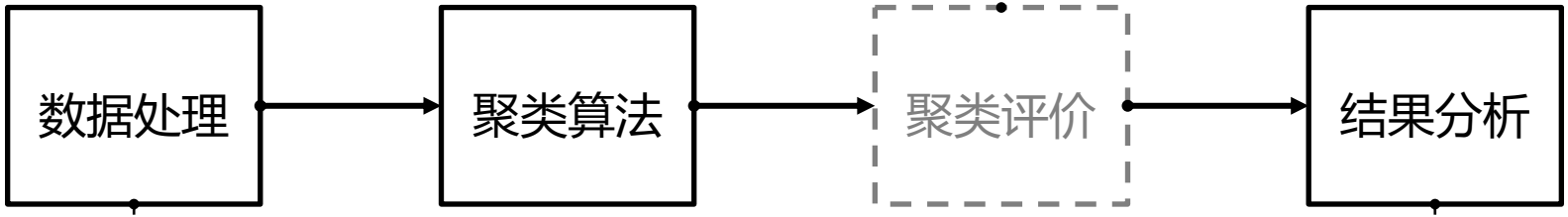
- 是一种探索性很强的统计分析方法。
- 无过多理论支持，完全依靠数据自身寻找内在的自然分组。
- 无论是否存在类别，都会得到一个类别划分方案。
- 对于同样的数据，使用不同的聚类算法，可能得到不同的分类方案。

结果检验

- 很难从理论上，论证聚类结果的正确性。
- 更多从应用上，依据类别的可解释性、类别之间的差异性、对理解问题的有用性等方面，判断聚类结果的好坏。

聚类分析概述

基本步骤



多大程度上满足了“类内尽可能相似、类间尽可能不同”的要求

对于结构化数据，确定用于聚类的多个变量，按需将各变量**标准化**，以防止量纲大的变量决定性过高。

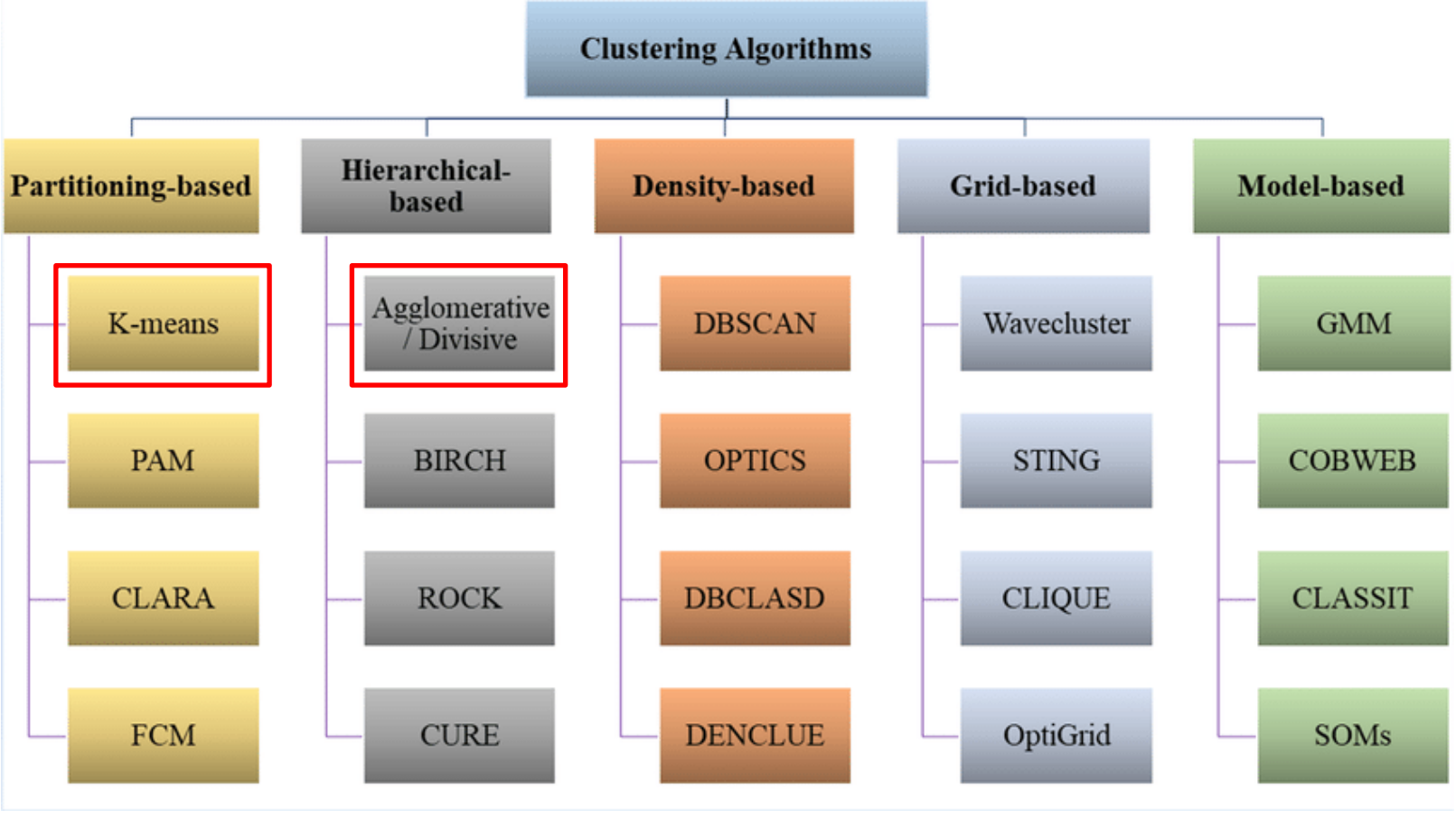
- 方法1：中心标准化。对于均值为μ，标准差为θ的数据x，中心标准化后的数据为$z = (x-\mu) / \theta$ 。
- 方法2：将数据的取值范围变为 0~1。

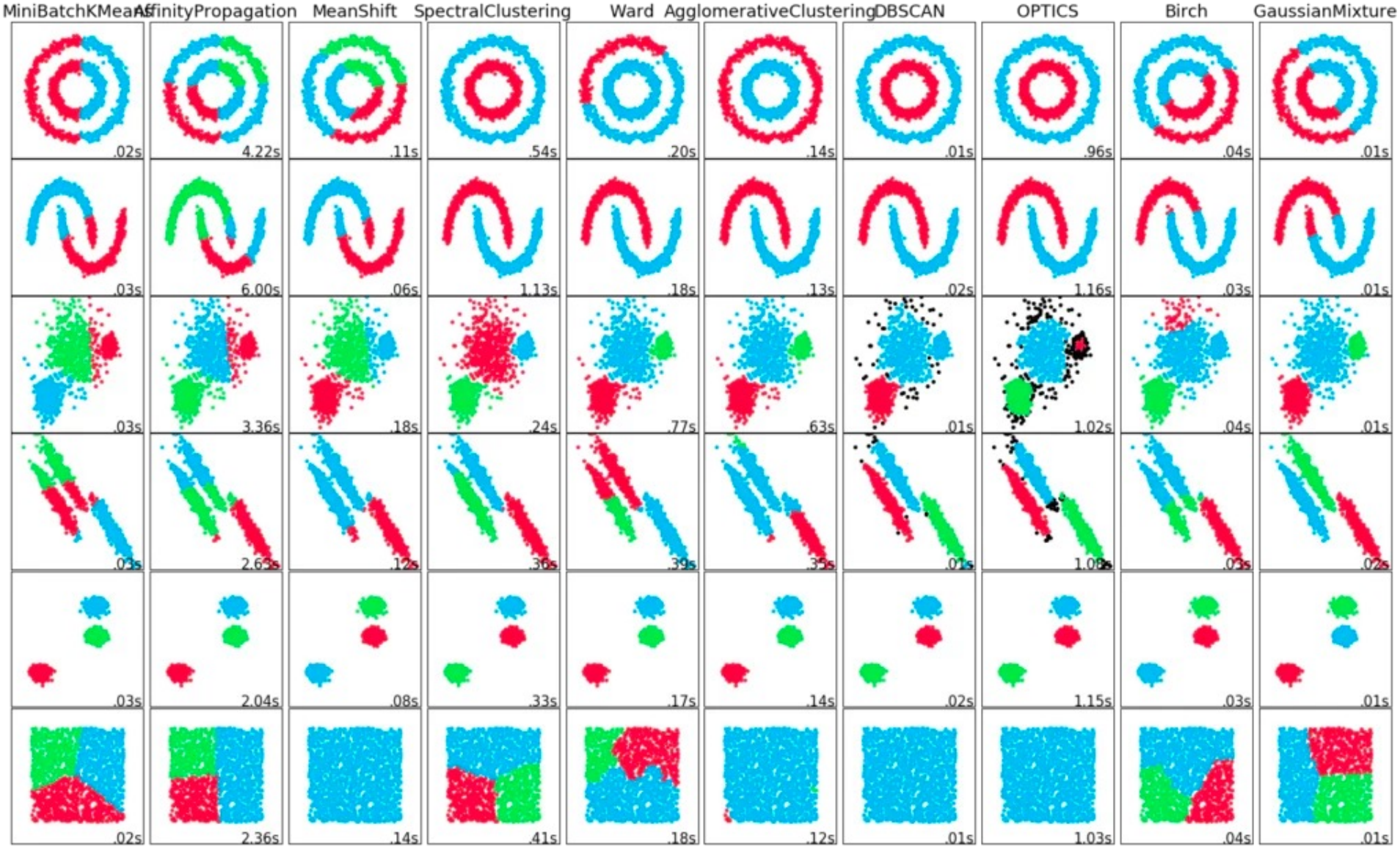
对于**非结构化数据**，一般需要更复杂的预处理。

分析不同类别的差异化特征，理解数据结构，进一步解决更复杂的问题。

聚类分析概述

聚类分析算法




MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
OPTICS
Birch
GaussianMixture
.02s
4.22s
.11s
.54s
.20s
.14s
.01s
.96s
.04s
.01s
.03s
6.00s
.06s
1.13s
.18s
.13s
.02s
1.16s
.03s
.01s
.03s
3.36s
.18s
.24s
.77s
.63s
.01s
1.02s
.04s
.01s
.03s
2.63s
.12s
.36s
.39s
.35s
.01s
1.08s
.03s
.02s
.03s
2.04s
.08s
.33s
.17s
.14s
.02s
1.15s
.03s
.01s
.02s
2.36s
.14s
.41s
.18s
.12s
.01s
1.03s
.04s
.01s

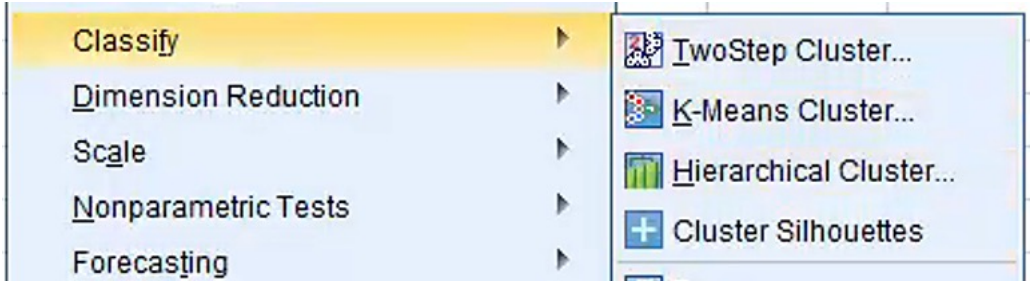
聚类分析概述

软件

层次聚类法
K-means聚类法 ┐ **SPSS**、Stata、Matlab、Python、R 等

两步聚类法 — **SPSS**

DBACAN — **ArcGIS**、Python、R、Matlab等

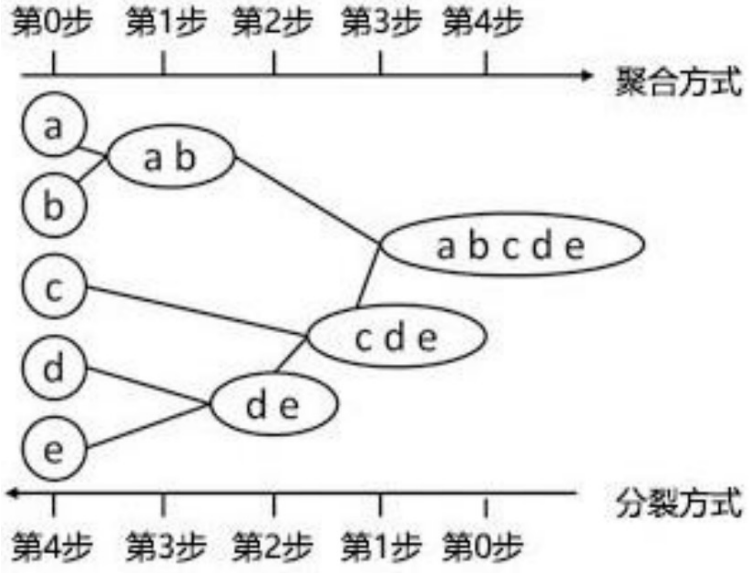


大纲

- 聚类分析概述
- **层次聚类法**
- K-means聚类法

层次聚类法

层次聚类法：通过逐步**聚合**或分裂数据点，形成一个由不同层次组成的树状结构。



聚合方式

① 每个样本分别视为一个独立的小类；

② 根据距离或相似性不断合并，每次将距离最近、相似性最高的两类合并为一个大类；

③ 直到所有样本属于一类。

分裂方式

① 所有样本视为一个大类；

② 根据距离不断分裂；

③ 直到每个样本是一个独立的小类。

层次聚类法

数据要求：全部为数值型，或全部为分类型。

关键：距离/相似性的测度。

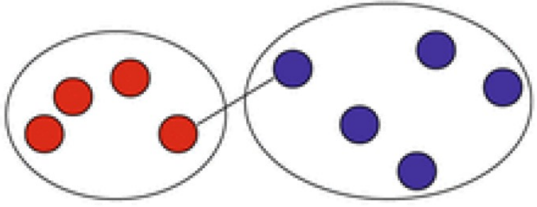
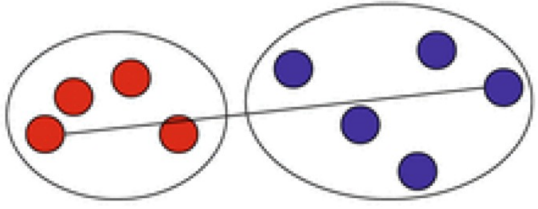
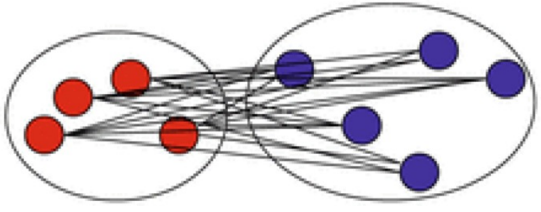
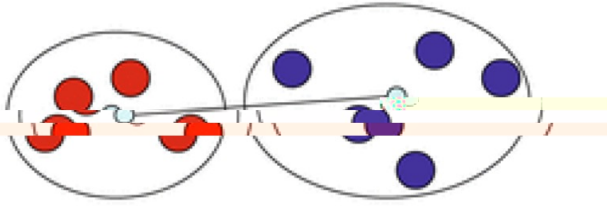
对于**样本点**之间的距离：

- 当全部为数值型变量时，最常用欧氏距离；
- 当全部为分类型变量时，可以改用卡方距离。

对于**类别**之间的距离：

- 不同的测度方法可能对结果影响很大；
- 组间平均距离法受异常点的影响较小，是较稳健的选择；
- 其他方法各有特质，适合特定结构的数据。

层次聚类法

方法	特征	图示
最短距离法(nearest neighbor，或称 single linkage)	以最近的一对样本点之间的距离代表两个类别之间的距离	
最长距离法(furthest neighbor，或称 complete linkage)	以最远的一对样本点之间的距离代表两个类别之间的距离	
组间平均距离法 (between-groups linkage)	以每一对样本点之间的距离的平均值代表两个类别之间的距离	
质心法 (centroids)	分别计算两个类别的质心（平均位置），以重心之间的距离代表类别之间的距离	

层次聚类法

数据：

以京沪高铁沿线21个设站城市为对象，收集了2006年和2012年高铁站周边 2km、4km、8km 三个圈层内的建成区面积，并分别计算了建成区面积占圈层圆环面积的比重。

问题：

发现这些城市在高铁站点周边**建成区扩张**的过程中有什么**模式**。

层次聚类法案例

编号	城市	2006年建成区面积比重（%）			2012年建成区面积比重（%）		
		2km圈层	4km圈层	8km圈层	2km圈层	4km圈层	8km圈层
1	北京	100.00	100.00	100.00	100.00	100.00	100.00
2	南京	100.00	94.99	81.04	100.00	97.91	92.33
3	上海	85.20	82.76	65.57	100.00	98.96	91.62
4	苏州	54.35	28.66	19.40	73.97	37.69	21.52
……	……	……	……	……	……	……	……
14	德州	0.00	0.00	3.25	0.00	0.00	7.50
15	曲阜	0.00	0.39	7.84	0.00	0.61	11.42
……	……	……	……	……	……	……	……
19	丹阳	0.00	0.31	0.86	0.00	0.21	15.86
20	滕州	2.39	5.65	11.18	12.15	6.95	15.64
21	镇江	13.68	18.00	24.44	44.83	38.84	35.21

层次聚类法案例

方法：

- 以**层次聚类法**，对表中6个变量（=2个年份*3个圈层）进行分析。
- 通过**聚合**方式，把21个城市依次合并，每次合并距离最近的2个城市，直至全部合为一类。
- 距离指标，两个城市之间采用**欧氏距离的平方**；

  两个类别之间采用**最长距离法**（complete linkage），

  即以最远一对城市之间的距离代表类别之间的距离。
- 因数据均为百分比，量纲基本相同，为保留原始数据信息，**未进行标准化。**

层次聚类过程

将每个城市视为一个独立类别，找到欧氏距离最近的两个城市。

- 以南京—上海为例：

编号	城市	2006年建成区面积比重（%）			2012年建成区面积比重（%）		
		2km圈层	4km圈层	8km圈层	2km圈层	4km圈层	8km圈层
2	南京	100.00	94.99	81.04	100.00	97.91	92.33
3	上海	85.20	82.76	65.57	100.00	98.96	91.62

欧氏距离的平方为：

$$d^2_{NJ\leftrightarrow SH} = (100-85.20)^2+(94.99-82.76)^2+(81.04-65.57)^2+(100-100)^2$$
$$+(97.91-98.96)^2+(92.33-91.62)^2$$
$$= 609.54$$

层次聚类过程

- 同理，得到每两个城市之间的距离矩阵（21×21）：

	北京	南京	上海	……	德州	曲阜	……	丹阳	滕州	镇江
北京	0.00									
南京	447.78	0.00								
上海	1772.99	609.54	0.00							
……	……	……	……	0.00						
德州	57916.81	51856.88	44861.30	……	0.00					
曲阜	56140.41	50321.12	43481.37	……	36.96	0.00				
……	……	……	……	……	……	……	0.00			
丹阳	56804.42	50786.09	43735.57	……	75.74	68.60	……	0.00		
滕州	49811.16	44262.47	37718.14	……	362.70	250.16	……	333.83	0.00	
镇江	30866.48	26377.85	20840.88	……	5246.28	4810.04	……	4932.52	2923.75	0.00

德州—曲阜之间的距离最小，将二者合并为一个新类：（德州、曲阜）

层次聚类过程

继续聚类，找到距离最近的两类城市。（欧氏距离的平方+最长距离法）

- 以丹阳—(德州、曲阜)为例：

	北京	南京	上海	……	德州	曲阜	……	丹阳	滕州	镇江
北京	0.00									
南京	447.78	0.00								
上海	1772.99	609.54	0.00							
……	……	……	……	0.00						
德州	57916.81	51856.88	44861.30	……	0.00					
曲阜	56140.41	50321.12	43481.37	……	36.96	0.00				
……	……	……	……	……	……	……	0.00			
丹阳	56804.42	50786.09	43735.57	……	75.74	68.60	……	0.00		
滕州	49811.16	44262.47	37718.14	……	362.70	250.16	……	333.83	0.00	
镇江	30866.48	26377.85	20840.88	……	5246.28	4810.04	……	4932.52	2923.75	0.00

层次聚类过程

继续聚类，找到距离最近的两类城市（欧氏距离的平方+最长距离法）。

- 按最长距离法，取75.74作为丹阳—(德州、曲阜)的距离

	北京	南京	上海	……	(德州、 曲阜)	……	丹阳	滕州	镇江
北京	0.00								
南京	447.78	0.00							
上海	1772.99	609.54	0.00						
……	……	……	……	0.00					
(德州、曲阜)	57916.81	51856.88	44861.30	……	0.00				
……	……	……	……	……	……	0.00			
丹阳	56804.42	50786.09	43735.57	……	75.74	……	0.00		
滕州	49811.16	44262.47	37718.14	……	362.70	……	333.83	0.00	
镇江	30866.48	26377.85	20840.88	……	5246.28	……	4932.52	2923.75	0.00

层次聚类法案例

层次聚类过程

	北京	南京	上海	……	(德州、 曲阜)	……	丹阳	滕州	镇江
北京	0.00								
南京	447.78	0.00							
上海	1772.99	609.54	0.00						
……	……	……	……	0.00					
(德州、曲阜)	57916.81	51856.88	44861.30	……	0.00				
……	……	……	……	……	……	0.00			
丹阳	56804.42	50786.09	43735.57	……	75.74	……	0.00		
滕州	49811.16	44262.47	37718.14	……	362.70	……	333.83	0.00	
镇江	30866.48	26377.85	20840.88	……	5246.28	……	4932.52	2923.75	0.00

同理，得到每两类城市之间的距离矩阵（20×20），结果发现，丹阳—(德州、曲阜)之间的距离最小，将二者合并为一个新类：（德州、曲阜、丹阳）。

层次聚类法案例

重复上述步骤，

直至所有的城市

都合并成一类。

每一步均为两两合并，

合并21个城市需20步。

步骤	合并的类别		系数	上一次出现的步骤		下一步骤
	类别1	类别2		类别1	类别2	
1	14	15	36.959	0	0	2
2	14	19	75.742	1	0	4
3	11	18	123.605	0	0	8
4	14	17	127.042	2	0	6
5	13	20	132.632	0	0	6
6	13	14	362.704	5	4	16
7	1	2	447.779	0	0	14
8	4	11	466.697	0	3	10
9	8	9	949.870	0	0	17
10	4	12	975.962	8	0	13
11	5	6	1038.637	0	0	15
12	16	21	1114.468	0	0	13
13	4	16	1717.824	10	12	16
14	1	3	1772.989	7	0	19
15	5	10	4980.461	11	0	18
16	4	13	5246.277	13	6	20
17	7	8	5806.168	0	9	18
18	5	7	9659.812	15	17	19
19	1	5	25613.833	14	18	20
20	1	4	57916.813	19	16	0

层次聚类法案例

重复上述步骤，
直至所有的城市
都合并成一类。

每一步均为两两合并，
合并21个城市需20步。

每一步被合并的2个小类的编号，以该类中首次出现的城市的编号表示。

步骤	合并的类别		系数	上一次出现的步骤		下一步骤
	类别1	类别2		类别1	类别2	
1	14	15	36.959	0	0	2
2	14	19	75.742	1	0	4
3	11	18	123.605	0	0	8
4	14	17	127.042	2	0	6
5	13	20	132.632	0	0	6
6	13	14	362.704	5	4	16
7	1	2	447.779	0	0	14
8	4	11	466.697	0	3	10
9	8	9	949.870	0	0	17
10	4	12	975.962	8	0	13
11	5	6	1038.637	0	0	15
12	16	21	1114.468	0	0	13
13	4	16	1717.824	10	12	16
14	1	3	1772.989	7	0	19
15	5	10	4980.461	11	0	18
16	4	13	5246.277	13	6	20
17	7	8	5806.168	0	9	18
18	5	7	9659.812	15	17	19
19	1	5	25613.833	14	18	20
20	1	4	57916.813	19	16	0

层次聚类法案例

重复上述步骤，
直至所有的城市
都合并成一类。

每一步均为两两合并，
合并21个城市需20步。

每一步被合并的2个小类的编号，以该类中首次出现的城市的编号表示。

每一次合并的前后关系：
上一步、下一步在哪？

步骤	合并的类别		系数	上一次出现的步骤		下一步骤
	类别1	类别2		类别1	类别2	
1	14	15	36.959	0	0	2
2	14	19	75.742	1	0	4
3	11	18	123.605	0	0	8
4	14	17	127.042	2	0	6
5	13	20	132.632	0	0	6
6	13	14	362.704	5	4	16
7	1	2	447.779	0	0	14
8	4	11	466.697	0	3	10
9	8	9	949.870	0	0	17
10	4	12	975.962	8	0	13
11	5	6	1038.637	0	0	15
12	16	21	1114.468	0	0	13
13	4	16	1717.824	10	12	16
14	1	3	1772.989	7	0	19
15	5	10	4980.461	11	0	18
16	4	13	5246.277	13	6	20
17	7	8	5806.168	0	9	18
18	5	7	9659.812	15	17	19
19	1	5	25613.833	14	18	20
20	1	4	57916.813	19	16	0

层次聚类法案例

重复上述步骤，
直至所有的城市
都合并成一类。

每一步均为两两合并，
合并21个城市需20步。

每一步被合并的2个小类的编号，以该类中首次出现的城市的编号表示。

每一次合并的前后关系：
上一步、下一步在哪？

步骤	合并的类别		系数	上一次出现的步骤		下一步骤
	类别1	类别2		类别1	类别2	
1	14 德州	15 曲阜	36.959	0	0	2
2	14	19	75.742	1	0	4
3	11	18	123.605	0	0	8
4	14	17	127.042	2	0	6
5	13	20	132.632	0	0	6
6	13	14	362.704	5	4	16
7	1	2	447.779	0	0	14
8	4	11	466.697	0	3	10
9	8	9	949.870	0	0	17
10	4	12	975.962	8	0	13
11	5	6	1038.637	0	0	15
12	16	21	1114.468	0	0	13
13	4	16	1717.824	10	12	16
14	1	3	1772.989	7	0	19
15	5	10	4980.461	11	0	18
16	4	13	5246.277	13	6	20
17	7	8	5806.168	0	9	18
18	5	7	9659.812	15	17	19
19	1	5	25613.833	14	18	20
20	1	4	57916.813	19	16	0

层次聚类法案例

重复上述步骤，

直至所有的城市

都合并成一类。

每一步均为两两合并，合并21个城市需20步。

每一步被合并的2个小类的编号，以该类中首次出现的城市的编号表示。

每一次合并的前后关系：上一步、下一步在哪？

步骤	合并的类别		系数	上一次出现的步骤		下一步骤
	类别1	类别2		类别1	类别2	
1	14 德州	15 曲阜	36.959	0	0	2
2	14 德州曲阜	19 丹阳	75.742	1	0	4
3	11	18	123.605	0	0	8
4	14	17	127.042	2	0	6
5	13	20	132.632	0	0	6
6	13	14	362.704	5	4	16
7	1	2	447.779	0	0	14
8	4	11	466.697	0	3	10
9	8	9	949.870	0	0	17
10	4	12	975.962	8	0	13
11	5	6	1038.637	0	0	15
12	16	21	1114.468	0	0	13
13	4	16	1717.824	10	12	16
14	1	3	1772.989	7	0	19
15	5	10	4980.461	11	0	18
16	4	13	5246.277	13	6	20
17	7	8	5806.168	0	9	18
18	5	7	9659.812	15	17	19
19	1	5	25613.833	14	18	20
20	1	4	57916.813	19	16	0

层次聚类法案例

重复上述步骤，

直至所有的城市

都合并成一类。

每一步均为两两合并，合并21个城市需20步。

每一步被合并的2个小类的编号，以该类中首次出现的城市的编号表示。

每一次合并的前后关系：上一步、下一步在哪？

步骤	合并的类别		系数	上一次出现的步骤		下一步骤
	类别1	类别2		类别1	类别2	
1	14 德州	15 曲阜	36.959	0	0	2
2	14 德州曲阜	19 丹阳	75.742	1	0	4
3	11	18	123.605	0	0	8
4	14 德州曲阜丹阳	17	127.042	2	0	6
5	13	20	132.632	0	0	6
6	13	14	362.704	5	4	16
7	1	2	447.779	0	0	14
8	4	11	466.697	0	3	10
9	8	9	949.870	0	0	17
10	4	12	975.962	8	0	13
11	5	6	1038.637	0	0	15
12	16	21	1114.468	0	0	13
13	4	16	1717.824	10	12	16
14	1	3	1772.989	7	0	19
15	5	10	4980.461	11	0	18
16	4	13	5246.277	13	6	20
17	7	8	5806.168	0	9	18
18	5	7	9659.812	15	17	19
19	1	5	25613.833	14	18	20
20	1	4	57916.813	19	16	0

层次聚类法案例

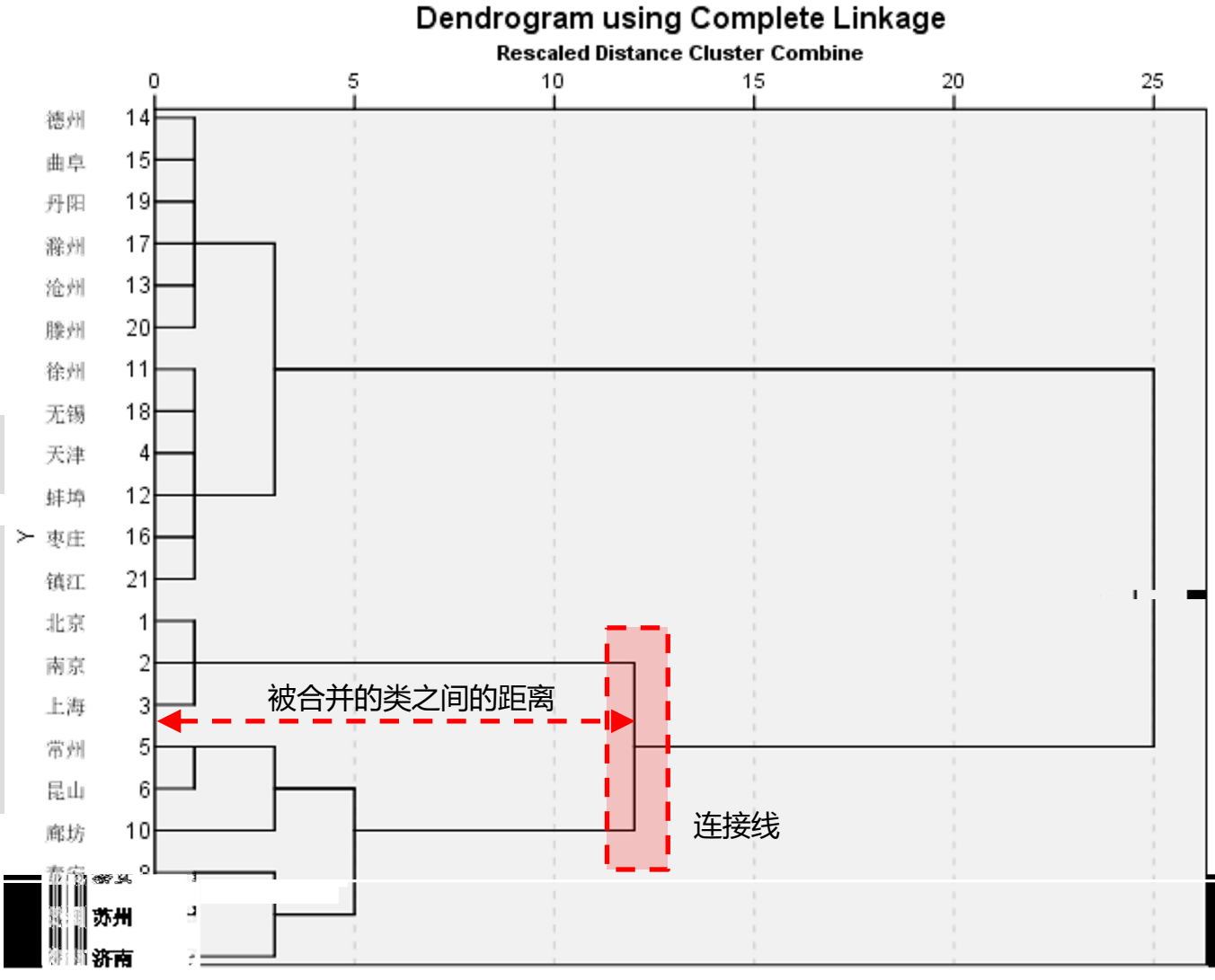
聚类谱系图

(dendrogram)

与前表等价，更直观。

21个样本

- 垂直：20条连接线代表20步合并
- 水平：代表小类之间的距离，缩放至0-25。



层次聚类法案例

划分类别

- 当类间距离**大幅增加**，即相似性大幅降低时，应**停止**聚合。
- 大幅增加这一判断**并无绝对标准**，可融入研究者的主观判断。

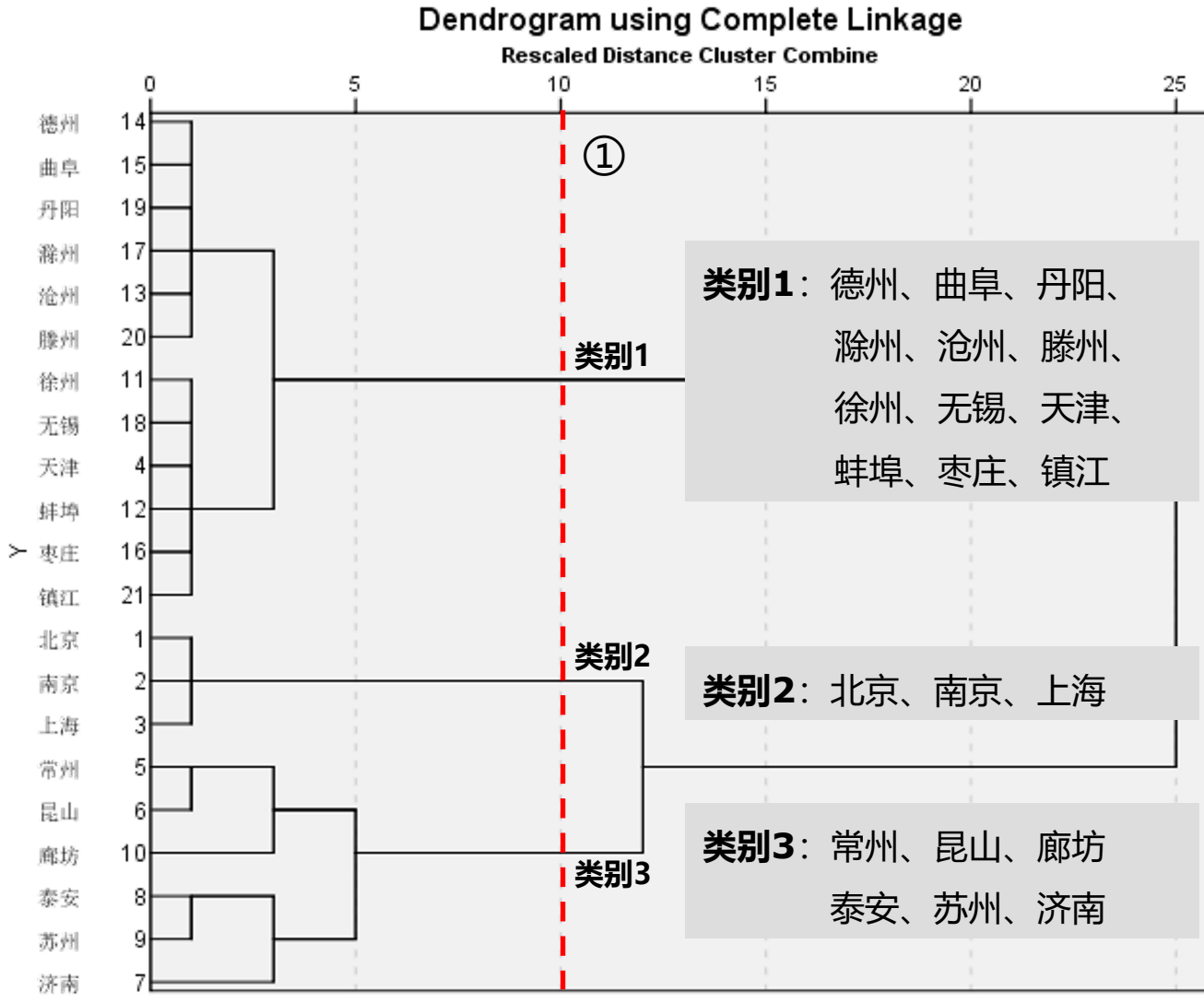
步骤	合并的类别		系数	上一次出现的步骤		下一步骤
	类别1	类别2		类别1	类别2	
1	14	15	36.959	0	0	2
2	14	19	75.742	1	0	4
3	11	18	123.605	0	0	8
4	14	17	127.042	2	0	6
5	13	20	132.632	0	0	6
6	13	14	362.704		4	16
7	1	2	447.779		0	14
8	4	11	466.697		3	10
9	8	9	949.870		0	17
10	4	12	975.962		0	13
11	5	6	1038.637		0	15
12	16	21	1114.468		0	13
13	4	16	1717.824		12	16
14	1	3	1772.989		0	19
15	5	10	4980.461		0	18
16	4	13	5246.277		6	20
17	7	8	5806.168		9	18
18	5	7	9659.812		17	19
19	1	5	25613.833	14	18	20
20	1	4	57916.813	19	16	0

相似性逐渐降低，距离逐渐增大

层次聚类法案例

划分类别

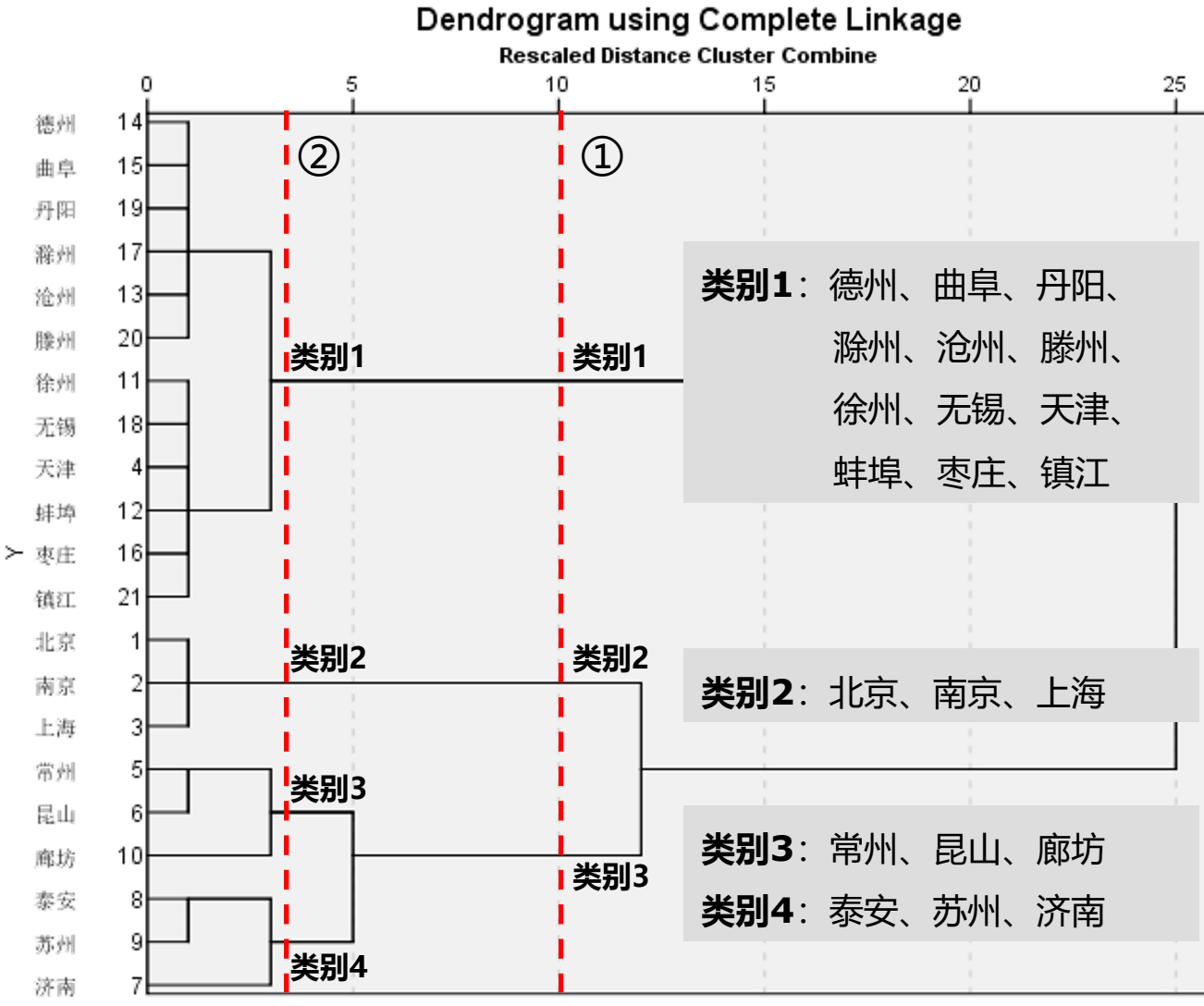
- 当类间距离**大幅增加**，即相似性大幅降低时，应**停止**聚合。



层次聚类法案例

划分类别

- 当类间距离**大幅增加**，即相似性大幅降低时，应**停止**聚合。
- 对于任意类别数k，在谱系图的合适位置画垂直线，使之与k条水平线相交即可。



层次聚类法案例

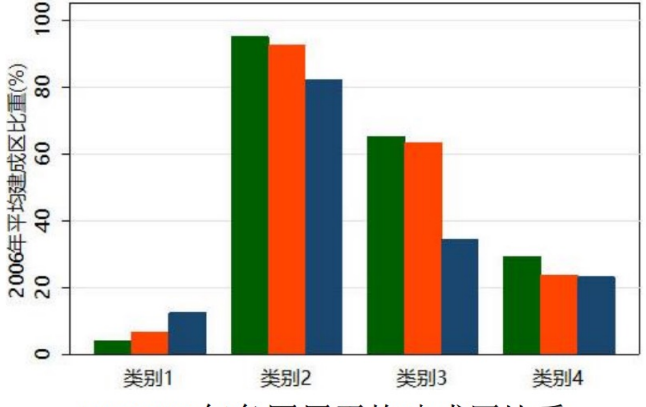
解释类别意义：按四类别方案，比较它们在六个变量上的均值差异。

类别1：德州、曲阜、丹阳、滁州、沧州、滕州、徐州、无锡、天津、蚌埠、枣庄、镇江

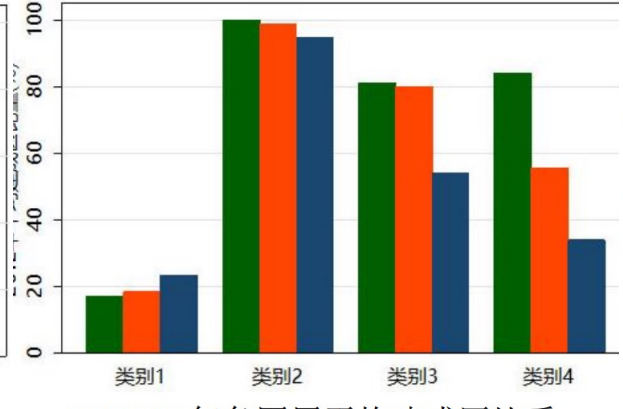
类别2：北京、南京、上海

类别3：常州、昆山、廊坊

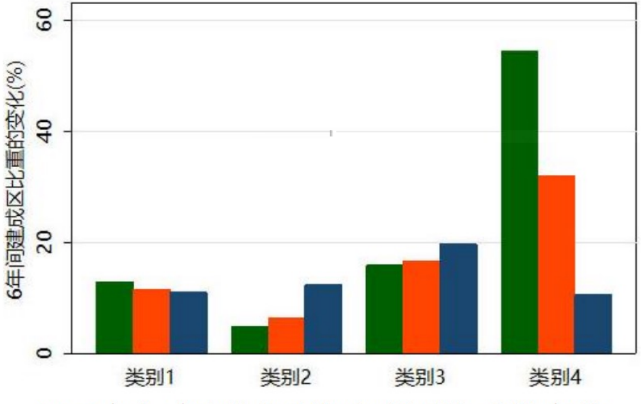
类别4：泰安、苏州、济南



(1) 2006 年各圈层平均建成区比重



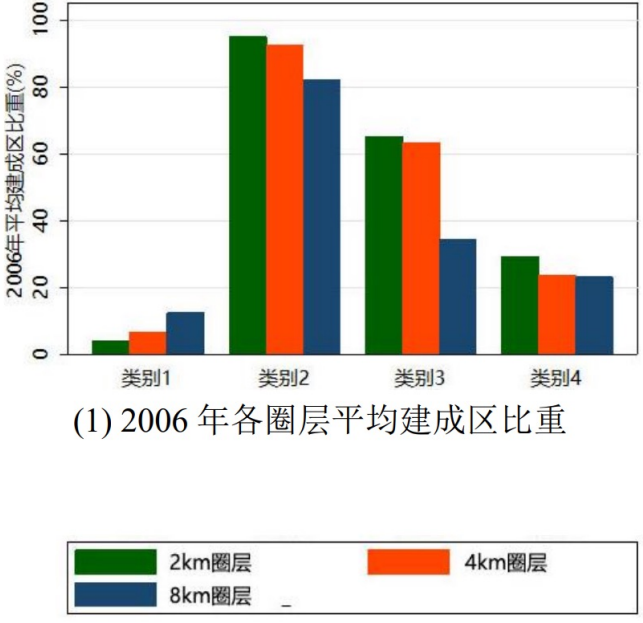
(2) 2012 年各圈层平均建成区比重



(3) 6 年间各圈层平均建成区比重的变化

层次聚类法案例

解释类别意义： 按四类别方案，比较它们在六个变量上的均值差异。



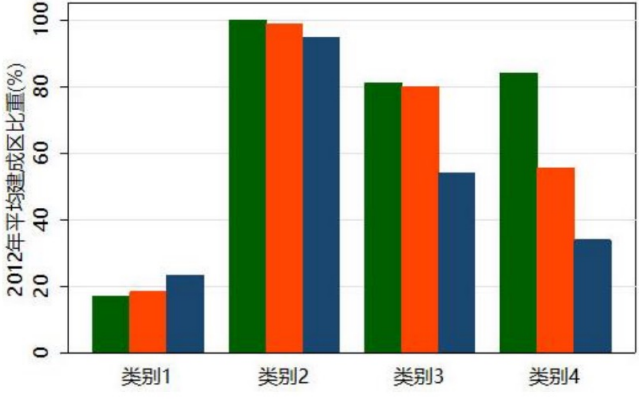
(1) 2006 年各圈层平均建成区比重

2006 年:

- 类别1——三个圈层建成水平均最**低**。
- 类别2——三个圈层建成水平已实现**高**水平**平衡**。
- 类别3——建成水平总体**较高**，但具有明显**不平衡**特征，8km圈层建成区比重偏低。
- 类别4——三个圈层建成水平居**中**且较**平衡**。

层次聚类法案例

解释类别意义： 按四类别方案，比较它们在六个变量上的均值差异。



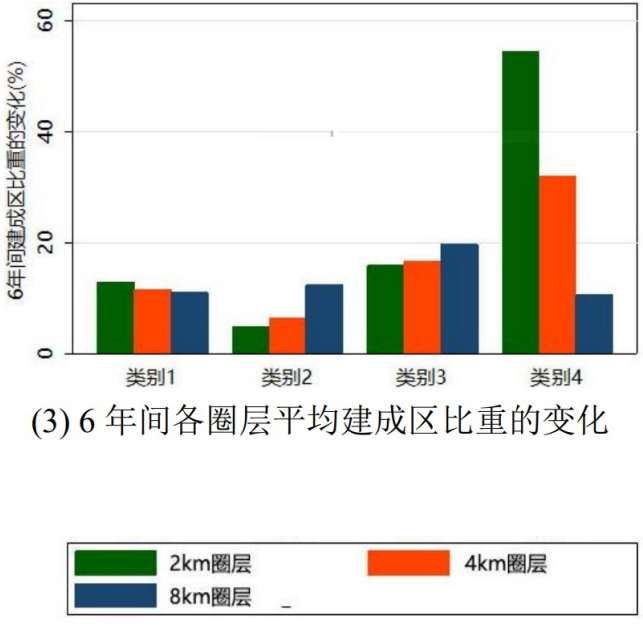
(2) 2012 年各圈层平均建成区比重

2012 年:

- 类别1——三个圈层建成水平仍处于**低**水平**平衡**状态。
- 类别2——三个圈层建成水平仍处于**高**水平**平衡**状态。
- 类别3——建成水平总体仍**较高**，但仍保持8km圈层建成区比重偏低的**不平衡**特征。
- 类别4——2km圈层建成区比重**较高**，三个圈层出现明显**不平衡**特征，显著拉开了差距。

层次聚类法案例

解释类别意义： 按四类别方案，比较它们在六个变量上的均值差异。



(3) 6 年间各圈层平均建成区比重的变化

2006 年 vs 2012年:

- 类别1、类别3的各圈层建成水平都有一定提升。
- 类别2由于起点高，所以变化有限且主要出现在8km圈层。
- 类别4各圈层建成区规模的增长快慢不一，2km圈层增长很快，4km居中，8km缓慢。

层次聚类法案例

聚类结论

类别	2006年特征		2012年特征		6年间变化特征	
	规模	圈层平衡	规模	圈层平衡	规模	圈层平衡
1	最小	不平衡，2km圈层水平较低	最小	低水平平衡	居中	均有明显增长
2	最大	高水平平衡	最大	高水平平衡	缓慢	主要在4km圈层
3	较大	不平衡：8km圈层水平明显偏低	较大	不平衡：8km圈层水平明显偏低	居中	均有明显增长
4	居中	较平衡：2km圈层水平相对较高	高低不一	不平衡：2km > 4km > 8km，三圈层明显拉开	快慢不一	不平衡：2km圈层增长很快，4km居中，8km缓慢

大纲

- 聚类分析概述
- 层次聚类法
- **K-means聚类法**

K-means聚类法

数据：

调查了某商业区内 350 名消费者的步行路径和活动记录，由此计算了区内 88 个商店的经过率和停留活动率。

商店编号	1	2	3	……	40	41	……	86	87	88
通过率	0.775	0.420	0.474	……	0.622	0.357	……	0.291	0.022	0.754
停留活动率	0.048	0.243	0.124	……	0.621	0.728	……	0.709	0.714	0.094

问题：

通过聚类分析考察经过率与停留活动率的不同组合模式， 将**商业空间**分为不同**类型**。

K-means聚类法

商店的经过率（pass ratio）：是受访的消费者总数中，从该商店经过的消费者所占的比例。

停留活动率（pass-enter ratio）：是从该商店经过的受访消费者中，进入该商店、发生停留活动的消费者所占的比例。

商店的客流量与经过率和停留活动率具有如下的正比例关系：

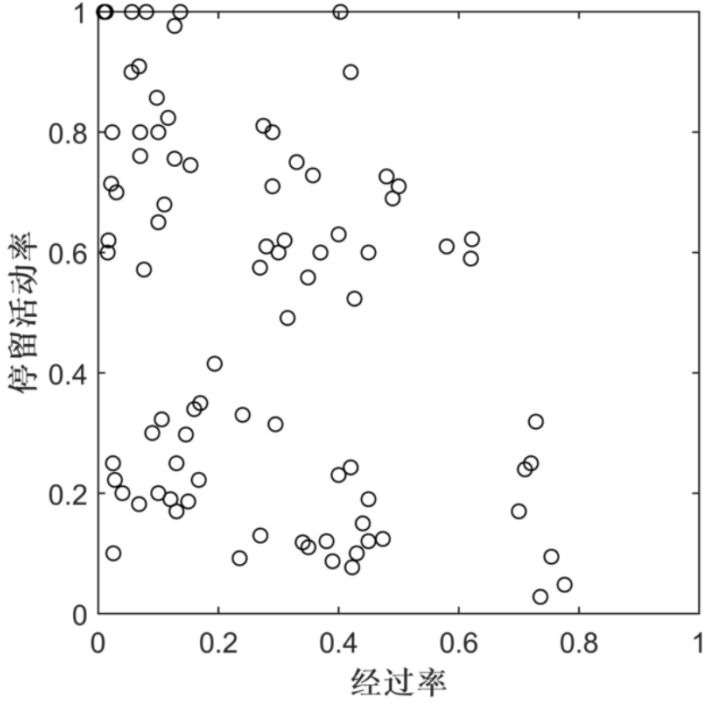
$$\text{客流量} = \text{总人数} \times \frac{\text{经过人数}}{\text{总人数}} \times \frac{\text{停留活动人数}}{\text{经过人数}} = \text{总人数} \times \text{经过率} \times \text{停留活动率}$$

K-means聚类法

步骤1： 预先设定类别的数量 k。

- 观察通过率—停留活动率的散点图，初步设定类别的数量 $k = 5$。

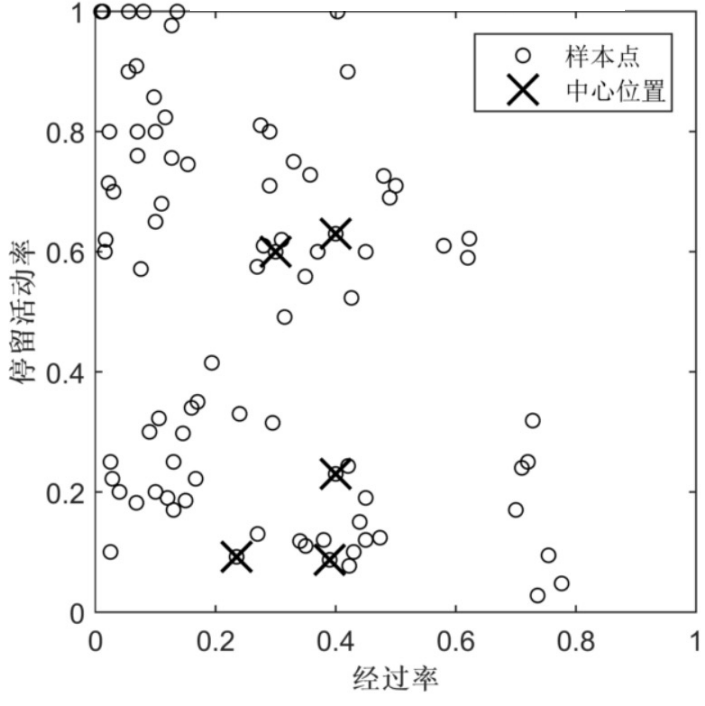
- 当变量个数较多时，难以通过这样的直观手段做出判断，可以先利用主成分分析技术将数据降至二维，再进行可视化。



K-means聚类法

步骤2： 初始化中心位置，将每个样本点分配给就近的中心。

- 随机选择 $k = 5$ 个样本点作为初始中心，每个中心对应一个类别。



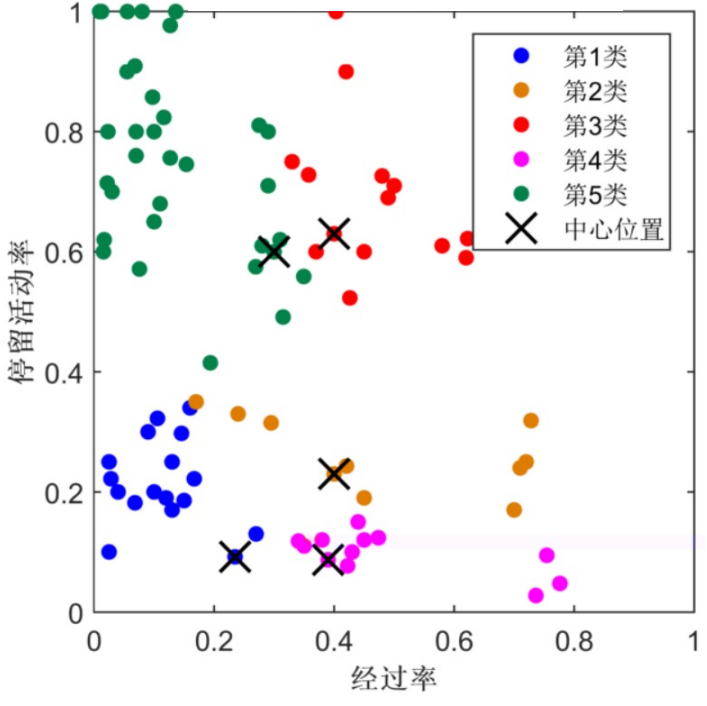
第 **1** 次迭代

K-means聚类法

步骤2： 初始化中心位置，将每个样本点分配给就近的中心。

- 随机选择 $k = 5$ 个样本点作为初始中心，每个中心对应一个类别。

- 计算每个样本点到 5 个中心的距离，并把该样本点的类别标记为最近中心的类别，由此得到每个类别的样本点集合。

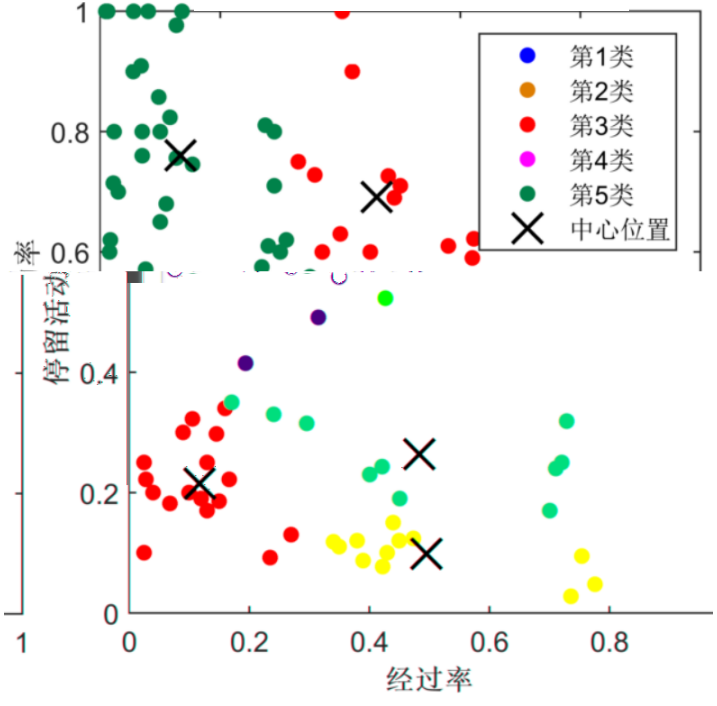


第 **1** 次迭代

K-means聚类法

步骤3： 更新中心位置，重新标记样本点的类别。

- 将5个中心移动到对应的样本点集合的平均位置。

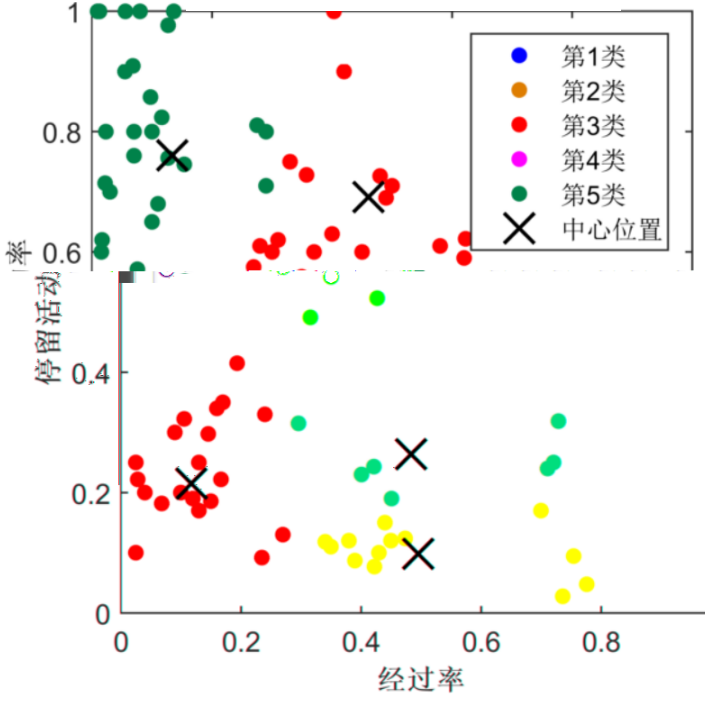


第 **2** 次迭代

K-means聚类法

步骤3： 更新中心位置，重新标记样本点的类别。

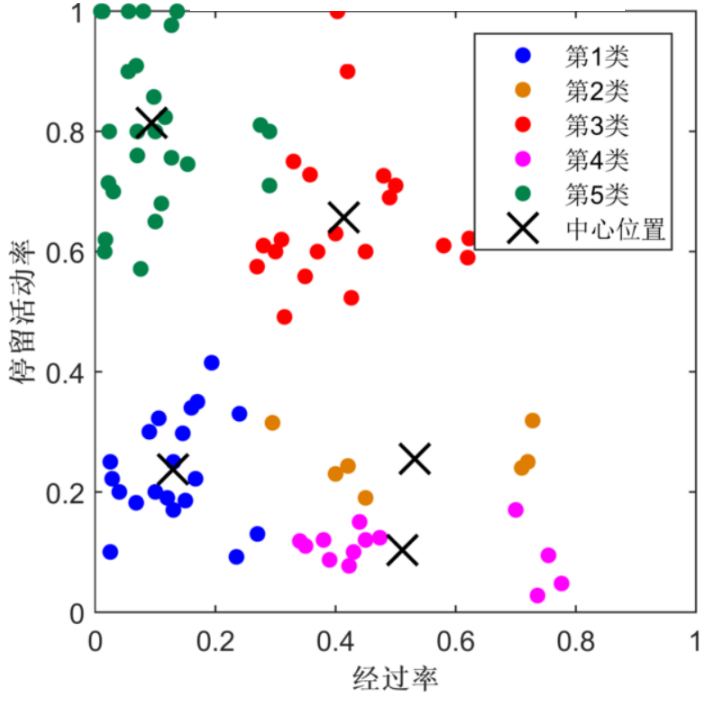
- 将5个中心移动到对应的样本点集合的平均位置。
- 重新计算样本点到中心的距离，并重新标记各样本点的类别。



第 **2** 次迭代

K-means聚类法

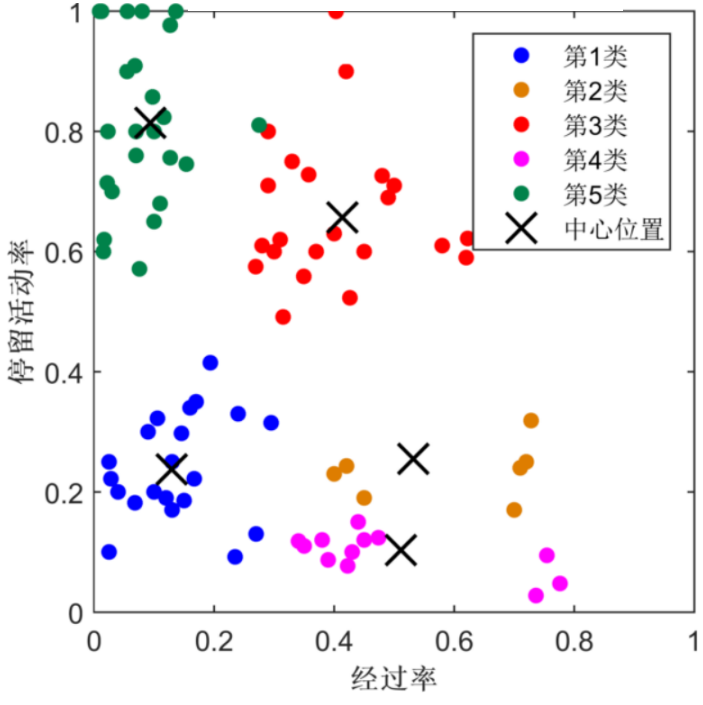
重复步骤3，依次更新中心位置和样本点的类别所属，直至不再发生变化，算法收敛。



第 **3** 次迭代

K-means聚类法

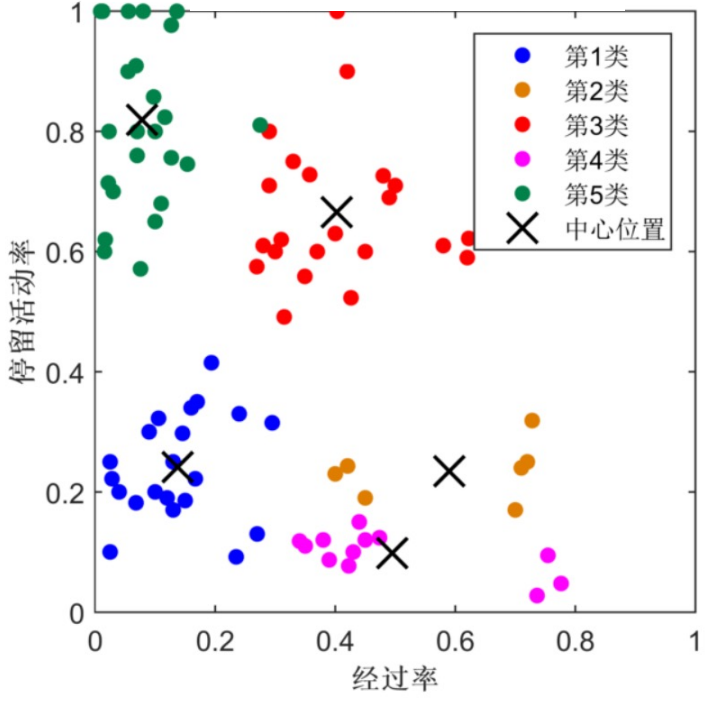
重复步骤3，依次更新中心位置和样本点的类别所属，直至不再发生变化，算法收敛。



第 **3** 次迭代

K-means聚类法

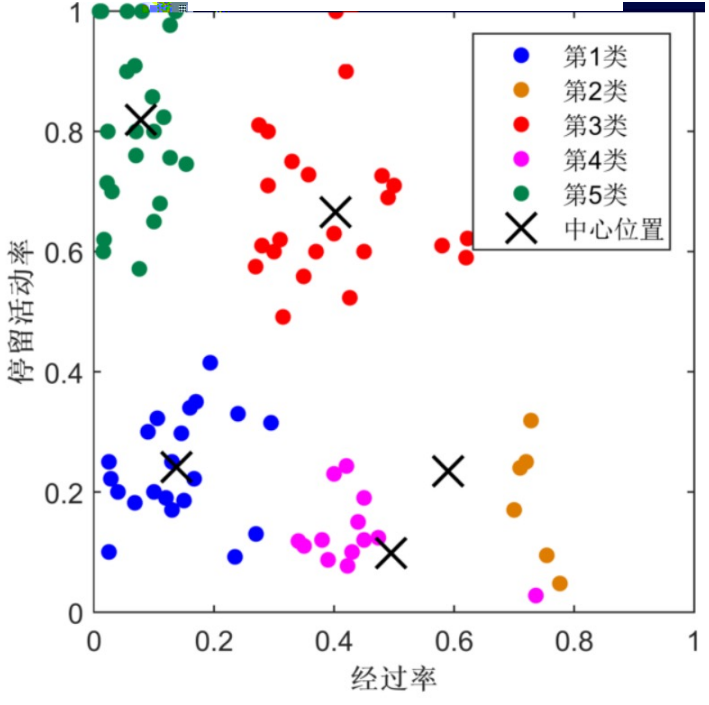
重复步骤3，依次更新中心位置和样本点的类别所属，直至不再发生变化，算法收敛。



第 **4** 次迭代

K-means聚类法

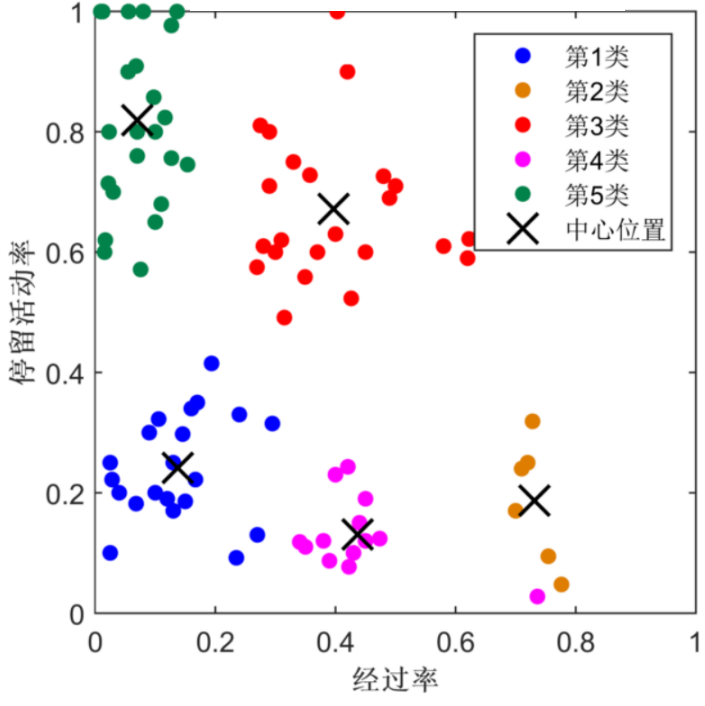
重复步骤3，依次更新中心位置和样本点的类别所属，直至不再发生变化，算法收敛。



第 **4** 次迭代

K-means聚类法

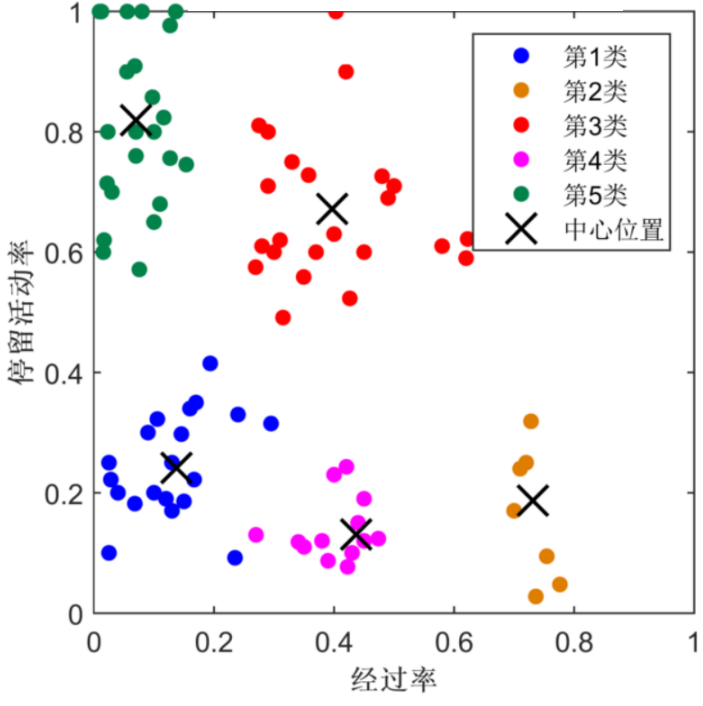
重复步骤3，依次更新中心位置和样本点的类别所属，直至不再发生变化，算法收敛。



第 **5** 次迭代

K-means聚类法

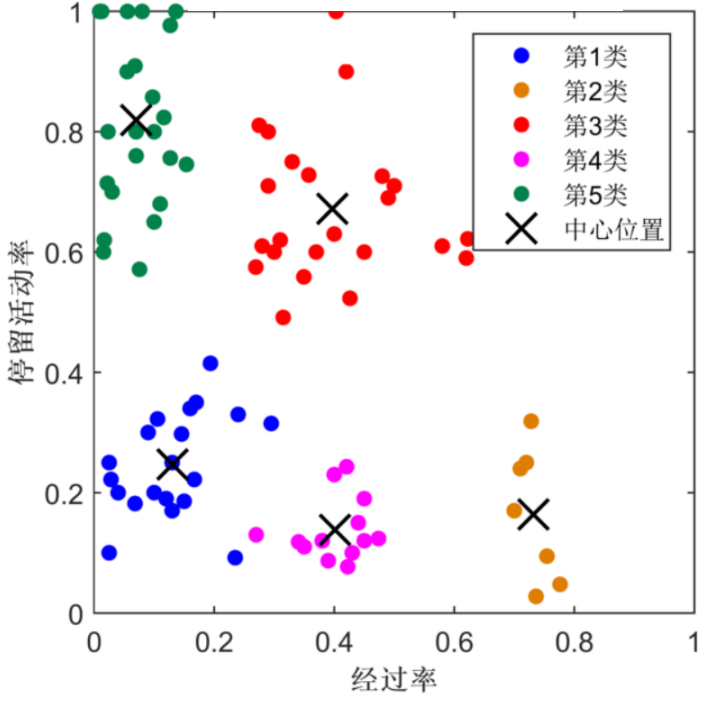
重复步骤3，依次更新中心位置和样本点的类别所属，直至不再发生变化，算法收敛。



第 **5** 次迭代

K-means聚类法

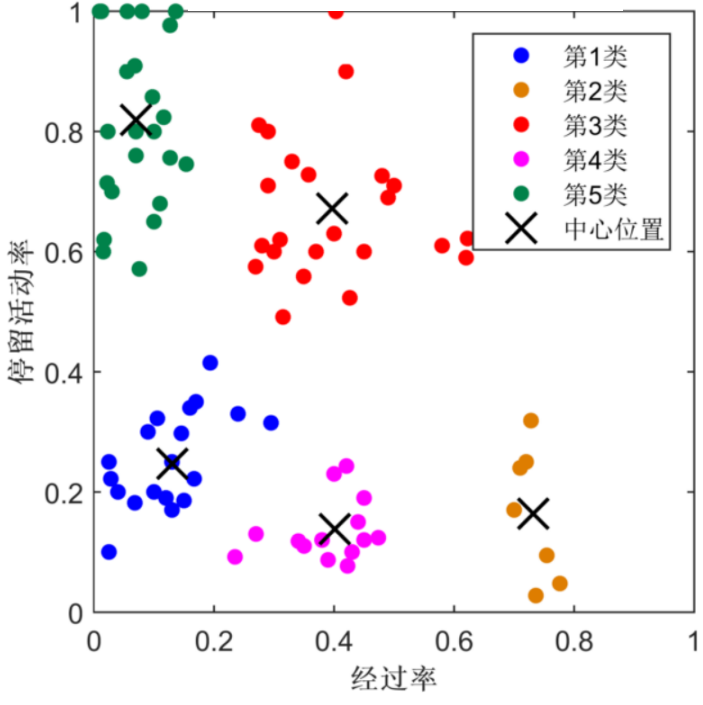
重复步骤3，依次更新中心位置和样本点的类别所属，直至不再发生变化，算法收敛。



第 **6** 次迭代

K-means聚类法

重复步骤3，依次更新中心位置和样本点的类别所属，直至不再发生变化，算法收敛。

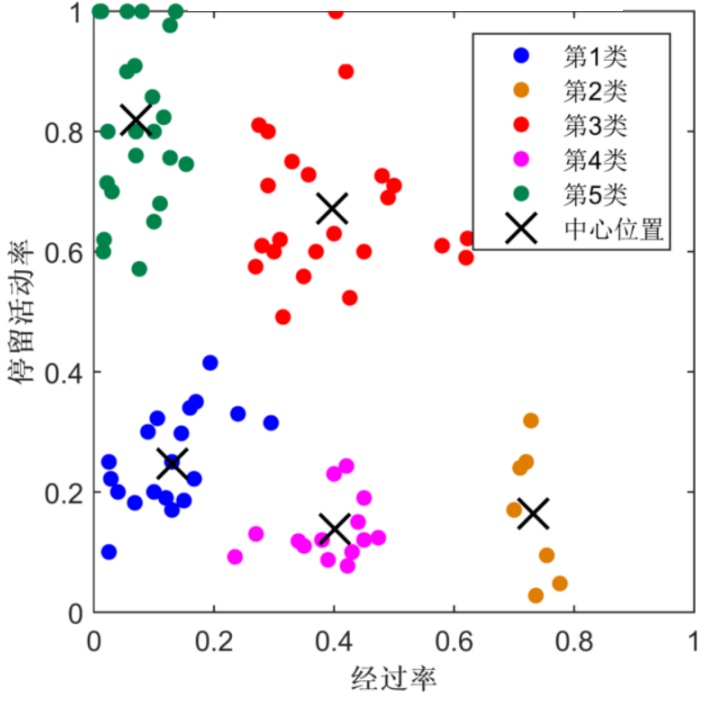


第 **6** 次迭代

K-means聚类法

重复步骤3，依次更新中心位置和样本点的类别所属，直至不再发生变化，算法收敛。

- 经过 6 步迭代，中心位置和类别所属不再发生变化，得到最终的聚类结果。
- k-means 算法对初始中心的位置较为敏感，多次运行可能会得到不同的结果。
- k 的不同取值会对聚类结果产生非常大的影响，对于“聚为几类最适宜”的问题，需要多次尝试和比较。



第 **6** 次迭代

K-means聚类法

评价聚类结果： “类内尽可能相似、类间尽可能不同”的程度。

- **Silhouette 值**：对于给定的聚类结果，样本i的Silhouette值s_i

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

式中，a_i是样本 i 到同一类别内其他样本的平均距离（intra-cluster distance），

b_i 是样本 i 到各个不同类别样本的平均距离的最小值（nearest-cluster distance）

s_i取值范围为 -1~1

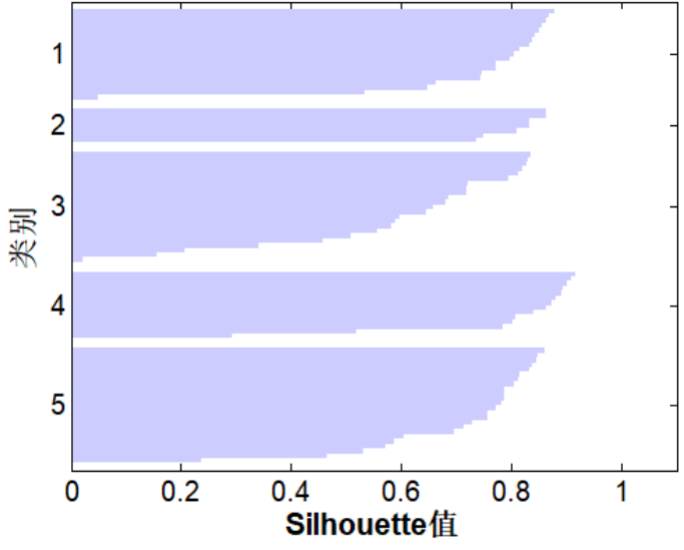
- >0时，值越高，表明 i 对自己所在分类的所属性越强，而对其他分类的所属性越弱.
- 接近0，表明 i 处在两个分类的边缘。
- <0时，表明 i 与其他类别的样本更相似，提示 i 可能被分到了错误的类别。

K-means聚类法

评价聚类结果：

- 本例中，各类别样本点的Silhouette 值均为正，且大多数>0.5，平均值为0.713。

证明了这一聚类结果表现理想，较好地实现了对数据的自然分割。



K-means聚类法

评价聚类结果：

- 取k = 2~10，分别运行 k-means 算法，并依次计算每一种类别数目下，聚类结果的Calinski-Harabasz伪F指标和平均Silhouette值。

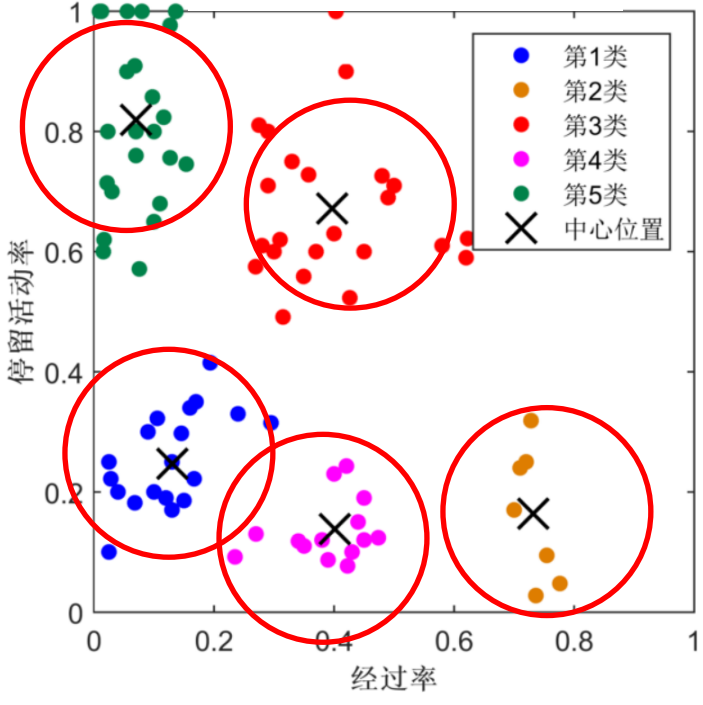
类别数量	2	3	4	5	6	7	8	9	10
伪 F	112.14	100.85	138.50	141.89	154.15	157.47	174.52	182.95	186.45
Sihouette	0.684	0.640	0.685	0.713	0.692	0.695	0.645	0.687	0.686

两项指标均是越高越好，前者的建议是十类最佳，而后者的建议是五类最佳。

- 当不同指标的结论不一致时，研究者应当更加关注聚类结果的实际意义。基于此，五分类方案具有更好的解释性。

K-means聚类法

解释类别意义：按五个类别的方案。



- 类别1——经过率低、停留率也低。形成区位瓶颈，空间活力不佳，客流量有限。
- 类别2——经过率最高、停留率最低。一般区位良好，但巨大的客流潜力未被充分利用。可视为潜力空间，着重优化。
- 类别3——经过率和停留率均较好。
- 类别4——与类别2类似，均无法留住消费者，区位不如类别2。
- 类别5——经过率低、停留率最高。大多区位不利，但功能吸引力很强，吸引消费者专程前往。

K-means聚类法

- **所属类别**：基于划分的聚类（partition-based methods）
- **适用数据**：全部为数值型变量
- **算法流程**：

  1. 预先设定类别的数量k；
  2. 算法随机选择k个样本点作为初始中心位置，并将每个样本点分配给就近的中心；
  3. 算法根据所得到的分组情况更新k个中心，将其分别移动到每一组样本点的平均位置，再以同样的方式重新分配样本，形成新的分组；
  4. 上述过程不断重复，直至中心位置不再移动、分组情况不再变化为止。

- **优势**：运算速度快，更适宜处理大规模数据。
- **注意点**：
  - 类别数 k 必须预先设定，但如何确定 k 的最佳值并无理论指导，需反复尝试比较。
  - k-means算法对初始中心位置较为敏感，两次运行可能得到不同的聚类结果。

总结

- 聚类分析：把样本划分为不同的类别（群、簇）， 使**类别内部相似性高**，**类别之间相似性低**。
- 目的： 发现数据内在的结构和模式。
- 探索式、无监督：没有因变量，预先不知道多少类，需注意结果的可解释性。
- 有许多不同的聚类算法，不同的算法会产生不同的结果。
  - 层次聚类法：构建嵌套的层次结构，适合较小的数据集。
  - K-means聚类法：速度快，简单有效，需预先选定类别数，初值影响大。
- 评价聚类效果：可解释性、Silhouette值（类内紧密程度，类间分离程度）