聚类分析

基本定义

什么是聚类分析

把一组对象分成若干个聚类(Cluster), 使得同一个聚类中的对象之间具有高的相似性(High intra-cluster similarity), 不同聚类中的对象之间具有低的相似性(Low inter-cluster similarity).

怎样度量聚类方法

聚类内对象的高度同质性(Homogeneity)
聚类间对象的高度分离性 (Separation)

一些常见的聚类准则及其计算复杂性

k-Center：最大半径最小化 （NP-hard）

k-Cluster：最大直径最小化 （NP-hard）

k-median：聚类内部距离之和的最小化 （若视k为输入变量，则NP-hard，对于固定的k，则是P问题）

k-means: 聚类内部距离平方之和的最小化（无论k是否固定，均为NP-hard）

Min-cut：最小割 （P）

Max-cut：最大割 （NP-hard）

Ncut：规范割 （NP-hard）

MRSD准则 （NP-hard）

k-means

- k-means算法如下:

1. 把对象划分成k 个非空子集;
2. 计算当前划分的每个聚类的中心;
3. 把每一个对象分配到离它最近的中心;
4. 返回到第2步, 当满足某种停止条件时停止。

- 停止条件:
  当分配不再发生变化时停止;
  当前后两次迭代的目标函数值小于某一给定的阈值时;
  当达到给定的迭代次数时。

层次方法（单链接与全链接）

这种方法不需要用户提供聚类的数目k 作为输入。

- 若定义两个聚类之间的距离为二者对象之间的最小距离，则该算法也称为单链接算法(Single-Linkage Algorithm，SLA)，也称为最小生成树算法。
- 若定义两个聚类之间的距离为二者对象之间的最大距离，则该算法也称为全链接算法(Complete-Linkage Algorithm，CLA)。

示例：

- 给定5个对象间的距离如下表

No	1	2	3	4	5
1	0				
2	6	0			
3	2	4	0		
4	3	4	5	0	
5	7	1	5	5	0

- **步骤1**: 每个对象当做一个聚类.
- **步骤 2**: 找出上述5个聚类中最近的两个聚类2和5，因为它们的距离最小: d_{25}=1. 所以, 2和5凝聚成一个新的聚类{2, 5}.
- **步骤3**. 计算聚类{2, 5}与聚类 {1}, {3}, {4}的距离
- $d_{\{2,5\}1}=\min\{d_{21}, d_{51}\}=\min\{6,7\}=6$
- $d_{\{2,5\}3}=\min\{d_{23}, d_{53}\}=\min\{4,5\}=4$
- $d_{\{2,5\}4}=\min\{d_{24}, d_{54}\}=\min\{4,5\}=4$

No	{2,5}	1	3	4
{2,5}	0			
1	6	0		
3	4	2	0	
4	4	3	5	0

星期六 Data Mining: Concepts and Techniques

– 4个聚类 {2,5}, {1}, {3}, {4}中最近的2个聚类是 {1}和{3}. 因此, 1和3凝聚成一个新的聚类. 现在, 我们有3个聚类: {1,3}, {2,5}, {4}.

- **步骤4.** 计算聚类 {1,3}与 {2,5}, {4}之间的距离
  - $d_{\{1,3\}\{2,5\}}=\min\{d_{1\{2,5\}}, d_{3\{2,5\}}\}=\min\{6,4\}=4$
  - $d_{\{1,3\}4}=\min\{d_{14}, d_{34}\}=\min\{3,5\}=3$
  - 因此, 聚类{1,3}和 {4}凝聚成一个新的聚类{1,3,4}.

No	{2,5}	{1,3}	4
{2,5}	0		
{1,3}	4	0	
4	4	3	0

– 现在, 我们得到2个聚类{1,3,4}和{2,5}

- **步骤5.** 计算{1, 3,4}的{2,5}聚类
  - $d_{\{2,5\}\{1,3,4\}}= \min\{d_{\{2,5\}\{1,3\}}, d_{\{2,5\}4}\}=\min\{4,4\}=4$

No	{2,5}	{1,3,4}
{2,5}	0	
{1,3,4}	4	0

– 聚类 {1, 3,4}和{2,5}凝聚成一个唯一的聚类{1,2,3,4,5}.