社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

2020年11月9日

CONTENTS

聚类分析经典算法

1 基于划分的算法 （迭代聚类法）

2 层次聚类法

(1) 聚合聚类（agglomerative clustering）

(2) 分裂聚类（divisive clustering）

1 相似度（similarity）或距离（distance）
   (1) Minkowski distance
   (2) Mahalanobis distance
   (3) cosine similarity
   (4) correlation coefficient
2 类或簇（cluster）：类与类之间的距离，也称为连接（linkage）
   (1) 最短距离或单连接（single linkage）
   (2) 最长距离或完全连接（complete linkage）
   (3) 中心距离
   (4) 平均距离

层次聚类算法有两种思路，分别是自下而上和自上而下。

1. 聚合层次聚类算法（agglomerative hierarchical clustering, AHC）
2. 分裂层次聚类算法（divisive hierarchical clustering, DHC）

1 聚合层次聚类算法

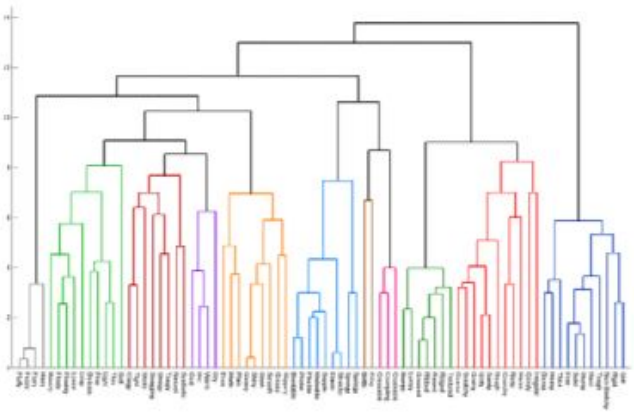
(1) 先将每个数据视为单独的类；
(2) 然后按照某种距离度量选择距离最近的两个或多个类进行合并；
(3) 重复合并的过程，直到最后只剩下一个类。

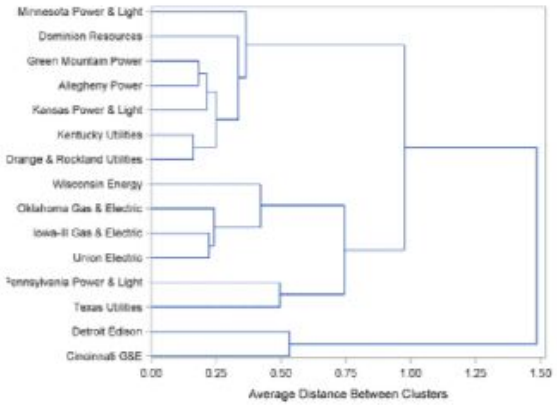
2 分裂层次聚类算法

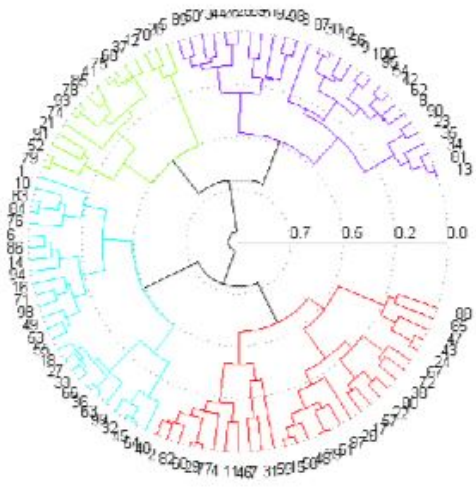
(1) 先将所有的数据视为一类；
(2) 然后基于某种准则，在已有类中选择一个类将其分割为两个类；
(3) 重复分割的过程，达到聚类的目的。

层次聚类法

层次聚类法(Hierarchical Clustering)：“一层一层地聚”——无需提前规定有几个类别，而是给出一个“路径”。先把每个个体单独看成一类，然后每次把“最相似”的个体聚在一起，直到最终只剩一个类别，或者反方向进行。这种路径通常用**系统树图**（Dendrogram）来表示。







聚合层次聚类算法需要预先确定下面三个要素：

1 距离或相似度
   (1) Minkowski distance
   (2) Mahalanobis distance
   (3) cosine similarity

2 合并规则
   - 类间距离最小，类间距离可以是最短距离、最长距离、中心距离、平均距离。

3 停止条件
   - 类的个数达到阈值。

1. cluster singlelinkage最短联结法；
2. cluster completelinkage最长联结法；
3. cluster averagelinkage平均联结法；
4. cluster waveragelinkage加权平均联结法；
5. cluster medianlinkage中位值联结法；
6. cluster centroidlinkage重心联结法；
7. cluster wardslinkage离差平方和（即Ward）联结法。

1 变量的标准化

(1) egen newx1=std(x1)

2 层次聚类分析，设定聚类方法

(1) cluster singlelinkage newx1 newx2 newx3 newx4 newx5

3 生成树状图 cluster dendrogram

在聚类分析结果的基础上，输出树状图，以便判断类数。

4 cluster generate type=group (3)

表示在聚类分析结果之上，建立一个新变量type（取值为1、2、3），通过该变量将每个观测案例按聚类分析结果归类到其前三个组别中去。

层次聚类法：自下而上

例

- 给定5个样本的集合，样本之间的欧氏距离由如下矩阵D表示

$$D = [d_{ij}]_{5\times 5} = \begin{bmatrix} 0 & 7 & 2 & 9 & 3 \\ 7 & 0 & 5 & 4 & 6 \\ 2 & 5 & 0 & 8 & 1 \\ 9 & 4 & 8 & 0 & 5 \\ 3 & 6 & 1 & 5 & 0 \end{bmatrix}$$

- 其中d_{ij}表示第i个样本与第j个样本之间的欧氏距离。
- 显然D为对称矩阵。应用聚合层次聚类法对这5个样本进行聚类。

- (1)
- 首先用5个样本构建5个类，$G_i = \{x_i\}$，$i = 1, 2, \cdots, 5$，
- 这样，样本之间的距离也就变成类之间的距离，所以5个类之间的距离矩阵亦为D

- (2)
- 由矩阵D可以看出，$D_{35} = D_{53} = 1$ 为最小，所以把G_3和G_5合并为一个新类，记作 $G_6 = \{x_3, x_5\}$。

- 计算G_6与G_1, G_2, G_4之间的最短距离，有

$$D_{61}=2, \quad D_{62}=5, \quad D_{64}=5$$

- 又注意到其余两类之间的距离是

$$D_{12}=7, \quad D_{14}=9, \quad D_{24}=4$$

- 显然，D_{61}=2最小，所以将G_1与G_6合并成一个新类，记作

$$G_7=\{x_1, x_3, x_5\}。$$

- 计算G_7与G_2, G_4之间的最短距离，

$$D_{72} = 5, \quad D_{74} = 5$$

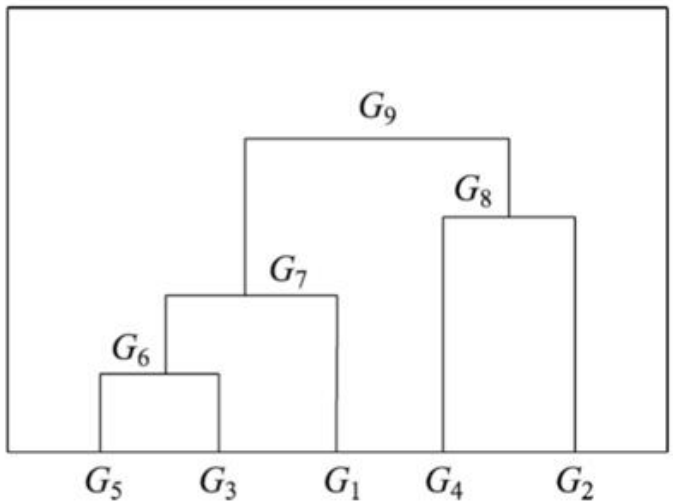
- 又注意到

$$D_{24} = 4$$

- 显然，其中D_{24}=4最小，所以将G_2与G_4合并成一个新类，记作

$$G_8 = \{x_2, x_4\}$$

- 将G_7与G_8合并成一个新类，记作 $G_9 = \{x_1, x_2, x_3, x_4, x_5\}$
- 即将全部样本聚成1类，聚类终止



1.曼哈顿距离（城市街区距离）（Manhattan distance）

```
d1 = np.sum(np.abs(x-y))
d2 = pdist([x,y],'cityblock')
print(d1,d2)
```

2.欧氏距离（Euclidean distance）

```
d1 = np.sqrt(np.sum(np.square(x-y)))
d2 = pdist([x,y])
print(d1,d2)
```

3.切比雪夫距离（Chebyshev distance）

```
d1 = np.max(np.abs(x-y))
d2 = pdist([x,y],'chebyshev')
print(d1,d2)
```

参考文献

1 李航，2019，《统计学习方法》（第2版），北京：清华大学出版社。
2 潘蕊等，2018，《数据思维实践：从零经验到数据英才》，北京：北京大学出版社。
3 张宪超，2018，《数据聚类》，北京：科学出版社。