第七章 聚类分析

7.1 聚类标准

7.2 系统聚类法

7.3 K均值聚类法

7.4 谱聚类法

7.5 基于密度的聚类法

第七章 聚类分析

将认识对象进行分类是人类认识世界的一种重要方法，如：在生物学中，为了研究生物的演变，需要对生物进行分类，生物学家根据各种生物的特征，将它们归属于不同的界、门、纲、目、科、属、种之中。事实上，分门别类地对事物进行研究，要远比在一个混杂多变的集合中研究更清晰、明了和细致，这是因为同一类事物会具有更多的近似特性。

第七章 聚类分析

在企业的经营管理中，为了确定其目标市场，首先要进行市场细分。因为无论一个企业多么庞大和成功它也无法满足整个市场的各种需求。而市场细分，可以帮助企业找到适合自己特色，并使企业具有竞争力的分市场，将其作为自己的重点开发目标。

通常，人们可以凭经验和专业知识来实现分类。而聚类分析（cluster analysis）作为一种定量方法，将从数据分析的角度，给出一个更准确、细致的分类工具。

第七章 聚类分析

聚类分析又称群分析，它是研究分类问题的一种多元统计分析。所谓类通俗地说，就是指相似元素的集合。要将相似元素聚为一类，通常选取元素的许多共同指标，然后通过分析元素的指标值来分辨元素间的差距，从而达到分类的目的。聚类分析可以分为 Q 型聚类（样本聚类）、R 型聚类（指标聚类）。

设有n个样品，每个样品测得p项指标（变量），原始数据阵为

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

其中x_{ij}（$i=1,\cdots,n$； $j=1,\cdots,p$）为第i个样品X_i的第j个指标的观测数据。

按行看，对n个样品进行聚类，称为 Q 型聚类

按列看，对p项指标（变量）进行聚类，称为 R 型聚类

第七章 聚类分析

7.1 聚类标准

1. 样本的相似性度量

要用数量化的方法对事物进行分类，就必须用数量化的方法描述事物之间的相似程度。一个事物常常需要用多个变量来刻画。如果对于一群有待分类的样本点需用p个变量描述，则每个样本点可以看成是R^p空间中的一个点。因此，很自然地想到可以用距离来度量样本点间的相似程度。

第七章 聚类分析

7.1 聚类标准

1. 样本的相似性度量

记Ω是样本点集，距离$d(\cdot,\cdot)$是$\Omega\times\Omega\to R^{+}$的一个函数，满足条件

（1）$d(x,y)\geq 0$，$x,y\in\Omega$；

（2）$d(x,y)=0$当且仅当$x=y$；

（3）$d(x,y)=d(y,x)$，$x,y\in\Omega$；

（4）$d(x,y)\leq d(x,z)+d(z,y)$，$x,y,z\in\Omega$。

第七章 聚类分析

7.1 聚类标准

1. 样本的相似性度量

在聚类分析中，对于两个样本相似性的度量，最常用的是 Minkowski 距离

$$d_q(x,y)=\left[\sum_{k=1}^{p}|x_k-y_k|^q\right]^{\frac{1}{q}},\quad q>0,$$

其中，$x=[x_1,x_2,\cdots,x_p]^T$，$y=[y_1,y_2,\cdots,y_p]^T$，当$q=1,2$或$q\to+\infty$时，则分别得到：

第七章 聚类分析

7.1 聚类标准

1. 样本的相似性度量

（1）绝对值距离

$$d_1(x,y)=\sum_{k=1}^{p}\left|x_k-y_k\right|,$$

（2）欧氏距离

$$d_2(x,y)=\left[\sum_{k=1}^{p}\left|x_k-y_k\right|^2\right]^{\frac{1}{2}},$$

（3）Chebyshev 距离

$$d_\infty(x,y)=\max_{1\le k\le p}\left|x_k-y_k\right|.$$

在Minkowski距离中，最常用的是欧氏距离，它的主要优点是当坐标轴进行正交旋转时，欧氏距离是保持不变的。因此，如果对原坐标系进行平移和旋转变换，则变换后样本点间的距离和变换前完全相同。

第七章 聚类分析

7.1 聚类标准

1. 样本的相似性度量

值得注意的是在采用 Minkowski 距离时,一定要采用相同量纲的变量。如果变量的量纲不同，测量值变异范围相差悬殊时，建议首先进行数据的标准化处理，然后再计算距离。在采用 Minkowski 距离时，还应尽可能地避免变量的多重相关性（multicollinearity)。多重相关性所造成的信息重叠，会片面强调某些变量的重要性。由于 Minkowski 距离的这些缺点，一种改进的距离就是马氏距离，定义如下

第七章 聚类分析

7.1 聚类标准

1. 样本的相似性度量

（4）马氏（Mahalanobis）距离

$$d(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)}\ ,$$

Σ为总体样本Z的协方差矩阵，实际中Σ往往是不知道的，常常需要用样本协方差来估计。马氏距离对一切线性变换是不变的，故不受量纲的影响。

第七章 聚类分析

7.1 聚类标准

2. 指标(变量)的相似性度量

在实际工作中，指标(变量)聚类法的应用也是十分重要的。在系统分析或评估过程中，为避免遗漏某些重要因素，往往在一开始选取指标时，尽可能多地考虑所有的相关因素。而这样做的结果，则是变量过多，变量间的相关度高，给系统分析与建模带来很大的不便。因此，人们常常希望能研究变量间的相似关系，按照变量的相似关系把它们聚合成若干类，进而找出影响系统的主要因素。

第七章 聚类分析

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mp} \end{bmatrix}$$

7.1 聚类标准

2. 指标(变量)的相似性度量

在对指标(变量)进行聚类分析时，首先要确定变量的相似性度量，常用的变量相似性度量有两种。

(1) 相关系数

记变量x_j的取值

$$[x_{1j},x_{2j},\cdots,x_{mj}]^T \in R^m (j=1,2,\cdots,p)。$$

则可以用两变量x_j与x_k的相关系数作为它们的相似性度量

$$r_{jk}=\frac{\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\left[\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)^2\sum_{i=1}^{m}(x_{ik}-\bar{x}_k)^2\right]^{\frac{1}{2}}},\quad j,k=1,2,\cdots,p$$

其中，$\bar{x}_j=\frac{1}{m}\sum_{i=1}^{m}x_{ij}$。对指标(变量)进行聚类分析时，利用相关系数矩阵$R=(r_{ij})_{p\times p}$是最多的。

第七章 聚类分析

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mp} \end{bmatrix}$$

7.1 聚类标准

2. 指标(变量)的相似性度量

（2）夹角余弦

也可以直接利用两变量x_j与x_k的夹角余弦r_{jk}来定义它们的相似性度量，有

$$r_{jk}=\frac{\sum_{i=1}^{m}x_{ij}x_{ik}}{\left(\sum_{i=1}^{m}x_{ij}^2\sum_{i=1}^{m}x_{ik}^2\right)^{\frac{1}{2}}},\quad j,k=1,2,\cdots,p.$$

7.1 聚类标准

2. 指标(变量)的相似性度量

上述2种对于指标（变量）的相似度量方式，均应具有以下两个性质：

ⅰ）$|r_{jk}| \le 1$，对于一切 j,k；

ⅱ）$r_{jk} = r_{kj}$,对于一切 j,k。

$|r_{jk}|$越接近1，x_j与x_k越相关或越相似。$|r_{jk}|$越接近零，x_j与x_k的相似性越弱。

第七章 聚类分析

7.1 聚类标准

3. 类与类之间的相似性度量——样本类G1和G2

对于两个样本类G_1和G_2，可以用下面的一系列方法度量它们间的距离

（1）最短距离法（nearest neighbor or single linkage method）

$$D(G_1,G_2)=\min_{\substack{x_i\in G_1\\ y_j\in G_2}}\{d(x_i,y_j)\},$$

它的直观意义为两个类中最近两点间的距离。

第七章 聚类分析

7.1 聚类标准

3. 类与类之间的相似性度量——样本类G1和G2

（2）最长距离法（farthest neighbor or complete linkage method）

$$D(G_1,G_2)=\max_{\substack{x_i\in G_1\\ y_j\in G_2}}\{d(x_i,y_j)\},$$

它的直观意义为两个类中最远两点间的距离。

（3）重心法（centroid method）

$$D(G_1,G_2)=d(\bar{x},\bar{y}),$$

其中$\bar{x},\bar{y}$分别为G_1,G_2的重心。

第七章 聚类分析

7.1 聚类标准

3. 类与类之间的相似性度量——样本类G1和G2

（4）类平均法（group average method）

$$D(G_1,G_2)=\frac{1}{n_1 n_2}\sum_{x_i\in G_1}\sum_{x_j\in G_2} d(x_i,x_j),$$

它等于G_1,G_2中两两样本点距离的平均，式中n_1,n_2分别为G_1,G_2中的样本点个数。

7.1 聚类标准

3. 类与类之间的相似性度量——样本类G1和G2

（5）离差平方和法（sum of squares method）

若记

$$D_1 = \sum_{x_i \in G_1} (x_i - \bar{x}_1)^T (x_i - \bar{x}_1),$$

$$D_2 = \sum_{x_j \in G_2} (x_j - \bar{x}_2)^T (x_j - \bar{x}_2),$$

$$D_{12} = \sum_{x_k \in G_1 \cup G_2} (x_k - \bar{x})^T (x_k - \bar{x}),$$

其中

$$\bar{x}_1 = \frac{1}{n_1} \sum_{x_i \in G_1} x_i, \ \bar{x}_2 = \frac{1}{n_2} \sum_{x_j \in G_2} x_j, \ \bar{x} = \frac{1}{n_1 + n_2} \sum_{x_k \in G_1 \cup G_2} x_k,$$

则定义

$$D(G_1, G_2) = D_{12} - D_1 - D_2.$$

第七章 聚类分析

7.1 聚类标准

3. 类与类之间的相似性度量——样本类G1和G2

事实上，若G_1, G_2内部点与点距离很小，则它们能很好地各自聚为一类，并且这两类又能够充分分离（即D_{12}很大），这时必然有$D = D_{12} - D_1 - D_2$很大。因此，按定义可以认为，两类G_1, G_2之间的距离很大。离差平方和法最初是由 Ward 在 1936 年提出，后经 Orloci 等人 1976 年发展起来的，故又称为 Ward 方法。

第七章 聚类分析

7.1 聚类标准

3. 类与类之间的相似性度量——指标类G1和G2

类似于样本聚类中，描述类与类间的相似程度最常用的最短距离法、最长距离法等，在指标(变量)聚类问题中，常用的有最长距离法、最短距离法等。

第七章 聚类分析

7.1 聚类标准

3. 类与类之间的相似性度量——指标类G1和G2

（1）指标类与类间的最长距离法

在最长距离法中，定义两类变量的距离为

$$R(G_1,G_2)=\max_{\substack{x_j\in G_1\\ x_k\in G_2}}\{d_{jk}\},$$

其中$d_{jk}=1-|r_{jk}|$或$d_{jk}^2=1-r_{jk}^2$，这时，$R(G_1,G_2)$与两类中相似性最小的两变量间的相似性度量值有关。

7.1 聚类标准

3. 类与类之间的相似性度量——指标类G1和G2

（2）指标类与类间的最短距离法

在最短距离法中，定义两类变量的距离为

$$R(G_1,G_2)=\min_{\substack{x_j\in G_1\\ x_k\in G_2}}\{d_{jk}\},$$

其中$d_{jk}=1-|r_{jk}|$或$d_{jk}^2=1-r_{jk}^2$，这时，$R(G_1,G_2)$与两类中相似性最大的两个变量间的相似性度量值有关。

第七章 聚类分析

7.2 系统聚类法

1. 系统聚类基本思想

系统聚类法是最常用的一种聚类方法，其基本思想是将样品各看成一类，然后定义类与类之间的距离，将距离最短的两类合并为一个新类，再计算新类与其它类之间的距离，将距离最短的两类合并为一个新类，如此下去，直到合并为一个大类为止。

7.2 系统聚类法

2. 系统聚类法的一般步骤

系统聚类法一般步骤如下：

（1）计算样品两两间的距离 d_{ij}，记 $D_{(0)}=(d_{ij})$；

（2）将每个样品各看成一类；

（3）将距离最近的两类合并为一个新类；

（4）计算新类与当前各类之间的距离。若类的个数等于 1，转（5），否则回到（3）；

（5）画聚类图；

（6）决定类的个数和类。

第七章 聚类分析

7.2 系统聚类法

例 7-1 设有四个样品，每个样品只有一个指标，分别是 3、4、7、9．试用最短距离法进行分类．样品之间采用绝对值距离．

解： 首先计算距离矩阵 $D_{(0)}$，四个样品依此记作 G_1，G_2，G_3，G_4．

$D_{(0)}$	G_1	G_2	G_3	G_4
G_1	0			
G_2	1	0		
G_3	4	3	0	
G_4	6	5	2	0

由 $D_{(0)}$ 可以看出 G_1 与 G_2 间距离最短为 1，因此将它们合并为一个新类，记作 $G_5=\{G_1,G_2\}$．

7.2 系统聚类法

计算G_5与G_3, G_4间的距离，得相应的$D_{(1)}$如下：

$D_{(1)}$	$G_5=\{3,4\}$	$G_3=\{7\}$	$G_4=\{9\}$
$G_5=G_1\cup G_2$	0		
G_3	3	0	
G_4	5	2	0

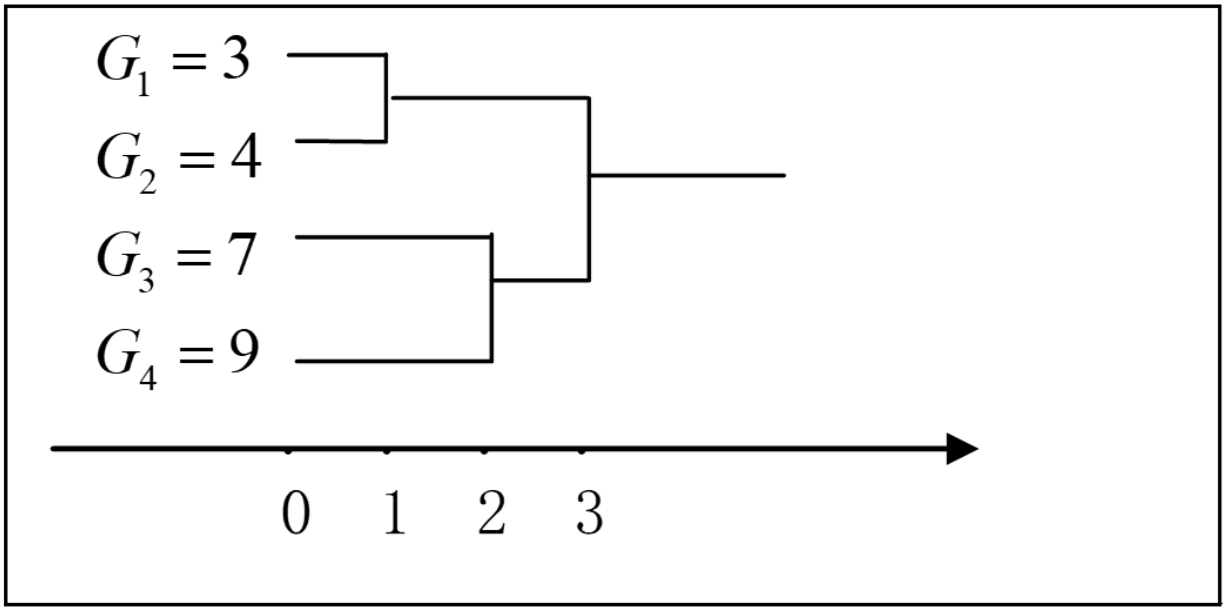
由$D_{(1)}$可以看出G_4与G_3间距离最短，因此可将G_4与G_3合并为一个新类．记作$G_6=\{G_3,G_4\}$．

计算G_6与G_5间的距离，得相应的$D_{(2)}$如下：

$D_{(2)}$	$G_5=\{3,4\}$	$G_6=\{7,9\}$
$G_5=G_1\cup G_2$	0	
$G_6=G_3\cup G_4$	3	0

7.2 系统聚类法

最后将G_6与G_5合并为一类．依上述计算过程画聚类图如下：



由此看来，将四个样品分为两类，即 3 与 4 为一类，7 与 9 为一类比较合适．

7.2 系统聚类法

3 Python 系统(层次)聚类

scipy.cluster.hierarchy 模块的层次聚类函数介绍：

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{bmatrix}.$$

1. distance.pdist

B=pdist(A, metric='euclidean')

用 metric 指定的方法计算$n\times p$矩阵 A（看作n个p维行向量，每行是一个对象的数据）中两两对象间的距离。输出 B 是包含距离信息的长度为$(n-1)\cdot n/2$的向量。可用 distance. squareform 函数将此向量转换为方阵，这样可使矩阵中的(i,j)元素对应原始数据集中对象i和j间的距离。

字符串	含 义
'euclidean'	欧氏距离（缺省值）
'cityblock'	绝对值距离
'minkowski'	Minkowski 距离
'chebychev'	Chebychev 距离
'mahalanobis'	Mahalanobis 距离

7.2 系统聚类法

2.linkage

Z=linkage(B, 'method')使用由'method'指定的算法计算生成聚类树，输入矩阵B为pdist函数输出的$(n-1)\cdot n/2$维距离行向量，'method'可取表字符串值。

字符串	含　义
'single'	最短距离（缺省值）
'average'	无权平均距离
'centroid'	重心距离
'complete'	最大距离
'ward'	离差平方和方法（Ward 方法）

3.fcluster

T = fcluster(Z, t)从 linkage 的输出 Z，根据给定的阈值 t 创建聚类。

4.H＝dendrogram(Z,p)

由linkage产生的数据矩阵Z画聚类树状图。p是结点数，默认值是30。

7.2 系统聚类法

例 7.2 在某地区有 7 个矽卡岩体，对 7 个岩体的三种元素 Cu、W、Mo 作分析的原始数据见表，对这 7 个样品进行聚类。

表　7 个矽卡岩体数据

	1	2	3	4	5	6	7
Cu	2.9909	3.2044	2.8392	2.5315	2.5897	2.9600	3.1184
W	0.3111	0.5348	0.5696	0.4528	0.3010	3.0480	2.8395
Mo	0.5324	0.7718	0.7614	0.4893	0.2735	1.4997	1.9850

按照最短距离聚类时，画出聚类图

7.2 系统聚类法

```
import numpy as np
from sklearn import preprocessing as pp
import scipy.cluster.hierarchy as sch
import matplotlib.pyplot as plt
a=np.loadtxt("data72.txt")
b=pp.minmax_scale(a.T)  #数据规格化
d = sch.distance.pdist(b,'euclidean')  #求对象之间的两两距离向量
dd = sch.distance.squareform(d)  #转换为矩阵格式
print(dd)
z=sch.linkage(d,'single');
s=[str(i+1) for i in range(7)];
plt.rc('font',size=16)
sch.dendrogram(z,labels=s);
plt.show()  #画聚类图
```