03

聚类分析

➢学习目标：

1.了解适合用聚类分析解决的问题;
2.理解对象之间的相似性是如何测量的;
3.区别不同的距离;
4.区分不同的聚类方法及其相应的应用;
5.理解如何选择类的个数;
6.简述聚类分析的局限。

3.1 聚类分析的基本思想

3.1.1 目的

- 聚类分析不仅可以用来对样品进行分类,而且可以用来对变量进行分类。对样品的分类常称为Q型聚类分析,对变量的分类常称为R型聚类分析。与多元分析的其他方法相比,聚类分析的方法还是比较粗糙的,理论上也不算完善,但由于它能解决许多实际问题,所以很受实际研究者重视,同回归分析、判别分析一起称为多元分析的三大方法。

目的

在一些社会、经济问题中,我们面临的往往是比较复杂的研究对象,如果能把相似的样品(或指标)归成类,处理起来就大为方便,如前所述,聚类分析的目的就是把相似的研究对象归成类。

3.1 聚类分析的基本思想

3.1.2 聚类的分类方法

分类方法

✓**系统聚类法：**首先,将n个样品看成n类,然后将性质最接近的两类合并成一个新类,得到n-1类,再从中找出最接近的两类加以合并,变成n-2类,最后所有的样品均在一类,将上述并类过程画成一张图(称为聚类图)便可决定分多少类。

✓**模糊聚类法：**将模糊数学的思想观点用到聚类分析中产生的方法。

✓ **K-均值法：**把样品聚集成k个类的集合，类的个数k可以预先给定或者在聚类过程中确定。该方法可应用于比系统聚类法适用的大得多的数据组。

✓**有序样品的聚类：**n个样品按某种原因(时间、地层深度等)排成次序，必须是次序相邻的样品才能聚成一类。

3.2 相似性度量

- 从一组复杂数据产生一个相当简单的类结构,必然要求进行相关性或相似性度量。当对样品进行聚类时,“靠近” 往往用某种距离来刻画。当对指标聚类时,根据相关系数或某种关联性度量来聚类。
- 每个样品有p个指标,故每个样品可以看成p维空间中的一个点,n个样品就组成p维空间中的n个点,用x_{ij}表示第i个样品的第j个指标,第j个指标的均值和标准差记作和S_j。用d_{ij}表示第i个样品与第j个样品之间的距离。

距离定义

绝对值距离：$d_{ij}(1)=\sum_{k=1}^{p}|x_{ik}-x_{jk}|$

欧氏距离：$d_{ij}(2)=[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2]^{\frac{1}{2}}$

切比雪夫距离：$d_{ij}(\infty)=\max_{1\le k\le p}|x_{ik}-x_{jk}|$

3.2 相似性度量

◆在聚类分析中不仅需要将样品分类,而且需要将指标分类。在指标之间也可以定义距离,更常用的是相似系数,用C_{ij}表示指标i和指标j之间的相似系数。 C_{ij}的绝对值越接近1,表示指标i和指标j的关系越密切; C_{ij}的绝对值越接近0,表示指标i和指标j的关系越疏远。对于间隔尺度,常用的相似系数有夹角余弦和相关系数。

间隔尺度定义

(1)夹角余弦：指标向量$(x_{1i}, x_{2i}, \dots, x_{ni})$和$(x_{1i}, x_{2i}, \dots, x_{ni})$之间的夹角余弦

$$C_{\mathrm{ij}}(1) = \frac{\sum_{k=1}^{n} x_{ki} x_{kj}}{\left[\left(\sum_{k=1}^{n} x_{ki}^2\right)\left(\sum_{k=1}^{n} x_{kj}^2\right)\right]^{\frac{1}{2}}}$$

(2)相关系数：将数据标准化后的夹角余弦

$$C_{\mathrm{ij}}(2) = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{X}_i)(x_{kj} - \bar{X}_j)}{\left[\sum_{k=1}^{n}(x_{ki} - \bar{X}_i)^2 \sum_{k=1}^{n}(x_{kj} - \bar{X}_j)^2\right]^{\frac{1}{2}}}$$

3.3 类和类的特征

3.3.1 类的定义

用G表示类,设G中有k个元素,这些元素用i,j等表示。

定义

定义3.1： T为一给定的阈值,如果对任意的$i,j \in G$,有$d_{ij} \leq T$ (d_{ij}为i,j的距离),则称G为一个类。

定义3.2　对阈值T,如果对每个$i \in G$,有$\frac{1}{k-1}\sum_{j\in G} d_{ij} \leq T$, 则称$G$为一个类。

定义3.3　对阈值T, V,如果$\frac{1}{k(k-1)}\sum_{i\in G}\sum_{j\in G} d_{ij} \leq T$, $d_{ij} \leq T$,对任意的$i,j \in G$,则称G为一个类。

定义3.4　对阈值T,若对任意一个$i \in G$,一定存在$j \in G$,使得$d_{ij} \leq T$,则称G为一个类。

3.3 类和类的特征

3.3.2 类的特征

◆现在类G的元素用$x_1, x_2, \ldots, x_m$表示,M为G内的样品数(或指标数),可以从不同的角度来刻画G的特征。常用的特征有下面三种。

特征

(1)均值$\bar{X}_G$ (或称为G的重心):

$$\bar{X}_G = \frac{1}{m}\sum_{i=1}^{m} X_i$$

(2)样本离差阵及协方差阵:

$$L_G = \sum_{i=1}^{m}(x_i - \bar{X}_G)(x_i - \bar{X}_G)' \qquad \Sigma_G = \frac{1}{n-1} L_G$$

(3)G的直径。它有多种定义,例如

$$D_G = \sum_{i=1}^{m}(x_i - \bar{X}_G)'(x_i - \bar{X}_G) = tr(L_G)$$

$$D_G = max_{i,j \in G} d_{ij}$$

3.3 类和类的特征

3.3.3 类的距离定义

◆在聚类分析中,不仅要考虑各个类的特征,而且要计算类与类之间的距离。令G_p和G_q中分别有k个和m个样品,它们的重心分别为$\bar{x}_p$和$\bar{x}_q$,它们之间的距离用$D(p.q)$表示。下面是一些常用的定义。

定义

(1)最短距离法：$D_k(p,q)=\min\{d_{jl}|j\in G_p,l\in G_q\}$，表示类$G_p$与类$G_q$最邻近的两个样本距离。

(2)最长距离法：$D_s(p,q)=\max\{d_{jl}|j\in G_p,l\in G_q\}$，表示类$G_p$与类$G_q$最邻近的两个样本距离。

(3)类平均法：$D_G(p,q)=\frac{1}{lk}\sum_{i\in G_p}\sum_{j\in G_p}d_{ij}$，表示类$G_p$与类$G_q$任两个样品距离的平均。

(4)重心法：$D_c(p,q)=d_{\bar{x}_p\bar{x}_q}$，表示两个重心$\bar{x}_p$和$\bar{x}_q$间距离。

(5)离差平方和法：$D_p=\sum_{i\in G_p}(x_i-\bar{X}_p)'(x_i-\bar{X}_G)$，$D_q=\sum_{j\in G_q}(x_j-\bar{X}_q)'(x_j-\bar{X}_q)$，

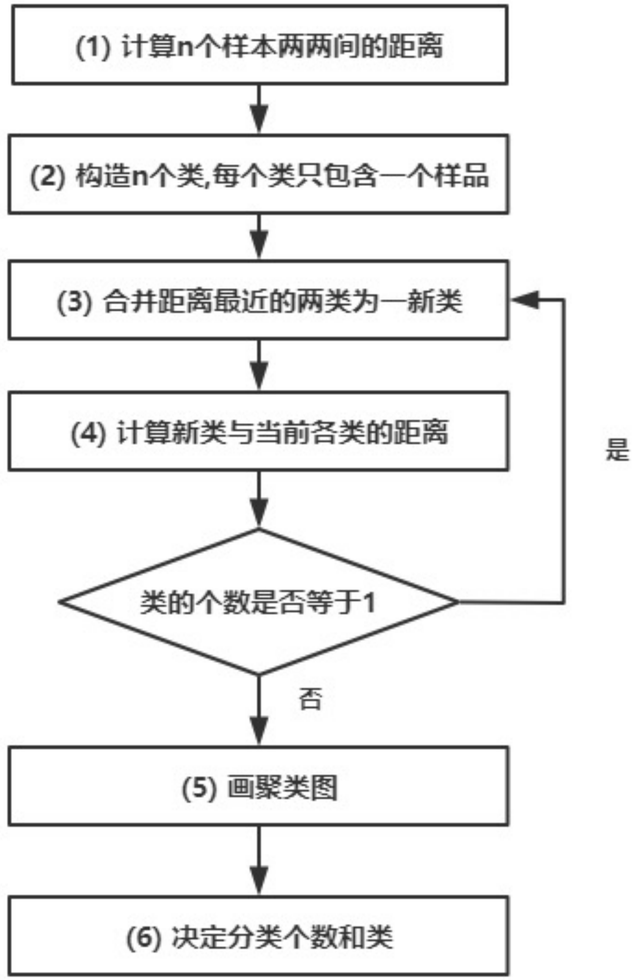
$D_{p+q}=\sum_{i\in G_p\sqcup G_q}(x_j-\bar{x})'(x_j-\bar{x})$，其中$\bar{x}=\frac{1}{k+m}\sum_{i\in G_p\cup G_q}X_i$。

用离差平方和法定义G_p和G_q之间的距离平方为：

$$D_w^2(p,q)=D_{p+q}-D_p-D_q$$

3.4 系统聚类法

系统聚类法是聚类分析诸方法中使用最多的。它包含下列步骤:



3.4 系统聚类法

3.4.1 最短距离法和最长距离法

(1)最短距离法：$D_k(p,q)=\min\{d_{jl}|j\in G_p,l\in G_q\}$

表示类G_p与类G_q最邻近的两个样本距离,计算各类之间的距离,最短的距离归为一类。

主要缺点：它有链接聚合的趋势，因为类与类之间的距离为所有距离中 的最短者，两类合并以后，它与其他类的距离缩小了，这样容易形成一个比较大的类，大部分样品都被聚在一类中，在树状聚类图中会看到一个延伸的链状结构。

(2)最长距离法： $D_s(p,q)=\max\{d_{jl}|j\in G_p,l\in G_q\},$

表示类G_p与类G_q最邻近的两个样本距离,计算各类之间的距离,最短的距离归为一类。

最长距离法克服了最短距离法链接聚合的缺陷，两类合并以后与其他类的距离是原来 两个类中的距离最大者，加大了合并后的类与其他类的距离 。

3.4 系统聚类法

3.4.2 重心法和类平均法

(1)重心法: 若样品之间采用欧氏距离,设某一步将类G_p与类G_q合并成G_r,它们各有$n_p, n_q, n_r(n_r = n_p + n_q)$个样品,它们的重心用$\bar{X}_p, \bar{X}_q, \bar{X}_r$表示,则

$$\bar{X}_r = \frac{1}{n_r}(n_p\bar{X}_p + n_q\bar{X}_q)$$

某一类G_k的中心为$\bar{X}_k$,它与新类G_r的距离为$D_c^2(k,r) = (\bar{X}_k - \bar{X}_r)'(\bar{X}_k - \bar{X}_r)$,经证明重心法的递推公式为:

$$D_c^2(k,r) = \frac{n_p}{n_r}D_c^2(k,p) + \frac{n_q}{n_r}D_c^2(k,q) - \frac{n_p}{n_r}\frac{n_q}{n_r}D_c^2(p,q)$$

(2)类平均法:将两类之间的距离平方定义为这两类元素两两之间的平均平方距离,即

$$D_G^2(k,r) = \frac{n_p}{n_r}D_G^2(k,p) + \frac{n_q}{n_r}D_G^2(k,q)$$

3.4 系统聚类法

3.4.3 离差平方和法

设将n个样品分成k类$G_1, G_2, \ldots, G_k$, 用X_{it}表示类G_t 中的第i个样品 (注意X_{it}是 p维向量),n_t表示类 G_t中的样品个数, $\bar{x}_t$是类G_t 的重心,则在类G_t中的样品的离差平方和为:

$$L_t = \sum_{i=1}^{n_t} (x_{it} - \bar{x}_t)^{'} (x_{it} - \bar{x}_t)$$

整个类内平方和为:

$$L = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (x_{it} - \bar{x}_t)^{'} (x_{it} - \bar{x}_t) = \sum_{t=1}^{k} L_t$$

若将类G_p与类G_q合并成G_r,则类G_k与新类G_r的距离递推公式为:

$$D_w^2(k,r) = \frac{n_p + n_k}{n_r + n_k} D_w^2(k,p) + \frac{n_q + n_k}{n_r + n_k} D_w^2(k,q) - \frac{n_k}{n_r + n_k} D_w^2(p,q)$$

3.4 系统聚类法

3.4.4 分类数的确定

分类准则

聚类分析的目的是要对研究对象进行分类,因此,如何选择分类数成为各种聚类方法中的主要问题之一。实际应用中人们主要根据研究的目的,从实用的角度出发,选择合适的分类数。德穆曼(Demirmen)曾提出根据树状结构图来分类的准则。

◆准则1:任何类都必须在邻近各类中是突出的,即各类重心之间距离必须大。
◆准则2:各类所包含的元素都不应过多。
◆准则3:分类的数目应该符合使用的目的。
◆准则4:若采用几种不同的聚类方法处理,则在各自的聚类图上应发现相同的类。

3.5 均值聚类和有序样品的聚类

3.5.1 均值法(快速聚类法)

K-均值法主要思想：把每个样品聚集到其最近形心(均值)类中

步骤：

(1)把样品粗略分成K个初始类。

(2)进行修改,逐个分派样品到其最近均值类中(通常用标准化数据或非标准化数据计算欧氏距离)。重新计算接受新样品的类和失去样品的类的形心(均值)。

(3)重复第2步,直到各类无元素进出。

注意：

样品的最终聚类在某种程度上依赖于最初的划分或种子点的选择。

为了检验聚类的稳定性,可用一个新的初始分类重新检验整个聚类算法。如果最终分类与原来一样,则不必再行计算;否则,须另行考虑聚类算法。

3.5 均值聚类和有序样品的聚类

3.5.2 有序样品的聚类

假设用$x_1, x_2, \dots, x_n$表示n个有顺序的样品,有序样品的分类结果要求每一类必须i呈:$\{x_i, x_{i+1}, \dots, x_{i+j}\}(i \geq 1, j \geq 0)$。增加了有序这个约束条件,分类方法也会发生改变。

1. 可能的分类数目

对于有序样品，n个样品分成k类的一切可能的分法有：$R'(n,k) = \binom{n-1}{k-1}$

2. 最优分割法(又称Fisher算法)

（1）定义类的直径

设某一类G_{ij}是$\{x_i, x_{i+1}, \dots, x_j\}(j > i)$，均值为$\overline{x}_{ij}$，$\overline{x}_{ij} = \frac{1}{j-i+1}\sum_{l=i}^{j} X_l$。

（2）定义目标函数

将n个样品分成k类,设某一种分法是

$$P(n,k): \{x_{i_1}, x_{i_1+1}, \dots, x_{i_2-1}\}, \{x_{i_2}, x_{i_2+1}, \dots, x_{i_3-1}\}, \dots, \{x_{i_k}, x_{i_{k+1}}, \dots, x_n\}$$

其中分点$1 = i_1 < i_2 < \cdots < i_k \leq i_{k+1} = n+1$。

3.5 均值聚类和有序样品的聚类

3.5.2 有序样品的聚类

则目标函数为：$e[P(n,k)] = \sum_{j=1}^{k} D(i_j, i_{j+1} - 1)$

当n, k固定时, $e[P(n,k)]$越小表示各类的离差平方和越小, 因此要寻 找一种分法$P(n,k)$使目标函数达到极小。

（3）精确最优解的求法

验证递推公式：　　$e[P(n,k)] = min_{2 \le j \le n}\{D(1, j-1)+\}$

$$e[P(n,k)] = min_{k \le j \le n}\{e[P(j-1, k-1)] + D(j, n)\}$$

当我们要分k类时，首先找j_k使上式达到最小，即

$$e[P(n,k)] = e[P(j_k - 1, k-1)] + D(j_k, n)$$

于是$G_k = \{j_k, j_k + 1, \dots, n\}$，然后找$j_k - 1$使它满足

$$e[P(j_k - 1, k-1)] = e[P(j_k - 1, k-2)] + D(j_{k-1}, j_k - 1)$$

得到类$G_k = \{j_{k-1}, \dots, j_k - 1\}$,

采用类似的方法得到所有类$G_1, G_2 \dots G_k$，这就是我们要求的最优解。

3.6 模糊聚类分析

3.6.1 模糊聚类的几个基本概念

(1)特征函数: 对于一个普通集合A,空间中任一元素x,要么$x \in A$,要么$x \notin A$,二者必居其一,这一特征用一个函数表示为:

$$A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

则称A(x)为集合A的特征函数。

(2)隶属函数:当用函数表示隶属度的变化规律时,就叫作隶属函数，即$0 \le A(x) \le 1$。

(3)模糊矩阵的运算法则：如果A和B是n×p和p×m的模糊矩阵,则乘积 C=A·B为n×m阵,其元素为:

$$C_{ij} = \bigvee_{k=1}^{p} (a_{ik} \bigwedge b_{kj}) \quad i = 1, 2, \dots, n ; j = 1, 2, \dots, m$$

符号“V”和“Λ”的含义为:

$$a \vee b = \max(a, b)$$

$$a \wedge b = \min(a, b)$$

3.6 模糊聚类分析

3.6.2 FCM聚类方法

假设将n个样本$\{x_1, x_2, .., x_n\}$划分为c类,其中$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'(i = 1, 2, \dots, n)$。记样本$x_i$属于g类的模糊隶属度为$u_{ig}$,由隶属度构成的矩阵记作$\mathrm{U} = (u_{ig})_{n*c}$,将各类的类中心记作$V = (v_1, v_2, \dots, v_c) \in R^{oc}$,其中矩阵U需要满足条件:

$$\mathrm{U} \sqsubset M = \{U | 0 \le u_{ig} \le 1, \sum_{g=1}^{c} u_{ig} = 1, \forall i\}$$

FCM聚类方法的核心是求解如下优化问题:

$$min_{U,V} \sum_{i=1}^{n} \sum_{g=1}^{c} (u_{ig})^m || x_i - v_g ||_A^2$$

$$m \ge 1; \sum_{g=1}^{c} u_{ig} = 1, \forall i; \left|| x_i - v_g \right||_A^2 = (x_i - v_g)' A (x_i - v_g)$$

以求得使目标函数达到最小的U和V。

3.6 模糊聚类分析

3.6.3 FCM聚类算法

(1)存储样本数据$X = \{x_1, x_2, \ldots, x_n\} \sqsubset R^p$。

(2)设定类的个数c和m的取值、最大迭代次数T、停止迭代的界限ε、矩阵A的形式,定义$E_t = ||V_t - V_{t-1}||$的范数形式。

(3)初始化矩阵V,即设定迭代的初始值$V_0 = (v_{1,0}, v_{2,0}, \ldots, v_{c,0}) \in R^{pc}$。

(4)迭代,从$t = 1$到$t = T$

a.由V_{t-1}计算$U_t = (u_{ig,t})$

$$u_{ig,t} = [\sum_{j=1}^{c} (\frac{||x_i - v_{g,t-1}||_A}{||x_i - v_{j,t-1}||_A})^{\frac{1}{m-1}}]^{-1} \ , i = 1, 2, \ldots, n \ ; g = 1, 2, \ldots, c$$

b.由U_t计算$V_t = \left(v_{1,t}, v_{2,t}, \ldots, v_{c,t}\right)$:

$$v_{g,t} = \frac{\sum_{i=1}^{n} (u_{ig,t})^m x_i}{\sum_{i=1}^{n} (u_{ig,t})^m} \ , g = 1, 2, \ldots, c$$

c. 如果$E_t = ||V_t - V_{t-1}|| \leq \varepsilon$, 停止迭代；否则继续。

(5)输出最后迭代的计算结果U_t,V_t。

3.7 计算步骤与上机实现

➢系统聚类,快速聚类,模糊聚类用R软件操作具体步骤如下:

(1)分析需要研究的问题,确定聚类分析所需的多元变量;

(2)选择对样品聚类还是对指标聚类;

(3)选择合适的聚类方法;

(4)选择所需的输出结果。

为了研究亚洲部分国家和地区的经济水平及相应的人口状况,并对亚洲部分国家和地区进行聚类分析,现选取人均国内生产总值、粗死亡率、粗出生率、城镇人口比重、平均预期寿命和65岁及以上人口比重作为衡量亚洲部分国家和地区经济水平及人口状况的指标,原始数据如表3-7所示(数据来源于世界银行)。

表 3-7　2015 年 15 个亚洲国家和地区经济水平及人口状况

国家和地区	人均国内生产总值(国际元/人)	粗死亡率(‰)	粗出生率(‰)	城镇人口比重(%)	平均预期寿命(年)	65 岁及以上人口比重(%)
阿富汗	1925.17	8.03	33.31	26.7	60.72	2.47
中国内地	14246.86	7.11	12.07	55.61	75.25	9.68
中国香港	56923.49	6.30	8.20	100.00	84.28	15.06
印度	6104.58	7.31	19.66	32.75	68.35	5.62
印度尼西亚	11057.56	7.17	19.58	53.74	69.07	5.17
以色列	36575.94	5.30	21.30	92.14	82.05	11.24
日本	40763.40	10.20	7.90	93.50	83.84	26.34
老挝	5691.26	6.63	26.27	38.61	66.54	3.81
中国澳门	111496.60	4.82	11.68	100.00	80.77	8.99
马来西亚	26950.34	4.98	16.79	74.71	74.88	5.89

续表

国家和地区	人均国内生产总值(国际元/人)	粗死亡率(‰)	粗出生率(‰)	城镇人口比重(%)	平均预期寿命(年)	65 岁及以上人口比重(%)
菲律宾	7387.32	6.77	23.32	44.37	68.41	4.58
沙特阿拉伯	53538.79	3.42	19.69	83.13	74.49	2.86
新加坡	85382.30	4.80	9.70	100.00	82.60	11.68
韩国	34647.07	5.40	8.60	82.47	82.16	13.13
泰国	16340.03	8.03	10.53	50.37	74.60	10.47

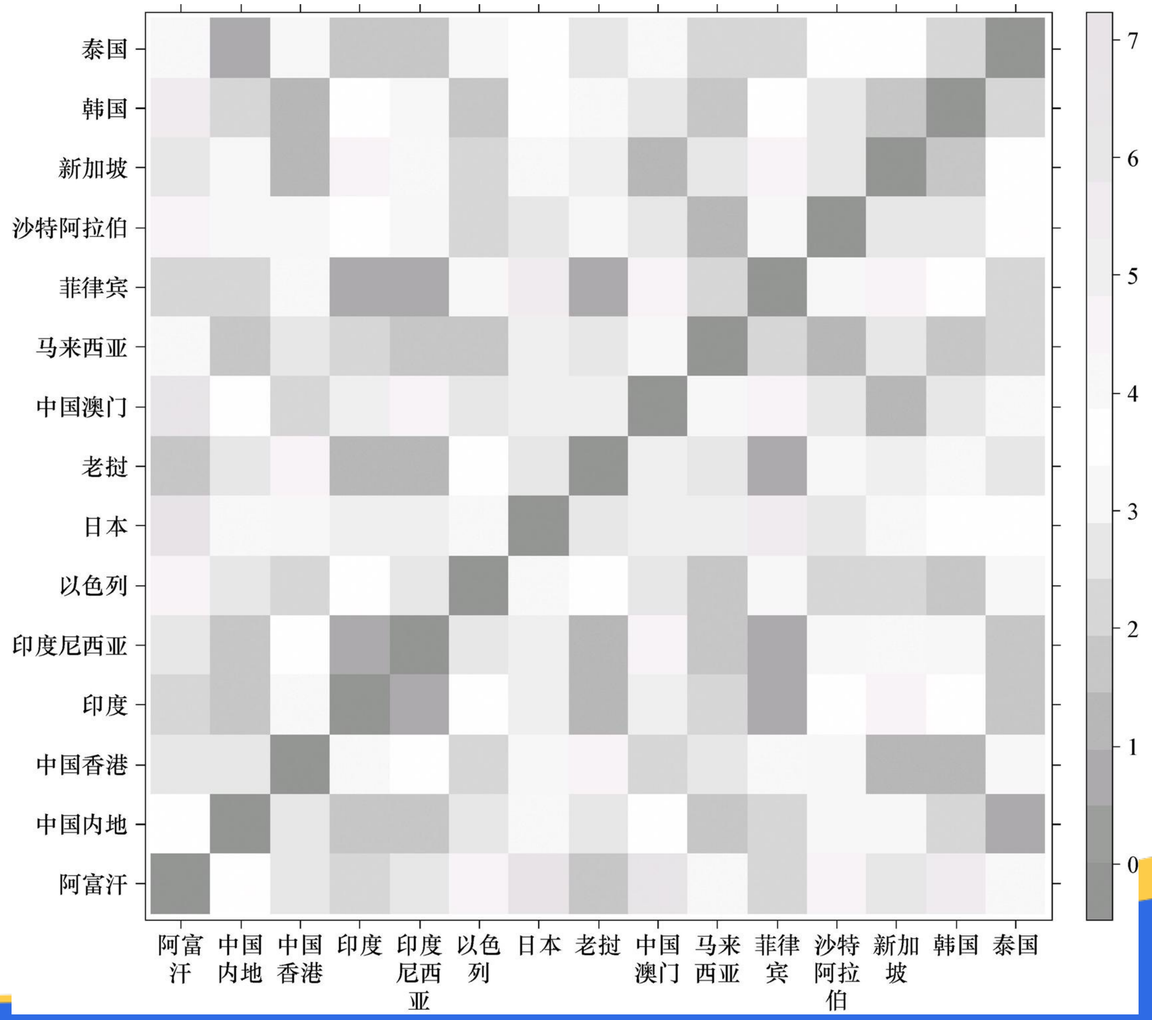
➢系统聚类R软件操作具体步骤如下:

第1步:读入数据。由于数据差异比较大,我们将数据进行标准化。

```
1.> rm(list=ls())
2.> ex3.7 <- read.table("表3-7.txt", head=TRUE, fileEncoding="utf8")
3.> dat37 <- ex3.7[, -1]
4.> rownames(dat37) <- ex3.7[, 1]
5.> head(dat37)
6.                          x1       x2       x3       x4       x5       x6
7.阿富汗                  1925.17    8.03    33.31    26.70    60.72    2.47
8.中国内地               14246.86    7.11    12.07    55.61    75.25    9.68
9.中国香港               56923.49    6.30     8.20   100.00    84.28   15.06
10.印度                   6104.58    7.31    19.66    32.75    68.35    5.62
11.印度尼西亚            11057.56    7.17    19.58    53.74    69.07    5.17
12.以色列                36575.94    5.30    21.30    92.14    82.05   11.24
13.>#数据标准化
14.> dat37_scale <- scale(dat37, center=TRUE, scale=TRUE)
```

第2步, 计算距离矩阵,并且用图显示。图形的右侧列出了一个柱状图,颜色越深说明两个地方的相似度越高。 距离越接近0,说明两个国家或地区越接近。印度和印度尼西亚比较相似。老挝和菲律宾之间的距离比较近。

1.>#计算欧氏矩阵
2.> dat37_dist <-dist(dat37_scale, method = "euclidean", diag = TRUE, upper = FALSE)
3.> library(lattice)
4.>#相似矩阵的图形
5.> lattice::levelplot(as.matrix(dat37_dist), xlab="", ylab="")

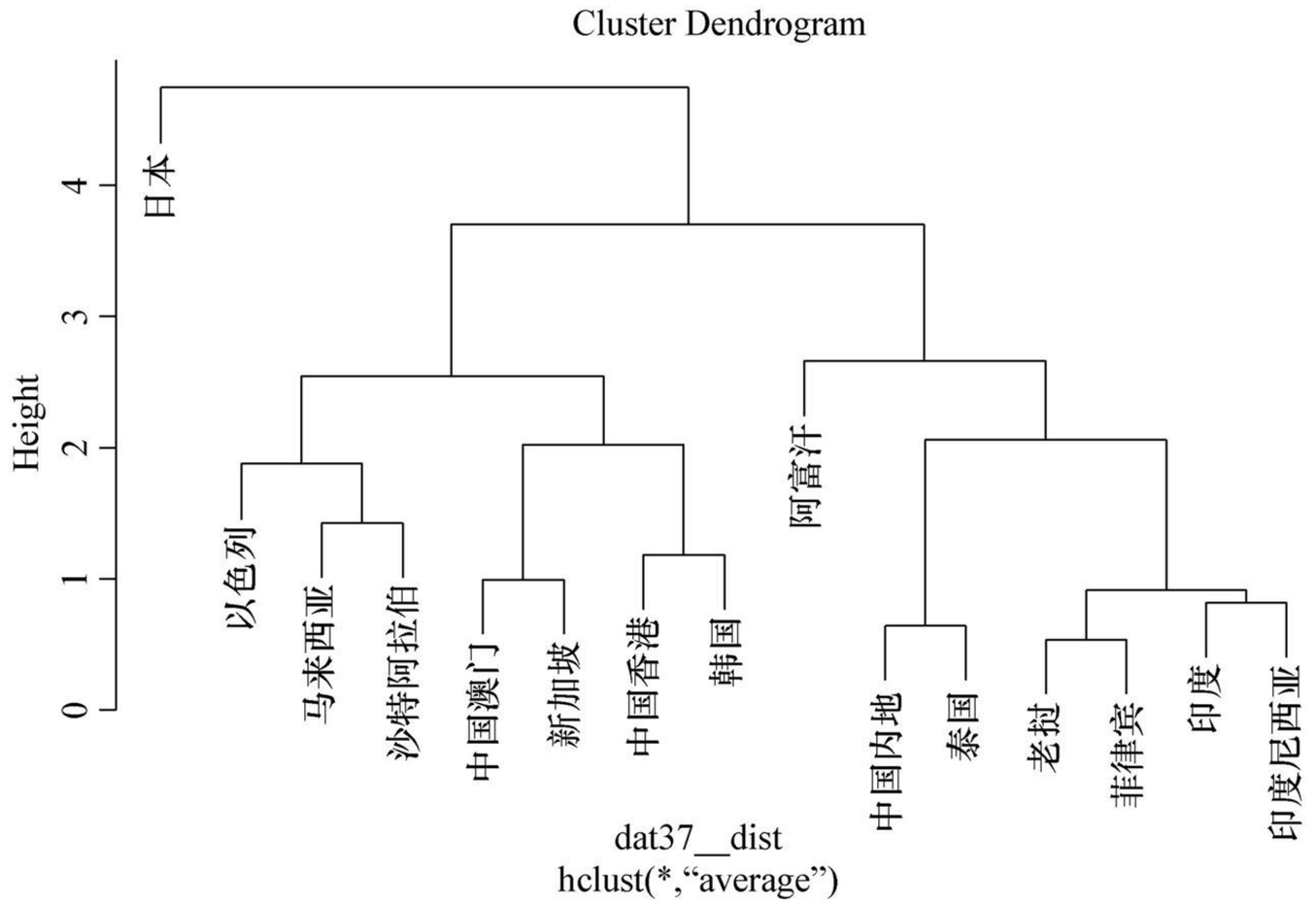


第3步,采用类平均法对数据进行分类,并且输出聚类图。

1.>#类平均法

2.> dat37_average <- hclust(dat37_dist, method="average")

3.> plot(dat37_average)



第4步,确定分类个数。分类数为3的频数最高,有12票。根据投票法原则,3类比较合适,并且给出了具体的结果。
第一类是阿富汗、中国内地、印度、印度尼西亚、老挝、菲律宾和泰国;
第二类是中国香港、以色列、中国澳门、马来西亚、沙特阿拉伯、新加坡和韩国;
第三类是日本。从经济水平和人口状况来看所做的分类,我们发现日本人口老龄化程度极其严重,因此被单独分为一类,第二类对应的国家和地区的经济水平较高,人口老龄化程度较严重,第三类经济水平相对较低,人口老龄化程度相对较轻。

```
1.>#分类的目数
2.> library(NbClust)
3.> fit <- NbClust::NbClust(dat37_scale, distance="euclidean", min.nc = 3,
4.+                                     max.nc = 6, method="average", index="alllong")
5.>#30个分类指标最优个数
6.> fit$Best.nc    #界面有限结束没有给出
7.> table(fit$Best.nc[1, ])
8. 0    3    4    5    6
9. 2   12    4    4    8
10.>#对应的最优分类
11.> fit$Best.partition
12.阿富汗    中国内地    中国香港    印度    印度尼西亚    以色列    日本    老挝    中国澳门    马来西亚
13.     1           1           2          1           1           2          3          1          2           2
14.菲律宾    沙特阿拉伯    新加坡    韩国    泰国
15.     1             2           2          2          1
```

➢快速聚类R软件操作具体步骤如下:

由于原始数据的量纲不一致,因此首先需要对数据进行标准化, 然后对标准化后的数据进行快速聚类。

根据上面系统聚类,我们将分类数设置为3(centers = 3)。

该方法将国家和地区分成三类,个数分别是1,7和7。
第一类是日本,
第二类是中国香港、以色列、中国澳门、马来西亚、沙特阿拉伯、新加坡和韩国;
第三类是阿富汗、中国内地、印度、印度尼西亚、老挝、菲律宾和泰国。这个结果和系统聚类一样。
Cluster means给出了最后分类的中心。

```
1.> fit.k <- kmeans(dat37_scale, centers = 3)
2.> fit.k
3.K-means clustering with 3 clusters of sizes 1, 7, 7
4.Cluster means:
5.          x1         x2         x3         x4         x5         x6
6.1  0.2143909  2.2364261 -1.1422235  0.9423414  1.1768453  2.7924641
7.2  0.7534202 -0.8368224 -0.3772712  0.8234161  0.6776983  0.1140929
8.3 -0.7840475  0.5173330  0.5404459 -0.9580363 -0.8458190 -0.5130164
9.Clustering vector:
10.阿富汗  中国内地  中国香港  印度  印度尼西亚  以色列  日本  老挝  中国澳门  马来西亚
11.    3        3        2      3         3        2     1     3       2        2
12.菲律宾  沙特阿拉伯  新加坡  韩国  泰国
13.    3          2        2     2     3
```

➢模糊聚类R软件操作具体步骤如下:

R软件中可以直接调用进行模糊聚类分析的函数有cmeans()和fanny()。若使用cmeans()函数,需要先加载包el071,使用fanny()函数需要先加载包cluster。

fanny()函数的调用格式为:

fanny(x,k,memb.exp=2,metric=c("euclidean","manhattan","SqEuclidean"),stand=FALSE,maxit=500,tol=1e-15,trace.lev=0,…)

k为类的个数;
memb.exp为隶属指数
metric提供了3种计算样本间差异的方法
stand为逻辑值,它等于TRUE时,将会在计算样本间差异前对数据进行标准化,其标准化方法为各样本值减去样本均值并除以对应变量的平均绝对偏差。

```
1.> library(cluster)   #加载cluster包
2.> fresult<-fanny(dat37_scale,3)   #使用fanny函数进行模糊聚类并将结果保存在fresult中
3.> summary(fresult)   #输出聚类结果
Closest hard clustering:

阿富汗   中国内地   中国香港   印度   印度尼西亚   以色列   日本   老挝   中国澳门   马来西亚
1        2         3         1        1          2        2      1      3          2
菲律宾   沙特阿拉伯   新加坡   韩国   泰国
  1          2          3       3      2
```