

中级应用统计学

聚类分析

张晨峰

华东理工大学商学院

2016年10月19日

4.1 简介



4 聚类分析

主要内容

- 简介
- 相似性度量
- 系统聚类法
- K-均值法

4.1 简介

聚类分析

聚类分析的目的就是把相似的研究对象归成类。

聚类分析的应用

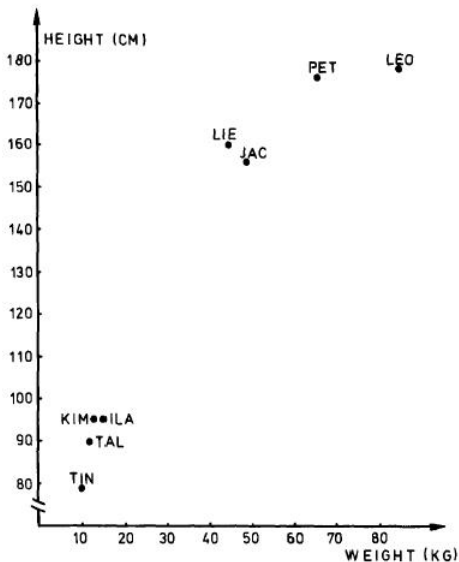
聚类分析广泛应用于自然科学、医学、经济学和市场学等领域。例如，在市场学中，通过对潜在消费者的调查，聚类分析对于建立和描述市场的不同细分是有用的。

4.1 简介

Table 1 Weight and Height of Eight People, Expressed in Kilograms and Centimeters

Name	Weight (kg)	Height (cm)
Ilan	15	95
Jacqueline	49	156
Kim	13	95
Lieve	45	160
Leon	85	178
Peter	66	176
Talia	12	90
Tina	10	78

4.1 简介



4.1 简介

聚类分析的分类

- Q型聚类分析——按照变量对观测值进行分类
- R型聚类分析——按照观测值对变量进行分类

常用的聚类分析方法

- 系统聚类法
- K-均值法

4.1 简介

聚类分析的特性

- 聚类分析前所有个体所属的类别是未知的，类别个数一般也是未知的，分析的依据只有原始数据，可能事先没有任何有关类别的信息可参考。
- 聚类分析主要用于探索性研究，其分析结果可提供多个可能的解，最终解的选择需要研究者的主观判断和后续分析。
- 聚类分析的解完全依赖于研究者所选择的聚类变量，增加或删除一些变量对最终解都可能产生实质性的影响。

4.2 相似性度量

衡量指标

- 距离：对象间差异程度的度量，距离越近，越相似。
- 相似系数：对象间相似系数越大，越相似。

4.2 相似性度量

假设

每个对象有 p 个指标，每个对象可以看成 p 维空间中的一个点， n 个对象就组成 p 维空间中的 n 个点。用 x_{ij} 表示第 i 个对象的第 j 个指标，第 j 个指标的均值和标准差记作 \bar{x}_j 和 S_j 。

4.2 相似性度量

定义

绝对值距离（曼哈顿距离）

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

定义

欧式距离

$$d_{ij} = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{1/2}$$


4.2 相似性度量

定义

切比雪夫距离

$$d_{ij} = \max_{1 \leq k \leq p} |x_{ik} - x_{jk}|$$

4.2 相似性度量

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

国际象棋棋盘上二个位置间的切比雪夫距离是指王从一个位子移至另一个位子需要走的步数。由于王可以往斜前或斜后方向移动一格，因此可以较有效率的到达目的格子。上图是棋盘上所有位置距e6位置的切比雪夫距离。

4.2 相似性度量

定义

明可夫斯基距离

$$d_{ij} = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^q \right]^{1/q}$$

上述距离度量的缺陷

- 距离的值与各指标的量纲有关，任何一个变量计量单位的改变都会使此距离的数值改变，从而使该距离的数值依赖于各变量计量单位的选择。
- 没有考虑各个变量之间的相关性和重要性。

4.2 相似性度量

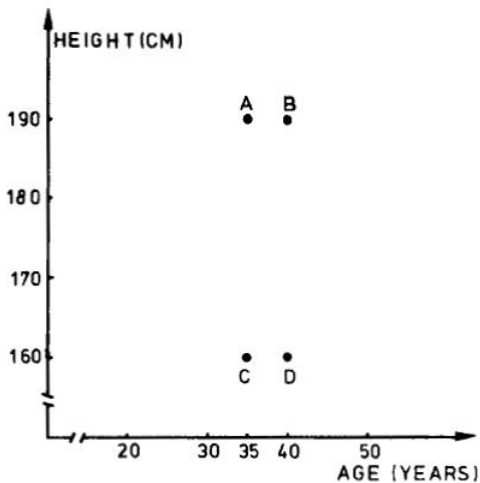


Figure 3 Plot of height (in centimeters) versus age for four people.

4.2 相似性度量

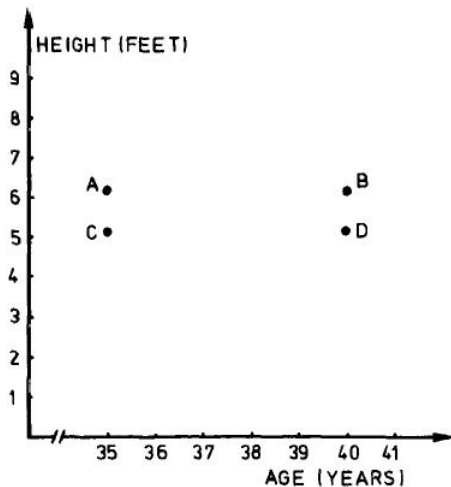


Figure 4 Plot of height (in feet) versus age for the same four people.

4.2 相似性度量

解决方案

- 先对数据标准化，然后计算距离
- 马氏距离

定义

马氏距离

$$d_{X,G}^2 = (X - \mu)' \Sigma^{-1} (X - \mu)$$

$$d_{X,Y}^2 = (X - Y)' \Sigma^{-1} (X - Y)$$

4.2 相似性度量

【例 3—2】 已知一个二维正态总体 G 的分布为：

$$N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right)$$

求点 $A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ 和点 $B = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ 至均值 $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ 的距离。

4.2 相似性度量

$$\Sigma^{-1} = \frac{1}{0.19} \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

从而

$$d_{A\mu}^2(M) = (1, 1) \Sigma^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 0.2/0.19$$

$$d_{B\mu}^2(M) = (1, -1) \Sigma^{-1} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 3.8/0.19$$

如果用欧氏距离，则有

$$d_{A\mu}^2(2) = 2, d_{B\mu}^2(2) = 2$$

4.2 相似性度量

相似系数

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

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

- 相关系数

4.2 相似性度量

表 12.3 11 种语言中的数词

英语 (E)	挪威语 (N)	丹麦语 (Da)	荷兰语 (Du)	德语 (G)	法语 (Fr)	西班牙语 (Sp)	意大利语 (I)	波兰语 (P)	匈牙利语 (H)	芬兰语 (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

4.2 相似性度量

表 12.4 11 种语言中数词的首字母配对频数

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du	3	5	4	10							
G	4	6	5	5	10						
Fr	4	4	4	1	3	10					
Sp	4	4	5	1	3	8	10				
I	4	4	5	1	3	9	9	10			
P	3	3	4	0	2	5	7	6	10		
H	1	2	2	2	1	0	0	0	0	10	
Fi	1	1	1	1	1	1	1	1	1	2	10

4.2 相似性度量

类的特征

类 G 的元素用 x_1, x_2, \dots, x_m 表示，常用的特征有以下三种。

定义

类 G 的均值（或称为重心）

$$\bar{x}_G = \frac{1}{m} \sum_{i=1}^m x_i$$

4.2 相似性度量

定义

类G的样本离差阵及协方差阵

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

$$\Sigma_G = \frac{1}{n-1} L_G$$

4.2 相似性度量

定义

类 G 的直径

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

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

4.2 相似性度量

类之间的距离

令 G_p 和 G_q 分别为两个类，它们分别有 k 个和 m 个对象，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。

- 最短距离法和最常距离法
- 类平均法
- 重心法
- 离差平方和法

4.2 相似性度量

定义

最短距离法和最长距离法

$$D_k(p, q) = \min\{d_{jl} | j \in G_p, l \in G_q\}$$

$$D_k(p, q) = \max\{d_{jl} | j \in G_p, l \in G_q\}$$

4.2 相似性度量

定义

类平均法

$$D_G(p, q) = \frac{1}{lk} \sum_{i \in G_p} \sum_{j \in G_q} d_{ij}$$

定义

重心法

$$D_c(p, q) = d_{\bar{x}_p \bar{x}_q}$$

定义

离差平方和法

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

4.3 分层聚类法(系统聚类法)

基本步骤

- 首先，将 n 个对象看成 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。
- 如此下去，最后所有的对象均在一类。

4.2 相似性度量

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \begin{array}{ccccc} & 1 & 2 & 3 & 4 & 5 \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \left[\begin{array}{ccccc} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{array} \right] \end{array}\end{array}$$

4.2 相似性度量

$$\begin{array}{c}
 (35) \\
 1 \quad 2 \quad 4 \\
 \begin{bmatrix}
 0 & & \\
 \textcircled{3} & 0 & \\
 7 & 9 & 0 \\
 8 & 6 & 5 & 0
 \end{bmatrix}
 \end{array}$$

$$\begin{array}{c}
 (135) \\
 2 \quad 4 \\
 \begin{bmatrix}
 0 & & \\
 7 & 0 & \\
 6 & \textcircled{5} & 0
 \end{bmatrix}
 \end{array}$$

$$\begin{array}{c}
 (135) \quad (24) \\
 (24) \quad \begin{bmatrix}
 0 & \\
 \textcircled{6} & 0
 \end{bmatrix}
 \end{array}$$

4.2 相似性度量

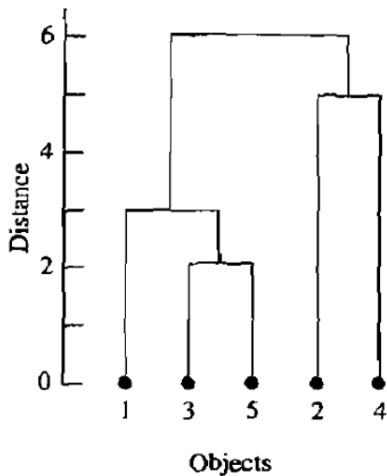


Figure 12.3 Single linkage dendrogram for distances between five objects.

4.2 相似性度量

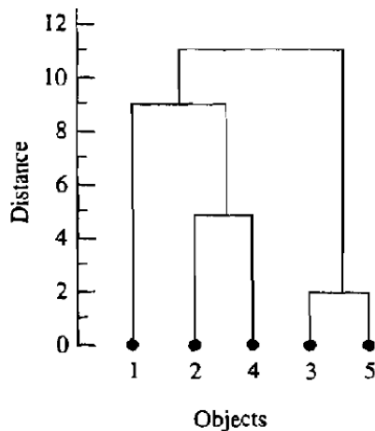


Figure 12.6 Complete linkage dendrogram for distances between five objects.

4.3 系统聚类法

分类的准则

- 各类重心之间的距离必须大
- 各类所包含的元素都不应过多
- 分类的数目应该符合使用的目的
- 若采用几种不同的聚类方法，则在各自的聚类图上应发现相同的类

4.3 系统聚类法

系统聚类法的性质

- 单调性，即 (D_r) 是严格单调上升的
- 空间的浓缩和扩张

定义

设有 A, B 两种系统聚类法，在第 k 步的距离阵记作 A_k 和 B_k ，若 $A_k \geq B_k$ ，则称 A 比 B 扩张或 B 比 A 浓缩。对系统聚类法有如下结论：

$$(K) \leq (G) \leq (S)$$

$$(C) \leq (G) \leq (W)$$

4.4 K-均值法

K-均值法

它是非分层聚类方法。这种聚类方法的思想是把每个对象聚集到其最近重心（均值）类中。

基本步骤

- 首先，把对象粗略分成 K 个初始类。
- 其次，进行修改，逐个分派对象到其最近均值的类中。重新计算接收新对象的类和失去对象的类的重心（均值）。
- 重复上一步，直到各类的元素不再变化。

4.2 相似性度量

Item	Observations	
	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

4.2 相似性度量

Cluster	Coordinates of centroid	
	\bar{x}_1	\bar{x}_2
(AB)	$\frac{5 + (-1)}{2} = 2$	$\frac{3 + 1}{2} = 2$
(CD)	$\frac{1 + (-3)}{2} = -1$	$\frac{-2 + (-2)}{2} = -2$

4.2 相似性度量

Returning to the initial groupings in Step 1, we compute the squared distances

$$d^2(A, (AB)) = (5 - 2)^2 + (3 - 2)^2 = 10$$

if A is not moved

$$d^2(A, (CD)) = (5 + 1)^2 + (3 + 2)^2 = 61$$

$$d^2(A, (B)) = (5 + 1)^2 + (3 - 1)^2 = 40$$

if A is moved to the (CD) group

$$d^2(A, (ACD)) = (5 - 1)^2 + (3 + .33)^2 = 27.09$$

Since A is closer to the center of (AB) than it is to the center of (ACD) , it is not reassigned.

4.2 相似性度量

Continuing, we consider reassigning B . We get

$$d^2(B, (AB)) = (-1 - 2)^2 + (1 - 2)^2 = 10$$

$$d^2(B, (CD)) = (-1 + 1)^2 + (1 + 2)^2 = 9$$

if B is not moved

$$d^2(B, (A)) = (-1 - 5)^2 + (1 - 3)^2 = 40$$

$$d^2(B, (BCD)) = (-1 + 1)^2 + (1 + 1)^2 = 4$$

if B is moved to the (CD) group

Since B is closer to the center of (BCD) than it is to the center of (AB) , B is reas-

4.2 相似性度量

signed to the (CD) group. We now have the clusters (A) and (BCD) with centroid coordinates $(5, 3)$ and $(-1, -1)$ respectively.

We check C for reassignment.

$$d^2(C, (A)) = (1 - 5)^2 + (-2 - 3)^2 = 41$$

if C is not moved

$$d^2(C, (BCD)) = (1 + 1)^2 + (-2 + 1)^2 = 5$$

$$d^2(C, (AC)) = (1 - 3)^2 + (-2 - .5)^2 = 10.25$$

if C is moved to the (A) group

$$d^2(C, (BD)) = (1 + 2)^2 + (-2 + 5)^2 = 11.25$$

Since C is closer to the center of the BCD group than it is to the center of the AC

4.2 相似性度量

Cluster	Squared distances to group centroids			
	Item			
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	0	40	41	89
(<i>BCD</i>)	52	4	5	5