聚類

非監督式學習

謝坤達

jumbokh@gmail.com

聚類

- 聚類分析是分群以找出各子聚類資料背後可能隱藏的特徵、樣型或關聯現象。
- 聚類分析事先並不知道聚類數目,而分群結果的特徵及其所代表的意義僅能事後加以解釋。

聚類應用

- 根據顧客基本資料和交易資料將顧客分群
- 定義並分析不同類型顧客的消費行為模式,以設計定制化的行銷方案
- 將信用卡使用行為分為不同群組樣型,以分析信用卡異常消費的情形,避盜刷所造成的損失。

聚類應用(續)

- 在製造業,可依據機台的特徵、功能等的相似程度,將機台分為可以相互替代和備援(backup)的聚類,以提升作業效率並維持良率(Chien&Hsu,2006)。
- 在網路行銷中,可將性質或特性相仿的網頁予以分類,增快網頁搜索速度,並根據流覽行為和客戶聚類分析作客戶消費行為預測和搭配行銷。

聚類分析

(1)資料準備與分群特徵選取:根據問題特性、資料類型及所選擇的分群演算法等,自搜集的變數中選取具代表性的變數作為分群特徵屬性。

(2)相似度計算:選擇衡量相似度的方式,如距離、相關係數等。

(3)分群演算法:利用分群演算法將資料分組

(4)分群結果評估與解釋:當分群結束後需檢查分群結果是否合理。

K-近鄰演算法 (K-NEAREST NEIGHBORS)

原理: 找到距離最近的K個鄰居→進行投票→決定類別

- 步驟1: 計算距離
- 步驟2: 進行投票
- 步驟3:決定類別

- Q.
- 1. 假設有3個聚類
- A={1,3,6}
- B={2,4}
- C={5,7}
- A與B之間有6個距離:
- $D_{1\&2}=233$, $D_{1\&4}=261$, $D_{3\&4}=169$
- $D_{6\&2}=80$, $D_{6\&4}=104$

觀察值	V1	V2
Y1	14	15
Y2	22	28
Y3	15	18
Y4	20	30
Y5	30	35
Y6	18	20
Y7	32	30

歐氏距離平方

序號	1	2	3	4	5	6	7
1	0	233	10	261	656	41	549
2	233	0	149	8	113	80	104
3	10	149	0	169	514	13	433
4	261	8	169	0	125	104	144
5	656	113	514	125	0	369	29
6	41	80	13	104	369	0	296
7	549	104	433	144	29	296	0

最小距離： $D_{min(C_A,C_B)}=D_{6\&2}=80$

最大距離： $D_{max(C_A,C_B)}=D_{1\&4}=261$

平均距離： $D_{average(C_A,C_B)}=\frac{D_{1\&2}+D_{1\&4}+D_{3\&2}+D_{3\&4}+D_{6\&2}+D_{6\&4}}{6}=166$

中心值距離： $聚類A的中心：\left(\frac{14+15+18}{3}\quad\frac{15+18+20}{3}\right)=\left(\frac{47}{3}\quad\frac{53}{3}\right)$

$聚類B的中心：\left(\frac{22+20}{2}\quad\frac{28+30}{2}\right)=(21,29)$

則，聚類 A 與聚類 B 的歐氏距離為：

$$D_{centroid(C_A,C_B)}=\left(21-\frac{47}{3}\right)^2+\left(29-\frac{53}{3}\right)^2=156.89$$

單一連結法，合併 2 和 4 後的歐氏距離

序號	1	2&4	3	5	6	7
1	0	233	10	656	41	549
2&4	233	0	149	113	80	104
3	10	149	0	514	13	433
5	656	113	514	0	369	29
6	41	80	13	369	0	296
7	549	104	433	29	296	0

最後聚類 AB 與聚類 C 在距離為 104 時合併為一群

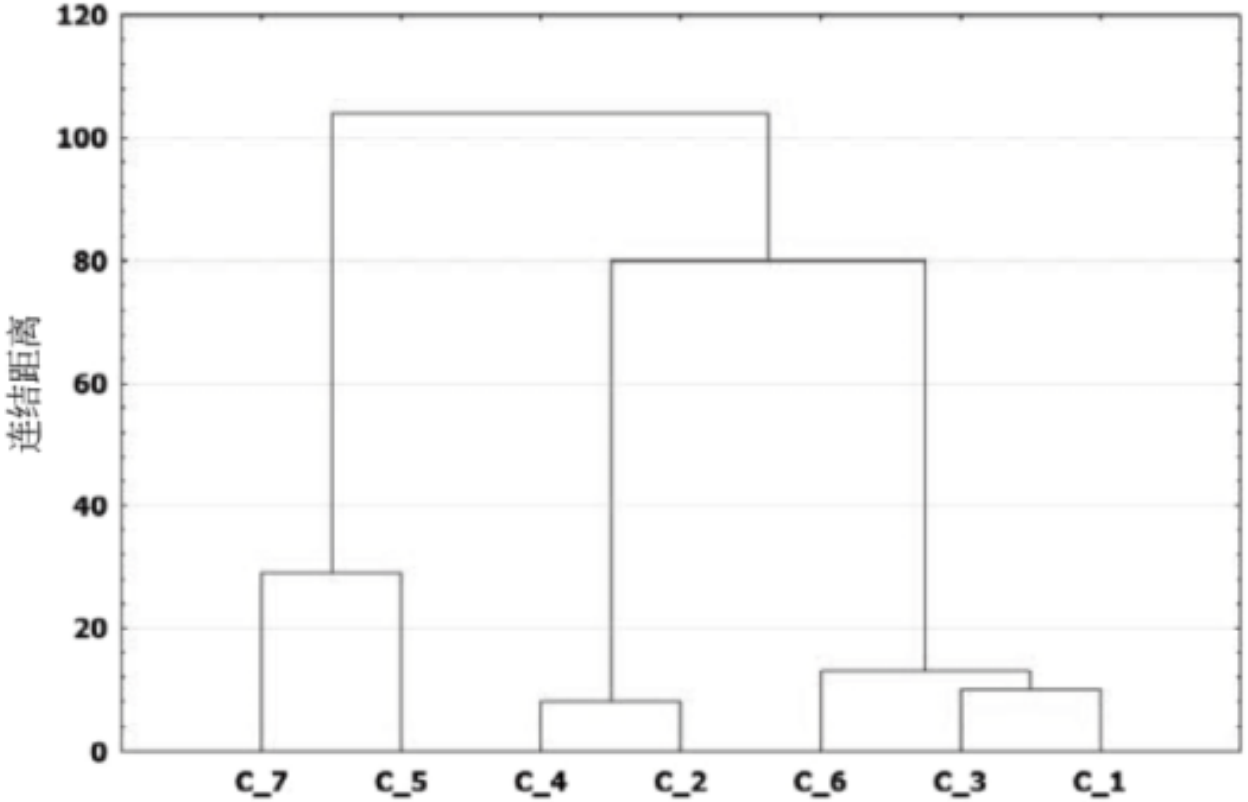

图 6.4　单一连结法树形图

表 6.6　起始聚类中心

聚类	V_1	V_2
A	14	15
B	20	30
C	18	20

表 6.7　K 平均法分群过程(初始重新分配)

序号	与聚类中心的距离			最小距离	分配的聚类
	聚类 A	聚类 B	聚类 C		
1	0	261	41	0	A
2	233	8	80	8	B
3	10	169	13	10	A
4	261	0	104	0	B
5	656	125	369	125	B
6	41	104	0	0	C
7	549	144	296	144	B

表 6.8　聚类中心(第一次重新分配)

聚类	V_1	V_2
A	14.5	16.5
B	26	30.75
C	18	20

表 6.9　K 平均法分群过程(第一次重新分配)

序号	与聚类中心的距离			最小距离	分配的聚类
	聚类 A	聚类 B	聚类 C		
1	2.5	392.06	41	2.5	A
2	188.5	23.56	80	23.56	B
3	2.5	283.56	13	2.5	A
4	212.5	36.56	104	36.56	B
5	582.5	34.06	369	34.06	B
6	24.5	179.56	0	0	C
7	488.5	36.56	296	36.56	B

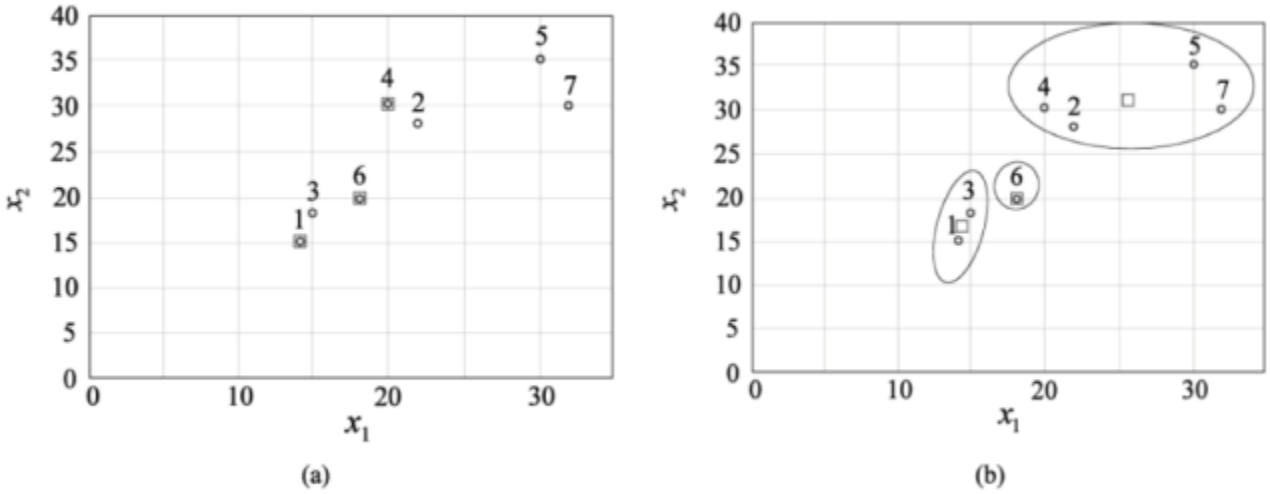

图 6.6　K 平均法对[范例 6.1]的分群过程

總結

- K值的選擇為預測準確程度的關鍵，且最好選擇奇數以避免投票平手的情況
- 交叉驗證
- 選擇合適的距離計算方式

優點:

1. 簡單易懂
2. 資料型態不受限
3. 在多種類別預測有較好的表現

缺點:

1. 計算成本高
2. 資料不平衡時容易產生預測不準確