

# 資料探勘： 概念與方法

---

## — 第七章 —

Jiawei Han

# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結



# 何謂分群分析?

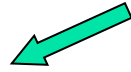
---

- 群組：資料物件集合
  - 相同群組彼此類似
  - 不同群組彼此不相似
- 分群分析
  - 根據資料特性找出像似性並將相似資料進行群組
- 無監督式學習：沒有預設類別
- 傳統應用
  - 作為了解資料分佈的工具
  - 作為其他方法的前處理步驟

# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結



# 資料結構

- 資料矩陣

- (兩種模式)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- 不相似矩陣

- (單種模式)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

# 分群分析資料類型

---

- 區間變數
- 二元變數
- 類別, 順序與比例變數
- 混合類別變數

# 區間變數

- 資料標準化

- 計算均值與絕對偏差：

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

當

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf}).$$

- 計算標準化指標(*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- 使用中值絕對差異比使用標準差來得好

# 物件相似與不相似

- 兩個資料物件的相似與不相似一般式使用距離當作指標
- 普片使用包含明可夫斯基距離(*Minkowski distance*):

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

$i = (x_{i1}, x_{i2}, \dots, x_{ip})$  與  $j = (x_{j1}, x_{j2}, \dots, x_{jp})$  為兩個  $p$ -維度資料物件,  $q$  為一正整數

- 當  $q = 1$ ,  $d$  為曼哈頓距離

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$



## 物件相似與不相似(Cont.)

- 當  $q = 2$ ,  $d$  為歐幾里得距離:

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

- 性質

- $d(i, j) \geq 0$
  - $d(i, i) = 0$
  - $d(i, j) = d(j, i)$
  - $d(i, j) \leq d(i, k) + d(k, j)$
- 也可適用加權距離, 參數式皮爾遜積差相關法, 或其他不相似指標

# 二元變數

- 二元資料列聯表

		物件 $j$		
		1	0	$sum$
物件 $i$	1	$a$	$b$	$a+b$
	0	$c$	$d$	$c+d$
$sum$		$a+c$	$b+d$	$p$

- 對稱二元變數距離指標:

$$d(i, j) = \frac{b+c}{a+b+c+d}$$

- 不對稱二元變數距離指標:

$$d(i, j) = \frac{b+c}{a+b+c}$$

- 賈噶係數( Jaccard coefficient)  
(用於評估不對稱二元變數):

$$sim_{Jaccard}(i, j) = \frac{a}{a+b+c}$$

# 二元變數間不相似

## ■ 範例

表7.2 用二元屬性表示的病患關聯資料表

姓名	性別	感冒	咳嗽	測試 1	測試 2	測試 3	測試 4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N

- 性別 為對稱屬性
- 剩下屬性為不對稱二元變數
- 當 Y 與 P 為 1, 且 N 值為 0

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

# 類別變數

---

- 通用二元變數, 可以包含兩個以上狀態, 例如 紅, 黃, 藍, 綠
- 方法 1: 簡單對應
  - $m$ : 正確個數,  $p$ : 變數全部個數

$$d(i, j) = \frac{p - m}{p}$$

- 方法 2: 使用大量二元變數
  - 對M個類別變數的每個類別變數建立一個新的二元變數

# 順序變數

- 順序變數可以為離散或連續值
- 順序是重要的, 例如, 名次
- 可被視為區間變數
  - 用  $x_{if}$  的名次  $r_{if} \in \{1, \dots, M_f\}$  取代  $x_{if}$
  - 透過取代第  $i$  個物件中的第  $f$  變數, 將變數範圍對應至  $[0, 1]$

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- 使用區間變數方法計算不相似

# 比例變數

- 比例變數：在非線性尺度中的有效度量，在冪次方尺度式是近似的，如  $Ae^{Bt}$  或  $Ae^{-Bt}$
- 方法：
  - 將其視為區間變數—**不是很好選擇!** (為什麼?—尺度會扭曲)
  - 套用邏輯斯轉換

$$y_{if} = \log(x_{if})$$

- 將其視為連續順序資料並將其名次視為區間

# 混合類別變數

- 資料庫會包含所有六種類型變數
  - 對稱二元, 不對稱二元, 類別, 順序, 區間與比例
- 可以使用權重公式來結合它們的效果

$$d(i, j) = \frac{\sum_{f=1}^P \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^P \delta_{ij}^{(f)}}$$

- $f$  為二元或類別:
  - $d_{ij}^{(f)} = 0$  當  $x_{if} = x_{jf}$ , 否則  $d_{ij}^{(f)} = 1$
- $f$  為區間變數: 使用正規化距離
- $f$  為順序或比例
  - 計算名次  $r_{if}$
  - 視  $z_{if}$  為區間變數  $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

# 向量個體

- 向量個體：文件關鍵字, 微陣列基因特性等
- 廣大應用：資訊檢索, 生物分類等
- 餘弦指標

$$s(\vec{X}, \vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{|\vec{X}| |\vec{Y}|},$$

- 變化：谷本系數(Tanimoto coefficient)

$\mathbf{x}^t$  為向量  $\mathbf{x}$  的轉置 (transposition),  $\|\mathbf{x}\|$  為向量  $\mathbf{x}$  的歐幾里得距離

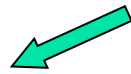
$$s(\vec{X}, \vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{\vec{X}^t \cdot \vec{X} + \vec{Y}^t \cdot \vec{Y} - \vec{X}^t \cdot \vec{Y}},$$



# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結



# 主要分群方法 (I)

---

- 分割方法:
  - 建立許多分割並利用某些條件進行評估, 例如 平方差和最小化
  - 傳統方法: k-means, k-medoids, CLARANS
- 階層方法:
  - 使用某些條件建立資料集階層分割
  - 傳統方法: Diana, Agnes, BIRCH, ROCK, CAMELEON
- 密度式方法:
  - 根據連接與密度函數
  - 傳統方法: DBSACN, OPTICS, DenClue

# 主要分群方法(II)

---

- 方格式方法:
  - 根據多層方格結構
  - 傳統方法: STING, WaveCluster, CLIQUE
- 模型式方法:
  - 對每個群組假設其模型並為該模型尋找最適資料
  - 傳統方法: EM, SOM, COBWEB
- 根據頻繁樣式:
  - 根據頻繁樣式分析
  - 傳統方法: pCluster
- 使用者引導或根據限制式:
  - 根據使用者設定或特定應用限制進行分群
  - 傳統方法: COD (障礙), 限制式分群

# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法 
5. 階層方法
6. 密度式方法
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結

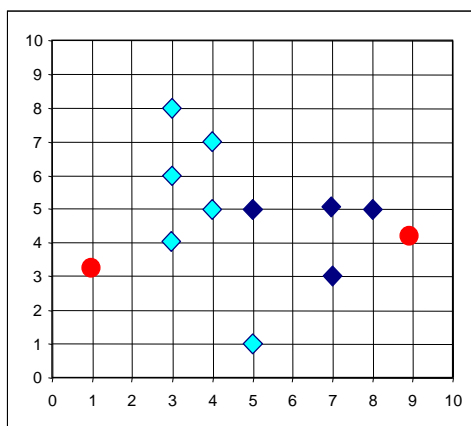
# *K-Means* 分群法

---

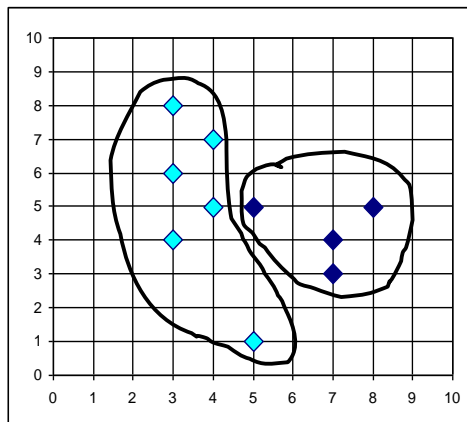
- 給定  $k$ , *k-means* 方法依照下列四個步驟進行分群:
  - 將個體分割成  $k$  個非空白子集合
  - 計算現在分割群組的種子點, 這些種子點為各個群組的中心點 (例如 群組均值點)
  - 將每個個體歸類於最接近的種子點
  - 回到步驟 2, 一直到每個群組個體沒有任何變化

# K-Means 分群法

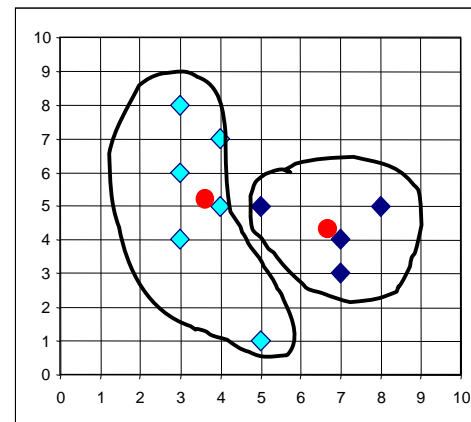
## ■ 範例



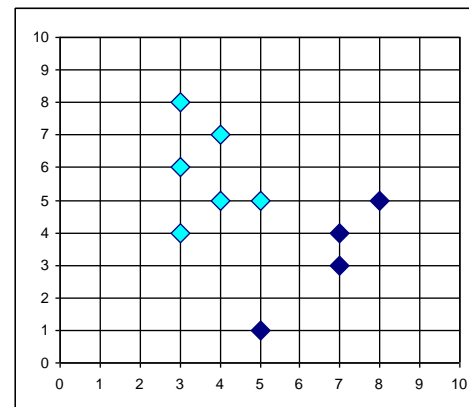
將每個  
個體分  
配至最  
接近中  
心



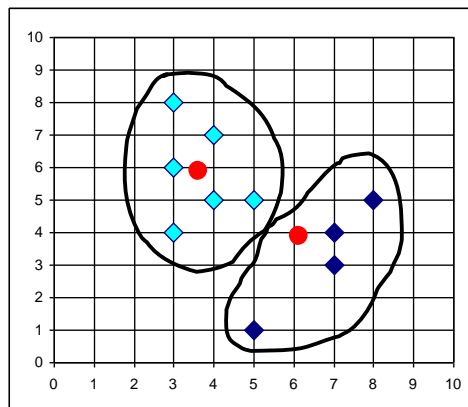
更新群  
組均值



重新分配



更新群  
組均值



重新分配

K=2

任意選擇 K 個體當作  
起始群組中心

# *K-Means* 法建議

- 優勢: 相當有效率:  $O(tkn)$ ,  $n$  為個體數目,  $k$  為群組數目,  $t$  為重複次數. 一般來說  $k, t \ll n$ .
  - 相較於其他方法: PAM:  $O(k(n-k)^2)$ , CLARA:  $O(ks^2 + k(n-k))$
- 建議: 經常找到區域極值. 全域極值可用下列方法找到: 絕對降溫法或基因運算法則
- 弱勢
  - 是用於均值可定義, 那類別資料呢?
  - 需要事先設定群組數目  $k$
  - 無法處理雜訊與離異值
  - 不適合發掘非凸面形狀群組

# *K-Means* 法變異

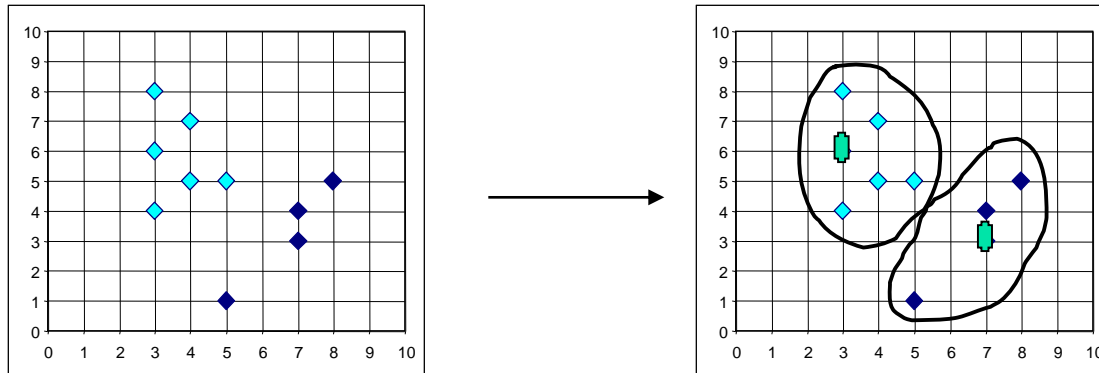
---

- 一些不同 *k-means* 主要差異在
  - 選擇起始  $k$  個均值
  - 不相似計算
  - 計算群組均值的方法
- 處理類別資料: *k-modes* (Huang'98)
  - 用模式取代群組均值
  - 對類別個體使用新的不相似指標
  - 使用頻率式方法來更新群組模式
  - 類別與數值資料混合: *k-prototype* 方法



# K-Means 方法的問題？

- k-means 對離異值非常敏感！
  - 因為具有極大值個體會扭曲資料分佈。
- K-Medoids: 不使用均值作為群組的參考點, 我們使用**medoids**, 它是群組最中心的個體



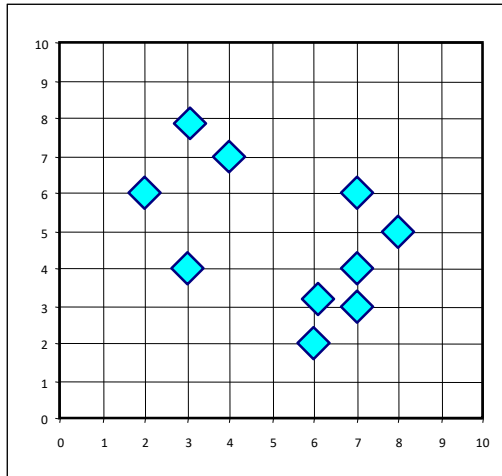
# *K-Medoids* 分群法

---

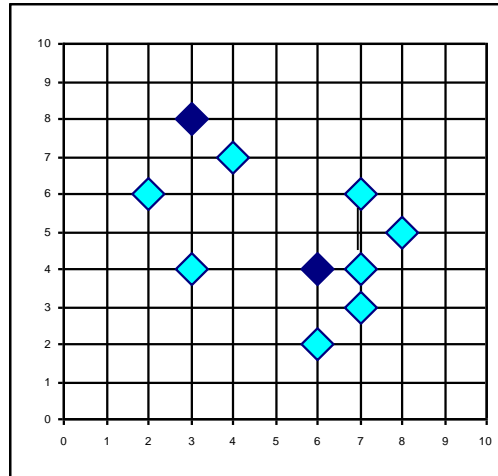
- 尋找代表性個體, 他們稱為群組的 medoids
- *PAM* (Partitioning Around Medoids, 1987)
  - 從一組起始 medoids 開始, 我們重複地用非 medoids 取代 medoids 如果整體群組的距離有改善
  - *PAM* 對小資料集有效, 但是擴展至大資料集時效率並不成等比例
- *CLARA* (Kaufmann & Rousseeuw, 1990)
- *CLARANS* (Ng & Han, 1994): 隨機取樣

# 傳統 K-Medoids 方法 (PAM)

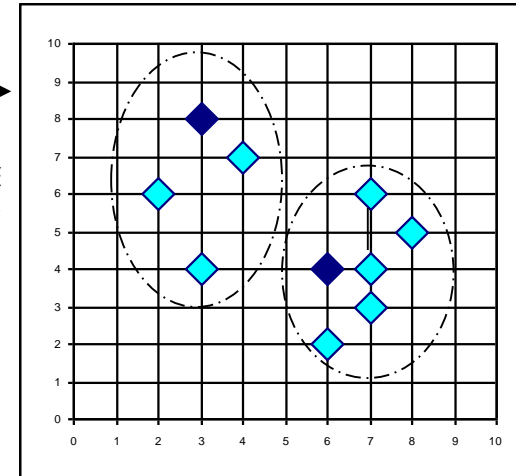
Total Cost = 20



任意選擇  
k 個體作  
為起始  
medoids



分配剩餘  
個體製作  
接近  
medoids

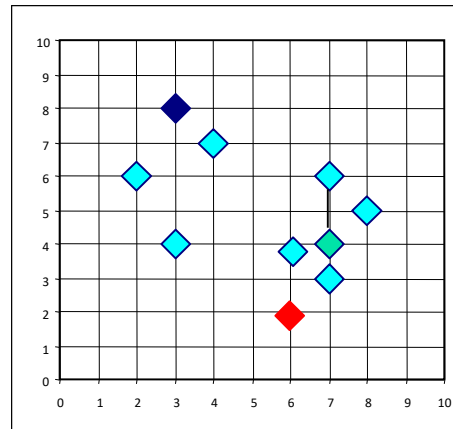


K=2

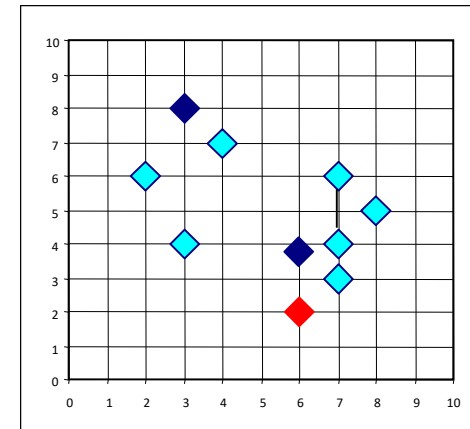
重複執行一直到  
沒有改變為止

當品質有改  
善, 交換  $O$  與  
 $O_{\text{random}}$

Total Cost = 26



計算交換後  
total cost

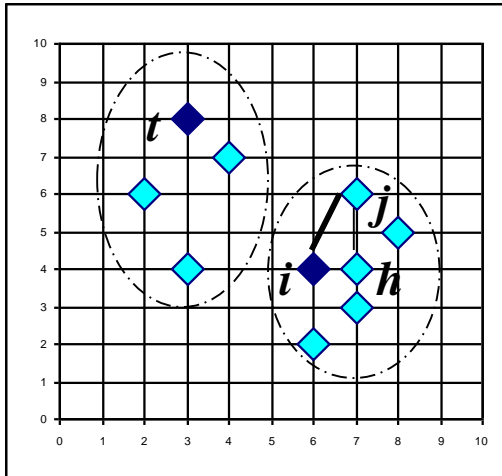


# PAM (Partitioning Around Medoids) (1987)

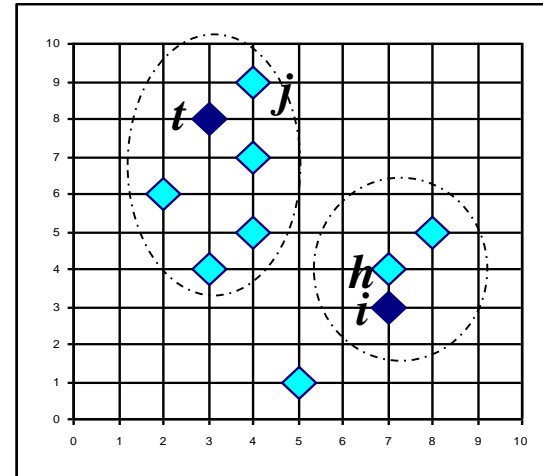
---

- PAM (Kaufman and Rousseeuw, 1987), 內建於 Splus
- 使用真實個體代表群組
  - 任意選擇  $k$  個代表個體
  - 對每一對未選個體  $h$  與已選擇個體  $i$ , 計算全部交換成本  $TC_{ih}$
  - 對每一對  $i$  與  $h$ ,
    - 如果  $TC_{ih} < 0$ ,  $i$  被  $h$  所取代
    - 將每個位選擇個體分配至最相似的代表個體
  - 重覆步驟 2-3 一直到沒有任何改變

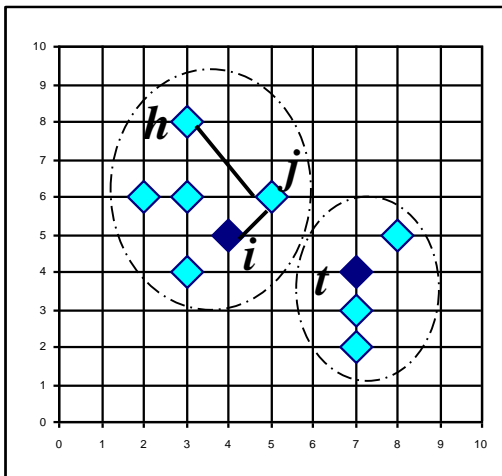
# PAM 分群: 全部交換成本 $TC_{ih} = \sum_j C_{jih}$



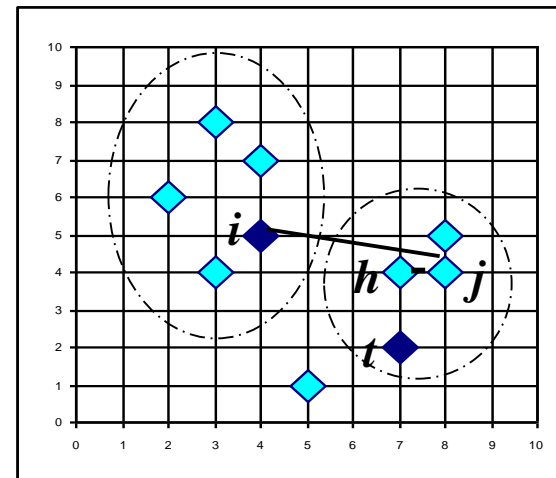
$$C_{jih} = d(j, h) - d(j, i)$$



$$C_{jih} = 0$$



$$C_{jih} = d(j, t) - d(j, i)$$



$$C_{jih} = d(j, h) - d(j, t)$$

# PAM問題?

- 在有雜訊與離異值時, Pam 比 k-means 更健全, 相較於均值, medoid 比較不受離異值或其他極值影響
- Pam 對小資料集有效率, 但擴展至大資料集時效率不成等比例.
  - 每次執行需 $O(k(n-k)^2)$

n 為資料筆數, k 為群組數目

➔ 取樣方法,

CLARA(Clustering LARge Applications)

# CLARA (Clustering Large Applications) (1990)

---

- *CLARA* (Kaufmann and Rousseeuw in 1990)
  - 內建統計分析套件, 如 *S+*
- 它從資料集取出多個樣本, 對每個樣本套用 *PAM*, 並用最最好的分群作為結果
- 優勢: 比 *PAM* 能處理更大資料
- 弱勢:
  - 效能取決於樣本大小
  - 當整體資料有偏差時, 根據樣本分群未必代表最好的分群

# *CLARANS* ("Randomized" CLARA) (1994)

---

- *CLARANS* (A Clustering Algorithm based on Randomized Search) (Ng and Han'94)
- *CLARANS* 隨機取鄰近樣本
- 群組過程像是在圖形進行搜尋, 每個節點都是一個可能答案 (一組  $k$  medoids)
- 如果找到區域極值, *CLARANS* 利用新隨機選取節點尋找新的區域極值
- 它比 *PAM* 與 *CLARA* 更有效並更具量度性
- 透過更專注方法與探索空間結構能進一步改善效能 (Ester et al.'95)



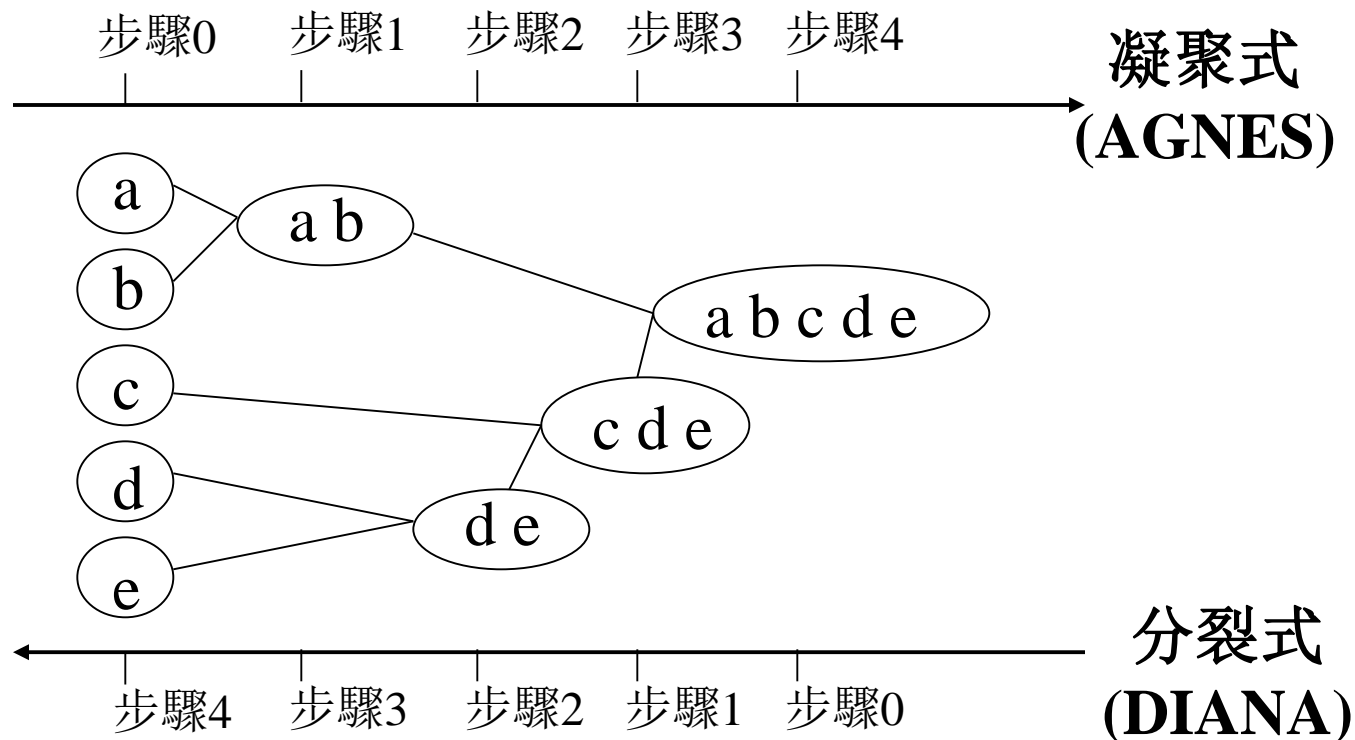
# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法 
6. 密度式方法
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結

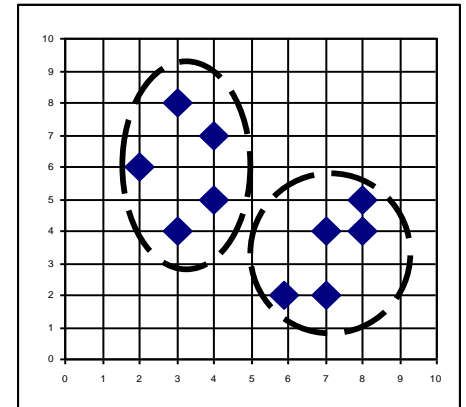
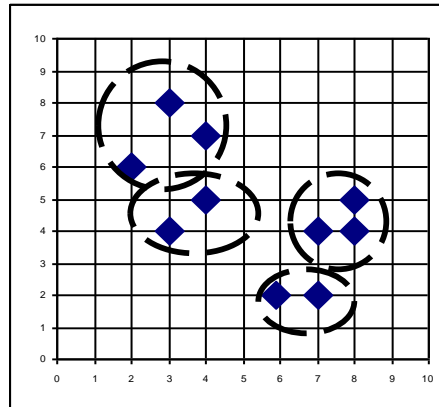
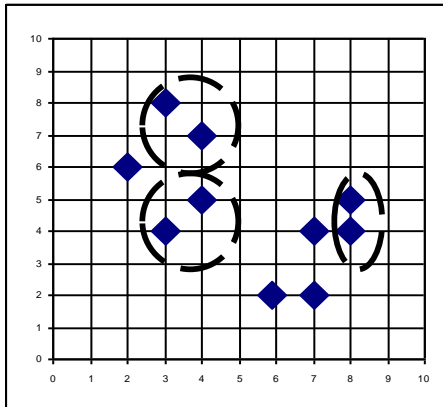
# 階層分群

- 使用距離矩陣作為分群條件。這個方法不需群組數目  $k$  作為輸入, 但是它需要一個結束條件



# AGNES (Agglomerative Nesting)

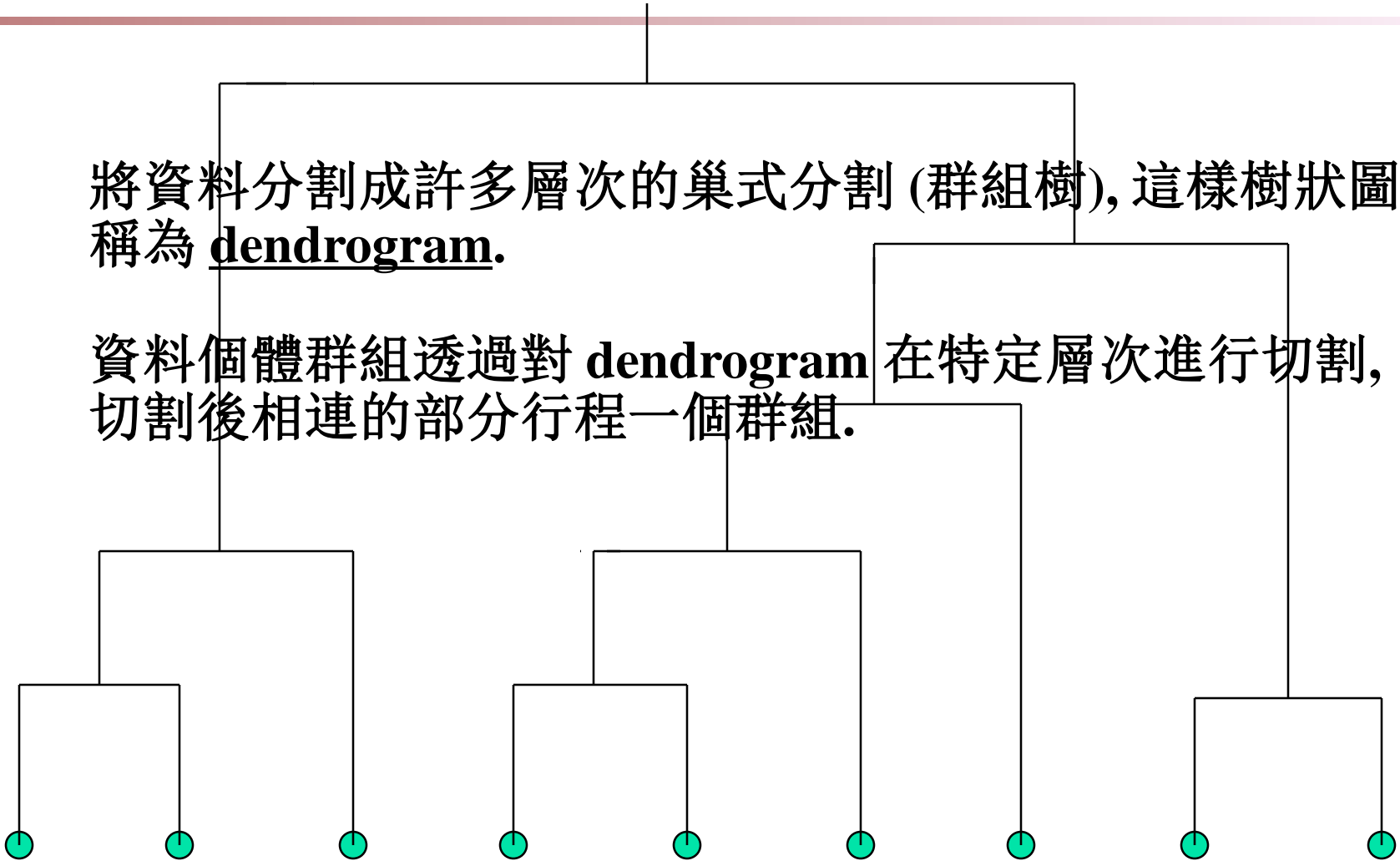
- 由 Kaufmann and Rousseeuw 提出(1990)
- 內建統計分析套件, 如 **Splus**
- 使用單一聯結方法與不相似矩陣.
- 將最相似節點進行合併
- 使用非遞減方式
- 最後所有節點會屬於相同群組



# *Dendrogram*: 顯示群組如何合併

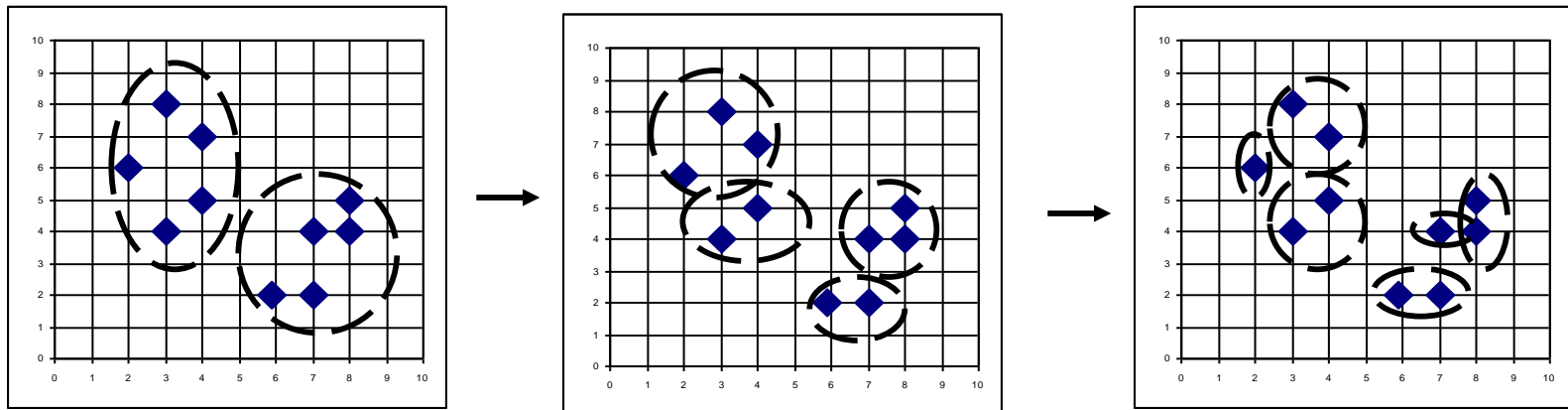
將資料分割成許多層次的巢式分割 (群組樹), 這樣樹狀圖稱為 **dendrogram**.

資料個體群組透過對 **dendrogram** 在特定層次進行切割, 切割後相連的部分行程一個群組.



# DIANA (Divisive Analysis)

- 由 Kaufmann and Rousseeuw 提出(1990)
- 內建統計分析套件, 如 **Splus**
- **AGNES** 的反向順序
- 最終每個節點形成一個群組



# 最近階層分群法

---

- 凝聚式分群法主要缺點
  - 量度性差: 最少需  $O(n^2)$ ,  $n$  為所有的個體
  - 先前執行結果無法回復
- 階層與距離式分群整合
  - BIRCH (1996): 使用 CF-樹與遞增調整子群組品質
  - ROCK (1999): 使用鄰居與連結分析對類別資料進行分群
  - CHAMELEON (1999): 使用動態模型進行階層分群

# BIRCH (1996)

---

- Birch: Balanced Iterative Reducing and Clustering using Hierarchies (Zhang, Ramakrishnan & Livny, SIGMOD'96)
- 遞增式建立 CF (Clustering Feature) 樹, 多階段分群的階層資料結構
  - 階段 1: 讀取資料庫, 在記憶體中建立起始 CF 樹 (一個多層次壓縮資料用於保存資料既有群組結構)
  - 階段 2: 使用任意分群方法來群組 CF-樹 的葉節點
- 具線性度量: 透過一次讀取可找到不錯分群結果, 並且透過額外讀取可以改善分群品質
- 弱勢: 僅能處理數值資料, 並且對資料順序很敏感.

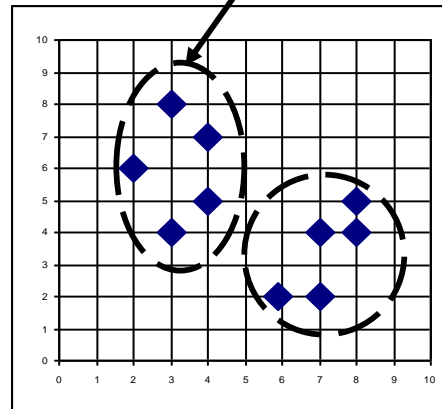
# BIRCH分群特點向量

分群特點:  $CF = (N, \vec{LS}, SS)$

$N$ : 資料點數目

$LS$ :  $\sum_{i=1}^N \vec{X}_i$

$SS$ :  $\sum_{i=1}^N \vec{X}_i^2$



$CF = (5, (16,30), (54,190))$

(3,4)

(2,6)

(4,5)

(4,7)

(3,8)

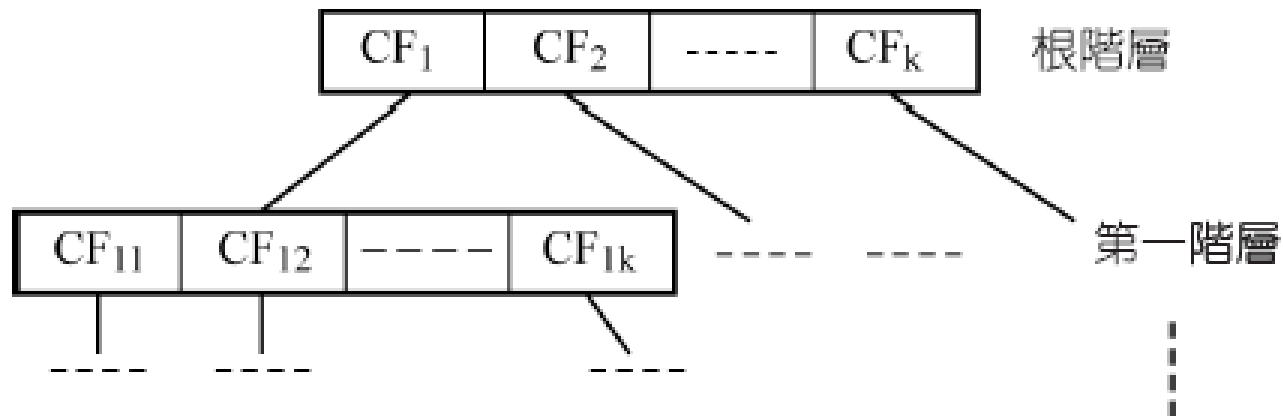


# BIRCH CF-樹

---

- 分群特點：
  - 一個群組的分群特點為該群組的統計總結：從統計的觀點為群組的第二百零、第一與第二動差.
  - 藉由分群特點來表示群組的個體，而這樣的方式能有效率的使用儲存空間
- CF 樹為一高度平衡樹, 它用於儲存一個階層分群的分群特點
  - 非葉節點有後裔或小孩
  - 非葉節點儲存它所有小孩的分群特點總合
- CF 樹有兩個參數
  - 分枝因素：設定最大小孩個數.
  - 界線值：設定葉節點子群組的直徑大小

# CF 樹結構



# 類別資料分群: ROCK 方法

- ROCK: RObust Clustering using linKs
  - S. Guha, R. Rastogi & K. Shim, ICDE'99
- 主要想法
  - 使用聯結來衡量 相似/類似
  - 非距離式
  - 計算複雜度:  $O(n^2 + nm_m m_a + n^2 \log n)$
- 方法: 樣本式分群
  - 隨機取樣
  - 利用連結進行分群
  - 將磁碟資料禁行標示
- 實驗
  - 參議院投票, 蘑菇資料

# ROCK 的相似指標

- 傳統類別指標並不是很好, 例如賈噶係數
- 範例: 兩組 (群) 交易
  - $C_1$ .  $\langle a, b, c, d, e \rangle$ :  $\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \{a, c, e\}, \{a, d, e\}, \{b, c, d\}, \{b, c, e\}, \{b, d, e\}, \{c, d, e\}$
  - $C_2$ .  $\langle a, b, f, g \rangle$ :  $\{a, b, f\}, \{a, b, g\}, \{a, f, g\}, \{b, f, g\}$
- 賈噶係數會導致錯誤分群結果
  - $C_1$ : 0.2 ( $\{a, b, c\}, \{b, d, e\}$ ) to 0.5 ( $\{a, b, c\}, \{a, b, d\}$ )
  - $C_1$  &  $C_2$ : could be as high as 0.5 ( $\{a, b, c\}, \{a, b, f\}$ )
- 根據賈噶係數相似函數:

- 例  $T_1 = \{a, b, c\}, T_2 = \{c, d, e\}$ 
$$Sim(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

$$Sim(T_1, T_2) = \frac{|\{c\}|}{|\{a, b, c, d, e\}|} = \frac{1}{5} = 0.2$$

# ROCK 連結指標

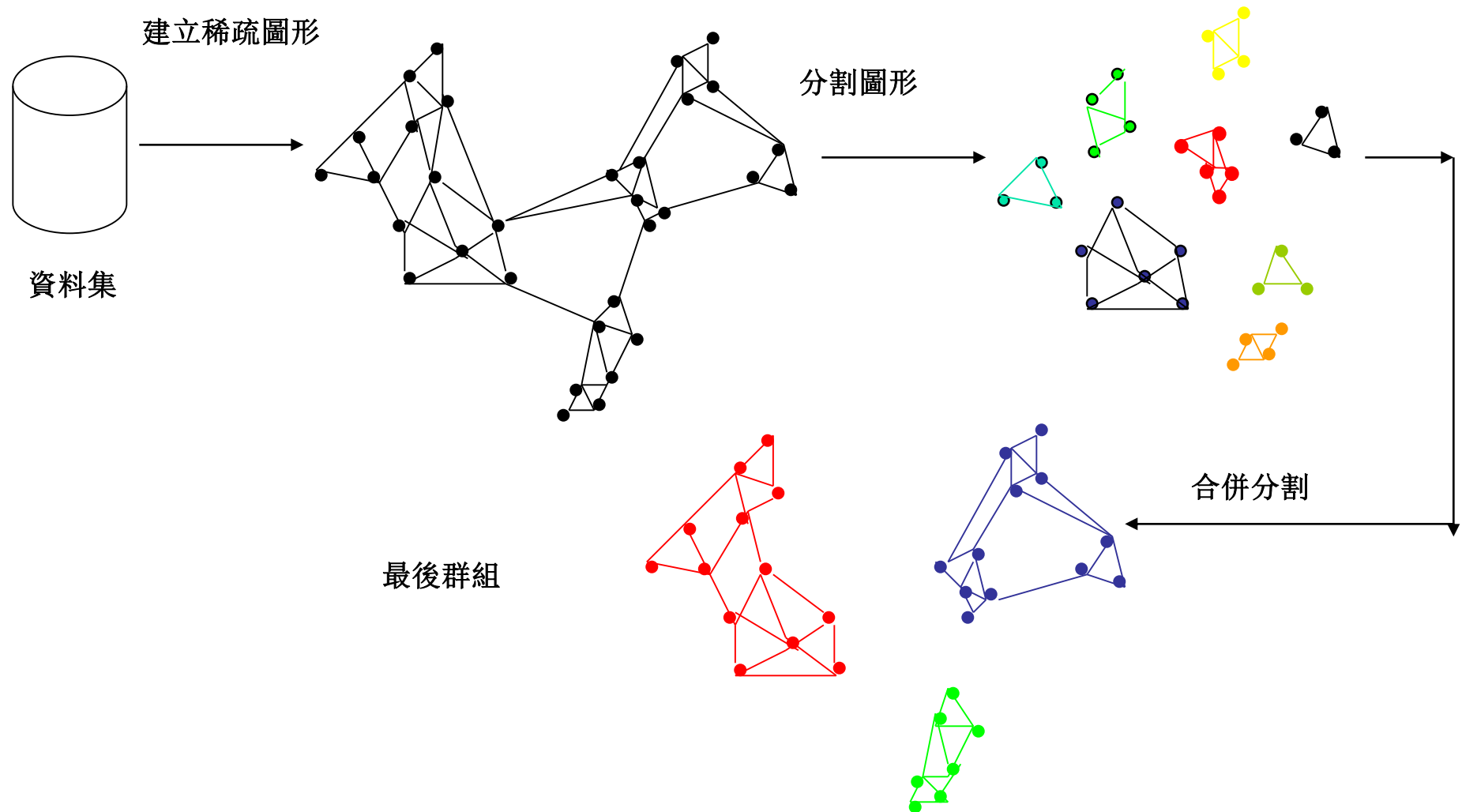
- link: 一般(共同)鄰居個數
  - $C_1 \langle a, b, c, d, e \rangle$ :  $\{a, b, c\}$ ,  $\{a, b, d\}$ ,  $\{a, b, e\}$ ,  $\{a, c, d\}$ ,  $\{a, c, e\}$ ,  $\{a, d, e\}$ ,  $\{b, c, d\}$ ,  $\{b, c, e\}$ ,  $\{b, d, e\}$ ,  $\{c, d, e\}$
  - $C_2 \langle a, b, f, g \rangle$ :  $\{a, b, f\}$ ,  $\{a, b, g\}$ ,  $\{a, f, g\}$ ,  $\{b, f, g\}$
- 當  $T_1 = \{a, b, c\}$ ,  $T_2 = \{c, d, e\}$ ,  $T_3 = \{a, b, f\}$ 
  - $\text{link}(T_1, T_2) = 4$ , 因為他們有 4 個共同鄰居
    - $\{a, c, d\}$ ,  $\{a, c, e\}$ ,  $\{b, c, d\}$ ,  $\{b, c, e\}$
  - $\text{link}(T_1, T_3) = 3$ , 因為他們有 3 個共同鄰居
    - $\{a, b, d\}$ ,  $\{a, b, e\}$ ,  $\{a, b, g\}$
- 因此連結指標比賈噶係數好

# CHAMELEON: Hierarchical Clustering Using Dynamic Modeling (1999)

---

- CHAMELEON: by G. Karypis, E.H. Han, and V. Kumar'99
- 根據動態模型衡量相似度
  - 兩個群組會合併當兩個群組間的互相連接與接近度高度相對於群組內部的互相連接與群組內項目的接近度
  - **Cure** 忽略個體互相連接資訊, **Rock** 忽略兩群組接近度資訊
- 兩階段方法
  1. 使用圖形分割方法: 將個體群組成大量小的子群組
  2. 使用凝聚式階層分群方法: 重複合併子群組來尋找真正群組

# CHAMELEON架構



# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法 
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結



# 密度式分群方法

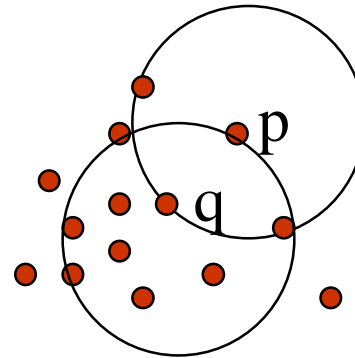
---

- 密度式分群 (區域分群條件), 如密度相連點
- 主要特色:
  - 發覺任意形狀群組
  - 處理雜訊
  - 讀取一次
  - 需要密度參數作為結束條件
- 有趣研究:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

# 密度式分群：基本概念

- 兩個參數：
  - **Eps**: 鄰進區域最大半徑
  - **MinPts**: 在 Eps-鄰近區域最少需要資料點
- $N_{Eps}(p)$ :  $\{q \in D \mid dist(p,q) \leq Eps\}$
- **直接密度可到達**: 相較於  $Eps, MinPts$ , 資料點  $p$  為資料點  $q$  的直接密度可到達當
  - $p$  屬於  $N_{Eps}(q)$
  - 核心點條件:

$$|N_{Eps}(q)| \geq MinPts$$



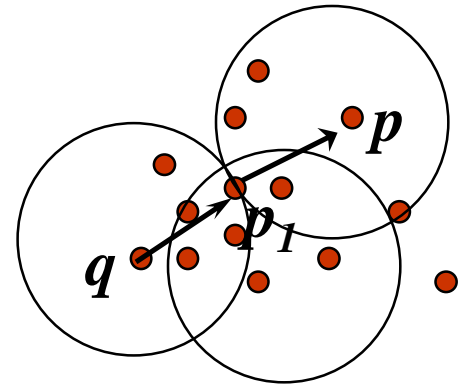
MinPts = 5

Eps = 1 cm

# 密度可到達與密度相連

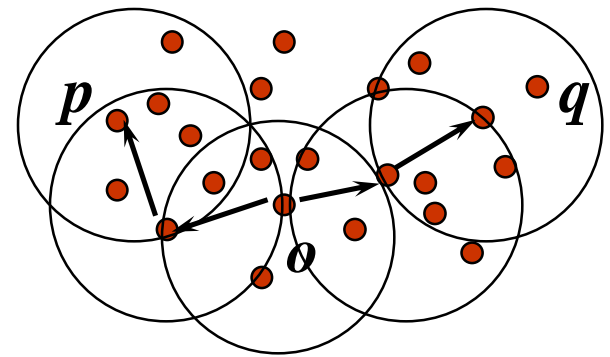
## ■ 密度可到達:

- 相對於  $Eps$ ,  $MinPts$ ,  $p$  為  $q$  的密度可到達當存在一系列點  $p_1, \dots, p_n$ ,  $p_1 = q$ ,  $p_n = p$  使得  $p_i$  可以直接密度到達  $p_{i+1}$



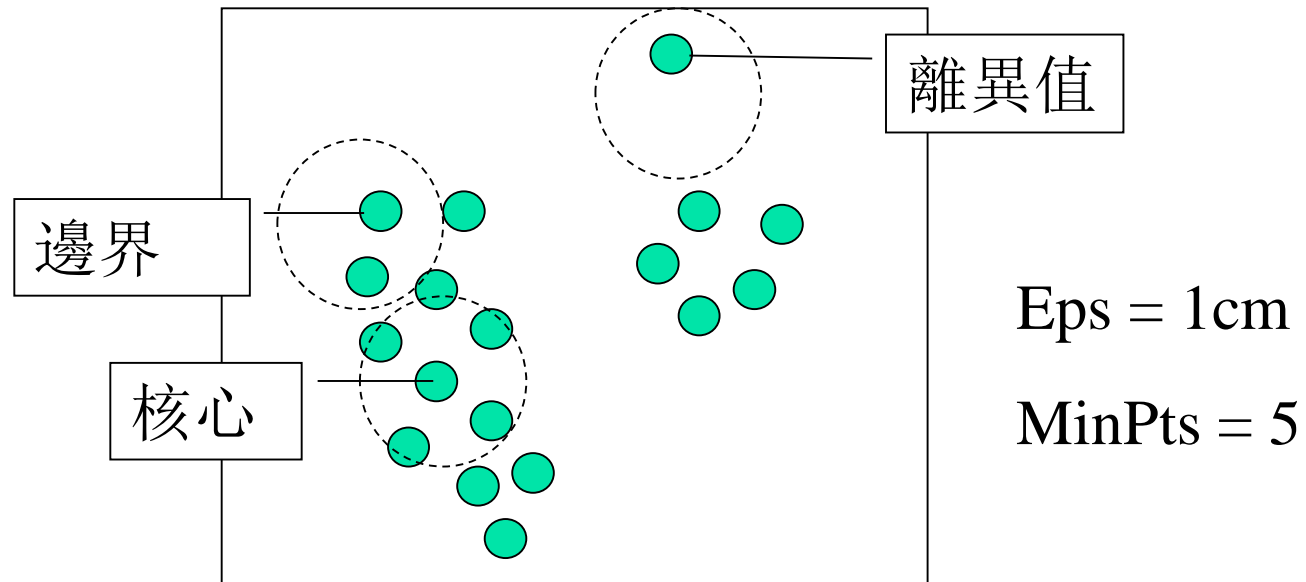
## ■ 密度連接

- $p$  與  $q$  密度相連當存在一個個體  $o$ , 並且在  $Eps$ ,  $MinPts$  條件下使得個體  $p$  與  $q$  為個體  $o$  的直接密度可到達



# DBSCAN: Density Based Spatial Clustering of Applications with Noise

- 由密度式連接分析來進行群組:群組為高密度區域



# DBSCAN: 運算法則

---

- 任意選擇一點  $p$
- 根據  $Eps$  與  $MinPts$  找出所有點  $p$  的密度可到達
- 如果  $p$  為核心, 則形成一個群組.
- 如果  $p$  為邊界點, 則沒有任何點為  $p$  密度可到達, 所以 DBSCAN 選擇資料庫中下一個資料點
- 重複動作直到所有的資料點都被處理過.

# OPTICS: 群組排序法 (1999)

---

- **OPTICS:**透過資料點排序來找出群組結構
  - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
  - 計算一個擴大的群組順序,這個順序代表資料中密度式群組的架構
  - 它所包含的訊息等同於透過許多參數所產生的密度式群組
  - 自動與互動式的群組分析
  - 可使用圖式技巧表示

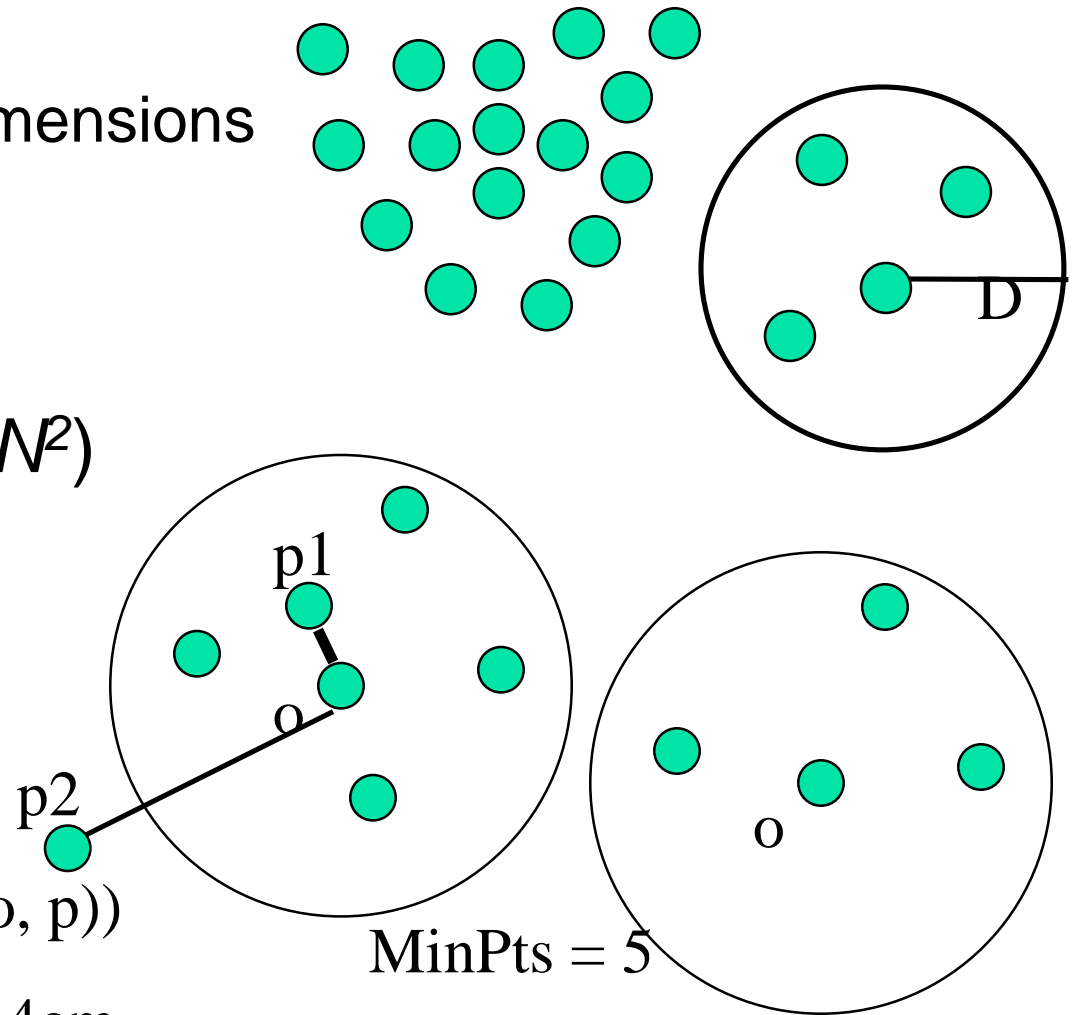
# OPTICS: Some Extension from DBSCAN

- Index-based:
  - $k$  = number of dimensions
  - $N = 20$
  - $p = 75\%$
  - $M = N(1-p) = 5$

■ Complexity:  $O(kN^2)$

- Core Distance

- Reachability Distance



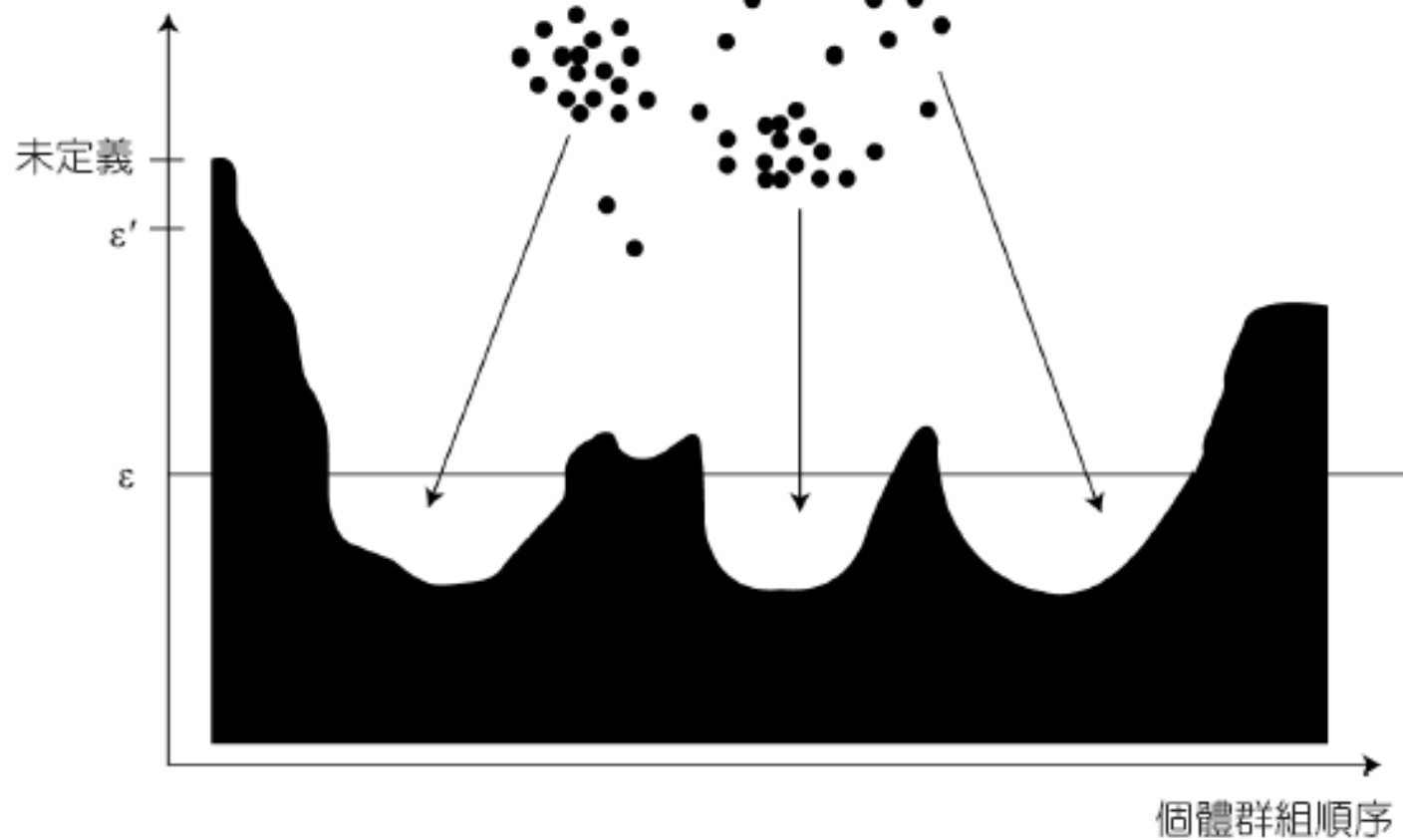
Max (core-distance (o), d (o, p))

$r(p1, o) = 2.8\text{cm}$ .  $r(p2, o) = 4\text{cm}$

MinPts = 5

$\epsilon = 3\text{ cm}$

可到達距離





# DENCLUE: 使用密度分佈函

- DENSity-based CLUstEring by Hinneburg & Keim (KDD'98)

- 使用密度分佈函數:

$$f_{Gaussian}(x, y) = e^{-\frac{d(x, y)^2}{2\sigma^2}}$$

$$f_{Gaussian}^D(x) = \sum_{i=1}^N e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$

- 主要特色

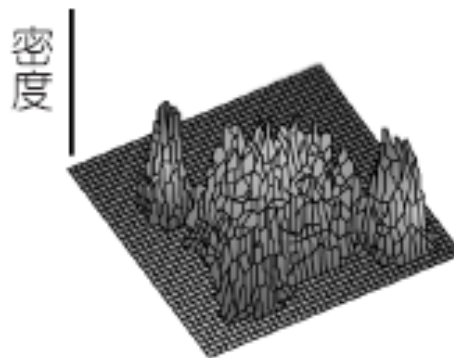
$$\nabla f_{Gaussian}^D(x, x_i) = \sum_{i=1}^N (x_i - x) \cdot e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$

- 具有紮實的數學基礎
- 對具大量雜訊的資料有好的群組特性
- 可以用簡約的數學來描述高維度任意形狀的群組
- 比現有方法快(e.g., DBSCAN)
- 需要大量參數

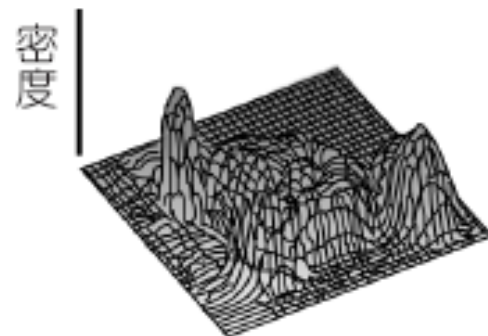
# 密度函數



(a) 資料集

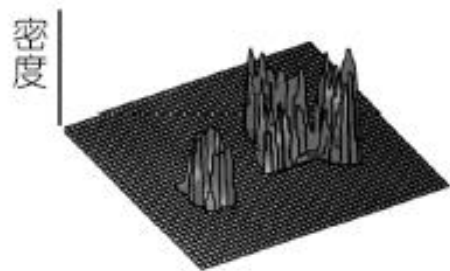


(b) 方波



(c) 高斯

# 中心定義與任意



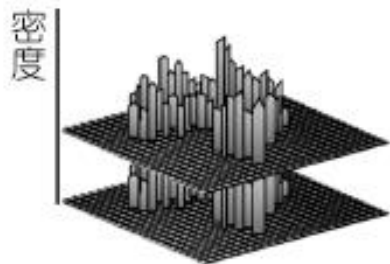
$\sigma = 0.2$



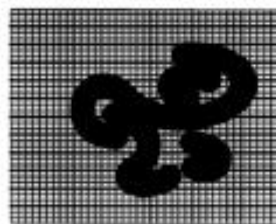
$\sigma = 0.6$



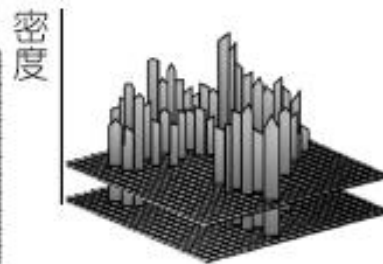
$\sigma = 1.5$



$\xi = 2$



$\xi = 2$



$\xi = 1$



$\xi = 1$

# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法
7. 方格式方法 
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結

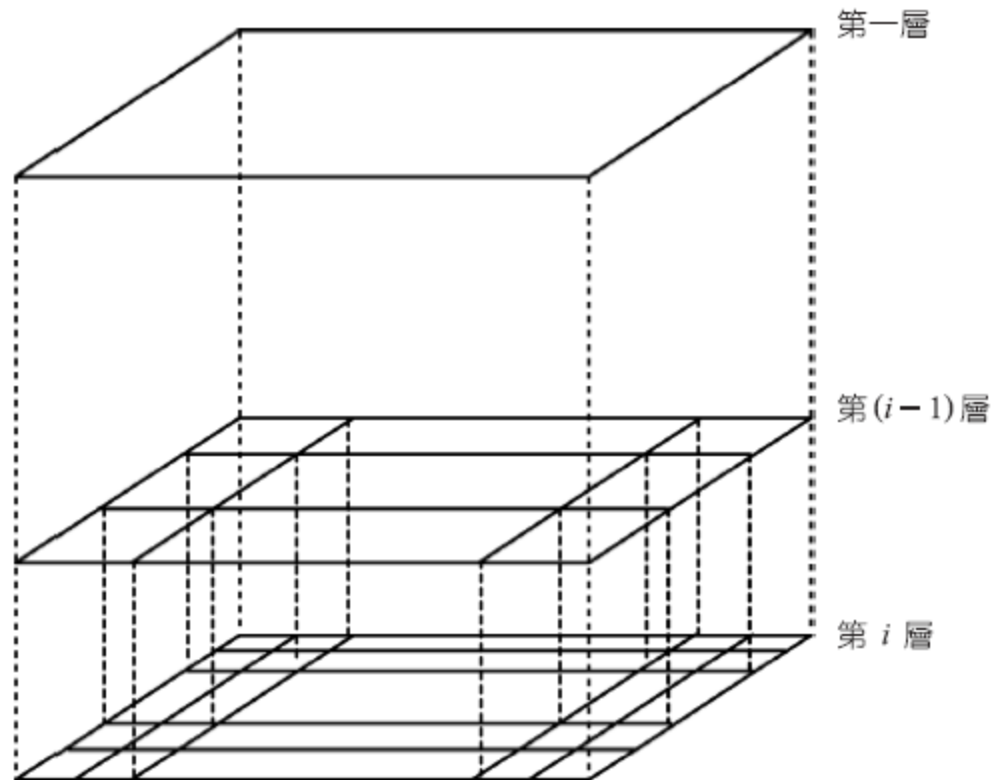
# Grid-Based Clustering Method

---

- 使用一個多重解析度方格資料結構
- 許多有趣方法
  - **STING** (a S**T**atistical **I**Nformation Grid approach) by Wang, Yang and Muntz (1997)
  - **WaveCluster** by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
    - 使用小波方法的多重解析度方格資料結構

# STING: A Statistical Information Grid Approach

- Wang, Yang and Muntz (VLDB'97)
- 將空間區域分成許多長方形儲存格
- 根據不同層次的解析度會產生許多層次的長方形儲存格



# STING 分群法

- 上層的儲存格會被分割進而產生下一層的儲存格
- 計算每個方格式的儲存格屬性的統計訊息
- 上層儲存格的統計參數可以從下層儲存格的統計參數計算得出
  - *count, mean, s, min, max*
  - 分佈類型—*normal, uniform*, 等.
- 透過由上而下的方法回答查詢
- 從階層式架構中的哪一層開始，一般來說是包含小量的儲存格
- 對這一層的每個儲存格計算其信賴區間

# STING建議

---

- 移除無關儲存格
- 當結束這一層, 繼續下一層
- 重複上述步驟直到最底層
- 優點:
  - 與查詢無關, 易平行處理, 可進行遞增式更新
  - $O(K)$ , 為最底層的儲存格數目
- 缺點:
  - 所有群組的邊界均為水平或垂直, 沒有對角邊界存在



# WaveCluster:使用小波轉換進行分群 (1998)

---

- Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
- 對原始的特性空間進行小波轉換 (wavelet transform)，並在轉換後的空間中尋找密集的区域
- 如何套用小波轉換進行分群
  - 藉由將多維度方格結構對應至資料空間來匯總資料
  - 這些多維度資料被表式成n維度特性空間
  - 對這個特性空間套用小波轉換來尋找密集区域
  - 套用多次小波轉換產生從細到粗的不同群組

# WaveCluster 方法

---

- 代入參數
  - 每個維度的方格數目
  - 小波, 小波轉換應用的數目
- 為何小波轉換適用於分群?
  - 使用帽型篩選 (hat-shaped filter) 弱化群組邊界外的資料點來分析群組
  - 自動移除離異值
- 主要特性:
  - 複雜度  $O(N)$
  - 可用於任意形狀的群組
  - 與輸入值順序無關, 不易受雜訊影響
  - 適用於低維度資料
- 同時是方格式與密度式

# 量化與轉換

- 首先將資料量化為D維度方格結構然後進行小波轉化
  - a) scale 1: 高解析度
  - b) scale 2: 中解析度
  - c) scale 3: 低解析度

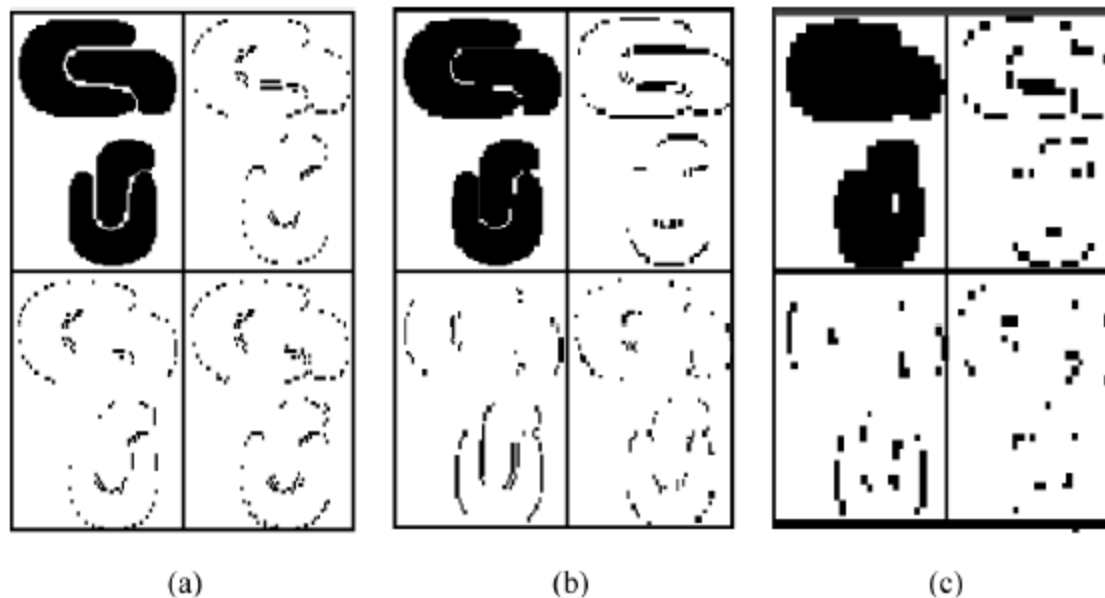
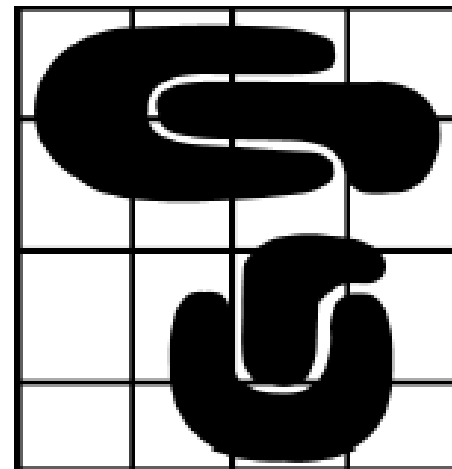


圖7.17 圖7.16特性空間的多重解析度：(a)scale = 1 (高解析度)；(b)scale = 2 (中解析度)；(c)scale = 3 (低解析度)。

# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法
7. 方格式方法
8. 模型式方法 
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結

# 模型式分群

---

- 何謂模型式分群?
  - 嘗試將原始資料與某些數學模型進行最適化
  - 根據假設:假設資料是根據某些特定機率分佈混合而成
- 典型方法
  - 統計方法
    - EM (Expectation maximization), AutoClass
  - 機器學習方法
    - COBWEB, CLASSIT
  - 類神經網路方法
    - SOM (Self-Organizing Feature Map)

# EM —期望最大化

- EM —一個尋找估計參數並且進行重複改善的方法
- k -means分割的延伸
  - 根據個體隸屬機率 (probability of membership) 的權重來進行分群
  - 依照權重來計算均值
- 概念
  - 從一組混合模型的起始參數開始
  - 利用混合模型的參數重複衡量個體權重
  - 新的個體權重會用於更新參數
  - 每個個體伴隨一個機率值，它用於說明某些屬性值屬於某個群組的機率
- 方法收斂很快但不一定是最佳解

# EM (Expectation Maximization) 方法

---

- 隨機選取k個個體代表群組的中心點
- 根據下列步驟進行變數改善
  - 期望步驟:計算個體 $X_i$ 屬於群組 $C_k$ 的機率

$$P(X_i \in C_k) = p(C_k|X_i) = \frac{p(C_k)p(X_i|C_k)}{p(X_i)},$$

- 最大化步驟:
  - 重新估計模型參數

$$m_k = \frac{1}{N} \sum_{i=1}^N \frac{X_i P(X_i \in C_k)}{\sum_j P(X_i \in C_j)}.$$

# 概念分群

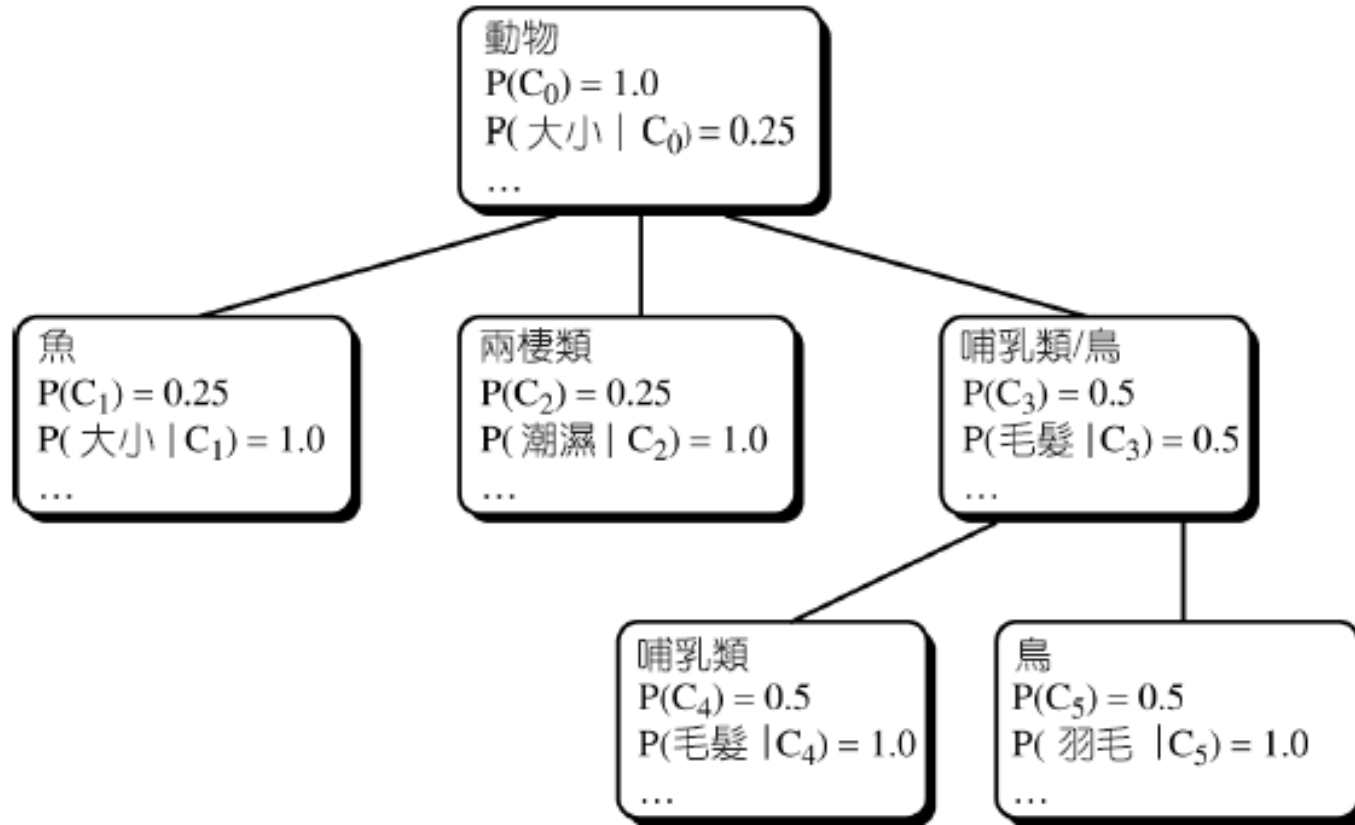
---

- 概念分群
  - 機器學習的分群方式
  - 對沒有類別的個體產生判別綱要 (classification scheme) 的分群方式
  - 會找出每個群組的特性
- COBWEB (Fisher'87)
  - 遞增式概念分群的方法
  - 一個類似判別樹的階層式分群
  - 每個節點代表一個概念，並包含對於該概念的機率描述



# COBWEB 分群法

## 判別樹



# 有關於概念分群

---

## ■ COBWEB限制

- 它假設不同屬性間的機率分佈是獨立的，這種假設並不是一定對，因為屬性間會存在相互關係
- 利用機率分佈表示群組在儲存與更新上是非常不便的，特別是當屬性擁有大量的值的時候

## ■ CLASSIT

- CLASSIT為COBWEB的延伸，應用於連續值資料
- 與COBWEB有相同問題

# 類神經網路方法

---

- 類神經網路方法
  - 用分類器 (**exemplar**) 代表群組
  - 當個體與某個分類器最相似，新的個體會分類至某個群組
- 典型方法
  - SOM (Soft-Organizing feature Map)

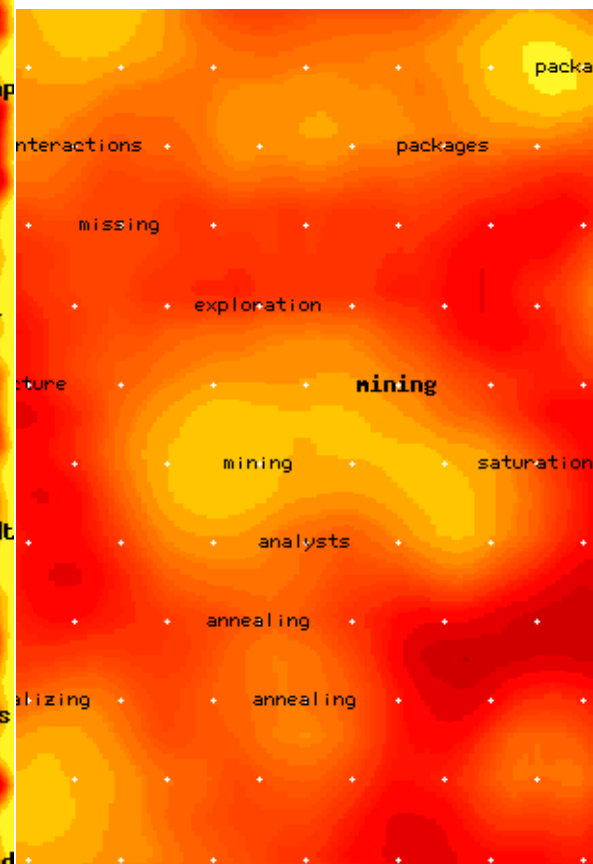
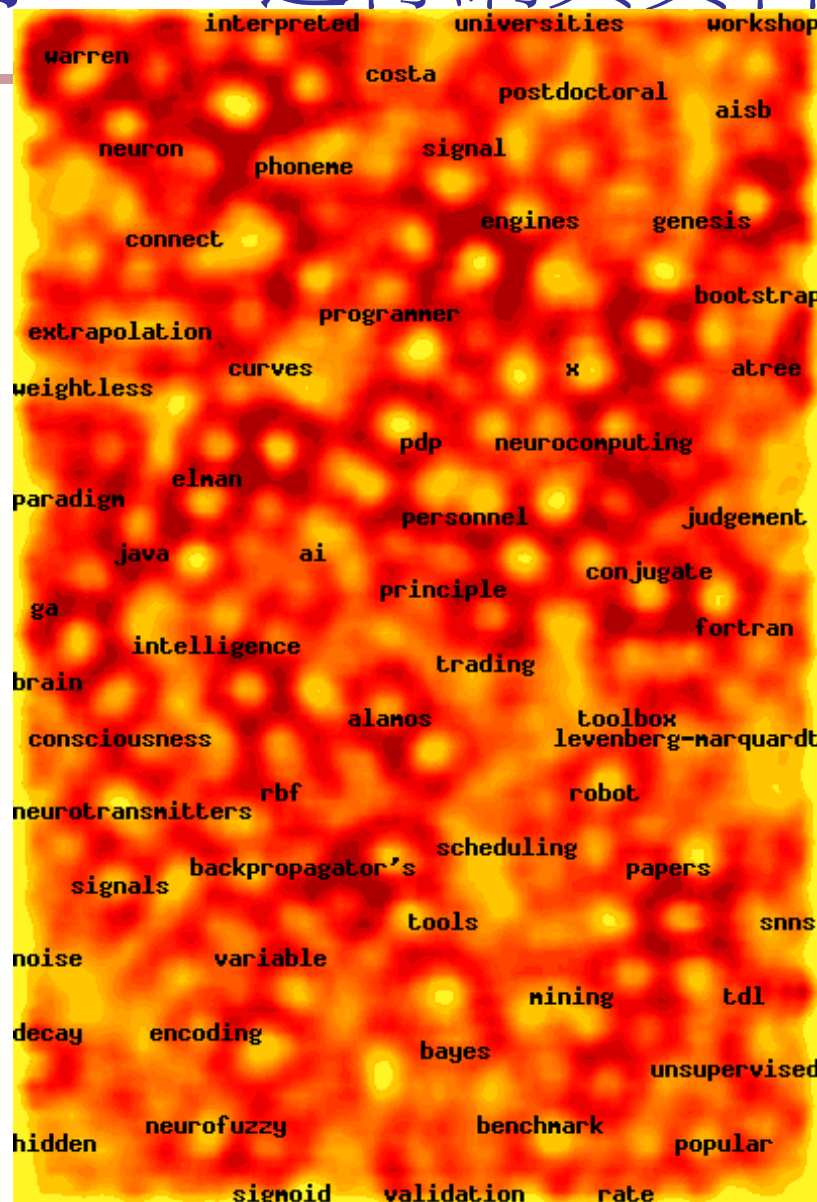
# 自我組織圖 (SOM)

---

- SOMs, 或稱為topological ordered maps, Kohonen Self-Organizing Feature Map (KSOMs)
- 將高維度資料對應至低維度 (2或3維) 目的空間,目的是在盡量維持資料的相似與距離
- 類似k-means:群組中心點大多位於低維度的轉換空間
- 群組是透過單元對新增個體的競爭而形成
  - 單元的權重與新增個體最接近者為獲勝單元
  - 獲勝單元與其最相近的單元進行權重的調整
- 被認為與人腦有相似的處理過程
- 在高維度資料的檢視是非常有用的

# 利用 SOM進行網頁資料分群

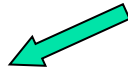
- 利用SOM對12088網頁文件分群結果
- 右圖為利用關鍵字mining尋找的結果
- 根據websom.hut.fi網頁



# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結



# 高維度資料分群

---

- 高維度資料分群
  - 許多應用: 文字文件, **DNA** 微陣列資料
  - 主要挑戰:
    - 不相關的維度會變成雜訊
    - 距離指標變得沒有意義
    - 僅有少數的維度與分群相關
- 方法
  - 特性轉換: 適用於大部分屬性是相關的資料
    - **PCA & SVD** 適用於當特性為高度相關或重複
  - 特性選擇: 包裝或過濾方法
    - 適用於當資料有良好的群組
  - 子空間分群: 對資料集的不同子空間進行群組
    - **CLIQUE, ProClus**, 與 頻繁樣式分群

# CLIQUE (Clustering In QUES)

---

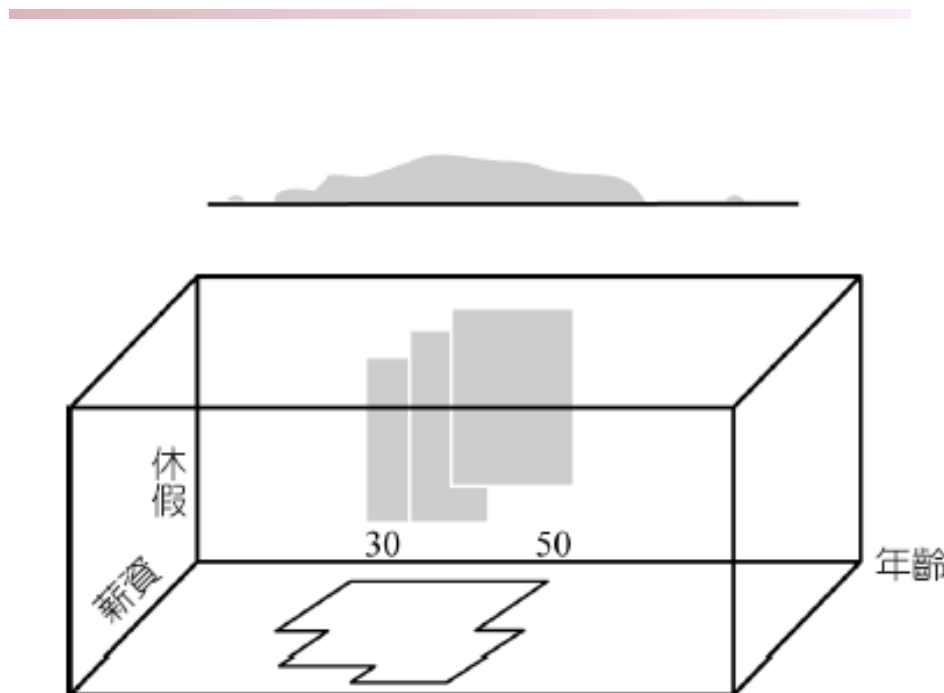
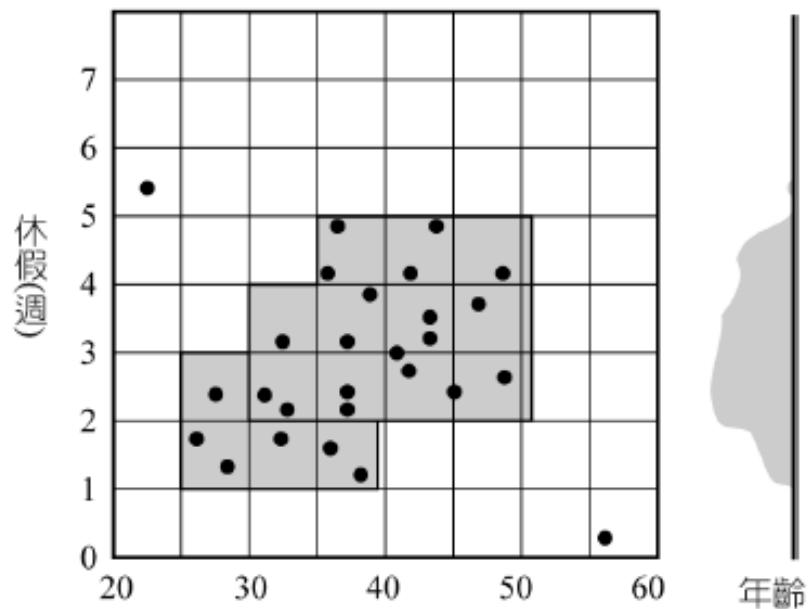
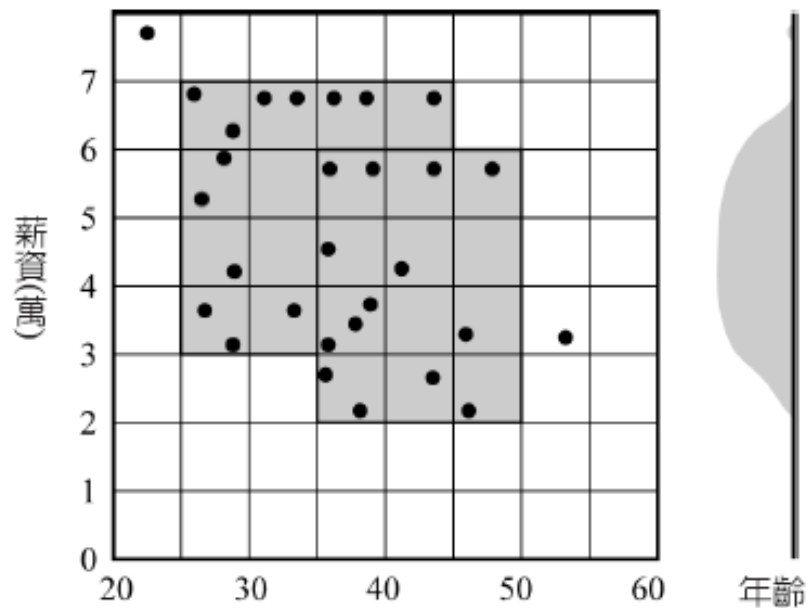
- Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98)
- 在高維度資料中找出子空間能對原始空間進行更好群組
- CLIQUE 被始為密度式與方格式
  - 將每個維度分割成相同數目等寬區塊
  - 分割 $m$ -維資料空間成不重疊的長方形單位
  - 一個單元是密集 (**dense**) 的，當它所包含資料點數超過使用者設定值
  - 群組是連接密集單元 (**connected dense units**) 的最大集合



# CLIQUE: 主要步驟

---

- 分割資料空間並找出每個分割儲存格中的資料點.
- 利用**Apriori**原則找出包含群組的子空間
- 判斷群組
  - 決定所有有趣子空間的密集單元
    - 決定所有有趣子空間的連結密集單元.
- 對每個群組產生最小描述
  - 透過找出連接密集單元的最大區域來形成群組
  - 透過找出連接密集單元的最小區域來形成群組



# CLIQUE 優缺點

---

## ■ 優點

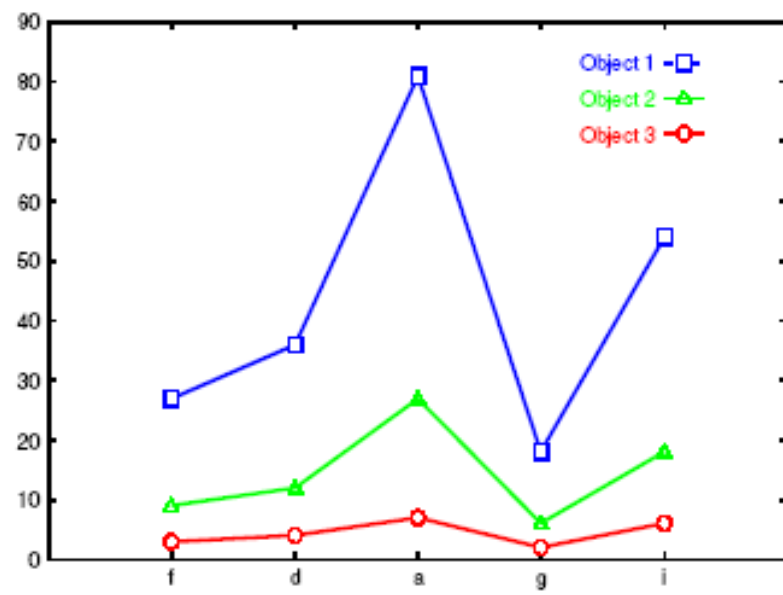
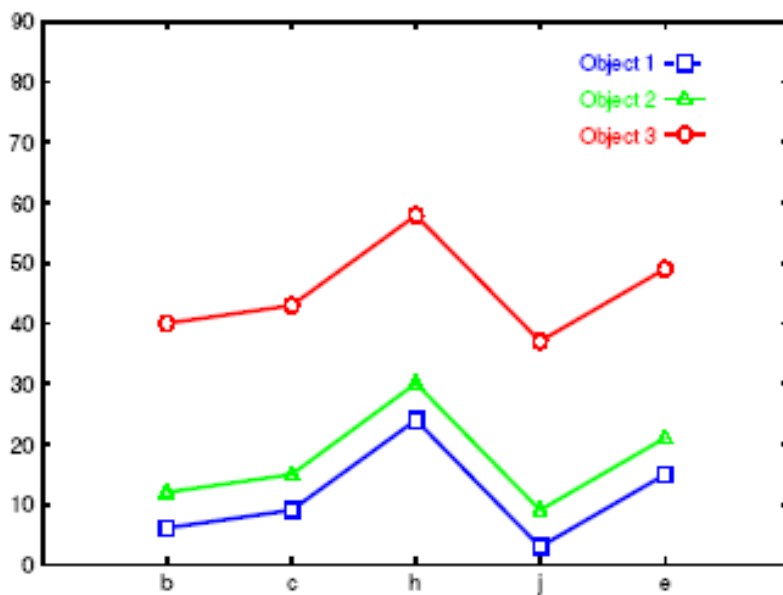
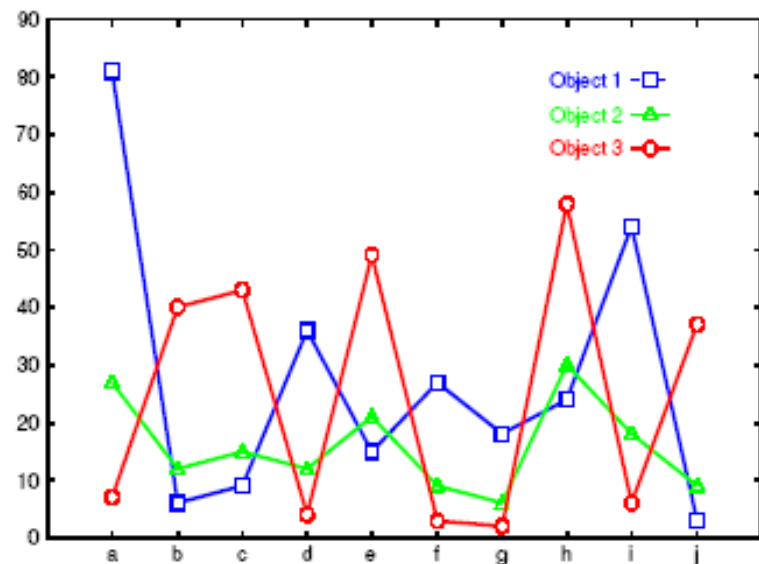
- 對高維度資料自動找出高密度的子空間
- 它與輸入個體的順序無關，同時不需假設任何資料分佈
- 它與輸入資料大小成線性關係，同時當資料維度變大時，它的可量度性也很好

## ■ 缺點

- 要產生有意義的群組取決於方格的大小與方格密度的設定值，這個是非常困難的，因為這兩個變數在所有的維度都有用到，因此分群的精準度會降低

# 樣式相似度分群( $p$ -Clustering)

- 右圖:顯示微陣列的片段，它包含三個基因(個體)與10個屬性(欄到)。三個個體從目視中並無明顯的樣式
  - 很難發現樣式
- 下圖:存在位移與大小的樣式

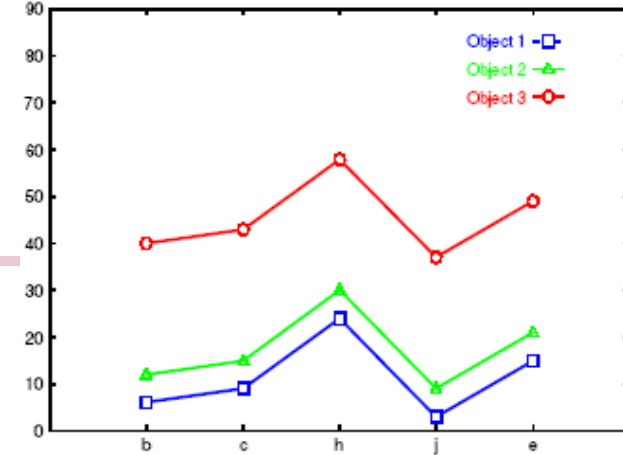


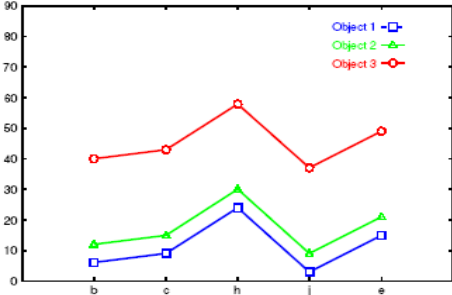
# 為何 $p$ -Clustering?

- 微陣列資料分析需要
  - 對數以千計維度進行分群
  - 找出位移與大小樣式
- 歐幾里得距離很難找位移樣式
- 對新屬性進行分群  $A_{ij} = a_i - a_j$ ? — 產生  $N(N-1)/2$  維度
- 利用新的平均平方殘餘分數指標衡量矩陣  $(I, J)$

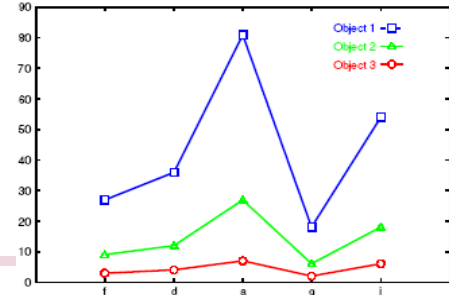
$$H(IJ) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (d_{ij} - d_{iJ} - d_{IJ} + d_{IJ})^2$$

- 當  $d_{ij} = \frac{1}{|J|} \sum_{j \in J} d_{ij}$      $d_{Ij} = \frac{1}{|I|} \sum_{i \in I} d_{ij}$      $d_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} d_{ij}$
- 子矩陣為  $\delta$ -cluster 當  $H(I, J) \leq \delta$  對於某個  $\delta > 0$
- 雙分群問題
  - bicluster的子矩陣並不一定是-bicluster ,
  - 由於平均的影響，即使 $\delta$ 值很小 $\delta$ -bicluster會包含離異值





# $p$ -Clustering: 樣式相似 度分群



- 假設  $x, y$  為  $O$  個體  $a, b$  為  $T$  中屬性,  $pCluster$  為一  $2 \times 2$  矩陣

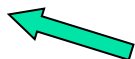
$$pScore\left(\begin{bmatrix} d_{xa} & d_{xb} \\ d_{ya} & d_{yb} \end{bmatrix}\right) = |(d_{xa} - d_{xb}) - (d_{ya} - d_{yb})|$$

- $(O, T)$  會形成  $\delta$ - $pCluster$  當  $pScore(X) \leq \delta$  對某個  $\delta > 0$ ,  $X$  為  $(O, T)$  中的任一個  $2 \times 2$  矩陣
- $\delta$ - $pCluster$  特性
  - 有向下特質
  - 比 **bicluster** 的方法更具同質性
- 可以使用樣式增長的方法來尋找這些樣式
- 要尋找大小樣式 (scaling pattern), 將方程式以對數表示

$$\frac{d_{xa} / d_{ya}}{d_{xb} / d_{yb}} < \delta$$

# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群 
11. 離異值分析
12. 總結

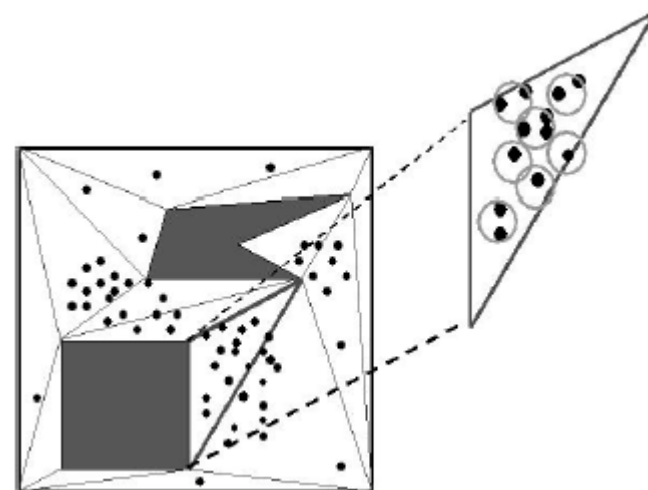
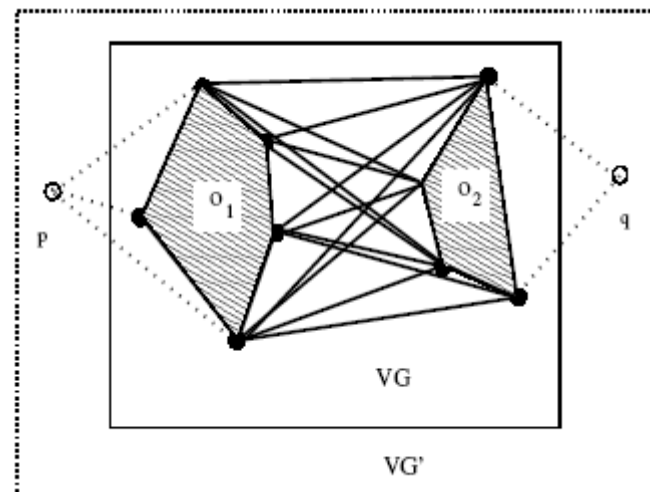
# 限制式分群分析

- 分群應用:依照使用者的喜好或限制進行分群
- 不同限制式分群:
  - 個體限制
    - 對房屋價格超過\$300K進行分群
  - 群組參數選擇的限制
    - 群組數目, MinPts, 等.
  - 距離與相似函數的限制
    - 權重函數, 障礙物 (例., 河, 湖)
  - 使用者對群組特性設定限制
    - 包含100個高價值顧客與5,000個普通顧客
  - 半監督式分群:透過弱化的監督式分群而得到改善

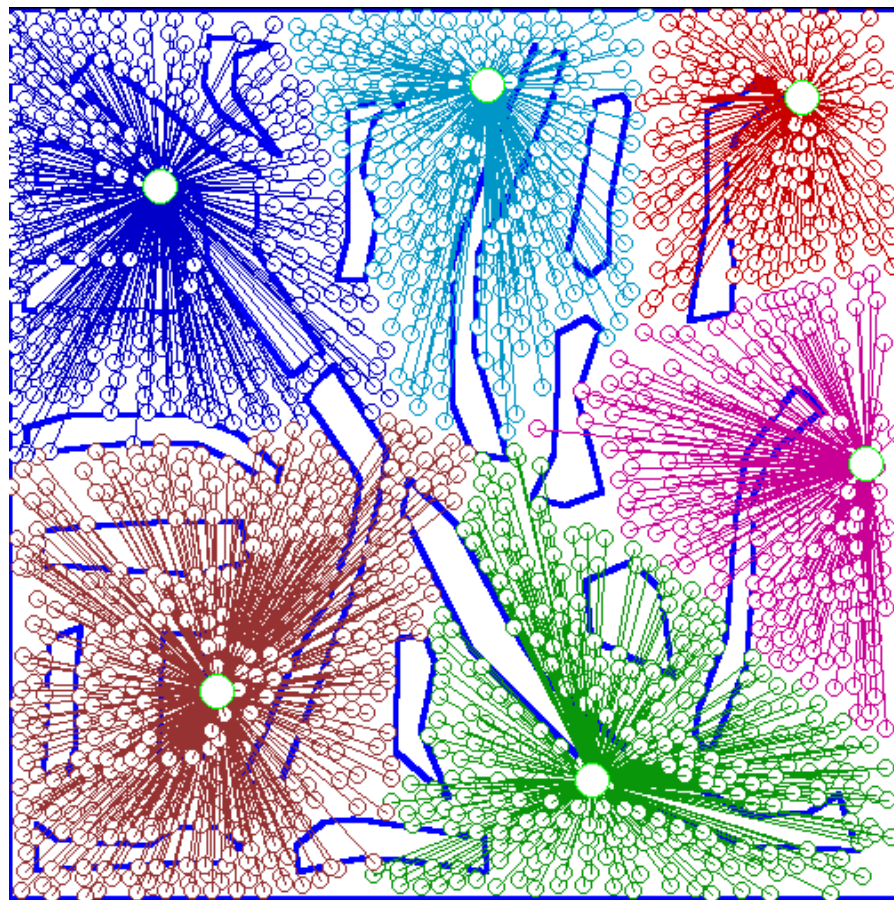


# 有障礙個體分群

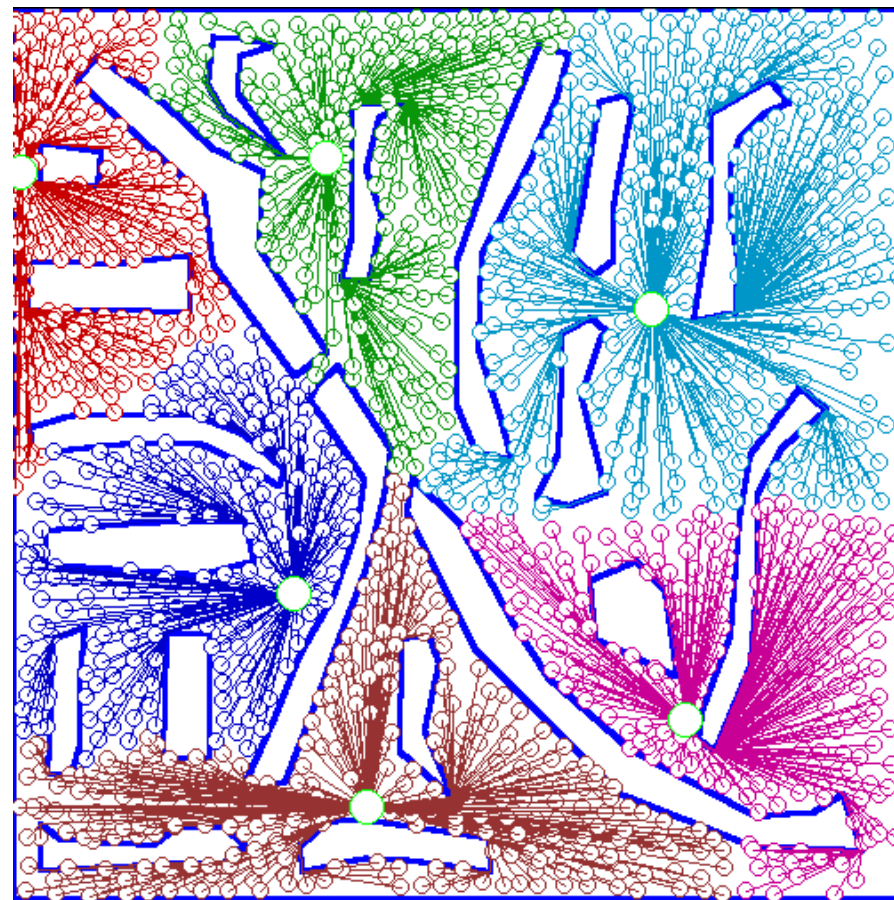
- K-medoids 比較好因為 k-means 會將 ATM 放在湖中央
- 可式圖與最短路徑
- 三角化與維群組
- 兩種類型索引可以由計算最短路徑中得出
  - **VV** 索引:用於任何成對的障礙個體
  - **MV** 索引:用於任何成對的微群組或障礙個體



## 範例:有障礙個體分群



沒有考慮障礙的分群




有考慮障礙的分群

# 使用者限制的群組

- 範例:決定 $k$ 個服務中心的位置，包含 $m$ 個高價值顧客與 $n$ 個普通顧客
- 提出方法
  - 將資料分割為滿足條件的 $k$ 個群組
  - 透過將個體從一個群組移到另一群組的方式，來同時滿足限制條件與改善群組
  - 為了增進處理的效率，資料會事先使用微群組的處理方式以避免要處理全部的資料

# 第七章. 分群分析

---

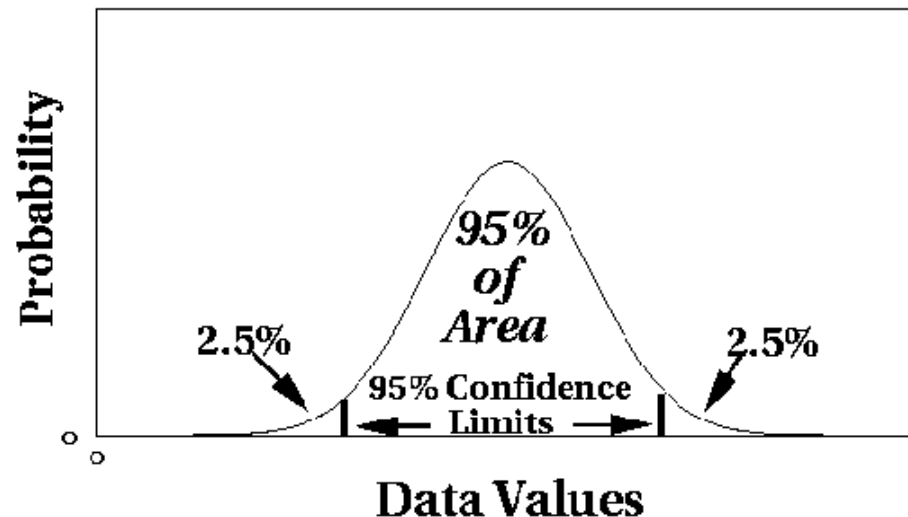
1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析 
12. 總結

# 何謂離異值分析?

---

- 何謂離異值?
  - 與一般資料極度不同的資料個體
- 問題：在大資料集中定義與發現離異值
- 應用：
  - 信用卡詐欺檢測
  - 電信詐欺檢測
  - 客戶區隔
  - 醫學分析

# 統計分佈式離異值檢測



假設資料集的機率分佈模型 (例. normal distribution)

- 透過模型不一致檢定
  - 資料分佈
  - 分佈參數 (例., mean, variance)
  - 變異值與期望離異值的個數
- 缺點
  - 大部分都是單根檢定
  - 許多例子分佈有可能是未知的

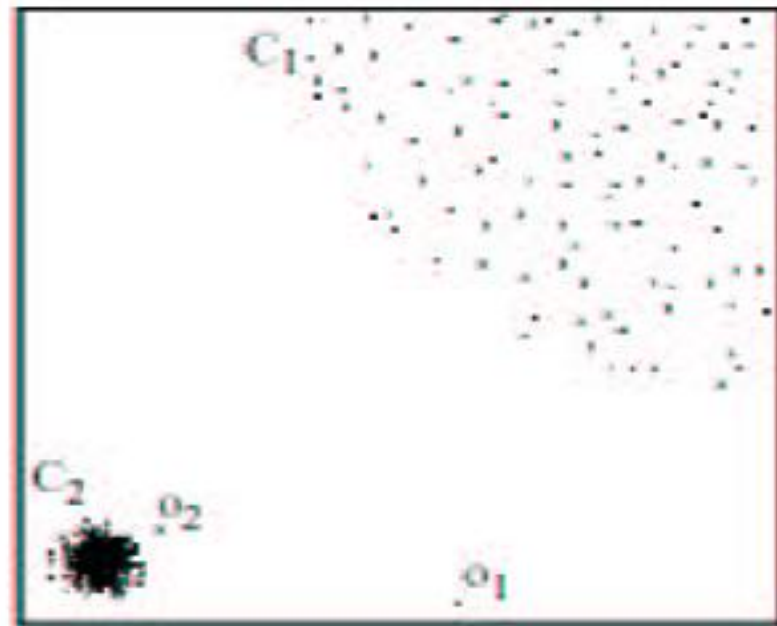
# 距離式離異值檢測

- 為了克服統計模型的限制
  - 需要一個不需了解資料分佈的多維度分析
- 距離式離異值: 一個個體 $o$ 在參數 $pct$ 與 $dmin$ 下為一個距離式離異值 $DB(pct, dmin)$ , 當資料集 $D$ 中至少有 $pct$ 部分的個體與個體 $o$ 的距離大於 $dmin$
- 探勘距離式離異值方法
  - 索引式方法
  - 巢狀式迴圈方法
  - 儲存格式方法



# 密度式區域離異值檢測

- 取決於資料集的全域 (global) 分佈
- 在分析密度相當不同的分佈時會有困難
- 例.  $C_1$  包含 400 稀疏分佈點,  $C_2$  包含 100 進密結合點, 2 離異點  $O_1, O_2$
- 距離式方法無法判斷  $O_2$  為離異值
- 需要區域離異值概念



- 區域離異因素 (LOF)
  - 假設離異值並不是一分為二
  - 每個點都有一個區域離異因素



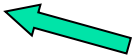
# 偏差式離異值檢測

---

- 使用個體群組的主要特性
- 偏離這個特性的個體會被認為是離異值
- 順序異常方法
  - 模擬人類從一組相似個體中發現不同個體的方式
- OLAP資料方塊方法
  - 使用資料方塊來找出大型多維度資料中異常的區域

# 第七章. 分群分析

---

1. 何謂分群分析?
2. 分群分析資料類型
3. 主要分群方法分類
4. 分割方法
5. 階層方法
6. 密度式方法
7. 方格式方法
8. 模型式方法
9. 高維度資料分群
10. 限制式分群
11. 離異值分析
12. 總結 

# 總結

---

- 群組是群組內相似、群組間不相似的個體集合,分群分析有廣泛應用
- 可對不同資料類型計算相似度
- 分群方法包含分割式方法、階層式方法、密度式方法、方格式方法、模型式方法、高維度資料方法與限制式方法
- 離異值檢測與分析對詐騙檢測、客製化行銷、醫療分析等非常有用
- 資料探勘研究上分群是活躍的課題