CH4 決策樹分析

授課老師：簡禎富 講座教授

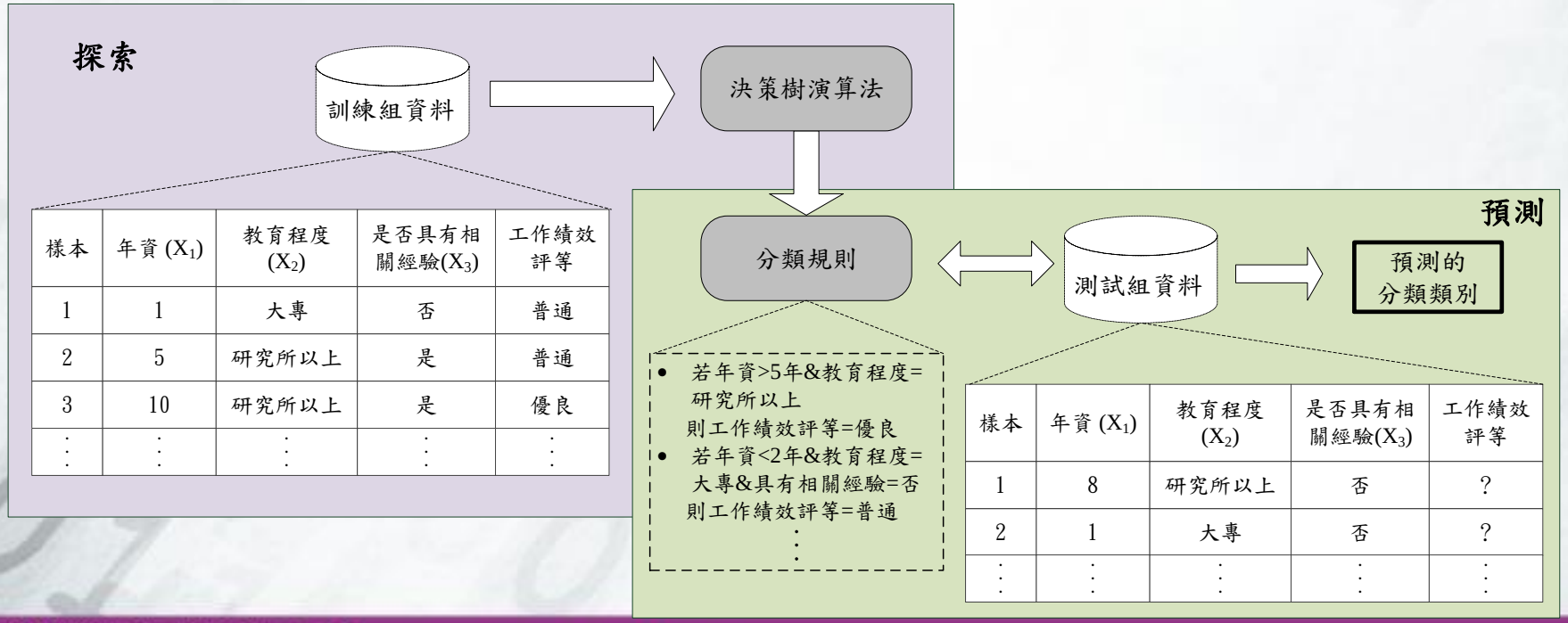
資料挖礦與大數據分析
Data Mining & Big Data Analytics

大綱

- 決策樹的建構
- 決策樹的演算法
- 分類模型評估
- 應用實例──建構cDNA 生物晶片之二元資料挖礦模式
- 結論

決策樹 (decision tree)

- 具有監督式的特徵萃取與描述的功能，將輸入變數根據目標設定來選擇分枝變數與分枝方式，並以樹枝狀的層級架構呈現，以萃取分類規則
- 目的：探索與預測

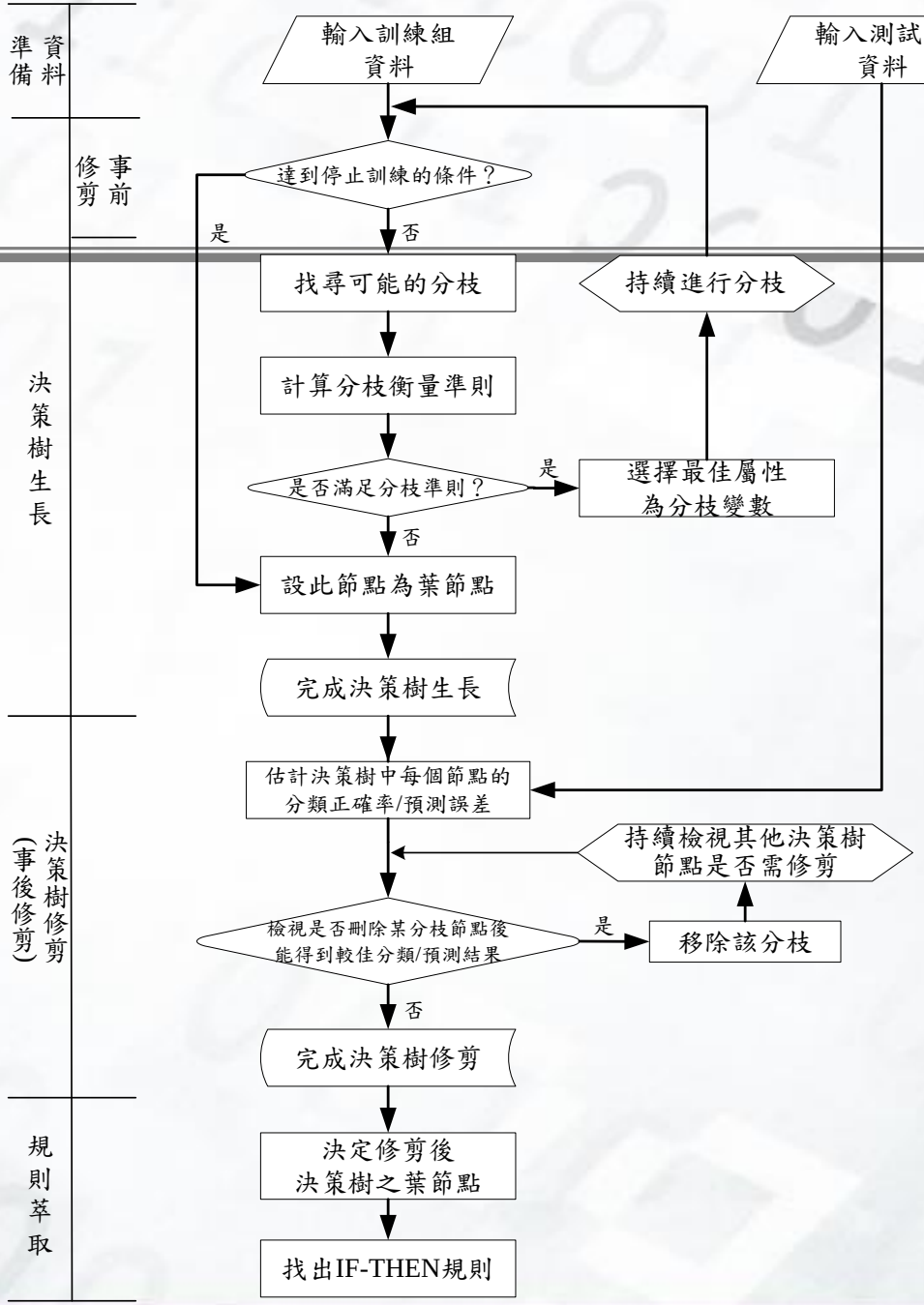


樣本	年資 (X_1)	教育程度 (X_2)	是否具有相關經驗(X_3)	工作績效評等
1	1	大專	否	普通
2	5	研究所以上	是	普通
3	10	研究所以上	是	優良
⋮	⋮	⋮	⋮	⋮

樣本	年資 (X_1)	教育程度 (X_2)	是否具有相關經驗(X_3)	工作績效評等
1	8	研究所以上	否	?
2	1	大專	否	?
⋮	⋮	⋮	⋮	⋮

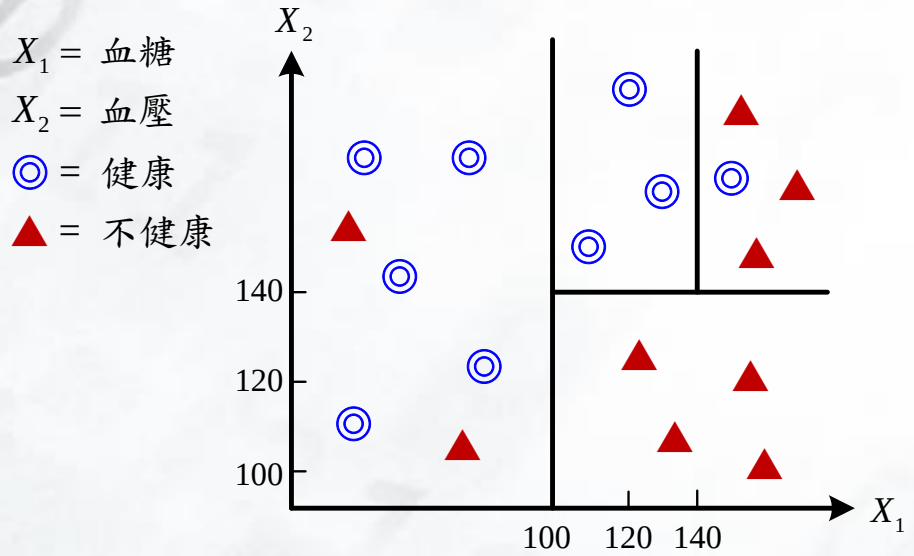
決策樹的建構

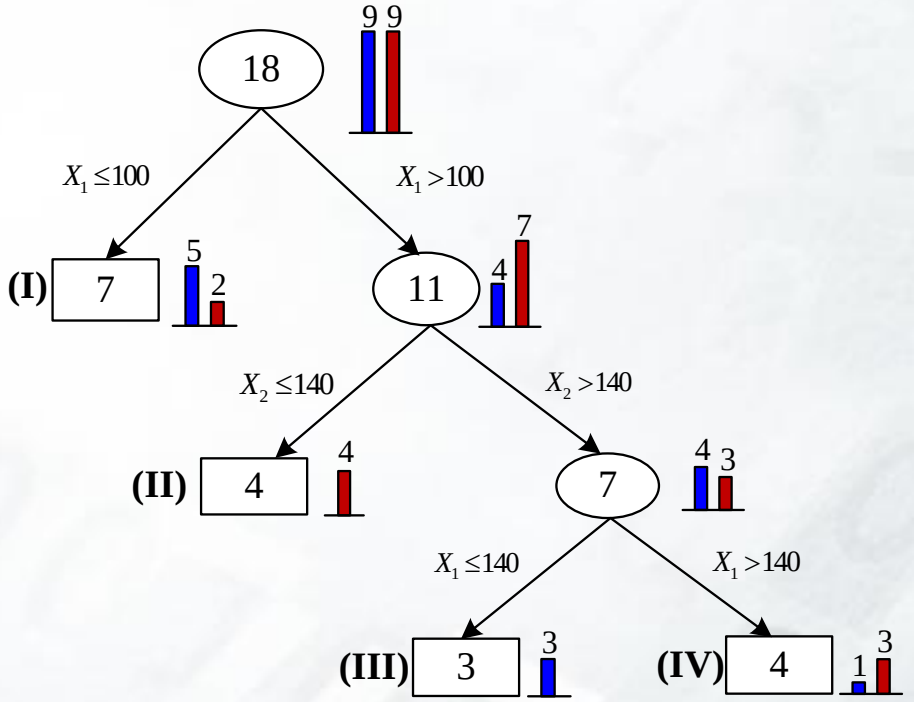
1. 資料準備
2. 決策樹生長
3. 修剪
4. 規則萃取



決策樹的建構範例

- 以健康為目標建立決策樹，衡量18位人員的血糖與血壓，並以血糖最低標準100與140以及血壓最低標準90來分類





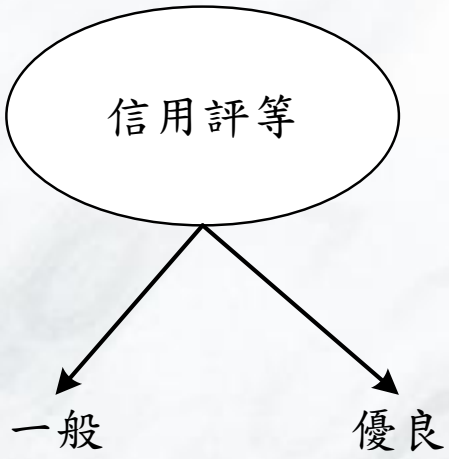
發現規則：

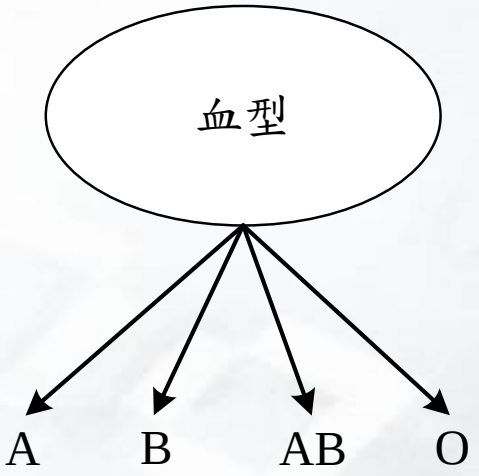
(I)若 $X_1<100$，則為健康(◎)
(II)若$X_1>100$且$X_2<140$，則為不健康(▲)
(III)若$100<X_1\le140$且$X_2>140$，則為健康(◎)
(IV)若$X_1>140$且$X_2>140$，則為不健康(▲)

資料準備 (1/2)

- 決策樹的分析資料包含兩種變數：
  1. 根據問題所決定的**目標變數**
  2. 根據問題背景與環境所選擇的各種屬性作為**分枝變數**

二元屬性	名目屬性
其測試條件可以產生兩種結果	結果可用不同屬性值來表示





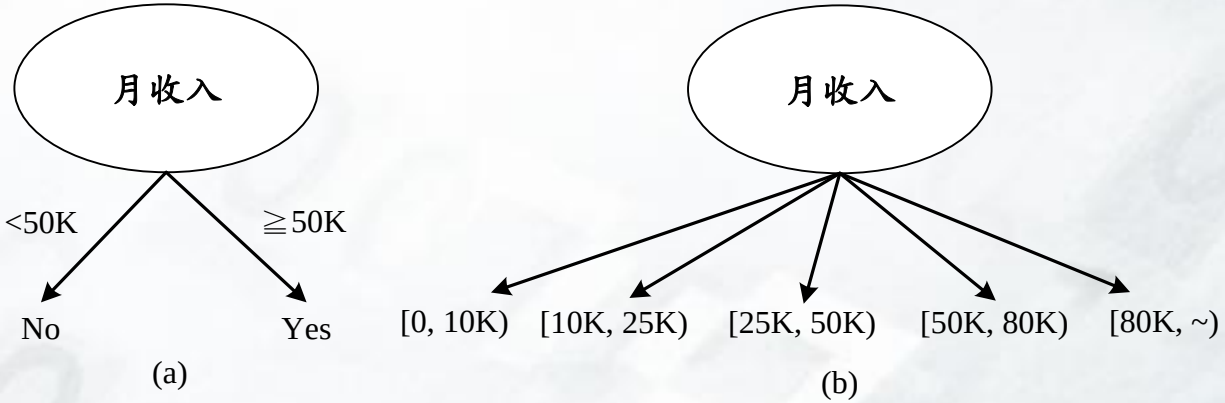
資料準備 (2/2)

順序屬性

- 可以生成二元或二元以上的分割，其屬性可以群組
- 群組必須不違反其屬性值特性

連續屬性

- 可表示成 X<a或X≧a的關係
- 須考慮到所有可能的分割點y，再選出最好的分割



分枝準則(splitting criteria)

- 決定樹的規模大小，包含樹的寬度以及深度
- 常見的分枝準則：
  - 資訊增益 (information gain)
  - Gini係數 (Gini index)
  - 卡方統計量 (Chi-square statistic)
  - 資訊增益比 (information gain ratio)
- 透過檢驗分枝屬性的顯著性後，分枝準則即能找出具有最佳分枝結果的屬性

決策樹分析資料表

- 假設訓練資料集合D中有*k*個類別，則 $C_j, j=1,2,\cdots,k$、屬性A有*l*種不同的資料值

決策樹分析資料表

類別 / 屬性 A 的資料值	C_1	C_2	…	C_k	總和
A_1	x_{11}	x_{12}	…	x_{1k}	$x_{1\cdot}$
A_2	x_{21}	x_{22}	…	x_{2k}	$x_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
A_l	x_{l1}	x_{l2}	…	x_{lk}	$x_{l\cdot}$
總和	$x_{\cdot 1}$	$x_{\cdot 2}$	…	$x_{\cdot k}$	N

資訊增益(Information Gain)(1/2)

- 資訊衡量(Information Measurement)是根據不同訊息的概似值或機率，以衡量不同條件下的資訊量
  - 若資料的各種訊息機率一致，則獲得的訊息量最大；反之，最小
  - 評估函數的價值亦取決於資料所帶來的訊息狀態個數

$$\text{Info(D)} = -\frac{x_{.1}}{N}\log_2\left(\frac{x_{.1}}{N}\right) - \frac{x_{.2}}{N}\log_2\left(\frac{x_{.2}}{N}\right) - \cdots - \frac{x_{.k}}{N}\log_2\left(\frac{x_{.k}}{N}\right)$$

$$= -\sum_{j=1}^{k} p_j \cdot \log_2(p_j)$$

- Info(D)又稱熵(entropy)
  - 衡量資料離散程度或亂度，作為評估訓練資料集合D下所有類別的期望訊息
  - 各類別出現的機率相等，則熵值為1，表示分類訊息雜亂度最高

資訊增益(Information Gain)(2/2)

- 假設該資料集合**D**要根據屬性**A**進行分割，產生共**L**個資料分割集合D_i ，其中 $x_{i.}$為各屬性值 A_i 下的分割資料總個數，x_{ij} 為屬性值 A_i 下且為類別 C_j 的個數，因此 $\text{Info}(\text{A}_\text{i})$ 為：

$$\text{Info}(\text{A}_\text{i}) = -\frac{x_{i1}}{x_{i.}}\log_2\left(\frac{x_{i1}}{x_{i.}}\right) - \frac{x_{i2}}{x_{i.}}\log_2\left(\frac{x_{i2}}{x_{i.}}\right) - \cdots - \frac{x_{ik}}{x_{i.}}\log_2\left(\frac{x_{ik}}{x_{i.}}\right)$$

- 屬性**A**的資訊則根據各屬性值下的資料個數多寡決定，如

$$\text{Info}_\text{A}(\text{D}) = \frac{x_{1.}}{N}\text{Info}(A_1) + \frac{x_{2.}}{N}\text{Info}(A_2) + \cdots + \frac{x_{l.}}{N}\text{Info}(A_l)$$

$$= \sum_{i=1}^{l}\frac{x_{i.}}{N}\text{Info}(A_i)$$

- 資訊增益可表示為： $Gain(\text{A}) = \text{Info}(D) - \text{Info}_\text{A}(D)$

範例[4.1]資訊增益

- 假設某公司人力資源部門欲瞭解職員的表現，抽取10位現職員工為樣本，將連續屬性離散化後如下表

職員	年資(A)	教育程度(B)	具備相關經驗(C)	員工表現
001	5年以下	研究所	是	優等
002	10年以上	研究所	否	普通
003	5年以下	研究所	是	優等
004	5年以下	大專	是	普通
005	5年以下	研究所	否	優等
006	10年以上	研究所	是	優等
007	5年至10年	大專	否	普通
008	5年至10年	研究所	是	優等
009	5年至10年	大專	否	普通
010	5年以下	研究所	是	普通

根據(A)、(B)、(C)三個屬性分別計算：

$Info(D)$

$Info(A_i)$

Gain

Gini係數

- 衡量資料集合對於所有類別的不純度(impurity)

$$Gini(D) = 1 - \sum_{j=1}^{k} p_j^2$$

- 各屬性值 A_i 下資料集合之不純度

$$Gini\,(A_i) = 1 - (\frac{X_{i1}}{X_{i.}})^2 - (\frac{X_{i2}}{X_{i.}})^2 - \cdots - (\frac{X_{ik}}{X_{i.}})^2 = 1 - \sum_{j=1}^{k} (\frac{X_{ij}}{X_{i.}})^2$$

- 屬性A的總資料不純度則等於所有屬性值分割下的期望平均

$$Gini_A(D) = \frac{X_{1.}}{N} Gini\,(A_1) + \frac{X_{2.}}{N} Gini\,(A_2) + \cdots + \frac{X_{l.}}{N} Gini\,(A_l)$$

- 計算其他屬性作為分枝變數所能帶來的純度

$$\Delta Gini\,(A) = Gini\,(D) - Gini_A(D)$$

[範例4.1] Gini係數(1/2)

$$Gini\,(\mathrm{D}) = 1-(0.5)^2-(0.5)^2 = 0.5$$

$$Gini_{\text{年資}}(\mathrm{D})$$

$$= \frac{5}{10}\left(1-(\frac{3}{5})^2-(\frac{2}{5})^2\right)+\frac{3}{10}\left(1-(\frac{1}{3})^2-(\frac{2}{3})^2\right)+\frac{2}{10}\left(1-(\frac{1}{2})^2-(\frac{1}{2})^2\right) = 0.473$$

$$Gini_{\text{教育程度}}(\mathrm{D}) = \frac{3}{10}\left(1-(\frac{0}{3})^2-(\frac{3}{3})^2\right)+\frac{7}{10}\left(1-(\frac{5}{7})^2-(\frac{2}{7})^2\right) = 0.286$$

$$Gini_{\text{有無相關經驗}}(\mathrm{D}) = \frac{6}{10}\left(1-(\frac{4}{6})^2-(\frac{2}{6})^2\right)+\frac{4}{10}\left(1-(\frac{1}{4})^2-(\frac{3}{4})^2\right) = 0.417$$

$$\Delta\text{Gini}(\text{年齡}) = Gini\,(\mathrm{D}) - Gini_{\text{年齡}}(\mathrm{D}) = 0.5-0.473 = 0.027$$

$$\Delta\text{Gini}(\text{教育程度}) = Gini\,(\mathrm{D}) - Gini_{\text{教育程度}}(\mathrm{D}) = 0.5-0.286 = 0.214$$

→以教育程度作為分枝屬性能得到較多資訊

$$\Delta\text{Gini}(\text{具備相關經驗}) = Gini\,(\mathrm{D}) - Gini_{\text{具備相關經驗}}(\mathrm{D}) = 0.5-0.417 = 0.083$$

[範例4.1] Gini係數(2/2)

- **若以年資與Gini係數為例說明連續屬性的分割過程**
  - 依據7個分割點計算其Gini係數，可得當年資以13.5年作為分割點時，其資訊增益為

$$\text{Gain(A)} = 0.500 - 0.444 = 0.056$$

年資	1	2	4	6	8	12	15

分割點		1.5		3		5		7		10		13.5		
評等		≦	>	≦	>	≦	>	≦	>	≦	>	≦	>	
優秀		1	4	1	4	3	2	3	2	4	1	4	1	
普通		1	4	2	3	2	3	3	2	4	1	5	0	
		1		0.965		0.971		1		1		0.892		
		0.500		0.476		0.480		0.500		0.500		0.444		

卡方統計量（χ^2 statistic）

- 以列聯表計算兩變數間的相依程度，當計算出的樣本卡方統計值越大，表示兩變數間的相依程度越高

年齡 \ 表現	優秀	普通	總和
(A_1)	3 (2.5)	2 (2.5)	5
(A_2)	1 (1.5)	2 (1.5)	3
(A_3)	1 (1.0)	1 (1.0)	2
總和	5	5	10

教育程度 \ 表現	優秀	普通	總和
(B_1)大專以下	0 (1.5)	3 (1.5)	3
(B_2)研究所以上	5 (3.5)	2 (3.5)	7
總和	5	5	10

年齡 \ 表現	優秀	普通	總和
(A_1) 是	4 (3)	2 (3)	6
(A_2) 否	1 (2)	3 (2)	4
總和	5	5	10

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{k}\frac{(x_{ij}-E_{ij})^2}{E_{ij}}, \qquad E_{ij} = \frac{x_{i.}\cdot x_{.j}}{N}$$

$$\chi^2(\text{年資}) = \frac{(3-2.5)^2}{2.5}+\frac{(2-2.5)^2}{2.5}+\frac{(1-2.5)^2}{2.5}+\frac{(2-2.5)^2}{2.5}+\frac{(1-2.5)^2}{2.5}+\frac{(1-2.5)^2}{2.5} = 0.533$$

$$\chi^2(\text{教育程度}) = \frac{(0-1.5)^2}{1.5}+\frac{(3-1.5)^2}{1.5}+\frac{(5-3.5)^2}{3.5}+\frac{(2-3.5)^2}{3.5} = 4.286$$

$$\chi^2(\text{具備相關經驗}) = \frac{(4-3)^2}{3}+\frac{(2-3)^2}{3}+\frac{(1-2)^2}{2}+\frac{(3-2)^2}{2} = 1.67$$

資訊增益比 (Information Gain-Ratio)

- 考慮候選屬性本身所攜帶的訊息，再將這些訊息轉換至決策樹，經由計算資訊增益與分枝屬性的資訊量之比值來找出最適合的分枝屬性

$$\mathrm{GR}(A) = \frac{Gain(A)}{Split\ Info(A)} \qquad Split\ Info(A) = -\sum_{i=1}^{l} \frac{x_{i.}}{N} \cdot \log_2\left(\frac{x_{i.}}{N}\right)$$

- 當分枝變數的屬性水準越多，表示使用該變數越容易得到較大的***Entropy***，同時亦代表該屬性分枝特性不顯著，因此會傾向選擇具有較小***Entropy***值之屬性為分枝變數

[範例4.1] 資訊增益比

$$\text{Split Info}(\text{年資}) = -\frac{5}{10}\log_2(\frac{5}{10}) - \frac{3}{10}\log_2(\frac{3}{10}) - \frac{2}{10}\log_2(\frac{2}{10}) = 1.485$$

$$\text{Split Info}(\text{教育程度}) = -\frac{3}{10}\log_2(\frac{3}{10}) - \frac{7}{10}\log_2(\frac{7}{10}) = 0.881$$

$$\text{Split Info}(\text{具有相關經驗}) = -\frac{6}{10}\log_2(\frac{6}{10}) - \frac{4}{10}\log_2(\frac{4}{10}) = 0.971$$

$$\text{GR}(\text{年資}) = \frac{0.039}{1.485} = 0.026$$

$$\text{GR}(\text{教育程度}) = \frac{0.396}{0.881} = 0.449$$

$$\text{GR}(\text{具備相關經驗}) = \frac{0.125}{0.971} = 0.129$$

變異降低 (variance reduction)

- 目標變數為**連續屬性**時，分枝準則可改用變異降低
  - 變異數是量測資料值與平均值的差異（即該節點內的各筆資料目標值與目標平均值之均方差）
  - 檢視其分枝節點內資料的變異程度是否有顯著縮減
  - 在評估完所有屬性進行分枝後所計算出的變異數後，最後再比較候選屬性的變異數縮減量，並選出具有最大變異數縮減量的屬性為分枝變數

$$S_t^2 = \frac{\sum_{i=1}^{N_t} (y_{i,t} - \bar{y}_t)^2}{N_t}$$

[範例4.1] 變異降低

・職員收入的資料

職員	年資(A)	教育程度(B)	具備相關經驗(C)	月收入(K)
001	5年以下	研究所	是	45
002	10年以上	研究所	否	60
003	5年以下	研究所	是	42
004	5年以下	大專	是	39
005	5年以下	研究所	否	42
006	10年以上	研究所	是	75
007	5年至10年	大專	否	40
008	5年至10年	研究所	是	45
009	5年至10年	大專	否	44
010	5年以下	研究所	是	38

・各屬性分枝後的變異數

	年資	教育程度	具備相關經驗
分枝點	[5年以下&5年至10年]、[10年以上]	[大專]、[研究所]	[否]、[是]
變異數	$0.8\times6.36+0.2\times56.25=16.34$	$0.3\times4.67+0.7\times149.39=105.97$	$0.4\times62.75+0.6\times160.22=121.23$

決策樹修剪

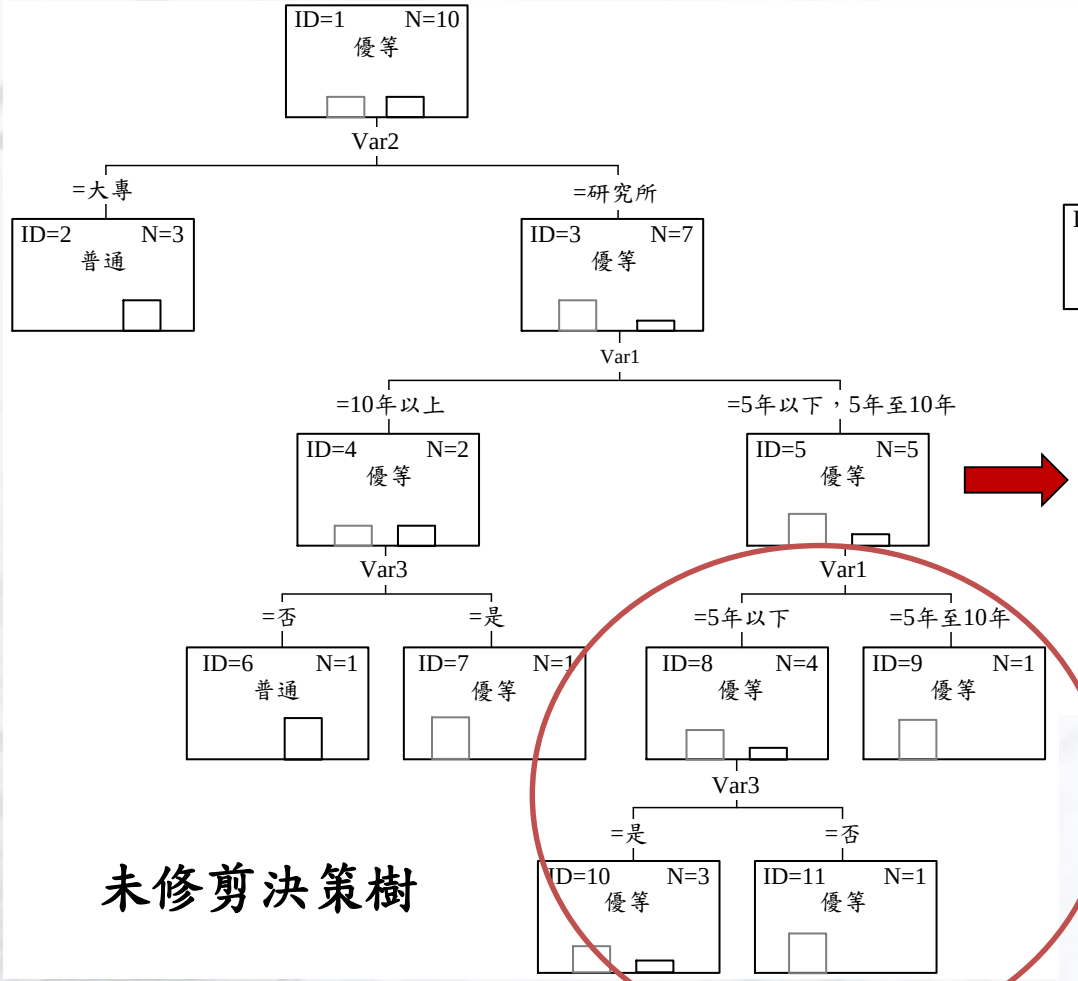
- **事先修剪(pre-pruning)**
  - 事先設定停止決策樹生長的門檻值，當分割的評估值未達此門檻值時，就會停止擴長
  - 優點：較具有執行效率
  - 缺點：可能過度修剪(over-pruning)、門檻值設定不易
- **事後修剪(post-pruning)**
  - 在樹完全長成後再修剪，引入測試組樣本來評估決策樹對於新輸入資料的分類與預測結果
  - 優點：可解決過度配適，避免產生稀少樣本樹的葉節點，以及加強對雜訊的忍受程度
  - 缺點：效率較低

最小成本複雜修剪
(Minimal Cost-Complexity Pruning)

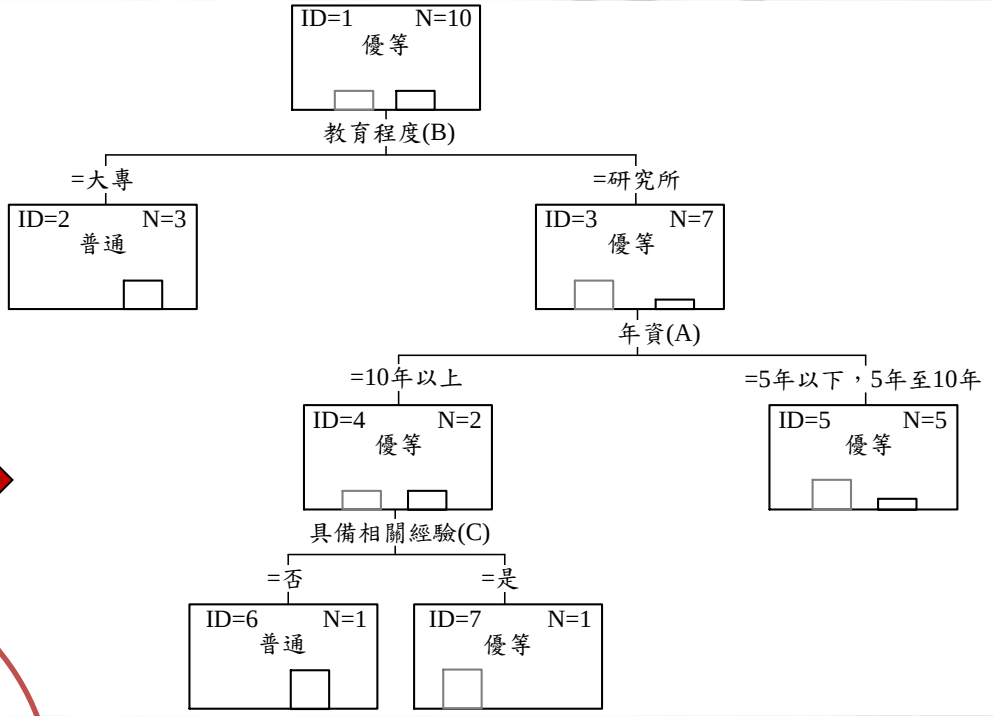
- 為事後修剪方法
- 同時考慮分類錯誤率以及決策樹的規模大小
  1. 以排列組合的方式列出數種修剪後的決策樹；
  2. 計算這些樹的分類錯誤率(classification error)與決策樹複雜度，即節點個數，並找出具有最小誤差的決策樹
- 分類錯誤率會隨著修剪分枝的數目呈正比遞增
- 對某一棵決策樹其成本—複雜性的定義為決策樹節點個數與分類錯誤率的函數

$$R_{\alpha}(t) = R(t) + \alpha \times N_{\text{leaf}}$$

[範例4.1] 決策樹修剪示例



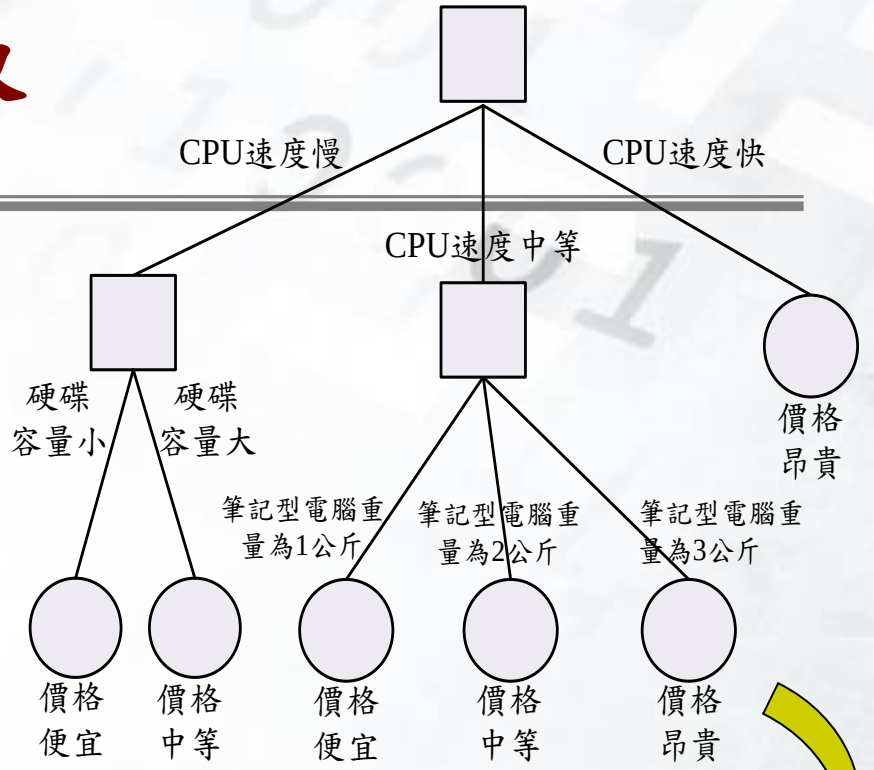
未修剪決策樹



修剪後決策樹

規則萃取

- 利用決策樹萃取資料中所隱含的資訊
- IF-THEN規則即為從根節點至葉節點的可能路徑(path)
- 沿著可能路徑可串連起作為分枝變數的屬性，形成一套具因果關係的分類模型，以分類資料並預測



IF	THEN
若「CPU速度慢」，且「硬碟容量小」	筆記型電腦的價格是「便宜」
若「CPU速度慢」，且「硬碟容量大」	筆記型電腦的價格是「中等」
若「CPU速度中等」，且「電腦重量為1公斤」	筆記型電腦的價格是「昂貴」
若「CPU速度中等」，且「電腦重量為2公斤」	筆記型電腦的價格是「中等」
若「CPU速度中等」，且「電腦重量為3公斤」	筆記型電腦的價格是「便宜」
若「CPU速度快」	筆記型電腦的價格是「昂貴」

屬性重複選取

- 重複選取屬性不僅產生多餘的規則，也會造成決策樹過於龐大而不容易解釋，適當地合併規則，可使決策樹更具效率
- 【範例】
  - IF「屬性*U*小於10、屬性*V*為*i*或*j*、屬性*U*小於5、屬性*V*為*j*」，THEN「被歸納於類別*A*」
  - IF「屬性*S*為*p*、屬性*T*大於30、屬性*W*為*l*或*m*、屬性*T*大於40」，THEN「被歸納於類別*C*」

透過合併的方式可形成以下的新規則

  - IF「屬性*U*小於5、屬性*V*為*j*」，THEN「被歸納於類別*A*」
  - IF「屬性*S*為*p*、屬性T大於40、屬性*W*為*l*或*m*」，THEN「被歸納於類別*C*」

決策樹的演算法

演算法		CART	C4.5/C5.0	CHAID
處理資料型態		離散、連續	離散、連續	離散
連續型資料分枝方式		只分2枝	不受限制	無法處理
分枝準則	類別型相依變數	Gini分散度指標	資訊增益比	卡方檢定
	連續型相依變數	變異數縮減	變異數縮減	卡方檢定或F檢定（需先轉化為類別變數）
分枝方法	類別型獨立變數	二元分枝	多元分枝	多元分枝
	連續型獨立變數	二元分枝	二元分枝	多元分枝（需先轉化為類別變數）
修剪方法		成本複雜性修剪	基於錯誤的修剪	無

CART

- 分類與迴歸樹演算法(Classification and regression tree, CART)以Gini係數作為決定分枝變數的準則，在每個分枝節點進行資料分隔，建立二分式的決策樹，以決定最佳分枝變數
- 給定節點 t，以Gini係數對分枝變數進行二元分割，假設屬性的分枝水準為 s，t_{left} 與 t_{right} 分別為節點 t 的左、右子節點，並比較分枝前後的純度差異

$$\Delta Gini(s,t) = Gini(t) - [Gini(t_{left}) + Gini(t_{right})]$$

  - $\Delta Gini(s,t) > 0$ 表示子節點的純度比其父節點高→不考慮分枝
  - $\Delta Gini(s,t) \leq 0$ 表示子節點的純度比其父節點低→候選分枝

- CART會選擇所有候選分枝變數中具有最大純度作為節點分枝

C4.5 / C5.0

- 以資訊增益比作為決定分枝變數的準則，為多元分枝決策樹，常用於處理類別型資料
- 提供較佳的準確性及資料解釋能力，C5.0是C4.5的進階版
- 假設給定一個節點 t，依據資訊增益比的結果，徹底搜尋並選擇具有最大資訊增益比的分枝變數，節點的資訊增益比>0，則繼續分枝，直到所有節點的資訊增益比均< 0

C4.5 決策樹修剪

- 採用基於錯誤的修剪(error-based pruning)以比較一個父節點和其子節點的純度
- 採用悲觀式估計分類錯誤率的概念，並直接用訓練資料的結果估計分類錯誤率
- 假設在某一個葉節點有*N*筆資料，其中有E筆資料分類錯誤，可能的分類錯誤率應該大於E/N，將E筆錯誤資料視為在*N*次試驗中可能發生的結果，可能發生錯誤的次數為 $0,1,\cdots,E$ ，給定一信心水準(confidence level, CL)下，則該葉節點預測錯誤的機率

$$CL = \sum_{x=0}^{E} C_x^N \, p^x (1-p)^{N-x}$$

CHAID

- **AID(Automatic Interaction Detection)演算法的延伸，根據卡方檢定統計量的顯著性檢定，決定最佳分枝屬性，為多元分枝決策樹演算法**
- **由使用者制訂合併的門檻值 α_1 與分割的門檻值 α_2、α_3，將每個屬性值視為不同群組，採用兩兩分枝檢定的方式，計算出用於檢定兩分枝是否有顯著差異的*p*-value值**
  - 若該 *p*-value >α_1，則合併此兩分枝成為新群組
  - 檢定結果為顯著且*p*-value < α_2時，將該節點中不同類別的樣本劃分至不同的分枝節點
  - 以Bonferroni調整*p*-value係數，*p*-value < α_3 的屬性中挑選最顯著的屬性作為分枝節點

分類模型評估 (1/3)

- 決策樹分類模式的分類結果，可從兩方面去評估：
  1. 以測試組資料的結果來客觀評估較佳的決策樹模型
  2. 由領域專家根據問題背景選出最適合的決策樹模型

預測類別 / 實際類別	Class 1	Class 2
Class 1	TP（true positive）	FN（false negative）
Class 2	FP（false positive）	TN（true negative）

- 根據分類結果，可計算出正確率或分類錯誤率

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad Error\ \text{rate} = 1 - Accuracy = \frac{FP + FN}{TP + TN + FP + FN}$$

分類模型評估 (2/3)

敏感度 (Sensitivity)

該類別確實正確被預測的比率

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$$

準確度 (Specificity)

為另一類別且確實被分為另一類別的比率

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}$$

精準率 (precision)

預測類別中，多少比率的資料剛好屬於該類別

$$p = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

回想率 (recall)

實際為某類別，同時被判為該類別的比率

$$r = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

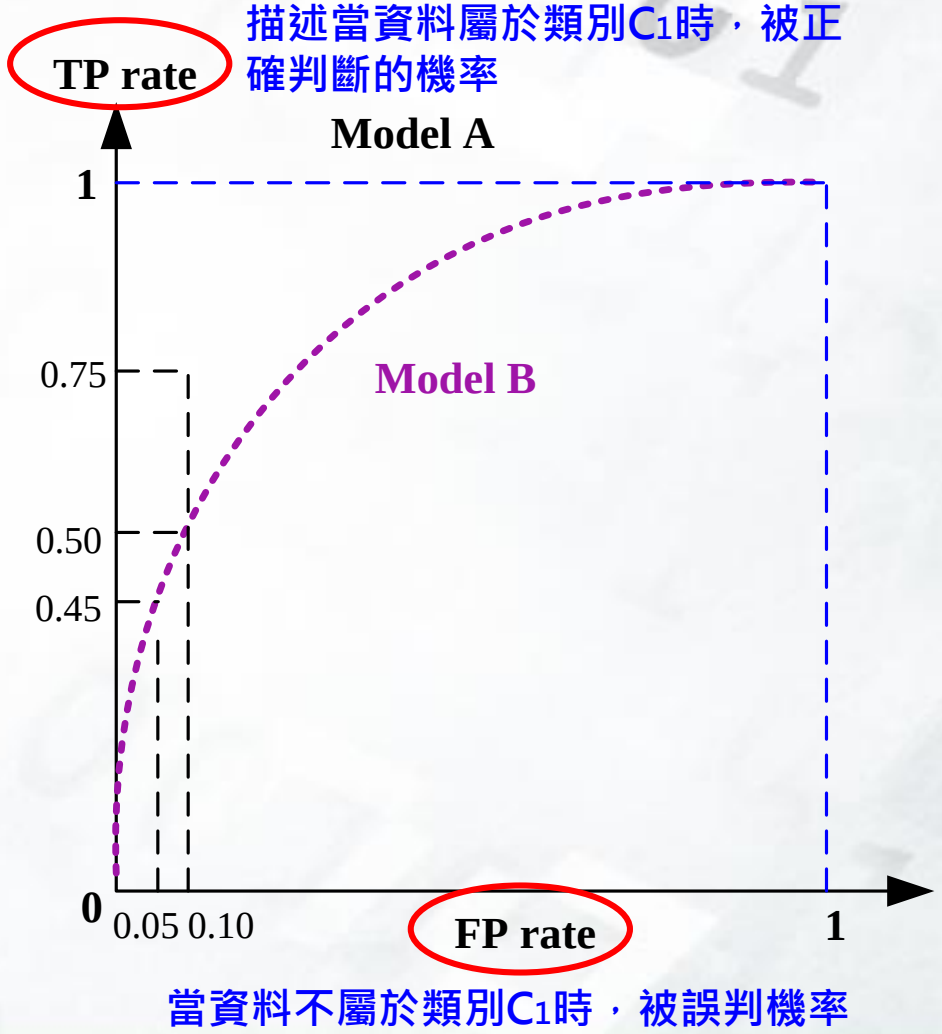
兩指標合併

$$\text{F}_1 = \frac{2}{\frac{1}{r} + \frac{1}{p}} = \frac{2rp}{r + p}$$

分類模型評估 (3/3)

- **ROC曲線:可作為衡量不同FP rate下TP rate的變化**
  - TP rate為越大越好
  - FP rate為越小越好
  - 準確度=1－FP rate，敏感度(TP rate)增加，準確度也會減少，即FP rate會增加

- **FP rate結果會根據分類門檻值變化，可根據ROC曲線的面積選擇最佳分類結果模式**
  →面積越大，模式分類效果越好



資料挖礦個案研究

建構cDNA 生物晶片之資料挖礦模式

PDF

Reference:

簡禎富、林國勝（2006），建構cDNA 生物晶片之二元資料挖礦模式及其實證研究，資訊管理學報，13(4)，133-159。

案例背景

- 針對生物晶片上cDNA資料發展資料挖礦理論模式，從中刪除不敏感之基因並擷取可能具影響力之基因，並從分析過程中獲得有意義之樣型或歸納出規則，搜尋出基因在正常人與病人不同的表徵，以及藉由瞭解基因與致病因子之間的關聯，並進一步萃取建立決策規則（簡禎富、林國勝，2006）

生物晶片資料之決策樹建構 (1/2)

- 本研究整合Significance Analysis of Microarrays(SAM)法及決策樹分析建構決策樹架構
- SAM法粗略篩選敏感(sense)基因
  - 根據基因表現值在不同狀態下之差異，利用重複量測及差異標準化的方式，給定各基因相對評分，超過門檻值之基因，判定該基因變化具統計顯著差異
  - 最後利用不同樣本之排列與重複量測的方式評估各基因之鑑別率(FDR)。由於基因表現值的變動對基因具專一性，首先定義各基因之相對差 ：

$$d_i = \frac{r_i}{s_i + s_0}, i = 1,2,...p \qquad r_i = \frac{\sum_j y_i (x_{ij} - \bar{x}_i)}{\sum_j (y_j - \bar{y}_j)^2}$$

生物晶片資料之決策樹建構 (2/2)

- **以降低變異為分枝準則進階篩選敏感(sense)基因**
  - 目標變數定義為病人(0)或正常人(1)，以SAM篩選後的顯著基因為分枝變數進行長樹
  - 若基因為連續尺度，則對基因區間進行分組，並計算各組檢定*p*-value值
  - 將所有顯著基因作為候選分枝變數，再從中選出可降低最多變異的基因作為分枝變數
- **長樹與規則萃取**
  - 考量在小樣本下，採用重複抽樣方法反覆進行交互驗證評估建構模式之效度，以求取最佳決策規則及顯著基因
  - 驗證作法為依研究個案分別訂定決策規則正確率、偽陰性(FN)及偽陽性(FP)門檻作為驗證標準

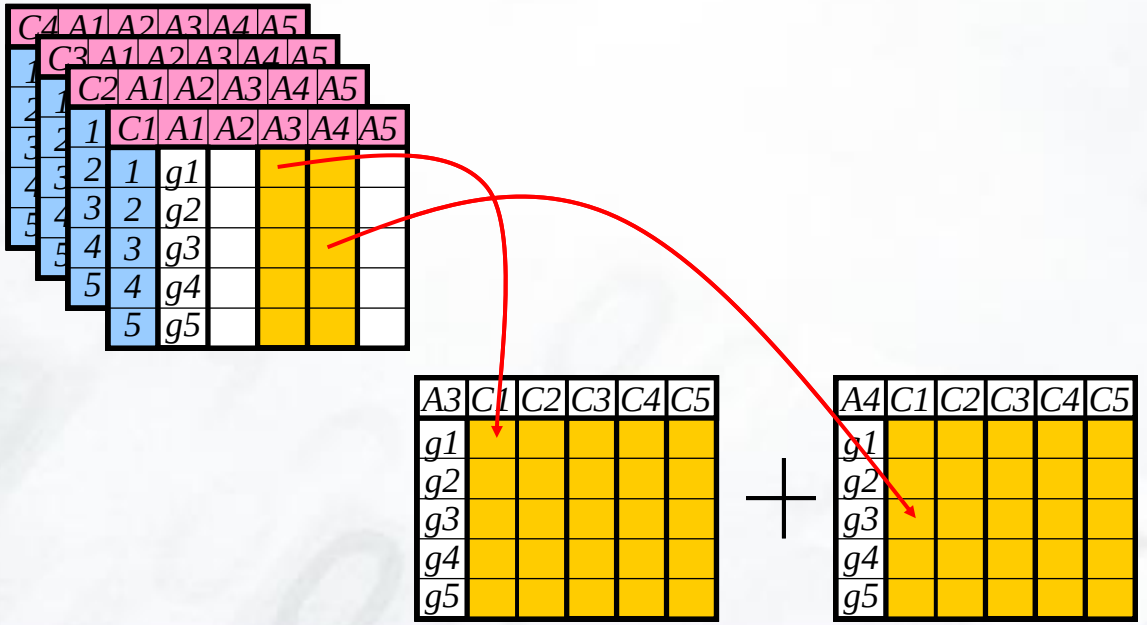
案例資料分析

- 本案例選用史丹佛大學的生物晶片資料庫(Stanford Microarray Database; SMD)(http://smd.princeton.edu/)中乳癌實驗晶片cDNA資料進行研究

	A	B	C	D	E	F	AN	AO	AP	AQ	AR	AS	AT
12	!Country=USA												
13	!SlideName=shbg054												
14	!Printname=SHBG												
15	!Channel 1 Description=Stratagene_aT												
16	!Channel 2 Description=BC-D-091_aT												
17	!Scanning Software=GenePix												
18	!Software version=Pro												
19	!Scanning parameters=PMTVolts=630650LaserPower=5.294.05												
20	Spot	Clone ID	Gene Symbol	Gene Name	Cluster ID	Accession	Channel 1	Ch1 Backg	Std Dev of	Ch1 Net (I	Ch2 Intens	Ch2 Net (I	Ch2 Net
21	1	IMAGE:34	ITGB2	integrin, beta	Hs.375957	W68291	252	235	102	3385	2488	2132	212
22	2	IMAGE:233721				H78560	292	256	138	5238	2820	2478	236
23	3	IMAGE:50	MTIF2	mitochondria	Hs.149894	H18070	271	252	106	10591	6245	5894	605
24	4	IMAGE:12	HMBS	hydroxymeth	Hs.82609	R06263	275	255	83	5514	3385	3034	283
25	5	IMAGE:23	GATA6	GATA bindi	Hs.50924	H77651	265	242	99	19392	9912	9564	992
26	6	IMAGE:18	ICAM5	intercellular	Hs.151250	R87763	293	267	137	4249	3498	3147	314
27	7	IMAGE:80	SEMA3B	sema domain	Hs.82222	AA455145	315	235	624	34020	22669	22333	2104
28	8	IMAGE:18	RIN2	Ras and Rab	Hs.446304	R83223	268	173	688	4744	5723	5462	570
29	9	IMAGE:29	CSE1L	CSE1 chrom	Hs.90073	N69204	139	70	109	1348	1177	1087	43
30	10	IMAGE:119384				T94272	180	188	124	419	1427	1155	6
45713	46069	IMAGE:36	CENTB5	centaurin, be	Hs.21446	AA024391	229	222	77	126	607	254	17
45714	46070	IMAGE:81	PIK3R2	phosphoinos	Hs.211586	AA485731	261	219	584	2029	2428	2059	203
45715	46071	IMAGE:78	MGC33630	hypothetical	Hs.359981	AA448270	277	215	682	300	1182	807	59
45716	46072	IMAGE:81	LOH12CR1	loss of heterc	Hs.105040	AA459384	235	218	143	1378	1351	952	85

資料準備

- 首先整理各別晶片資料去除冗餘及不需要之名目欄位
- 再去除不需要及無效之數值欄位，如Accession No.名稱遺失。
- 再去除個別Accession No.遺失值過多者（20%遺失值，本資料即遺失值超過25個），不予做後續分析



生物晶片資料之決策樹建構

- 將處理過之病人與非病人各64筆資料所彙整之資料表，任意成對挑選出各50筆、共100筆作為訓練資料，剩餘各14筆、共28筆做為最後驗證用資料，並重複抽樣五次
- 設定門檻值△=3

$$\left| d_{(i)} - \bar{d}_{(i)} \right| \geq \Delta$$

List of Significant Genes for Delta = 3

Positive genes (1295)

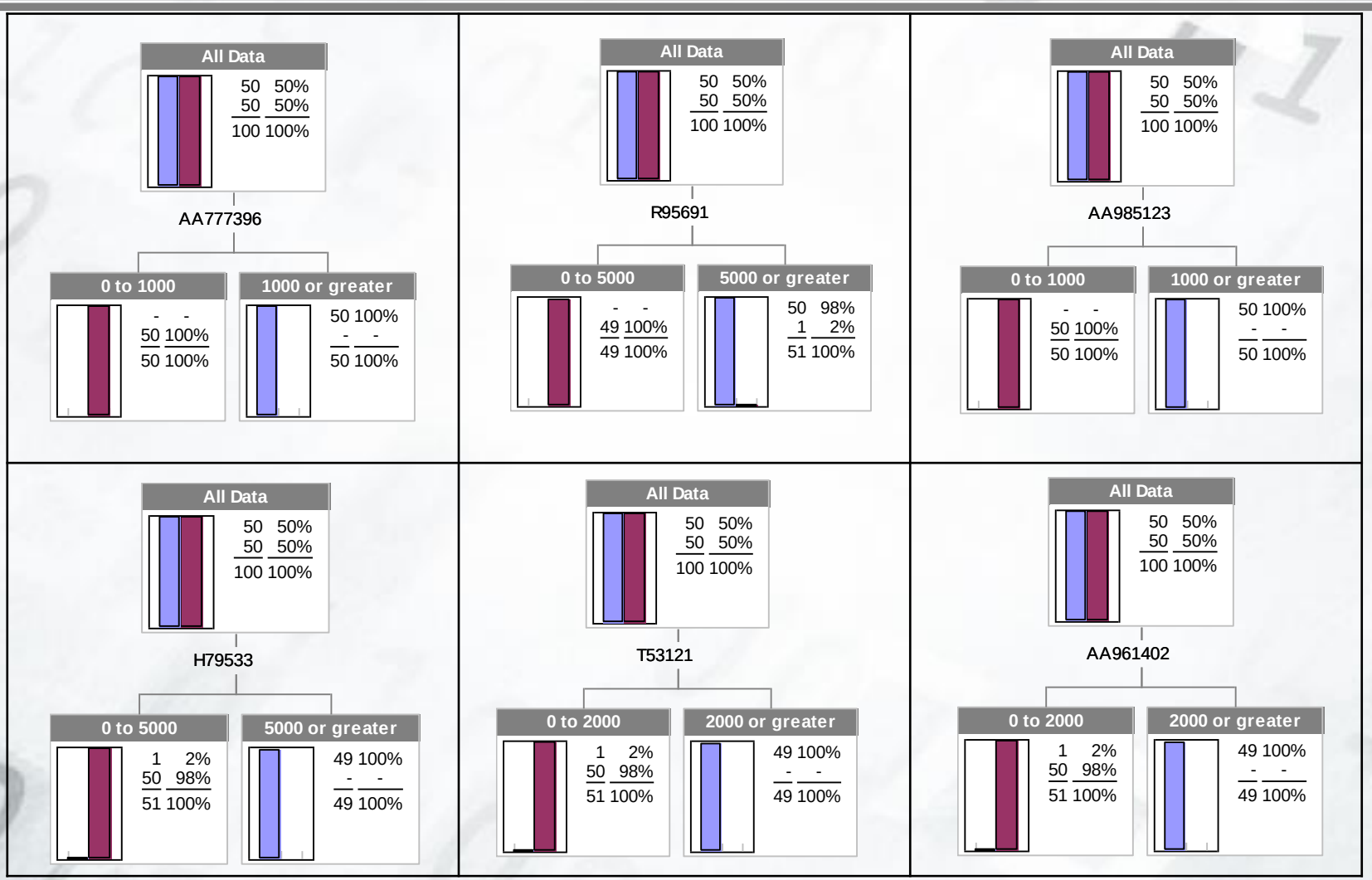
Row	Gene ID	Gene Name	Score(d)	Numera	Denominato	Fold Change	q-value	localfdr(%)
9249	AA490471	9248	11.68782239	7929.85	678.471124	-14.39202537	0	0.00618
990	AA044829	989	10.89221312	21239.1	1949.93178	-41.13201974	0	0.05313
3335	AA232837	3334	10.65158735	2309.83	216.853125	-2.35542618	0	0.06679
9465	AA496741	9464	10.34652123	3194.67	308.767549	-9.347879281	0	0.08293
2441	AA155913	2440	10.24891263	15161.8	1479.35986	65.78418633	0	0.08765
8758	AA486362	8757	10.11791467	7443.35	735.660483	-18.18078571	0	0.09349
8207	AA476918	8206	9.983305843	18902.7	1893.43393	-10.37560848	0	0.09873
1474	AA082747	1473	9.637078379	4599.03	477.222434	3.042728346	0	0.10741
2461	AA156571	2460	9.574580254	4453.19	465.105507	2.088080434	0	0.10801
3362	AA233805	3361	9.534615191	981.73	102.964827	-1.497639757	0	0.1082
5126	AA421258	5125	9.485455034	1667.21	175.764894	-4.076386858	0	0.10821
####	AA598653	10309	9.415108893	13221.5	1404.28859	68.6539458	0	0.10779
6107	AA442984	6106	9.17529274	4425.43	482.320306	-44.4454479	0	0.1018
8646	AA485739	8645	8.960838471	10783.1	1203.35279	-51.21086673	0	0.08924
4435	AA402920	4434	8.857941956	10553.3	1191.39751	-54.64247389	0	0.08029
733	AA031287	732	8.85792581	4150.83	468.600222	13.18695136	0	0.08028
7931	AA464246	7930	8.854319481	11811	1333.92408	3.847989778	0	0.07993
2516	AA158584	2515	8.72843308	6137.33	703.142241	3.928536536	0	0.0659
7034	AA455026	7033	8.647998638	2550.51	294.924885	11.41879499	0	0.05505
8783	AA486532	8782	8.580556798	7848.83	914.72269	2.332697647	0	0.04473
898	AA041382	897	8.552551284	4799.59	561.188099	-34.76479947	0	0.0401

Sep 1 (Imputed) / SAM Work (Do not edit!) / SAM Sample Size Plotsheet / SAM Plot / SAM Output /

決策樹規則 (1/2)

- 彙整If Then規則及其分枝正確率(判定為乳癌病人之分枝)、平均正確率及模式解釋力等資訊
- 如50/50(5)為此規則50人為判定乳癌患者，實際患者亦為50人，此規則在五次分析中出現五次；整體正確率為計算所有正確判別情形；
  - **Rule 1:** IF (AA777396 < 1,000) THEN patients
    （若基因AA777396 < 1,000，則判定為患有乳癌）
  - **Rule 2:** IF (AA985123<1,000) THEN patients
    （若基因AA985123<1,000，則判定為患有乳癌）
  - **Rule 3:** IF (AA961402<2,000) THEN patients
    （若基因AA961402<2,000，則判定為患有乳癌）

決策樹規則 (2/2)


All Data
50 50%
50 50%
100 100%
AA777396
0 to 1000
- -
50 100%
50 100%
1000 or greater
50 100%
- -
50 100%
All Data
50 50%
50 50%
100 100%
R95691
0 to 5000
- -
49 100%
49 100%
5000 or greater
50 98%
1 2%
51 100%
All Data
50 50%
50 50%
100 100%
AA985123
0 to 1000
- -
50 100%
50 100%
1000 or greater
50 100%
- -
50 100%
All Data
50 50%
50 50%
100 100%
H79533
0 to 5000
1 2%
50 98%
51 100%
5000 or greater
49 100%
- -
49 100%
All Data
50 50%
50 50%
100 100%
T53121
0 to 2000
1 2%
50 98%
51 100%
2000 or greater
49 100%
- -
49 100%
All Data
50 50%
50 50%
100 100%
AA961402
0 to 2000
1 2%
50 98%
51 100%
2000 or greater
49 100%
- -
49 100%

模式驗證 (1/2)

- **重複抽樣所剩餘的28筆資料當作測試集進行模式規則驗證**
  - 醫學上偽陰性較偽陽性顯得重要
  - 根據生物晶片與生物資訊領域知識，若偽陰性高於10%則該規則予以刪除，若偽陽性高於20%時則刪除
- **將測試集資料分別帶入各次分析中所挖掘出的決策規則中，Category為正常人 / 病人（0/1）之分類**

AA777396 <1000	AA985123 <1000	AA961402 <2000	H79533 <5000	T53121 <2000	R95691 <5000	AA701996 <2000	T98611 >5000	AI380522 <2000	W56522 <2000	AI001134 <2000	AI923787 >2000	Category
6646	3270	6919	19470	7150	10679	9590	462	5181	21011	5169	194	0
3635	6582	559	11294	8057	21355	9877	280	5688	19294	8149	199	0
7611	3548	1298	12609	12873	18847	10252	390	4537	18459	3109	274	0
6375	5971	2626	11609	12154	34794	9945	377	4028	4844	6719	325	0
8428	13564	3968	23179	16591	45387	20716	515	7210	16203	10294	558	0
4311	8384	4765	29194	17105	37955	22682	965	11827	26230	9488	412	0
4678	4479	2901	14241	7664	27040	12798	649	7653	12416	8884	243	0
5235	7010	661	16918	6467	38116	25514	638	11762	20313	8371	776	0
1638	6245	1822	12946	4927	34315	17677	345	13167	15963	6345	286	0
1255	5606	561	13747	7625	29581	15286	317	10496	16971	6195	221	0
1120	5057	1190	11276	8080	24317	17162	258	10774	13370	7040	276	0

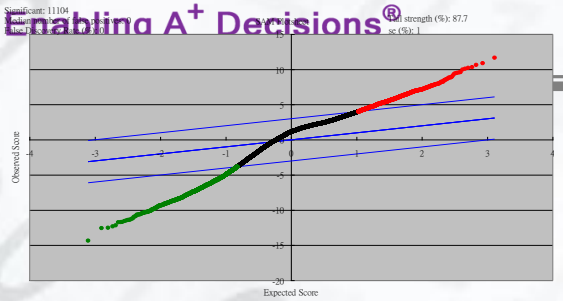
模式驗證 (2/2)

- 當基因AA777396檢測值小於1,000時，則判定為患有乳癌，大於1,000時則為正常人；
- 經過五次驗證資料測試後刪除下列四條規則（偽陰性高於10%）
  - IF (T53121 < 2,000) THEN patients
  - IF (AA961402 < 2,000) THEN patients
  - IF (AA938940 < 1,000) THEN patients
  - IF (AA913206 < 5,000) THEN patients

項次	決策規則	分枝正確分類率（次數）	平均整體正確率	平均模式解釋力
1	IF (AA777396 < 1,000) THEN patients	50/50(5)	1.00	100%
2	IF (AA985123 < 1,000) THEN patients	50/50(5)	1.00	100%
3	IF (R95691 < 5,000) THEN patients	49/49(4) 50/50(1)	0.99	97%
4	IF (H79533 < 5,000) THEN patients	50/51(5)	0.99	96%
9	IF (AA701996 < 2,000) THEN patients	48/48(4) 49/49(1)	0.98	93%
10	IF (AI380522 < 2,000) THEN patients	48/48(4) 49/49(1)	0.98	93%
11	IF (T98611 > 5,000) THEN patients	48/48(5)	0.98	92%
12	IF (AI679372 > 5,000) THEN patients	48/48(1)	0.98	92%
13	IF (AA233079 < 5,000) THEN patients	50/52(3)	0.98	92%
14	IF (W56522 < 2,000) THEN patients	50/52(5)	0.98	92%
15	IF (AI001134 < 2,000) THEN patients	50/52(4) 48/48(1)	0.98	92%
16	IF (AI375135 < 4,000) THEN patients	50/52(2)	0.98	92%
17	IF (AI923787 > 2,000) THEN patients	49/50(4) 48/48(1)	0.98(5)	92%
18	IF (H52245 > 2,000) THEN patients	48/48(2)	0.98(2)	92%
19	IF (W01204 < 2,000) THEN patients	50/52(3)	0.98(3)	92%
20	IF (AA486362> 2,000) THEN patients	49/50(1)	0.98	92%
21	IF (H12338> 1,000) THEN patients	49/50(1)	0.98	92%

基因選取與規則建立模式

Enabling A+ Decisions®

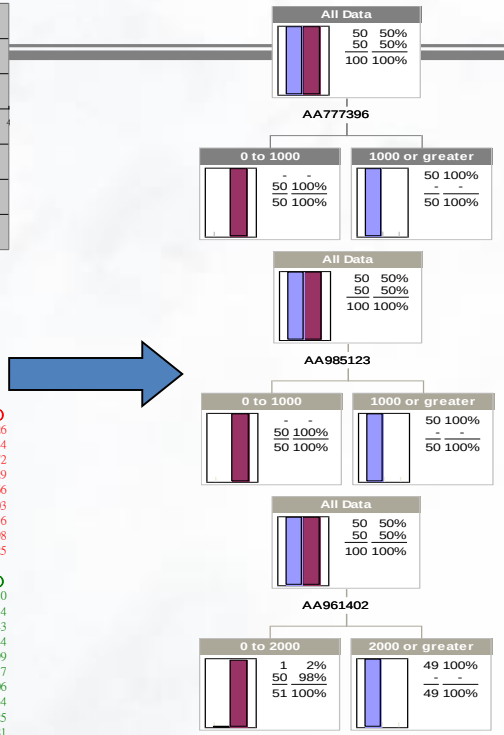


Positive genes

Row	Gene ID	Gene Name	Score(d)	Numerator(r)	Denomina	Fold Change	q-value(%)	localfdr(%)
9249	AA490471	9248	11.68782	7929.85	678.4711	-14.3920254	0	0.0061826
990	AA044829	989	10.89221	21239.0725	1949.932	-41.1320197	0	0.05312944
3335	AA232837	3334	10.65159	2309.83	216.8531	-2.35542618	0	0.06678572
9465	AA496741	9464	10.34652	3194.67	308.7675	-9.34787928	0	0.08293329
2441	AA155913	2440	10.24891	15161.83	1479.36	65.78418633	0	0.08765466
8758	AA486362	8757	10.11791	7443.35	735.6605	-18.1807857	0	0.09348803
8207	AA476918	8206	9.983306	18902.73	1893.434	-10.3756085	0	0.09872816
1474	AA082747	1473	9.637078	4599.03	477.2224	3.042728346	0	0.1074098
2461	AA156571	2460	9.57458	4453.19	465.1055	2.088080434	0	0.10800825

Negative genes

Row	Gene ID	Gene Name	Score(d)	Numerator(r)	Denomina	Fold Change	q-value(%)	localfdr(%)
2741	AA173619	2740	-14.3232	-7197.25	502.4893	0.063547829	0	0
9069	AA489080	9068	-12.5495	-2896.97	230.8434	0.270494599	0	0.00805814
9281	AA490594	9280	-12.4941	-5781.39	462.7298	-0.04048957	0	0.00922343
7580	AA459663	7579	-12.3155	-7059.03	573.184	0.313167519	0	0.01330544
1708	AA115761	1707	-12.155	-2640.65	217.2485	-0.18978784	0	0.01742699
7897	AA464019	7896	-11.7179	-2392.75	204.1956	0.108561802	0	0.03102317
10449	AA599175	10448	-11.7041	-11775.61	1006.107	0.270747364	0	0.03150906
6801	AA453335	6800	-11.6773	-5422.37	464.3498	0.250893608	0	0.03246234
3346	AA233079	3345	-11.5563	-19815.27	1714.672	0.008910178	0	0.03692225
7046	AA455067	7045	-11.4871	-3268.87	284.5686	-0.19965295	0	0.03958281



整理規則正確率90%以上者

偽陽性、偽陰性驗證

AA777396 <1000	AA985123 <1000	AA961402 <2000	H79533 <5000	T53121 <2000	R95691 <5000	AA701996 <2000	T98611 >5000	A1380522 <2000	W56522 <2000	AI001134 <2000	AI923787 >2000	Category
6646	3270	6919	19470	7150	10679	9590	462	5181	21011	5169	194	0
3635	6582	559	11294	8057	21355	9877	280	5688	19294	8149	199	0
7611	3548	1298	12609	12873	18847	10252	390	4537	18459	3109	274	0
6375	5971	2626	11609	12154	34794	9945	377	4028	4844	6719	325	0
8428	13564	3968	23179	16591	45387	20716	515	7210	16203	10294	558	0
4311	8384	4765	29194	17105	37955	22682	965	11827	26230	9488	412	0
4678	4479	2901	14241	7664	27040	12798	649	7653	12416	8884	243	0
5235	7010	661	16918	6467	38116	25514	638	11762	20313	8371	776	0
1638	6245	1822	12946	4927	34315	17677	345	13167	15963	6345	286	0
1255	5606	561	13747	7625	29581	15286	317	10496	16971	6195	221	0
1129	5952	1189	11276	8080	24317	17162	358	10774	13379	7049	276	0
3044	5519	3754	15935	14117	43556	14981	718	19371	21673	7015	682	0
5075	1799	8528	33827	2861	24515	3565	297	10760	25192	3157	264	0
2723	2033	10423	36631	3197	20740	5154	360	9620	25991	3554	329	0
157	131	354	355	410	385	291	32638	433	362	248	3957	1
-164	-27	-25	800	177	727	103	18508	466	241	-34	11488	1
366	36	100	291	432	800	279	29981	128	240	4	7377	1
-34	79	7	86	306	1360	149	36920	25	15	51	7616	1
-391	-147	-80	40	325	1943	54	40426	302	31	-149	21041	1
43	-50	-22	323	216	791	121	21914	268	56	40	7864	1
36	-5	306	174	190	1038	202	24192	79	19	88	4853	1
-78	20	-26	221	145	1544	389	40903	299	83	11	22389	1
-165	-10	-24	14	-145	1565	92	30584	208	52	-35	9050	1
-40	-28	12	133	59	989	122	19105	342	109	27	8084	1
-105	256	-88	212	162	876	422	21235	166	92	158	6685	1
-13	239	365	589	631	1390	553	37699	1074	236	415	11384	1
470	655	835	2722	457	1802	269	14051	864	694	785	6378	1
221	709	818	3309	396	1297	439	11675	735	742	557	10540	2

SAM: 五次分別11,104, 12,829, 13,219, 12,770, 13,745個顯著基因

Decision tree: 五次分別12, 14, 18, 14, 16個，聯集21個

自21個中刪除4個規則

規則解釋與評估

- 本研究選取模式解釋能力90%以上的21個基因為醫療檢測參考因子並建立其個別決策規則
- 本案例的決策規則，係純以晶片資料進行分析，後續可整合相關病歷資料，以更深入探討病人基因表現值與不同病人特性之關係
- 本案例所提出之生物晶片決策樹分析提供一個有效的方法，由乳癌實驗晶片cDNA資料分析結果也驗證其模式效度

結論

- 決策樹常扮演特徵萃取與描述的角色，其透過變數的選擇與目標的指定對資料進行分類而成樹枝狀的架構，經常用於解決分類的問題，並作探索與預測
- 決策樹分析對於高維度的資料也可快速學習，並建構層級式的樹狀結構，而挖掘所得的結果也可轉換為一系列容易瞭解的IF-THEN規則，適合用來挖掘未知的樣型