目錄

表目錄

圖目錄

第一章　緒論

1.1　研究背景與動機

人類在整理大量資料的時候，會很自然的做一個動作：分門別類。把某些特徵相近的資料歸在同一類別，改以類別為單位來處理，使得資料的辨識度更高；一則可以處理雜訊的問題，二則可以簡化資料處理的複雜度，使得管理起來更為容易。例如在整理數十萬筆市民資料的時候，可能就會依照每個市民的國籍、年齡、職業、婚姻狀況等等屬性來分類。

為了解決這個問題，自1960 年起資料分群（data clustering）的相關理論與演算法便開始發展，以期能根據每筆資料已確定的部份屬性，將整組資料自動分成幾種類別，又稱做群集（cluster）。即使事前對不同屬性之間的關係及屬性所代表的訊息了解不多，也能結合一些數學上的測量方法，找出資料數量上的相關性。

資料分群的結果，除了能幫助簡化資料、建立資料的分類規則，更可以發掘原本未知的假設；因而應用十分廣泛，包含生命科學、醫學、行為科學，都可以用資料分群的方法來分析資料，以做進一步的評估。例如某疾病的形成原因不明，研究人員就可以試著對該疾病的病患資料分群，由分群結果中歸於同一類的病患共通點，得到一些新的假設方向。

然而目前的群聚演算法，在群聚的過程中，均針對資料本身呈現的特性，按照演算步驟計算。雖然可以將資料分成數類群集，但仍然會有錯誤的產生。歸根究底，資料本身僅能顯現量上的相關性，無法表現質的因果關係；因而正確的群集，除了資料上的表現，更需要的是如何將質的因果關係，即領域知識放入其中。本研究之目的即發展一種方法，可將領域知識放入群聚演算法，使得群聚的結果更為正確。

1.2 研究問題定義

資料群聚的過程中，判定哪幾筆資料屬於同一群集，一直是重要的研究議題。較常使用的方法之一，便是計算各筆資料之間的相似程度。基本概念是：相似的屬於同一群集。多數群聚演算法，僅用資料本身的特徵項目(feature)，計算各筆資料間的相似度；並且對於資料中所有的特徵項目，均視為一樣重要的。但是不同的特徵項目，對於相似度的判斷應該佔有的權重(weight)不同。這時候，領域知識便可以彌補資料本身的不足。

在此將本研究所要處理的問題，以條列說明如下：

1. 檢討並設計新的資料相似度量測標準。使得不同的特徵項目，對相似度的權重不同。
2. 一般的群聚方法，無法將領域知識放入群聚的過程。本研究藉由權重的設計，將專家知識納入群聚機制。
3. 檢驗待修改後的相似度量測標準，以及能放入專家知識的機制建立後，是否可以幫助原本的群聚方法增進正確性。

1.3 研究方法與步驟

本研究將以下列四個階段進行，分別為文獻探討、機制建立、實例驗證及結論，詳細說明如下：

1. 文獻探討部分

   首先說明目前的群聚演算法；不同的群聚方法，適合的資料特性。接著探討相關研究中對於各筆資料的相似度量測方法，並舉例討論各種相似度量測的特性和問題。

2. 機制建立

   依據「不同的特徵項目，對於相似度的判斷應該佔有的權重(weight)不同」，這樣的概念，定義出能反映權重的相似度量測標準。並在演算法當中，設計一能讓專家放入領域知識的機制。

3. 實例驗證

經過實際案例，檢驗根據新定義出來的相似度量測標準之下；並能考慮領域知識所建構之群聚演算法，分群的結果是否正確。並與其他的演算法比較。

4. 結論

經過前面幾個步驟，最後運用系統化的方式，討論其結果以及未來研究方向。

1.4 論文架構

依據以上方法，本篇論文架構共分為五章，如圖 1.1。第一章為緒論，說明本研究期望建立群聚方法機制之研究背景與動機，研究目的及範圍，並概要的說明本研究方法。第二章為文獻探討，針對於目前的群聚演算法進行文獻探討回顧，說明各類的群聚方法適用的範疇。第三章則為研究方法，說明本研究的研究方法及架構。第四章為實證研究，將以實例來驗證本論文設計出來的群聚方法，結果是否正確。第五章為結論與未來發展方向。

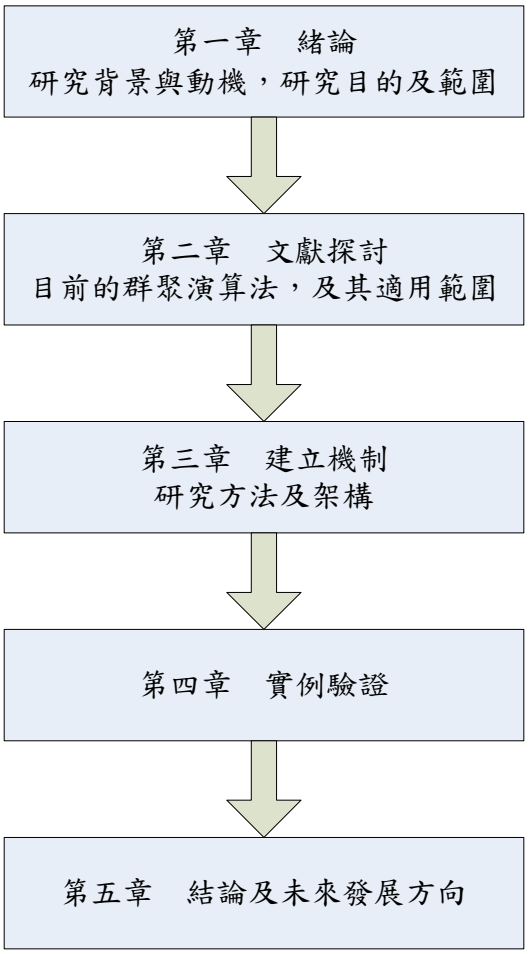
第一章　緒論
研究背景與動機，研究目的及範圍
第二章　文獻探討
目前的群聚演算法，及其適用範圍
第三章　建立機制
研究方法及架構
第四章　實例驗證
第五章　結論及未來發展方向

圖 1.1 研究架構

第二章　文獻探討

群聚技術（clustering technology）在資料探勘（data mining）領域中，是一項非常重要的技術，它可以在大量的資料中，找出資料的分布狀況並找到其隱藏的意義，例如當使用者面臨要分析處理龐大的資料時，往往無法輕易的獲知這些資料所代表的意義，而利用群聚技術可以先將這些資料分成若干個群聚，再針對不同的群聚加以分析，如此，便可以簡化使用者分析資料時的複雜性。

群聚技術的目的是在分析資料的內容，將性質相似的資料群聚在一起，而讓不同的群集間資料相異性大。

在選擇資料群聚演算法時，通常是以資料的型態、數量和分布做為考量的標準，而當我們希望發掘資料集不同層面的訊息時，更可以多參考比較幾種演算法的分群結果。

群聚技術與分類方法(classification)最大不同是，群聚不預先知道資料有哪些類別，而根據資料的內容，將資料歸類成群[7]。類別是根據群聚的結果產生的。而分類方法則是在類別已知的情形下，判斷每筆資料分別屬於哪一群集類別。

目前已有的資料群聚演算法大約可分階層（hierarchical clustering）、分割（partition clustering）、密度基礎（density-based）、格子基礎（grid-based）四種不同種類。以下簡單介紹四者的主要特性及優缺點。

2.1　階層群聚演算法

階層式演算法會將資料分群的過程，用樹狀圖（dendrogram）中不同的層級（level）記錄下來，又可分聚合型（agglomerative）與分解型（divisive）兩種不同的角度，前者屬於 bottom-up 特性，從每個資料物件開始一步步融合，最後全數融合成最後一個群集，而後者正好相反是採 top-down 的方式，兩者的差別可見圖2.1。[8]

由圖2.1 也可以看出，階層式演算法主要的特色是，在某一層級一旦兩資料物件被分在同個群集，一直往上到最後一層級為止，這兩

個資料物件都不會再被分開到其他群集去。

這類演算法的優點是很容易看出資料間相關性的完整架構，早期較著名的聚合型分群演算法有Single-link、Complete-link、Group average、Centroid、Ward，都是以測量兩個不同群集間一對對資料的距離，做為融合的群集的標準。

然而上述方法共同的問題是無法改正先前步驟中的錯誤決定，於是後來有一些演算法試圖改進這一點，例如使用反覆定位（iterative reallocation）的方法修正結果的BIRCH[16]，或加強分析物件間鏈結關係的方法CURE[5]、Chameleon[10]。

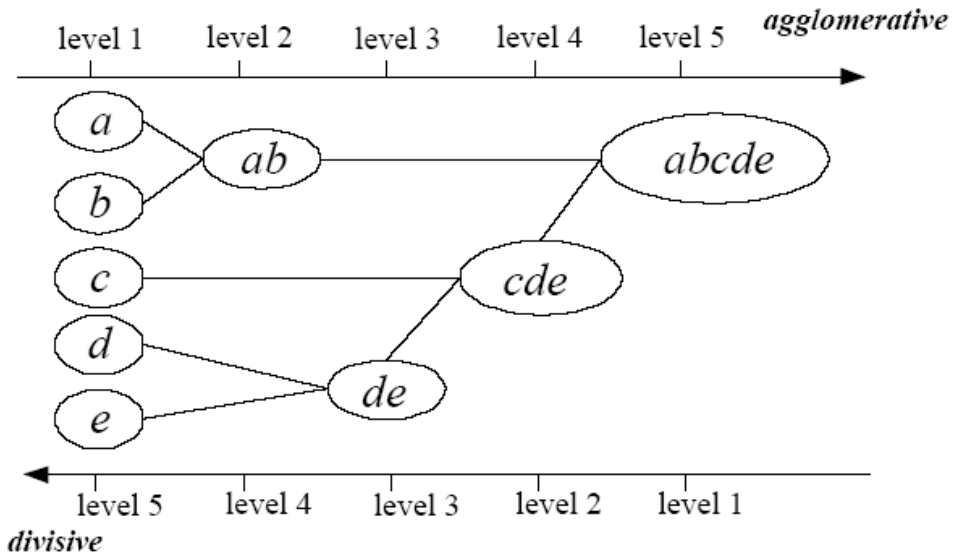

圖 2. 1　聚合型與分解型的演算法，對資料物件$\{a, b, c, d, e\}$分群過程

由於群聚演算法多以兩點之距離為基礎，但此種演算法並不適用於資料屬性為布林值型與種類型之資料。ROCK之群聚方式屬於階層法，以link與相似度概念為群聚之基礎。其群聚成效不僅優於傳統的群聚法，更可以較好的方式呈現群聚結果。[4]

2.1.1 ROCK演算法

ROCK 演算法之 Criterion Function 可確保於群聚到 k 個群組時，群組內之差異為最小，而群組間之差異為最大 (式 2.1)。以正規化因子 (任兩群組間之交叉鏈結數 (the numbers of cross link)除以交叉鏈結數之期望值)評估兩群組合併之適合度。

$$E_l = \sum_{i=1}^{k} n_i * \sum_{p_q, p_r \in C_i} \frac{link(p_q, p_r)}{n_i^{1+2f(\theta)}} \qquad \text{(式 2.1)}$$

ROCK 演算法有 2 個參數，分別為：θ 與 k。θ (thresholds)鄰近點門檻值，依據資料集中任兩筆記錄之相似度判斷兩點是否為鄰近點；k，欲群聚之最少組數。可依應用領域特性選擇計算任兩點之相似度。ROCK 之群聚程序如下述：

最初之群組總數為資料包含之記錄數。

1. 計算任意兩點的相似度。若相似度大於或等於門檻值，則判定此兩點互為鄰近點 (neighbors)。相似性的計算使用 Jaccard 係數：

$$sim(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

例：若 θ=0.5；A={a, c, d}；B={a, b, d, e}；C={a, c, d, e}；D={a, c, e}。使用 Jaccard 係數來評量兩點之間相似度，則

$$sim(i,j) = sim(j,i); \quad i, j = A, B, C, D$$
$$sim(A,B) = |A \cap B| / |A \cup B| = 2/5 = 0.4 < 0.5;$$
$$sim(A,C) = 3/4 = 0.75 > 0.5;$$
$$sim(A,D) = 2/4 = 0.5 = 0.5;$$
$$sim(B,C) = 3/5 = 0.6 > 0.5;$$
$$sim(B,D) = 3/5 = 0.6 > 0.5;$$
$$sim(C,D) = 3/4 = 0.75 > 0.5;$$

互與鄰近點之組合有 (A, C) 、(A, D)、 (B, C)、(B, D)與 (C, D)互為鄰近點。

資料點	鄰近點 1	鄰近點 2	鄰近點 3
A	C	D	
B	C	D	
C	A	B	D
D	A	B	C

2. 計算任意兩點之間鏈結的 (link) 個數。當任意兩點 p_i , p_j 有 n 個共同鄰近點時，其 $links(p_i , p_j)$=n。

   例：由步驟 1 之鄰近點結果計算得

   $link(A,B)$=2；$link(A,C)$=1；$link(A,D)$=1；

   $link(B,C)$=1；$link(B,D)$=1；$link(C,D)$=2；

3. 計算將任兩群組間群聚於同一群族適合度 (goodness measure)。兩群組 C_i 與 C_j 之適合度以 $g(C_i, C_j)$表示(式 2.2)。$link[C_i, C_j]$ (the numbers of cross link between C_i and C_j)表群組 i 與群組 j 之間的鏈結數，將群組 i 之任一點與群組 j 之任一點間的鏈結個數總後得到 $link[C_i, C_j]$。

$$g(C_i, C_j) = \frac{link[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}} \qquad \text{(式 2.2)}$$

其中 $f(\theta) = \dfrac{1-\theta}{1+\theta}$

例：if k=3，

$f(\theta)$=(0.5/1.5)=0.667；1+2 $f(\theta)$=2.334；

$g(A,B)=2/(2^{2.334} - 1^{2.334} - 1^{2.334})=0.6575$；

$g(A,C)=1/(2^{2.334}-1^{2.334}-1^{2.334})=0.3287$；

$g(A,D)=1/(2^{2.334}-1^{2.334}-1^{2.334})=0.3287$；

$g(B,C)=1/(2^{2.334}-1^{2.334}-1^{2.334})=0.3287$；

$g(B,D)=1/(2^{2.334}-1^{2.334}-1^{2.334})=0.3287$；

$g(C,D)=2/(2^{2.334}-1^{2.334}-1^{2.334})=0.6575$；

4. 選取群聚適合度最大值之兩群組，將其合併為一個新群組，並刪去原有兩個群組。若群聚適合度之最大值為零時，停止演算程序。

   例：群聚適合度最大值之兩群組為(A, B)或(C, D)，若選擇合併(A, B)則群組數為 3 組：(A, B)，(C)，(D)。

5. 重複步驟 1 至 4，直至群組總數為 k，則停止演算程序。

ROCK之演算時間與所需記憶體之複雜性為$O(n^2 + nm_m m_a + n^2 logn)$，$m_m$為鄰近點個數之最大值；$m_a$為鄰近點個數之平均值；$n$為輸入資料之記錄筆數。

2.1.2 Jaccard coefficient

ROCK演算法使用Jaccard coefficient，本研究機制設計為修改相似度量測標準，於此小節介紹其發展。Jaccard coefficient是由Jaccard, P.於1901年所提出，並以此命名。Jaccard的發展，主要與simple matching coefficient相比較。種類型的資料類別，可以轉換為布林資料。而一般布林變數，都有呈現「不對稱」(asymmetric)的現象：0與1所代表的重要性是不同的。對於較重要的結果，都會放在1的布林值，而較為不重要的則分配為0。舉例來說，若有一個布林變數表示HIV test的結果，我們會將HIV positive設定為value=1，而HIV negitive則設定為value=0。計算相似度的時候，考慮同時出現布林value=1(positive match)會相對同時出現布林value=0(negative match)來的顯著。這類的相似度量測標準，稱為「noninvariant similarity」。其中最常見的noninvariant similarity，便是Jaccard coefficient[8]。

2.2 分割群聚演算法

切割式群聚演算法是發展最早的群聚技術，這一類的演算法使用者必須先決定所要分割的群聚數目，再以重心點基礎（centroid-based）或中心點基礎（medoid-based）的方式進行分群。切割式群聚演算法是以距離（distance）作為評估標準，通常評估的方法有：曼哈頓距離（Manhattan distance）與歐幾里得距離（Euclidean distance），目前較重要的方法有K-means[12]、PAM[11]、CLARA[11]以及CLARANS[13]。

K-means是最典型的以重心基礎的切割式群聚演算法，它是以群聚的重心作為群聚的代表點，但因為代表點不一定要是群聚中的一點，所以可以找到最佳的群聚。然而，此方法所得的群聚的品質很容易受到雜訊（noises）或是離群值（outliers）所影響。另一種方法是以中心點作為代表點（如PAM演算法），這些群聚技術對於小型的資料集合（data sets）有著不錯的處理能力，但是隨著資料集合的增加，處理的效率也越來越差，所以，通常處理大型資料庫是採用取樣的方式來解決（如CLARA演算法）。然而取樣的演算法會受到樣本的數量以及取樣方法所影響，倘若樣本數量太少，則群聚的結果不足以代表整個資料庫資料分布的狀況及意義；若取樣的方法不佳，則會影響到群聚的品質。CLARANS是架構於PAM與CLARA上的中心點基礎的切割式演算法，它是第一個針對於空間資料庫所設計的切割式群聚演算法，但只能發掘簡易的資料點分佈形狀（object shapes），對於呈凸多邊形（convex shapes）或巢狀（nested）分佈的資料處理效果不佳，也無法有效率地針對高維度資料進行群聚分析。

2.3 密度基礎群聚演算法

在一個資料集合內，假設有某些資料點分佈密度相當密集，則這些資料點形成一個群聚，換句話說，在群聚內資料分佈的密度應該大於群聚外資料分佈的密度，而密度基礎群聚演算法，便是基於以上的觀念

所發展出來的群聚方法，目前較重要的方法有DBSCAN [3]、OPTICS [1]。

2.4 格子基礎群聚演算法

格子基礎群聚演算法將資料空間量化成許多格子，大量的減少群聚的時間。在這類演算法中，較具代表性的STING [15]，Wave Cluster [14]。 STING演算法是將資料空間切割成格子狀，其群聚方式是由上而下的，利用廣度搜尋將格子內的群聚作合併。STING探索存在格子的統計資訊，然後群聚，其缺點是分群邊緣的形狀不是水平就是垂直，儘管有快速的群聚處理時間，但會有損及其品質及正確率。此外，STING是呈現階層式架構，較高的階層會儲存較低階層資訊的總合，所以查詢的速度非常快速，其時間複雜度為O(k)（k為最底層格子的數）。

第三章　研究方法

3.1 研究資料範圍

本研究探討在群聚演算法當中，新的相似度量測標準，以及領域知識放入機制的建立。由於 ROCK 以 link 與相似度概念為群聚之基礎，其群聚成效優於傳統的群聚法，更可以較好的方式呈現群聚結果；因此本研究選用 ROCK 為基礎，修改相似度量測標準，並設計領域知識放入的機制。在此先定義處理資料的類型：

資料型態為種類型(categorical)之資料。由於 ROCK 演算法不能處理數值型態的資料，所以要面對的資料集，必須先行轉換為種類型態。舉例來說，在 PET 資料當中，如 tumor marker 值原本為 8.35 的數值，必須轉換為「小於 10」這類的種類型資料(資料轉換的部份需要有領域專家協助，確保資料轉換後的正確性)。

3.2 研究工具

本研究原始資料為文字檔格式、Microsoft Access 和 Microsoft Excel 的資料。在此先以 Microsoft Excel 作為轉換格式的工具，將處理過的資料輸入資料庫中，再匯入 MATLAB 做為各項研究步驟的輸入資料。程式平台架設於 Windows XP Professional，以 MATLAB 作為整合介面的工具。MATLAB 應用軟體為一套應用於數值計算、數據視覺化及動態模擬的軟體，故本研究程式當中群聚演算法的機制將以 MATLAB 的程式來撰寫。

3.3 研究機制建立

本研究使用 ROCK 演算法為基礎，修改原演算法當中使用的 Jaccard coefficient 相似度量測，並且提出加入領域知識的機制。3.3.1 小節說明 Jaccard coefficient 與其他數種較常為人所用的相似度量測標準。3.3.2 小節說明修改相似度量測的理由，以及加入領域知識的機制設計。3.3.3 小節根據本研究提出的概念，實際建構一個群聚演算法系統。

3.3.1 幾種常見的相似度量測標準

要有一個好的分群結果，最關鍵的就是要找到物件之間最適合的相似度量測（similarity measure），用以區分出兩物件的相似度是高還是低。對於不同特性的資料，最適合的相似度量測可能不同。若是相似度量測不適合，真正符合期望、相似度高的兩物件，相似度可能會被不適合的相似度量測定得不夠高，而沒有分在同一群，反而和其他不相似的物件混在一起，分群的結果就不是我們期望的。

目前相關研究中，有採用向量空間量測的符合係數(matching coefficient)、Jaccard 係數(Jaccard coefficient)等利用物件的共通性來評估物件間的相似度，或透過機率計算，量測物件間的共同相關性（correlation）來評估物件間的相似度的相互資訊(mutual information)等等。而共同相關性與相互資訊法，均適合使用在數值型資料；雖然兩者的使用範圍均相當廣，但對於種類型的資料，應用上有其困難。表3.1整理幾種適合種類型資料的向量空間相似度量測定義。

表 3. 1　幾種向量空間的相似度量測定義

相似度量測(Similarity measure)	定義(Definition)
Jaccard 係數(Jaccard coefficient)	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$
符合係數(matching coefficient)	$\lvert X \cap Y \rvert$
Dice 係數(Dice coefficient)	$\frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
重疊係數(Overlap coefficient)	$\frac{\lvert X \cap Y \rvert}{\max(\lvert X \rvert, \lvert Y \rvert)}$
cosine	$\frac{\lvert X \cap Y \rvert}{\sqrt{(\lvert X \rvert \times \lvert Y \rvert)}}$

其中對於種類型問題，較常見到的，還是 Jaccard 係數跟符合係數。以一個例子來比較，符合係數、Jaccard 係數之間的差異。

假設資料庫當中有 10 筆資料，如表 3.2 所示：

表 3. 2　範例資料

資料	Item lists				
D1	A	B	C		E
D2		B	C		
D3	A	B	C		
D4	A		C		
D5	A	B			
D6		B	C		
D7	A	B		D	E
D8		B			
D9	A		C		
D10		B	C	D	E

若是使用符合係數當作相似度的評斷標準，它比較兩筆資料之間，相同的項目有幾種：

$sim(X,Y) = |X \cap Y|$

sim(D1,D2)=2

sim(D2,D3)=2

sim(D1,D4)=2

sim(D1,D7)=3

若是使用 Jaccard 係數當作相似度的評斷標準，它比較兩筆資料之間，「聯集分之交集」：

$$sim(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

sim(D1,D2)=1/2=0.5

sim(D2,D3)=2/3=0.67

sim(D1,D4)=1/2=0.5

sim(D1,D7)=3/5=0.6

因為Jaccard係數這種相似度量測比符合係數這種相似度量測多了正規化（normalize）的考量，相對於符合係數只比較相同項目，是比較適合本研究的相似度量測方式。也是目前最常被使用的相似度量測。

將兩種不同的相似係數做成表格比較：

表 3. 3　符合係數與 Jaccard 的比較

	符合係數	Jaccard 係數
sim(D1,D2)	2	0.5
sim(D2,D3)	2	0.67
sim(D1,D4)	2	0.5
sim(D1,D7)	3	0.6

可以發現：

符合係數不及 Jaccard 係數，因為它僅比較兩兩之間的相同項目個數，而在這種相似度衡量之下，完全無法區分出 D1、D2、D3 與 D4 之間的差異。

3.3.2 修改相似度量測標準，並提出加入領域知識的機制

Jaccard 係數是目前最常被使用的相似度量測，本研究參考的 ROCK 演算法便是使用 Jaccard 作為相似度量測的標準。

但是 Jaccard 係數有其缺點。當某一項目在資料當中出現的機會很高時，會在估計相似度時，往往有「高估」的情形發生。這是應該要解決的部份。舉例來說：若研究超商的顧客，是否有數種顧客群；針對不同顧客做不同促銷方案。將顧客在超商購買商品的種類，蒐集成為資料集。顧客 A 購買「烏龍茶」，順便買了包「免洗杯」。顧客 B 購買「雪碧汽水」，也買了「免洗杯」。顧客 C…凡是購買瓶裝飲料的顧客，多數都購買了「免洗杯」。「免洗杯」這個項目，很明顯的增加顧客間購買商品的相似度；也許購買「烏龍茶」與購買「雪碧汽水」

的顧客，之間的相似度並不強烈，卻因「免洗杯」這類「較不重要」的商品，被歸為同一類顧客。這是因為「烏龍茶」、「雪碧汽水」與「免洗杯」對於相似度的計算上，佔有相同的權重。

為此，本研究提出不同的相似度測量標準—不同項目的權重應該要有所區別。對於區分不同群集的重要性較強的項目，計算相似度時應該有較大的權重；反之對於區分不同群集的重要性較弱的項目，計算相似度時應該有較小的權重。

這裡有一個強烈，但符合一般經驗的基本假設是：出現頻率很高的項目，對於各筆資料的區隔程度便很低。相反地，如果有一個項目出現的次數很低，那麼若有兩筆資料同時都有這項目，則此項目對兩筆資料之相似性就有很高貢獻。亦即這樣的項目，對各筆資料間的區隔程度很高。

根據這樣的原則，可以依照「項目發生頻率」來判別單一項目的重要與否。若是發生頻率較低的，對於群聚過程的相似度判斷，便屬於「較重要項目」。而發生頻率較高的，對於群聚過程的相似度判斷，便屬於「較不重要項目」。如此，便可以區別出不同項目對於相似度判斷的權重。

因此新的相似度計算：

定義：

$a = |X \cap Y|$

$b = |X \cup Y|$

f=共同含有的「頻繁」項目之個數

r=共同含有的「特殊」項目之個數

相似度 $sim(X,Y) = \dfrac{a-f}{(b-f)-r}$

再以 3.3.1 小節當中的例子說明：

假設我們定義出現頻率大於等於 80%的項目，對於相似度判斷是不那麼重要的。則項目 B 是較不重要的項目，屬於「f」類(門檻值 $\hat{f}$=0.8)。

出現頻率小於等於 30%的項目，對於相似度判斷是更為重要的。則項目 D 與項目 E 是較重要的項目，屬於「r」類(門檻值$\hat{r}$=0.3)。如表 3.4 所示：

表 3. 4　範例資料

資料	Item lists				
D1	A	B	C		E
D2		B	C		
D3	A	B	C		
D4	A		C		
D5	A	B			
D6		B	C		
D7	A	B		D	E
D8		B			
D9	A		C		
D10		B	C	D	E
類別		f		r	r

sim(D1,D2)=(2-1)/(4-1)=0.33

sim(D2,D3)=(2-1)/(3-1)=0.5

sim(D1,D4)=2/4=0.5

sim(D1,D7)=(3-1)/(5-1-1)=0.67

將計算結果與 Jaccard 相比較

表 3. 5　Jaccard 與新定義的相似度測量標準比較

	Jaccard 係數	新的相似度係數
sim(D1,D2)	0.5	0.33
sim(D2,D3)	0.67	0.5
sim(D1,D4)	0.5	0.5
sim(D1,D7)	0.6	0.67

從反映不同權重對於相似程度的貢獻上，新定義的相似係數又比 Jaccard 更好。以 *sim*(D1,D2)與 *sim*(D2,D3)為例，其中包含了對相似度貢獻不高的項目「B」，而反映出來，其相似度均比 Jaccard 來的低。在 *sim*(D1,D7)的情形，因為包含了對相似度貢獻較高的項目「E」，而反映出來，其相似度比 Jaccard 來的高。

依照修改過後的相似度計算標準，演算步驟如下：

定義參數 θ 為相似度門檻值。k 為最小群組數。

1. 計算任意兩點的相似度。若相似度大於或等於門檻值，則判定此兩點互為鄰近點 (neighbors)。

   相似度 $sim(X,Y)=\dfrac{a-f}{(b-f)-r}$

   $a=|X\cap Y|$

   $b=|X\cup Y|$

   f=共同含有的「頻繁」項目之個數

   r=共同含有的「特殊」項目之個數

   例：若 θ=0.5，頻繁項目之門檻值為 0.7，稀少項目之門檻值為 0.3；資料庫中有 7 筆資料：A={a, c, d}；B={a, b, d, e}；C={a, c, d, e}；D={ b, c, d, e }；E={ a, c, e }；F={a, d}；G={a, c, e, f}。以 A、B、C、D 四資料點為例，評量兩筆資料之間相似度，則：

   i. 根據頻繁項目與稀少項目之門檻值，a 為頻繁項目，b, f 為稀少項目，其餘均為一般項目。

   ii. 分別計算兩筆資料相似度如下：

   A,B 相似度如下：

   a=2, b=5, f=1, r=0

   $sim(A,B)=\dfrac{2-1}{(5-1)-0}=0.25$

   A,C 相似度如下：

a=3, b=4, f=1, r=0

$$sim(A,C)=\frac{3-1}{(4-1)-0}=0.67>0.5$$

A,D 相似度如下：

a=2, b=5, f=0, r=0

$$sim(A,D)=\frac{2-0}{(5-0)-0}=0.4$$

B,C 相似度如下：

a=3, b=5, f=1, r=0

$$sim(B,C)=\frac{3-1}{(5-1)-0}=0.5$$

B,D 相似度如下：

a=3, b=5, f=0, r=1

$$sim(B,D)=\frac{3-0}{(5-0)-1}=0.75>0.5$$

C,D 相似度如下：

a=3, b=5, f=0, r=0

$$sim(C,D)=\frac{3-0}{(5-0)-0}=0.6>0.5$$

互與鄰近點之組合有 (A, C) 、(A, D)、 (B, C)、(B, D)與 (C, D)互為鄰近點。

資料點	鄰近點 1	鄰近點 2	鄰近點 3
A	C	D	
B	C	D	
C	A	B	D
D	A	B	C

2. 計算任意兩點之間鏈結的 (link) 個數。當任意兩點 p_i, p_j 有 1 個共同鄰近點時，其 $links(p_i, p_j)$=1。

   例：由步驟 1 之鄰近點結果計算得

   $link(A,B)$=2；$link(A,C)$=1；$link(A,D)$=1；

   $link(B,C)$=1；$link(B,D)$=1；$link(C,D)$=2；

3. 計算將任兩群組間群聚於同一群族適合度 (goodness measure)。兩群組 C_i 與 C_j 之適合度以 $g(C_i, C_j)$ 表示。$link[C_i, C_j]$ (the numbers of cross link between C_i and C_j) 表群組 i 與群組 j 之間的鏈結數，將群組 i 之任一點與群組 j 之任一點間的鏈結個數總後得到 $link[C_i, C_j]$。

$$g(C_i, C_j) = \frac{link[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$$

其中 $f(\theta) = \frac{1-\theta}{1+\theta}$

例：if k=3，

$f(\theta)$=(0.5/1.5)=0.667；1+2$f(\theta)$=2.334；

$g(A,B)=2/(2^{2.334} - 1^{2.334} - 1^{2.334})=0.6575$；

$g(A,C)=1/(2^{2.334} - 1^{2.334} - 1^{2.334})=0.3287$；

$g(A,D)=1/(2^{2.334} - 1^{2.334} - 1^{2.334})=0.3287$；

$g(B,C)=1/(2^{2.334} - 1^{2.334} - 1^{2.334})=0.3287$；

$g(B,D)=1/(2^{2.334} - 1^{2.334} - 1^{2.334})=0.3287$；

$g(C,D)=2/(2^{2.334} - 1^{2.334} - 1^{2.334})=0.6575$；

4. 選取群聚適合度最大值之兩群組，將其合併為一個新群組，並刪去原有兩個群組。若群聚適合度之最大值為零時，停止演算程序。

   例：群聚適合度最大值之兩群組為(A, B)或(C, D)，若選擇合併(A, B)則群組數為 3 組：(A, B)，(C)，(D)。

5. 重複步驟 1 至 4，直至群組總數為 k，則停止演算程序。

由於 Jaccard coefficient 對於每一項目均判斷為相同權重，本研究對此做出修改。修改後的 coefficient 根據各項目出現頻率的不同，對於各項目有不同的權重設定。

在實際的資料處理上，若單純的依各項目「出現頻率」，判斷相似度的計算權重，也有可改進之處。出現頻率很低的項目，若非對相似度判斷極為重要，就是對於相似度的判斷極為不重要；亦即無法避免有「noise」的發生。這時便需導入專家知識，作為各項目權重判斷的建議。請領域專家挑選出哪些項目是「較重要項目」，哪些項目又屬於「較不重要項目」；以去除根據頻率判斷結果的 noise。這邊假設提供知識的領域專家保證為專業，不會發生專家領域知識不足的情形。

應用修該過後的相似度量測係數，並且設計可導入專家知識之機制後，演算法的流程為：

1. 依照各項目出現頻率，區分為較重要項目、普通項目以及較不重要項目。
2. 依照各項目的權重，計算每筆資料之間的相似度。
3. 依據資料相似度，計算連結數。
4. 依據連結數計算群組適合度函數。
5. 將適合度最大的兩群集合併為一群，直到適合度最大值等於零，或已達到最少群集數 k。終止演算程序。
6. 若結果不合理，以專家知識為建議，修改較重要項目、普通項目以及較不重要項目的判斷。再重新開始演算流程，直到結果合理。圖 3.1 為此群聚方法的流程。

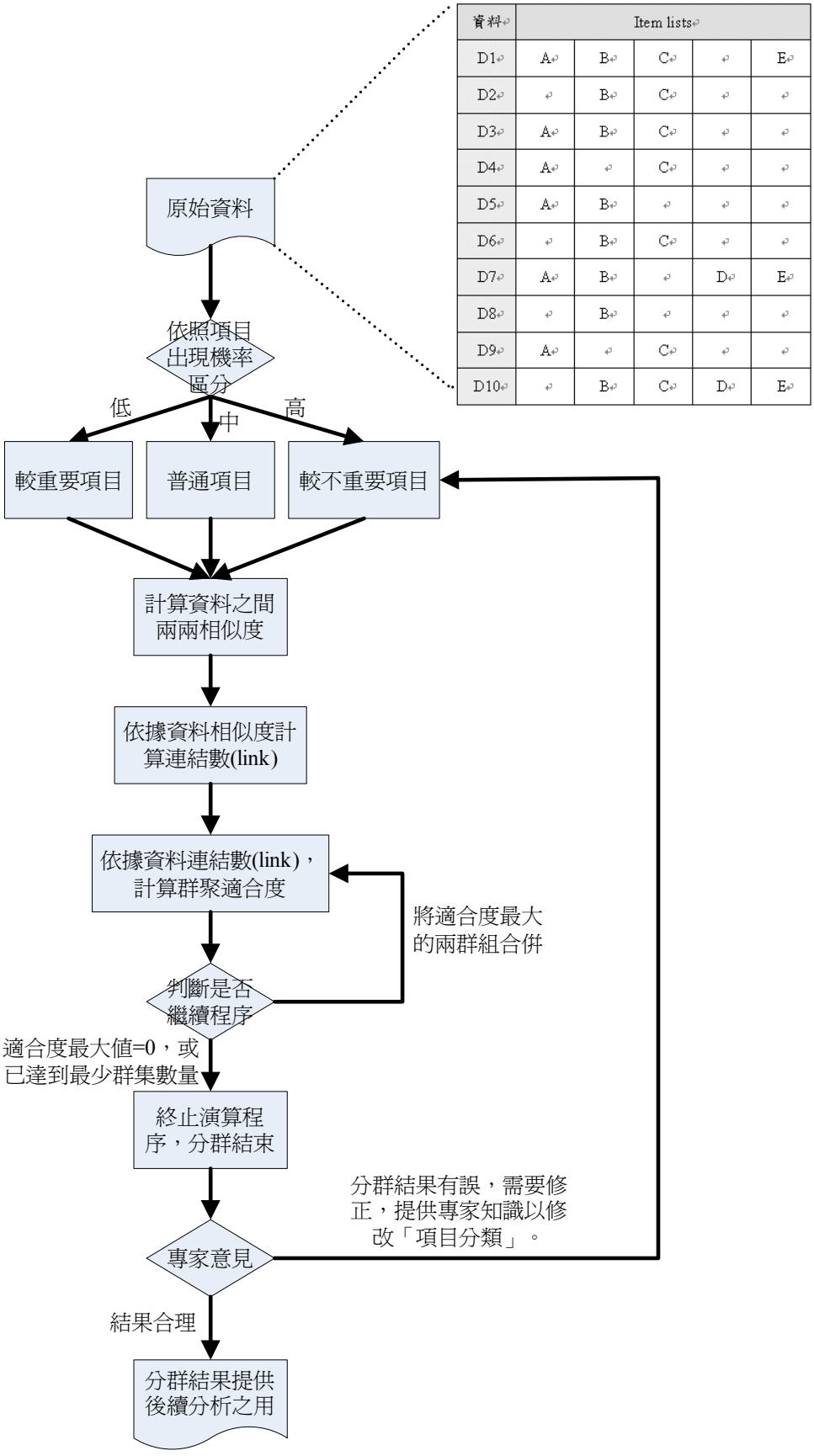
資料
Item lists
D1 A B C E
D2 B C
D3 A B C
D4 A C
D5 A B
D6 B C
D7 A B D E
D8 B
D9 A C
D10 B C D E
原始資料
依照項目出現機率區分
低
中
高
較重要項目
普通項目
較不重要項目
計算資料之間兩兩相似度
依據資料相似度計算連結數(link)
依據資料連結數(link)，計算群聚適合度
將適合度最大的兩群組合併
判斷是否繼續程序
適合度最大值=0，或已達到最少群集數量
終止演算程序，分群結束
分群結果有誤，需要修正，提供專家知識以修改「項目分類」。
專家意見
結果合理
分群結果提供後續分析之用

圖 3. 1 群聚方法流程圖

第四章　實例驗證

檢驗修改後的相似度量測標準，以及導入專家知識的機制，對於群聚結果的影響，本章以兩個資料集驗證群聚方法的合理性。分別為 ZOO 資料集，以及 PET 資料集。

ZOO 資料集僅有 101 筆資料，為小規模資料集，小規模資料集易於分析影響群聚結果正確性的影響。相對 ZOO 資料集，PET 資料集規模較大，資料欄位複雜。以此資料集檢驗本研究提出之方法在複雜之資料集是否適用。

4.1 ZOO 資料集

ZOO 資料集由 Richard Forsyth 在 1990 年整理而成。

這個資料庫裡面包含了 101 個種類的動物，而每種的動物都有 15 個 Boolean-valued attributes，2 個 Numeric-valued attributes。表 4.1 為 ZOO 資料集的基本資料表：

表 4. 1　ZOO Database 基本資料表

資料集名稱	ZOO Database
資料筆數	101 筆
整理人	Richard Forsyth
群數	7 群

4.1.1 資料特徵值說明

表 4. 2 ZOO Database 特徵值列表

順序	內容說明	資料屬性
1	animal name	Unique for each instance
2	Hair	Boolean
3	Feathers	Boolean
4	Eggs	Boolean
5	Milk	Boolean
6	airborne	Boolean
7	aquatic	Boolean
8	predator	Boolean
9	toothed	Boolean
10	backbone	Boolean
11	breathes	Boolean
12	venomous	Boolean
13	fins	Boolean
14	legs	Numeric (set of values: {0,2,4,5,6,8})
15	tail	Boolean
16	domestic	Boolean
17	catsize	Boolean
18	type	Numeric (integer values in range [1,7])

在這個 Database 裡面，將這 101 種動物分為 7 個種類[1]。這 7 個種類，在資料庫當中並沒有特別取名，但查了生物學上的資料之後，分別將這 7 群命名為：(1)哺乳類(有 41 種)，(2)鳥類(有 20 種)，(3)爬蟲類(有 5 種)，(4)魚類(有 13 種)，(5)兩棲類(有 4 種)，(6)昆蟲(有 8 種)，以及(7)其他(有 10 種)。

4.1.2 資料前置分析及處理

[1] 基本上，這個資料庫裡面分為 7 類的 101 種動物，大致是按照生物學上的分類；但是後來因為跟我使用演算法分類出來的結果有所出入，我特別去查了這些動物在生物學上的分類方法，發現這個資料庫在有些地方設計的不太好。後面會有更詳細的說明。

原始資料檔如下所示(節錄)：

aardvark,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,1

antelope,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1

…

必須將所有的資料型態轉換為種類型(categorical)之資料。由於本研究的方法不能處理數值型態的資料，所以要面對的資料集，必須先行轉換。

所有 Boolean 的特徵項目，其值為 1 時，轉換成以屬性名稱的種類型資料。其值為 0 時，不處理該筆資料。舉例說明：

動物「aardvark」，原先該筆資料值「1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,1」；轉換過後的資料為「hair, eggs, aquatic, predator , toothed, backbone, 4_legs, catsize」，輸入群聚系統分群之後，再將分群的結果與原先資料集的群集相比較。

4.1.3 試驗說明

對這個資料庫，做了幾次不同的實驗，測試分群之後的結果。

(1) 使用 ROCK 演算法(相似度量測為 Jaccard coefficient)進行群聚。

在演算法的參數設定上，k 均假設為 5，表示這比資料集的群數最少要 5 群以上。對不同的 θ 門檻值測試，101 筆資料分群結果，正確個數如表 4.3 所示。

表 4. 3　使用 Jaccard coefficient 的正確個數

θ 值	0.4	0.5	0.6	0.7	0.8	0.9
Jaccard	55	55	53	74	73	28

平均正確個數 56.33 個，正確率為 55.77%。

(2) 使用新定義的 $sim(X,Y)=\frac{a-f}{(b-f)-r}$，作為相似度的量測標準。其中「較重要項目」$r$，與「較不重要項目」$f$，是由頻率計算而來的。

設計了幾組不同的 r 與 f 門檻值參數，分別為($\hat{r}$=0.2, $\hat{f}$=0.8)，($\hat{r}$=0.3, $\hat{f}$=0.7)，($\hat{r}$=0.4, $\hat{f}$=0.6)；k 同樣是 5 群；對不同的 θ 門檻值測試，101 筆資料分群結果，正確個數如表 4.4 所示。

表 4. 4 $sim(X,Y)=\frac{a-f}{(b-f)-r}$ 的正確個數

θ 值	0.4	0.5	0.6	0.7	0.8	0.9
r=0.2, f=0.8	55	70	71	70	78	28
r=0.3, f=0.7	71	65	50	52	54	43
r=0.4, f=0.6	81	67	73	41	59	78

平均正確個數為 61.45 個，正確率為 60.83%。

這樣的結果不很理想。舉例來說：哺乳類、魚類、以及爬蟲類幾乎不能區分開來。以致於有時候最大群裡面會有 60~70 種不同的動物被分為同一群，而其他的群集區分的也不正確。這樣的結果，與資料集原先的分群結果相比較，計算之後大約都只有 60%左右。 (實際正確率會因為不同的參數調整，而有所差異。但是正確率都不會相差太遠。) 這樣的正確率實在是太低了！所以必定要找到是哪種原因造成分群結果不理想。

(3)將相似度量測標準修改為 $sim(X,Y)=\frac{a-f+r}{(b-f)-r}$。

上面的測試分類結果，全部都不很理想。經過多次修改參數值，交叉比較不同參數設定(例如：相似度門檻值、判定是否為「較重要」，

「較不重要」項目等門檻值)之後，直覺的認為是「相似度量測」的部份，沒有辦法突顯出「較重要項目」(r)的重要性：原先的想法裡面，認為在「分母」的地方減去 r 的項目，便可以增加權重。但是實驗之後覺得增加的不夠多；若在「分子」部分增加 r 的部份，便可明顯的增加相對權重。對於相似度的量測應該會更能精確的表現出「群聚」的效果。

設計了幾組不同的 r 與 f 門檻值參數，分別為($\hat{r}$=0.2, $\hat{f}$=0.8)，($\hat{r}$=0.3, $\hat{f}$=0.7)，($\hat{r}$=0.4, $\hat{f}$=0.6)；k 同樣是 5 群；對不同的 θ 門檻值測試，101 筆資料分群結果，正確個數如表 4.5 所示。

表 4. 5 $sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ 的正確個數

θ 值	0.4	0.5	0.6	0.7	0.8	0.9
r=0.2, f=0.8	53	69	71	71	76	53
r=0.3, f=0.7	44	63	71	74	62	53
r=0.4, f=0.6	64	71	65	74	78	74

平均正確個數為 65.89 個，正確率為 65.24%。

實驗結果發現，使用修改過後的相似度量測，經過分群後，比未修改的好一些。跟上面的相似度量測標準 $sim(X,Y)=\frac{a-f}{(b-f)-r}$ 比較起來，明顯看的出來的差別就是：「爬蟲類」可以與「哺乳類」和「魚類」區分開來；但是「哺乳類」和「魚類」還是一樣無法區分開。這樣的結果，與資料庫分群的「正確答案」做比較，正確率大約在 60~70% 左右。仍然不是很理想。

(4)加入專家的領域知識

因為分群結果不理想，推測原因是出在「較重要項目」，與「較不重要項目」的判斷上。必定有些「attribute」因為出現頻率的關係，被錯誤判別了計算相似度時的權重。因此導入「專家知識」，由專家知識判斷哪些「attribute」應該被區分為「較重要」(*r*)，而哪些「attribute」應該被區分為「較不重要」(*f*)；去除由出現頻率而來的「noise」。

檢視在這個資料集當中，究竟哪些「attribute」被區分為「較重要」(*r*)，而哪些「attribute」被區分為「較不重要」(*f*)。依照出現頻率來說，較為重要的有：5_legs、6_legs、8_legs、domestic、venomous... 這種 attribute；而較不重要的則是 backbone 這類的項目。這與一般的生物知識不符合。

經由專家的領域知識，另外擬了一份「較重要項目」，與「較不重要項目」的判斷。在「較重要項目」的部份，依照下列理由選擇了 6 個項目：

Milk：只有哺乳類會哺乳，這是區分出「哺乳類」最重要的特徵。

Eggs：非哺乳類大多為卵生，這也是一個重要的特徵。

Feathers：只有鳥類有羽毛。

Fins：只有魚類有魚鰭。

6_legs：昆蟲一定是 6 隻腳的。

backbone：資料庫裡的第 6 類「昆蟲」跟第 7 類「其他」都沒有脊椎骨，這個特徵對於是否能區分出它們相當重要。所以將這個 attribute 轉換成「no_backbone」。

另外，將 toothed 與 venomous 當作「較不重要項目」。因為這兩個特徵對於區分物種並無多大的用處；有時候反而會造成誤導。比如說：有毒與沒毒的青蛙，都屬於兩棲類…等。

將這份「項目清單」取代了由頻率判斷的「項目清單」，存入 MATLAB 再重新計算一次。

舉例來說，Calf 與 Carp 的資料如下。

Calf：hair, milk, toothed , breathes, 4_legs, tail, domestic, catsize

Carp：eggs, aquatic, toothed, fins, 0_legs, tail, domestic

以出現頻率區分($\hat{r}$=0.3, $\hat{f}$=0.7)，較為重要的項目有：0_legs, 2_legs, 5_legs, 6_legs, 8_legs, airborne domestic, feathers, fins, venomous；較不重要項目有：backbone, breathes, tail。以兩個不同判斷 r 與 f 的方式計算相似度，這兩種動物的相似度如下：

以出現頻率計算—a=2, b=12, r=1, f=0, sim(Calf, Carp)=0.33

以專家知識計算—a=2, b=12, r=0, f=0, sim(Calf, Carp)=0.17

可以看出若以頻率為判斷 r 與 f 的方式，相似度與專家知識判斷的不同。這便是「noise」對結果產生的影響之處。

導入專家知識後，對不同的 θ 門檻值測試，101 筆資料分群結果，正確個數如表 4.6 所示。

表 4. 6 放入專家知識的分群結果

θ 值	0.4	0.5	0.6	0.7	0.8	0.9
專家知識	75	70	75	71	71	84

平均正確個數為 74.33 個，正確率為 73.59%。

4.1.4 試驗結果記錄

將表 4.3~表 4.6 合併，如表 4.7 所示：

表 4. 7 不同相似度量測標準的比較(1)

θ 值	0.4	0.5	0.6	0.7	0.8	0.9
Jaccard	55	55	53	74	73	28
$sim(X,Y)=\frac{a-f}{(b-f)-r}$	69	67.33	64.67	54.33	63.67	49.67
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	53.67	67.67	69	73	72	60
專家知識	75	70	75	71	71	84

使用變異數分析檢定：

不同相似度量測：虛無假設H_0為不同相似度量測對分群正確率有影響，其對立假設H_1為不同相似度量測無影響。

θ門檻值：虛無假設H_0為不同θ門檻值對分群正確率有影響，其對立假設H_1為不同θ門檻值無影響。以顯著水準α為 0.05 來檢定。

變異數分析：集區設計

摘要	個數	總和	平均	變異數
Jaccard	6	338	56.33333	281.4667
$sim(X,Y)=\frac{a-f}{(b-f)-r}$	6	368.67	61.445	59.26727
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	6	395.34	65.89	56.94904
專家知識	6	446	74.33333	27.06667
θ=0.4	4	252.67	63.1675	110.3122
θ=0.5	4	260	65	45.8526
θ=0.6	4	261.67	65.4175	86.47056
θ=0.7	4	272.33	68.0825	85.61389
θ=0.8	4	279.67	69.9175	18.01389
θ=0.9	4	221.67	55.4175	540.9039

ANOVA

變源	SS	自由度	MS	F	P-值	臨界值
相似度量測	1047.926	3	349.3086	3.247108	0.051714	3.287382
θ值	510.2222	5	102.0444	0.948586	0.478643	2.901295
錯誤	1613.63	15	107.5753			
總和	3171.778	23				

雖然表面數字上看起來加上領域專家知識的正確程度會比較高，但是統計上並不是顯著差異。

4.1.5 試驗結果討論

以生物學家所做的分類，與分群演算法的結果做一個比較。生物上，分類的方法為「界門綱目科屬種」。如圖 4.1 所示：

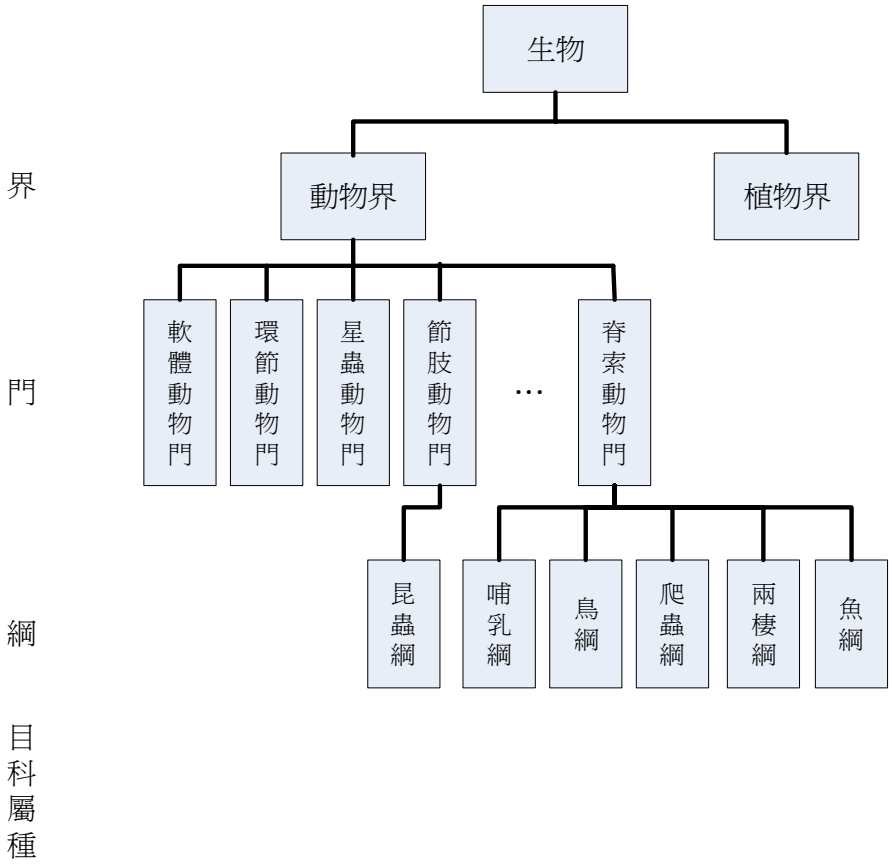

圖 4.1　生物學上的分類圖

而在 ZOO 這個資料庫裡面，它對動物的分類是這樣的：

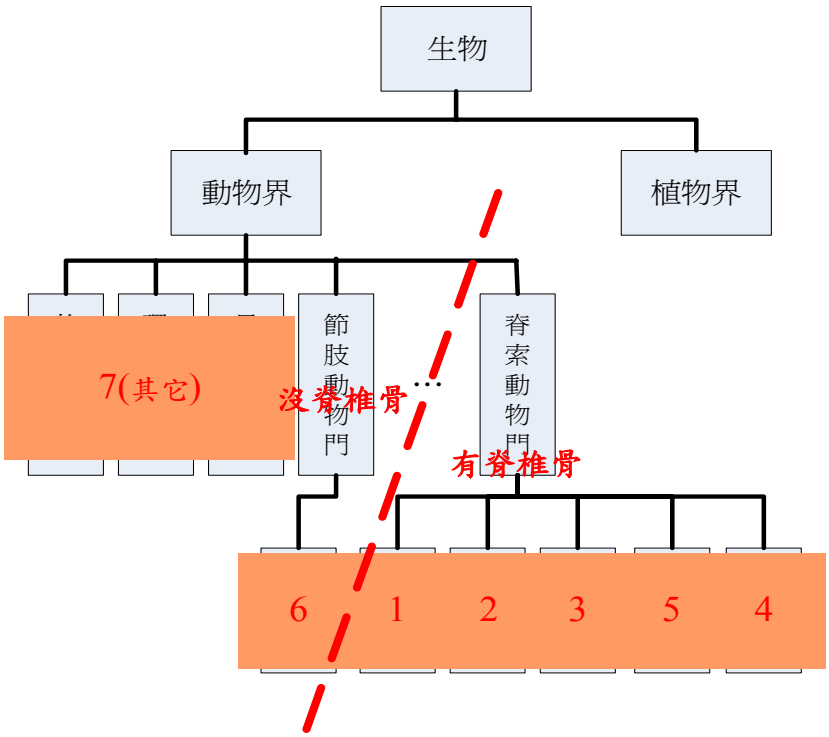

圖 4.2 資料集的分類方式

但是，分群演算法由資料上來區分，卻只能先區分出「脊索動物門」與的「非脊索動物門」的動物；「昆蟲」以及「其他」都屬於「非脊索動物門」。由各種動物的屬性來區分，並沒有辦法明顯區別「昆蟲」以及「其他」類別，因為它們的特徵實在是太接近了！

另外，「脊索動物門」中，演算法也不能區分出「兩棲」與「爬蟲」。推論原因，應該是區別這兩類別最大的特徵--「是否兩棲」這個屬性的缺乏(attrubute 裡面沒有「是否在陸上活動」這個屬性，所以無法推估)。從資料上來看，「兩棲」與「爬蟲」的差異也的確很小。所以，以生物學上的角度重新定義本資料集的群集，將動物分為 5 種群集：

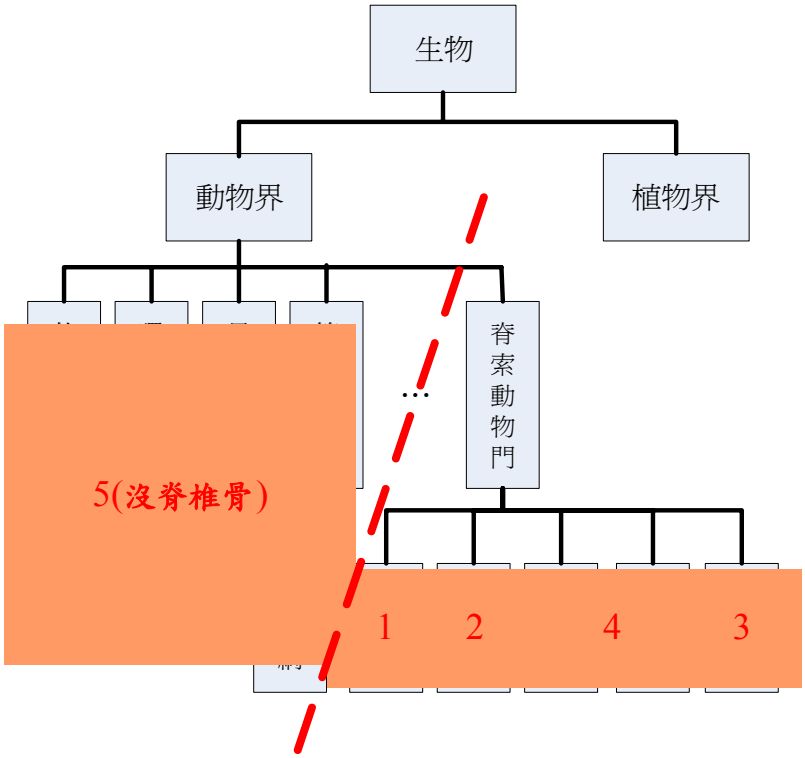

圖 4.3 重新定義資料集的分類方式

重新定義資料集當中的類別後，再次計算正確次數。

表 4.8 不同相似度量測標準的比較(2)

θ 值	0.4	0.5	0.6	0.7	0.8	0.9
Jaccard	55	55	53	77	73	28
$sim(X,Y)=\frac{a-f}{(b-f)-r}$	69.67	67.67	67.67	54.67	64.33	49.67
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	53.67	68.67	70	73	72	60
專家知識	78	76	78	79	77	94

使用變異數分析檢定：

不同相似度量測：虛無假設 H_0 為不同相似度量測對分群正確率有影響，其對立假設 H_1 為不同相似度量測無影響。

θ 門檻值：虛無假設 H_0 為不同 θ 門檻值對分群正確率有影響，其

對立假設H_1為不同θ門檻值無影響。以顯著水準α為 0.05 來檢定。

變異數分析：集區設計

摘要	個數	總和	平均	變異數
Jaccard	6	341	56.83333	304.1667
$sim(X,Y)=\frac{a-f}{(b-f)-r}$	6	373.68	62.28	66.7686
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	6	397.34	66.22333	59.17171
專家知識	6	482	80.33333	45.86667
θ=0.4	4	256.34	64.085	138.6096
θ=0.5	4	267.34	66.835	76.04297
θ=0.6	4	268.67	67.1675	108.7789
θ=0.7	4	283.67	70.9175	123.5472
θ=0.8	4	286.33	71.5825	28.04389
θ=0.9	4	231.67	57.9175	756.4539

ANOVA

變源	SS	自由度	MS	F	P-值	臨界值
相似度量測	1816.093	3	605.3642	4.833986	0.01506	3.287382
θ值	501.5	5	100.3	0.800921	0.566112	2.901295
錯誤	1878.463	15	125.2309			
總和	4196.056	23				

檢驗結果，不同相似度量測對分群正確率有影響。

專家知識放入分群演算法之後的正確率，顯著的比原本的高。在這個資料集當中，正確率最高的，分群正確率高達 93.1%。如表 4.9(錯誤部份 Mark 起來)。

表 4.9 加入專家知識的分群結果

動物編號	動物名稱	分群結果	資料集群集	正確屬性
1	aardvark	1	1	哺乳類
2	antelope	1	1	哺乳類
6	buffalo	1	1	哺乳類
37	hare	1	1	哺乳類
95	vole	1	1	哺乳類
18	deer	1	1	哺乳類
29	giraffe	1	1	哺乳類
23	elephant	1	1	哺乳類
56	oryx	1	1	哺乳類
55	opossum	1	1	哺乳類
50	mole	1	1	哺乳類
32	goat	1	1	哺乳類
66	pony	1	1	哺乳類
7	calf	1	1	哺乳類
71	reindeer	1	1	哺乳類
69	pussycat	1	1	哺乳類
5	boar	1	1	哺乳類
70	raccoon	1	1	哺乳類
65	polecat	1	1	哺乳類
48	lynx	1	1	哺乳類
11	cheetah	1	1	哺乳類
51	mongoose	1	1	哺乳類
99	wolf	1	1	哺乳類
68	puma	1	1	哺乳類
45	leopard	1	1	哺乳類
46	lion	1	1	哺乳類
4	bear	1	1	哺乳類
49	mink	1	1	哺乳類
36	hamster	1	1	哺乳類
97	wallaby	1	1	哺乳類
10	cavy	1	1	哺乳類
85	squirrel	1	1	哺乳類
64	platypus	1	1	哺乳類
33	gorilla	1	1	哺乳類

動物編號	動物名稱	分群結果	資料集群集	正確屬性
30	girl	1	1	哺乳類
94	vampire	1	1	哺乳類
28	fruitbat	1	1	哺乳類
76	sealion	1	1	哺乳類
3	bass	2	4	魚類
9	catfish	2	4	魚類
19	dogfish	2	4	魚類
93	tuna	2	4	魚類
61	pike	2	4	魚類
35	haddock	2	4	魚類
62	piranha	2	4	魚類
39	herring	2	4	魚類
13	chub	2	4	魚類
83	sole	2	4	魚類
87	stingray	2	4	魚類
74	seahorse	2	4	魚類
8	carp	2	4	魚類
81	slowworm	2	3	爬蟲類
20	dolphin	2	1	哺乳類
67	porpoise	2	1	哺乳類
75	seal	2	1	哺乳類
12	chicken	3	2	鳥類
17	crow	3	2	鳥類
22	duck	3	2	鳥類
84	sparrow	3	2	鳥類
101	wren	3	2	鳥類
80	skua	3	2	鳥類
38	hawk	3	2	鳥類
34	gull	3	2	鳥類
88	swan	3	2	鳥類
44	lark	3	2	鳥類
60	pheasant	3	2	鳥類
79	skimmer	3	2	鳥類
96	vulture	3	2	鳥類
59	penguin	3	2	鳥類
21	dove	3	2	鳥類

動物編號	動物名稱	分群結果	資料集群集	正確屬性
72	rhea	3	2	鳥類
58	parakeet	3	2	鳥類
24	flamingo	3	2	鳥類
57	ostrich	3	2	鳥類
42	kiwi	3	2	鳥類
91	tortoise	3	3	爬蟲類
14	clam	4	7	其他
16	crayfish	4	7	其他
47	lobster	4	7	其他
89	termite	4	6	昆蟲
43	ladybird	4	6	昆蟲
25	flea	4	6	昆蟲
100	worm	4	7	其他
82	slug	4	7	其他
31	gnat	4	6	昆蟲
52	moth	4	6	昆蟲
98	wasp	4	6	昆蟲
41	housefly	4	6	昆蟲
86	starfish	4	7	其他
15	crab	4	7	其他
40	honeybee	4	6	昆蟲
78	seawasp	4	7	其他
54	octopus	4	7	其他
26	frog	5	5	兩棲類
27	frog	5	5	兩棲類
53	newt	5	5	兩棲類
92	tuatara	5	3	爬蟲類
90	toad	5	5	兩棲類
63	pitviper	5	3	爬蟲類
73	scorpion	6	7	其他
77	seasnake	7	3	爬蟲類

使用 Jaccard 的錯誤率大約跟相似度標準 $sim(X,Y)=\frac{a-f}{(b-f)-r}$ 的結果差不多(差異大多來自參數的調整)。我認為是因為在一開始的相似

度標準：$sim(X,Y)=\frac{a-f}{(b-f)-r}$，差距沒有與 Jaccard 非常的明顯。所以表現在分群上，一開始的相似度標準並沒有明顯的比 ROCK 使用 Jaccard coefficient 好。而相似度標準改為 $sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ 之後，便看的出分群的效果有進步。但最大的差異還是在於：專家知識的提供。

4.2 PET 資料集

本資料集為中國醫藥大學附設醫院核子醫學科高嘉鴻主任與新光醫院正子中心合作所收集之癌症病患資料。希望結合正子斷層造影原理 (Positron Emission Tomography，PET)的結果，醫生的經驗，以及客觀的受檢者的數據(如受檢者基本資料、病史、家族史等)，能提供判斷是否罹患癌症的一項參考指標。

4.2.1 資料說明

資料結構設計

資料來源為新光醫院正子中心的歷史資料為基準。本資料集一共有 2897 筆資料，其資料項目如下（以下列出欄位及其說明）：

輸入資料區塊分為

(a)受檢者個人資料：

受檢者編號	年齡	性別
自動編碼	數值	男或女

(b)PET 檢查：

PET 結果	SUV	部位
數值	數值	文字

欄位內容設定：

若該 PET 檢查結果若判斷為正則值為 1，反之則為-1，不清楚則為 0。

SUV[2]為檢驗值。

部位為檢驗呈陽性反應之部位(此項不納入資料分析處理)。

(c)tumor marker[3]

AFP	CEA	CA125	CA153	CA199	PSA
數值	數值	數值	數值	數值	數值

欄位內容設定：為其 tunor marker 測量值。

(d)受檢者個人病史：

目前認定可能與癌症高度相關的疾病。

欄位名稱定義：

B 型肝炎	C 型肝炎
Y/N	Y/N

(e)受檢者家族病史：

目前認定家族病史與癌症高度相關的疾病。直系血親（祖父母以上、父母）、旁系血親（兄弟姊妹）的紀錄。

欄位名稱定義

大腸直腸癌	乳癌
Y/N	Y/N

[2]SUV 為 standardized uptake value 的縮寫。藉由計算 SUV 的方式，評估比較檢驗部位攝取 FDG(一種檢定癌症的放射藥物)的量，以鑑別診斷腫瘤的良惡性。惡性腫瘤攝取 FDG 的量比良性腫瘤攝取的量高。

[3] tunor marker (血液腫瘤標記)的偵測，可作為篩檢與追蹤的工具。血液腫瘤標記檢查做為輔助診斷惡性腫瘤有其很大的功用。

(f)生活習慣：

目前認定可能會致癌的生活習慣並同時紀錄頻率和維持習慣時間。

欄位名稱定義：

抽煙習慣	抽煙(年)	喝酒習慣	喝酒(年)	吃檳榔習慣	吃檳榔(年)
Y/N	數值	Y/N	數值	Y/N	數值

欄位內容設定：

若該受檢者目前或過去曾經有的致癌生活習慣之測量值。

(g)其他檢查：

胸部 x-ray	腹部超音波	內視鏡	婦科超音波	子宮頸抹片	大便潛血
數值	數值	數值	數值	數值	數值

欄位內容設定：

若該檢查結果若判斷為正則值為 1，反之則為-1，

不清楚則為 0。

(h)Notes

相關檢查的結果

欄位名稱定義：

胸部 x-ray	腹部超音波	婦科超音波	子宮頸抹片	大便潛血
結果概述	結果概述	結果概述	結果概述	Y/N

欄位內容設定：

該受檢者相關檢驗的結果。例如：腹部超音波檢查結果可能為肝水泡、膽結石等症狀。

(i)Cancer 判別：

Cancer
數值

欄位內容設定：

若該檢查結果若判斷為正則值為 1，反之則為 0

將以依照資料類型與名稱整理為表 4. 10 所示。

表 4. 10　正子中心資料集特徵項目表

編號	項目	資料型態	編號	項目	資料型態
1	Cancer	0 或 1	16	抽煙習慣	0 或 1
2	年齡	數值	17	抽煙(年)	數值
3	性別	0 或 1	18	抽煙(頻率)	數值
4	PET 結果	0,1,-1	19	喝酒習慣	0 或 1
5	SUV	數值	20	喝酒(年)	數值
6	AFP	數值	21	喝酒(頻率)	數值
7	CEA	數值	22	吃檳榔習慣	0 或 1
8	CA125	數值	23	吃檳榔(年)	數值
9	CA153	數值	24	胸部 x-ray	0 或 1
10	CA199	數值	25	腹部超音波	0 或 1
11	PSA	數值	26	內視鏡	0 或 1
12	B 型肝炎	0 或 1	27	婦科超音波	0 或 1
13	C 型肝炎	0 或 1	28	子宮頸抹片	0 或 1
14	大腸直腸癌	0 或 1	29	大便潛血	0 或 1
15	乳癌	0 或 1	30	Notes	結果概述

4.2.2 資料前置分析及處理

必須將所有的資料型態轉換為種類型之資料。由於本研究的方法不能處理數值型態的資料，所以要面對的資料集，必須先行轉換。

所有 Boolean 的特徵項目，其值為 1 時，轉換成以屬性名稱的種類型資料。其值為 0 時，轉換成相對屬性名稱的種類型資料。例如：B 型肝炎=1，轉換成「有 B 型肝炎」；B 型肝炎=0，轉換成「無 B 型肝炎」。

「年齡」轉換為區間的種類型資料，例如 33 歲就轉換成「30~35 歲」。tunormarker 測量值的部份，則根據醫師提供的門檻值(表 4.11)，將數值轉換為「正常」，或是「異常」的種類型資料。

例如 AFP=8，轉換為「AFP 正常」；AFP=50，轉換為「AFP 異常」。

表 4. 11 tunormarker 測量值臨床指標

	AFP	Free β hCG	CA 125	CA 15-3	CA 19-9	CA 72-4	Calcitonin	CEA
NORMAL	<10 ng/ml	<0.1 ng/ml	<35 U/ml	<30 U/ml	<37 U/ml	<6 U/ml	<10 pg/ml	<5 ng/ml
SUSPECT	10～ 200 ng/ml		35 ～65 U/ml	30 ～ 50 U/ml	37 ～120 U/ml	6 ～ 10 U/ml	10 ～ 100 pg/ml	5 ～ 10 ng/ml
Pathological	>200 ng/ml	>0.1 ng/ml	>65 U/ml	>50 U/ml	>120 U/ml	>10 U/ml	>100 pg/ml	>10 ng/ml
臨床指標	肝癌、肝硬化、少數生殖細胞瘤、70%其他睪丸瘤、100%卵黃囊瘤	睪丸癌 生殖細胞癌 30% 其他類癌症 60% 絨毛膜癌 水囊狀胎塊	卵巢癌診斷率爲 90%、 CEA 與 CA 72-4 對於黏液性腺癌特異性很高	如果>50 U/ml 預後不好，建立長期參考指標^	胰臟與膽道癌、大腸直腸癌：CEA + CA 19-9 胃癌：CA 19-9、CEA、CA 72-4	胃癌 消化道癌 卵巢黏液性癌	高危險性家族群髓性甲狀腺癌。 SCLC 肺癌	消化道癌 大腸直腸癌
臨床用途	高危險群常規檢查，診斷性檢查，臨床上與治療後追蹤檢查	診斷檢查、預後檢查、臨床上與治療後追蹤檢查、癌症復發檢查	追蹤檢查，手術不乾淨癌症復發診斷，治療後檢查與預後評估，子宮內膜異位檢查	臨床追蹤檢查、早期復發診斷與 CEA 合併追蹤檢查，其敏感度>80%	臨床上與治療後追蹤檢查。 胰腺炎鑑別診斷	臨床追蹤檢查價值++	診斷檢查 臨床上與治療後追蹤檢查。	診斷，鑑別、治療、早期復發診斷。胰臟、小腸、胃、乳房、卵巢、肺輔助檢查
健保給付	是	否	是	是	是	否	是	是
註	神經管缺陷產前診斷	懷孕時升高,亦是唐氏症篩檢標記	配合 CEA 使用	配合 CEA 使用	膽囊炎與急性胰囊炎值會高	非腫瘤疾病其值極少升高	少數白血病，腎，骨骼疾病會高	腎衰竭會高
	Thyroglobulin	NSE	PAP	PSA	free PSA	SCC	TPA	CYFRA 21-
NORMAL	<25 ng/ml	<12.5 ng/ml	<3 ng/ml	<2.5 ng/ml	F/T ratio >18%	<1.5 ng/ml	<100 U/L	<1.8 ng/ml
SUSPECT		12.5～ 25 ng/ml	3 ～ 5 ng/ml	2.5 ～ 5 ng/ml	<18%	1.5 ～ 2.5 ng/ml	100 ～ 150 U/L	1.8 ～ 3.3 ng/ml
Pathological	>25 ng/ml	>25 ng/ml	>5 ng/ml	>5 ng/ml	<18%	>2.5 ng/ml	>150 U/L	>3.3 ng/ml
臨床指標	甲狀腺癌	神經母細胞瘤 診斷率 30-60% 、SCLC(小細胞肺臟癌)	前列腺癌	前列腺癌診斷率極高建議與 free PSA 配套使用	當 Total PSA 值介於 4-10(or 20)ng/ml 時以 F/T ratio 區分 CANCER /BPH	子宮頸癌 食道癌 肺臟鱗狀細胞癌	膀胱癌， 乳房、消化道、肺、卵巢癌等輔助標記	NSCLC(NON-SMALL CELL LUNG CA. 肺
臨床用途	早期復發診斷 >5 ng/ml	診斷檢查 臨床上與治療後追蹤檢查	診斷檢查 臨床上與治療後追蹤檢查	比 PAP 敏感 癌症復發檢查	區分診斷提高特異性	臨床上與治療後追蹤檢查	非特異腫瘤標記	臨床診斷與治療後追蹤檢查
健保給付	是	是	是	是	是(re: PSA)	是	是	是(re: SCC)
註		配合 CEA,Cyfra21-1 做肺癌診斷輔助	配合 PSA 使用	高敏感度適篩檢正常值與年齡相關	不建議單獨使用	偽陽性較高	特異性低	Squamous cell carcinoma 高敏感性

(資料來源：中國醫藥大學附設醫院核子醫學科)

4.2.3 試驗說明

在完整資料集當中，以抽樣的方式，隨機抽出兩個 500 筆資料，作實驗設計。第一個 500 筆資料當中，包含 53 名癌症病患。第二個 500 筆資料當中，包含 19 名癌症病患。

(1) 使用 ROCK 演算法(相似度量測為 Jaccard coefficient)進行群聚分析。

在演算法的參數設定上，k 均假設為 10，表示這筆資料集的群數最少要 10 群以上 (不同的 k 值設定，對結果的影響不大，所有的 k 均設相同值，減少一個變數的可能影響) 。對不同的 θ 門檻值測試，第一個 500 筆資料分群結果，正確區分出癌症個數如表 4.12 所示。

表 4. 12　ROCK 演算法對第一個 500 筆正確判斷的癌症個數

θ 值	0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
Jaccard	3	3	2	3	3	3	6	18	0	0

對不同的 θ 門檻值測試，第二個 500 筆資料分群結果，正確區分出癌症個數如表 4.13 所示。

表 4. 13　ROCK 演算法對第二個 500 筆正確判斷的癌症個數

θ 值	0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
Jaccard	0	0	0	1	1	1	1	3	4	12

(2)相似度量測標準為 $sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ 。

設計了幾組不同的 r 與 f 門檻值參數，分別為($\hat{r}$=0.2, $\hat{f}$=0.8)，($\hat{r}$=0.3, $\hat{f}$=0.7)，($\hat{r}$=0.4, $\hat{f}$=0.6)；k 同樣是 10 群；對不同的 θ 門檻值測試，第一個 500 筆資料分群結果，正確區分出癌症個數如表 4.14 所示。

表 4. 14　$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ 對第一個 500 筆正確判斷的癌症個數

θ 值		0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	r=0.3 f=0.7	8	5	7	6	3	5	8	15	3	6
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	r=0.3 f=0.7	4	3	3	3	4	6	10	13	8	24
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	r=0.4 f=0.6	4	5	6	4	5	3	5	5	6	10

第二個 500 筆資料分群結果，正確區分出癌症個數如表 4.15 所示。

表 4. 15　$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ 對第二個 500 筆正確判斷的癌症個數

θ 值		0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	r=0.2 f=0.8	2	2	1	3	2	3	1	5	4	13
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	r=0.3 f=0.7	1	1	1	2	2	2	3	6	3	3
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$	r=0.4 f=0.6	3	3	2	2	4	6	5	5	5	8

(3)加入專家的領域知識

因為分群結果不理想，推測原因是出在「較重要項目」，與「較不重要項目」的判斷上。必定有些「attribute」因為出現頻率的關係，被錯誤判別了計算相似度時的權重。因此導入「專家知識」，由專家知識判斷哪些「attribute」應該被區分為「較重要」(r)，而哪些「attribute」應該被區分為「較不重要」(f)；去除由出現頻率而來的「noise」。

檢視在這個資料集當中，究竟哪些「attribute」被區分為「較重要」(r)，而哪些「attribute」被區分為「較不重要」(f)。若依照出現頻率來說，較為重要的有：肝右葉水泡、腸息肉...，以及 20 歲以下，70 歲以上…這種 attribute(表 4.16)；而較不重要的反而是 AFP 正常這類的項目。

表 4. 16 根據頻率計算得到的 r 項目(第一個 500 筆，$\hat{r}$=0.2)

20 歲	右腎水泡	肝水泡.慢性膽囊病變	喝酒 2 年
20 歲以下	右腎結石	肝右葉水泡	脾內鈣化疑似右腎結石
30 歲	右腎結節	肝右葉血管瘤	腸炎
40 歲	左側卵巢水泡	肝右葉局部病變	腸息肉
50 歲	左側卵巢水泡切除	肝右葉局部鈣化	腸息肉切除
60 歲	左腎水泡	肝右葉結節	腹部超音波正常
70 歲以上	左腎水腫	肝右葉結節疑似血管瘤	慢性肝實質病變
AFP_unnormal	左腎結石	肝右葉腫瘤疑似血管瘤	慢性肝實質病變.肝水泡
CA125_unnormal	左腎結節	肝左葉水泡	慢性肝實質病變.膽切除
CA153_unnormal	吃檳榔 10 年	肝左葉肥大	慢性結石性膽囊炎
CA199_unnormal	吃檳榔 20 年	肝左葉結節	慢性腎病變
CEA_unnormal	多囊腎	肝局部病變	疑似左腎結石
PSA_normal	早期肝硬化	肝硬化	疑似肝內血管瘤
十二指腸潰瘍	有 B 型肝炎	肝結節疑似血管瘤	疑似肝臟血管瘤
大便潛血異常	有 C 型肝炎	肝腫瘤疑似血管瘤	總膽管擴大
大腸息肉	有吃檳榔習慣	肝臟水泡 10 公分	總膽管擴張
大腸息肉切除	有抽煙習慣	抽菸 1 年	膽切除
子宮下垂行子宮卵巢切除	有家族病史_大腸直腸癌	抽菸 2 年	膽息肉
子宮內膜異位	有家族病史_乳癌	直腸癌手術後	膽息肉.右腎水泡
子宮肌瘤	有喝酒習慣	胃出血	膽息肉.脾腫大
子宮肌瘤切除	卵巢水泡	胃炎	膽息肉.總膽管擴大
子宮肌瘤手術後	卵巢水泡.卵巢一側切除	胃息肉	膽息肉切除
子宮肌瘤行子宮切除	卵巢水泡切除	胃發炎	膽結石
子宮肌瘤行子宮及單側卵巢切除	卵巢巧克力囊腫	胃潰瘍	膽結石合併慢性膽囊病變
子宮肌瘤行子宮卵巢切除	卵巢囊腫切除	胃潰瘍手術後	膽囊切除
子宮肌瘤行子宮單側卵巢切除	每天 1 包菸	食道潰瘍	膽囊息肉
子宮卵巢切除	每天 1 杯酒	脂肪肝	膽囊息肉.右腎結節
子宮頸抹片異常	每天 2 包菸	胸部 x-ray 異常	膽囊疑似腺肌瘤
子宮頸癌手術後	每天 2 杯酒	婦科超音波正常	膽囊擴張
小腸粘黏手術後	肝小水泡囊	婦科超音波異常	
內視鏡異常	肝水泡	痔瘡	
右腎切除	肝水泡.右腎水泡	喝酒 1 年	

經由醫師的領域知識，另外擬了一份「較重要項目」，與「較不重要項目」的判斷。在「較重要項目」的部份，選擇了 6 個項目：

AFP、CEA、CA125、CA153、CA199 以及 PSA 等 6 項 tumor marker 檢驗值異常的狀況。

另外，將「沒有 B 型肝炎」、「沒有 C 型肝炎」、「沒有吃檳榔習慣」...等共 6 個項目，對於判定是否為癌症病無顯著效果的屬性，當作「較不重要項目」(表 4.17)。

表 4. 17 專家判斷的項目判斷清單

較重要項目	較不重要項目
AFP_unnormal	沒有 B 型肝炎
CA125_unnormal	沒有 C 型肝炎
CA153_unnormal	沒有吃檳榔習慣
CA199_unnormal	沒有家族病史_大腸直腸癌
CEA_unnormal	沒有家族病史_乳癌
PSA_unnormal	胸部 x-ray 正常

將這份「項目清單」取代了由頻率判斷的「項目清單」，存入 MATLAB 再重新計算一次。

對不同的 θ 門檻值測試，第一個 500 筆資料分群結果，正確區分出癌症個數如表 4.18 所示。

表 4. 18 第一個 500 筆資料放入專家知識的正確判斷癌症個數

θ 值	0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
專家知識	7	8	5	4	8	12	16	25	25	40

對不同的 θ 門檻值測試，第二個 500 筆資料分群結果，正確區分出癌症個數如表 4.19 所示。

表 4. 19 第二個 500 筆資料放入專家知識的正確判斷癌症個數

θ 值	0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
專家知識	1	1	1	1	2	1	5	11	12	16

4.2.4 試驗結果記錄

表 4.12~表 4.19 合併成下面兩個表：

表 4. 20 不同方法應用在第一個 500 筆資料正確判斷的癌症個數

PET資料集--第一個500筆

θ 值	0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
Jaccard	3	3	2	3	3	3	6	18	0	0
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ r=0.2 f=0.8	8	5	7	6	3	5	8	15	3	6
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ r=0.3 f=0.7	4	3	3	3	4	6	10	13	8	24
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ r=0.4 f=0.6	4	5	6	4	5	3	5	5	6	10
專家知識	7	8	5	4	8	12	16	25	25	40

表 4. 21 不同方法應用在第二個 500 筆資料正確判斷的癌症個數

PET資料集--第二個500筆

θ 值	0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
Jaccard	0	0	0	1	1	1	1	3	4	12
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ r=0.2 f=0.8	2	2	1	3	2	3	1	5	4	13
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ r=0.3 f=0.7	1	1	1	2	2	2	3	6	3	3
$sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ r=0.4 f=0.6	3	3	2	2	4	6	5	5	5	8
專家知識	1	1	1	1	2	1	5	11	12	16

以這兩筆資料分別做無母數檢定：

Wilcoxon 符號等級檢定(第一個500筆)

等級

		個數	等級平均數	等級總和
R0.2F0.8 - ROCK	負等級	1[a]	5.00	5.00
	正等級	8[b]	5.00	40.00
	等值結	1[c]		
	總和	10		
R0.3F0.7 - ROCK	負等級	1[d]	6.00	6.00
	正等級	7[e]	4.29	30.00
	等值結	2[f]		
	總和	10		
R0.4F0.6 - ROCK	負等級	2[g]	5.50	11.00
	正等級	7[h]	4.86	34.00
	等值結	1[i]		
	總和	10		
DOMAIN - ROCK	負等級	0[j]	.00	.00
	正等級	10[k]	5.50	55.00
	等值結	0[l]		
	總和	10		
R0.3F0.7 - R0.2F0.8	負等級	5[m]	5.80	29.00
	正等級	5[n]	5.20	26.00
	等值結	0[o]		
	總和	10		
R0.4F0.6 - R0.2F0.8	負等級	6[p]	4.83	29.00
	正等級	3[q]	5.33	16.00
	等值結	1[r]		
	總和	10		
DOMAIN - R0.2F0.8	負等級	3[s]	2.00	6.00
	正等級	7[t]	7.00	49.00
	等值結	0[u]		
	總和	10		
R0.4F0.6 - R0.3F0.7	負等級	5[v]	6.60	33.00
	正等級	4[w]	3.00	12.00
	等值結	1[x]		
	總和	10		
DOMAIN - R0.3F0.7	負等級	0[y]	.00	.00
	正等級	10[z]	5.50	55.00
	等值結	0[aa]		
	總和	10		
DOMAIN - R0.4F0.6	負等級	1[bb]	1.00	1.00
	正等級	8[cc]	5.50	44.00
	等值結	1[dd]		
	總和	10		

a. R0.2F0.8 < ROCK
b. R0.2F0.8 > ROCK
c. ROCK = R0.2F0.8
d. R0.3F0.7 < ROCK
e. R0.3F0.7 > ROCK
f. ROCK = R0.3F0.7
g. R0.4F0.6 < ROCK
h. R0.4F0.6 > ROCK
i. ROCK = R0.4F0.6
j. DOMAIN < ROCK
k. DOMAIN > ROCK
l. ROCK = DOMAIN
m. R0.3F0.7 < R0.2F0.8
n. R0.3F0.7 > R0.2F0.8
o. R0.2F0.8 = R0.3F0.7
p. R0.4F0.6 < R0.2F0.8
q. R0.4F0.6 > R0.2F0.8
r. R0.2F0.8 = R0.4F0.6
s. DOMAIN < R0.2F0.8
t. DOMAIN > R0.2F0.8
u. R0.2F0.8 = DOMAIN
v. R0.4F0.6 < R0.3F0.7
w. R0.4F0.6 > R0.3F0.7
x. R0.3F0.7 = R0.4F0.6
y. DOMAIN < R0.3F0.7
z. DOMAIN > R0.3F0.7
aa. R0.3F0.7 = DOMAIN
bb. DOMAIN < R0.4F0.6
cc. DOMAIN > R0.4F0.6
dd. R0.4F0.6 = DOMAIN

檢定統計量[c]

	R0.2F0.8 - ROCK	R0.3F0.7 - ROCK	R0.4F0.6 - ROCK	DOMAIN - ROCK	R0.3F0.7 - R0.2F0.8
Z 檢定	-2.090[a]	-1.689[a]	-1.368[a]	-2.805[a]	-.153[b]
漸近顯著性 (雙尾)	.037	.091	.171	.005	.878

R0.4F0.6 - R0.2F0.8	DOMAIN - R0.2F0.8	R0.4F0.6 - R0.3F0.7	DOMAIN - R0.3F0.7	DOMAIN - R0.4F0.6
-.774[b]	-2.193[a]	-1.247[b]	-2.805[a]	-2.556[a]
.439	.028	.212	.005	.011

a.以負等級爲基礎。
b.以正等級爲基礎。
c.Wilcoxon 符號等級檢定

所以，由檢定及果可以得知(α=0.05，雙尾檢定)：

1. 專家知識導入演算法的結果，優於 ROCK 演算法使用 Jaccard 係數的結果 (DOMAIN-ROCK 的漸近顯著性=0.005<0.025)。
2. r 與 f 由頻率判定時，結果與 ROCK 演算法使用 Jaccard 係數並無顯著差異。

Wilcoxon 符號等級檢定(第二個 500 筆)

等級

		個數	等級平均數	等級總和
R0.2F0.8 - ROCK	負等級	0[a]	.00	.00
	正等級	8[b]	4.50	36.00
	等值結	2[c]		
	總和	10		
R0.3F0.7 - ROCK	負等級	2[d]	7.00	14.00
	正等級	8[e]	5.13	41.00
	等值結	0[f]		
	總和	10		
R0.4F0.6 - ROCK	負等級	1[g]	8.50	8.50
	正等級	9[h]	5.17	46.50
	等值結	0[i]		
	總和	10		
DOMAIN - ROCK	負等級	0[j]	.00	.00
	正等級	8[k]	4.50	36.00
	等值結	2[l]		
	總和	10		
R0.3F0.7 - R0.2F0.8	負等級	6[m]	4.25	25.50
	正等級	2[n]	5.25	10.50
	等值結	2[o]		
	總和	10		
R0.4F0.6 - R0.2F0.8	負等級	2[p]	6.00	12.00
	正等級	7[q]	4.71	33.00
	等值結	1[r]		
	總和	10		
DOMAIN - R0.2F0.8	負等級	4[s]	2.50	10.00
	正等級	4[t]	6.50	26.00
	等值結	2[u]		
	總和	10		
R0.4F0.6 - R0.3F0.7	負等級	1[v]	1.50	1.50
	正等級	8[w]	5.44	43.50
	等值結	1[x]		
	總和	10		
DOMAIN - R0.3F0.7	負等級	2[y]	1.50	3.00
	正等級	4[z]	4.50	18.00
	等值結	4[aa]		
	總和	10		
DOMAIN - R0.4F0.6	負等級	6[bb]	3.50	21.00
	正等級	3[cc]	8.00	24.00
	等值結	1[dd]		
	總和	10		

a. R0.2F0.8 < ROCK
b. R0.2F0.8 > ROCK
c. ROCK = R0.2F0.8
d. R0.3F0.7 < ROCK
e. R0.3F0.7 > ROCK
f. ROCK = R0.3F0.7
g. R0.4F0.6 < ROCK
h. R0.4F0.6 > ROCK
i. ROCK = R0.4F0.6
j. DOMAIN < ROCK
k. DOMAIN > ROCK
l. ROCK = DOMAIN
m. R0.3F0.7 < R0.2F0.8
n. R0.3F0.7 > R0.2F0.8
o. R0.2F0.8 = R0.3F0.7
p. R0.4F0.6 < R0.2F0.8
q. R0.4F0.6 > R0.2F0.8
r. R0.2F0.8 = R0.4F0.6
s. DOMAIN < R0.2F0.8
t. DOMAIN > R0.2F0.8
u. R0.2F0.8 = DOMAIN
v. R0.4F0.6 < R0.3F0.7
w. R0.4F0.6 > R0.3F0.7
x. R0.3F0.7 = R0.4F0.6
y. DOMAIN < R0.3F0.7
z. DOMAIN > R0.3F0.7
aa. R0.3F0.7 = DOMAIN
bb. DOMAIN < R0.4F0.6
cc. DOMAIN > R0.4F0.6
dd. R0.4F0.6 = DOMAIN

檢定統計量[c]

	R0.2F0.8 - ROCK	R0.3F0.7 - ROCK	R0.4F0.6 - ROCK	DOMAIN - ROCK	R0.3F0.7 - R0.2F0.8
Z 檢定	-2.598[a]	-1.429[a]	-1.946[a]	-2.558[a]	-1.098[b]
漸近顯著性 (雙尾)	.009	.153	.052	.011	.272

R0.4F0.6 - R0.2F0.8	DOMAIN - R0.2F0.8	R0.4F0.6 - R0.3F0.7	DOMAIN - R0.3F0.7	DOMAIN - R0.4F0.6
-1.266[a]	-1.123[a]	-2.535[a]	-1.577[a]	-.178[a]
.205	.261	.011	.115	.858

a. 以負等級爲基礎。
b. 以正等級爲基礎。
c. Wilcoxon 符號等級檢定

所以，由檢定及果可以得知(α=0.05，雙尾檢定)：

1. 專家知識放入演算法的結果，優於 ROCK 演算法使用 Jaccard 係數的結果(DOMAIN-ROCK 的漸近顯著性=0.005<0.025)。

2. r 與 f 由頻率判定時，除了 r=0.2,f=0.8 外，結果與 ROCK 演算法使用 Jaccard 係數的結果並無顯著差異。

4.2.5 試驗結果討論

1. ROCK 演算法使用 Jaccard 係數的結果，以及修改後的量測標準比較：

相似度量測標準由原先的 Jaccard 換成 $sim(X,Y)=\frac{a-f+r}{(b-f)-r}$ 之後，r 與 f 的判定，若是由頻率來決定，得到的結果，在統計上沒有顯著的與 ROCK 使用 Jaccard 的方法有差別。但是從平均值來看，正確的數量是有所進步的。換句話說，修改後的相似度標準，至少跟 Jaccard 一樣好，沒有顯著的贏 Jaccard，很可能是在 r 與 f 的判定上，有太多「noise」的緣故。與專家判定的 r 與 f 相比較，便可以發現其中有相當大的差異。

將專家的知識導入演算法之後，分群的效果就有明顯的變好了。統計上檢定的結果，結合專家知識的分群方法，顯著的比 ROCK 演算法使用 Jaccard 係數的方法來的好。

2. 相似度門檻值的討論：

比較 ZOO 資料集與 PET 資料集可以發現，PET 資料集的結果，不論是何種方法，在門檻值越高的部份，似乎正確率會越高；但是 ZOO 資料集就沒有這樣明顯的現象。比較兩資料集的差異，ZOO 資料集的幾個群集，群集大小都相差不多，而 PET 的資料，第一個 500 筆資料含有 53 個癌症病患，第二個 500 筆含有 19 個癌症病患。比較起來，這樣的比例相對很低。似乎在想要區分的群集比例差異較大的時候，選用大一點相似度門檻值，對於群聚的效果會比較正確。

第五章　結論及未來研究方向

本章共分為兩部分，將提出本研究之結論，最後提出本研究後續可再深入探究之議題與建議。

5.1 結論

本研究設計新的資料相似度量測標準。可以使得不同的特徵項目，對相似度的權重不同。在 ZOO 資料集的實例驗證中，雖然群聚正確率有些微提高，但在統計上並無顯著差異。PET 資料集的情形也與 ZOO 資料集相同。

一般的群聚方法，無法將領域知識放入群聚的過程。本研究設計能放入專家知識的機制。將領域知識放入群聚演算法，使得群聚的結果更為正確。在 ZOO 資料集，以及 PET 資料集的實例驗證中，群聚正確率均有提高，在統計上有顯著差異。群聚方法結合專家知識的機制建立後，可以幫助原本的群聚方法增進正確性。

5.2 未來發展方向

以條列式來說明本研究未來之發展方向

1. 領域知識的擷取

目前以知將領域知識放入群聚演算法，可使得群聚的結果更為正確；未來可以應用更多擷取知識的機制，再加入其他的群聚演算法。

2. 門檻值的探討

階層式群聚演算法的相似度門檻值一直是個問題。如何選取適當的相似度門檻值，目前仍然沒有一個系統化的方法訂定。雖然本研究推論，在群集差異大時，選用較大的門檻值效果較好，但仍需後續研究，有更進一步的證據才能確定。

參考文獻

[1] Ankerst, M., M. Breunig, H.P. Kriegel and J. Sander, “OPTICS: Ordering points to identify the clustering structure,” In Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD’99), pp. 49-60, Philadelphia, PA, June 1999.

[2] Dorian, Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann, California, 1999.

[3] Ester, M., H.P. Kriegel, Sander J. and X. Xu, “Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” In Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD’96), pp. 226-231, Portland, OR, Aug. 1996.

[4] Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim, “ROCK: A Robust Clustering Algorithm for Categorical Attributes,” IEEE Conference on Data Eng., 1999

[5] Guha, Sudipto, R. Rastogi, and K. Shim. “Cure: An efficient clustering algorithm for large databases.” SIGMOD'98

[6] Guha, Sudipto, R. Rastogik, and K. Shim, “Clustering Algorithm for Categorical Attributes,” Technical report, Bell Laboratories, Murray Hill, 1997.

[7] Han, J., “From Data Mining to Web Mining: An Overview,” Conference tutorial, 2000 Int. Database Systems Conf. (IDS’2000), Hong Kong, June 2000.

[8] Han, J. and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001

[9] Jain, A.K., M.N. Murty and P. J. Flynn. *Data Clustering: A Review.* ACM Computing Surveys, vol. 31, no. 3, pp.264-323, 1999.

[10] Karypis, George, Han Eui-Hong, Kumar Vipin "Chameleon: Hierarchical Clustering Using Dynamic Modeling," Computer pp.68-75 1999 IEEE

[11] Kaufman, L. and P.J. Rousseeuw, "*Finding Groups in Data: an Introduction to Cluster Analysis*," John Wiley & Sons, 1990

[12] MacQueen, J., "Some Methods for Classification and Analysis of Multivariate Observations," In Proc. 5th Berkeley Symp. Math. Stat. and Prob., Vol. 1, pp. 281-297, 1967

[13] Ng, R. T. and J. Han, , "Efficient and Effective Clustering Methods for Spatial Data Mining", Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 144-155

[14] Sheikholeslami G., S. Chatterjee, and A. Zhang, "WaveCluster: A multi-resolution clustering approach for very large spatial databases," In Proc. 1998 Int. Conf. Very Large Databases (VLDB'98), pp. 428-439, New York, Aug. 1998

[15] Wang, W., Yang and R. Muntz, "STING: A Statistical Information grid Approach to Spatial Data Mining," In Proc. 1997 Int. Conf. Very Large Data Bases(VLDB'97), pp. 186-195, Athens, Greece, Aug. 1997

[16] Tian, Z., Raghu Ramakrishnan, Miron Livny," BIRCH: An Efficient Data Clustering Method for Very Large Databases, " SIGMOD Conf. 1996: pp. 103-114 .

[17] 張智星，MATLAB 程式設計與應用，清蔚科技，民國八十九年。