# 第四章 SOM 分群整合應用

本章將從 4.1 節的分群理論為出發點，介紹關於分群的定義、背景說明與相關應用範疇，再配合分群法的分類及分群各階段的主要任務，最後並針對如何評估與解釋分群結果等主題加以闡述，以呈現群集分析的主要概念。接著於 4.2~4.6 節，便針對目前各種常與 SOM 結合的分群方法加以闡述，並舉出相關之整合應用。而 4.7 節針對以上介紹之分群法進行評估分析後，最後 4.8 節將介紹 SOM 於資訊生物學的基因分群應用，至於其他相關分群應用請參考第 3.2.1 節中兩篇 SOM 參考目錄研究【Kaski et al., 1998; Oja et al., 2002】與所列相關目錄，在此不加贅述。

本章各子章節以敘述性介紹為主，省略演算法探討，並將文獻檢索與文獻滾雪球法所得到本章相關文獻共約近百篇文獻一併於對應章節中陳述。如此建立之 SOM 分群文獻參考目錄不僅滿足本研究之研究目的，了解 SOM 與其他分群法之整合應用，更可進一步提供作為後續相關研究者完善便利的索引資源。

## 4.1 分群理論簡介

### 4.1.1 分群理論定義

集群(或稱聚類、群聚、叢集)分析(cluster analysis)，簡單地說，是一門在資料中找尋群組(groups)的學問【Kaufman, 1990】，也是將類似的目標對象歸聚成一群之行為【Hartigan, 1975】。也可進一步定義成對個體間的相似性加以量化之評估方法，使資料分類成有秩序之組織系統，並歸屬於統計學中的一種多變量分析法。若用這些方法來詮釋的話，則集群分析可說是各種用來找出資料集合中相似資料的數學方法之通稱【Romesburg, 1984】。然而就欲分析的資料對象而言，集群分析是一種將眾多個體或目標對象歸類為若干未知的分群，但與分類不同

的是，分群的數量及其特性必須從資料中獲取而無法事先得知【Afifi, 1984】。換句話說，群集分析嘗試將資料點歸類成同質性的群集，並假設無法事先得知群集資訊。而分析的第一步就是檢查資料點彼此之間的鄰近性(proximity)，因此亦可視為一種探索資料的分析技術。若是以分析結果而言，群集分析是一種將觀察資料結合成群類(groups)或群集(clusters)的技術，使其可達到以下兩個目標。其一是每個群類或群集之內，就某種特性而言，具有高度同質性或表現出緊密結實的分布狀態；其次是每個群組或群集之間，就某種特性而言，具有高度異質性，而不同群組或群集中的觀察值彼此相異【Sharma, 1996】。因此整體而言，群集分析是一種能根據資料變數之相似性與相異性，客觀地進行分類分群的邏輯程序，其目的在根據某種特性而劃分成的多個群集中，同一群集具有高度同質性(homogeneity)，而不同群集間則具有明顯的異質性(heterogeneity)【張紘愷, 2004】。而現今群集分析一詞較常指尋找資料中隱含群集的程序之通稱。

SOM 分群(clustering)便是將樣式(patterns)，包括觀察值、資料項目或是特徵向量等，進行非監督式學習(unsupervised learning)，並分類(classification)成若干群組或群集(clusters)【Jain et al., 1999】。進一步地說，分群是將資料有系統地分割為適當數量且具有類似元素(components)之群組，而每個群組內的元素將比相異群組間的元素具有更高的同質性【Pavel, 2002】。一個分群問題便是將一給定的資料集合分割成若干群組或群集，使得同一群資料點比不同群之間的資料點更具相似性【Guha et al., 1998】。若就使用方法而言，「分群法」是一種最普遍將資料分類成群的方法，其主要目的在於找出資料中相似的群組，並找出各群組之代表點，藉以達到降低資料數目之目的。這些完成分類的資料群組便泛稱為「群集」(或稱叢聚)。

值得注意的是，在分群程序中並沒有預先定義類別(categories)，也就是沒有預設對應規則，這也符合類神經網路非監督式學習的精神

【Berry & Linoff, 1996】。另一方面，分類則是將一筆資料項目歸類至一事先定義的類別中【Fayyad et al., 1996】。兩者之間的關聯為分群將產生分類程序中適合該資料集合的類別初始值。

### 4.1.2 分群背景說明

分群在資料探勘程序中為最有用的任務之一，其目的為發掘資料結構，並針對其中值得關注的資料分配(data distributions)與樣式型態加以鑑別與歸類於群集中，讓使用者可依據群集的相似性與相異性加以歸納並作出有用的結論。截至目前為止，關於集群、群組或類別的應用完全是以直觀法而不採用正式量化的數學定義，事實亦證明正規化定義於不同資料分配不僅難有統一規則，甚至可能會誤置或不合使用。因此 Bonner(1964)便提出最根本評估上述名詞定義之方法就是使用者的價值判斷(value judgment)。也就是說，如何使用這項名詞將會顯出調查者之價值觀。不過 Bonner 的論述並不全盤令人信服，因而許多學者仍致力於探討資料群集的量化評估法，Cormack(1971)與 Gordon(1999)便嘗試使用代表群集內部凝聚力的「同質性」(homogeneity)與象徵群集外部隔離度的「分離性」(separation)來定義一個群集。儘管如此，不可否認的在某些資料分布下直觀判斷遠比正式定義有效，況且許多實例亦證明沒有單一定義足以適用於所有情況。因此，以上論述或許可說明一觀念，便是若欲設計一些數量指標(indices)來定義資料群集的同質性或分離性，使其精確地符合數理原理之做法，將會產生許多不同之分群規則【Everitt, 2001】。

### 4.1.3 分群應用簡介

分群觀念與方法已經成功應用於許多領域中，諸如生命科學、醫藥科學與工程科學等，並且在不同背景有其不同專有名稱。例如人工智慧中樣式辨認的非監督式學習(unsupervised learning)，生物學與生態學的數值分類學(taxonomy)，社會科學的拓樸(topology)，圖論中的分

割(partition)、心理學的 Q 分析與行銷研究中常提到的市場區隔(segmentation)等【Halkidi, 2001；Everitt, 2001】。

以下整理出一些將分群作為其核心步驟之重要應用：以分群的應用方向而言可分為資料縮減(data reduction)、假說產生(hypothesis generation)、假說檢定(hypothesis testing)與群組預測(prediction based on groups)【Theodoridis & Koutroubas, 1999】。若以較詳細的特定領域應用便可分為商業(business)的市場調查與顧客消費行為分析、生物學(biology)的分類法定義、空間性的資料分析(spatial data analysis)如衛星影像、醫療設備資料或地理資訊系統(Geographical Information Systems, GIS)與網頁探勘(web mining)【Han & Kamber, 2001】；影像切割(image segmentation)、物體與字元辨識(object and character recognition)、知識擷取(information retrieval)與資料探勘(data mining)【Jain et al., 1999】；天文學(astronomy)、精神病學(psychiatry)、考古學(archaeology)、市場調查(market research)與作物分類(crop classification)等【Everitt, 2001】。

綜合以上所述，群集分析技術關切的焦點是探索目標資料集合，並評估是否可用數量較少的群組或群集產生意義的摘要，同時群集內的資料必須符合內在相似性與外在差異性的基本精神。

### 4.1.4 分群法分類

在過去 30 年以來，因運算效能的大幅提升而使得各種分析技術日益精進，造成許多分群方法亦隨之產生。從以下一些代表性文獻對分群法之分類，不難看出其演進脈絡及分析趨勢。傳統分群法基本上可分為階層式(hierarchical)與分割式(partitional)兩大類，且各自分別有其多元化的演算法，如圖 4.1 所示。其中最具代表性的是階層式的凝聚分群法與分割式的 K-means 分群法【Vesanto, 2000】。隨著人工智慧與軟式計算科學(softcomputing)的興起，遂產生模糊分群(fuzzy clustering)、類神經網路分群與演化式分群法(evolutionary clustering)，其中類神經

網路中最具代表性者即為自組織映射圖網路，而演化式分群法則為遺傳演算法【Jain et al., 1999】。在對分群的回顧研究中，共計列舉介紹 11 項分群相關技術，如表 4.1 所示。除上述所列之外，尚包括以搜尋法為基礎(search-based)的分群法，如模擬退火法(simulated annealing)；接近直觀式的最鄰近者分群法(nearest neighbor clustering)；為處理現今大型資料型態，如 CLARANS 與 BIRCH 分群法；以及因應分群使用者之實際需求而加入局部性限制條件之分群法。

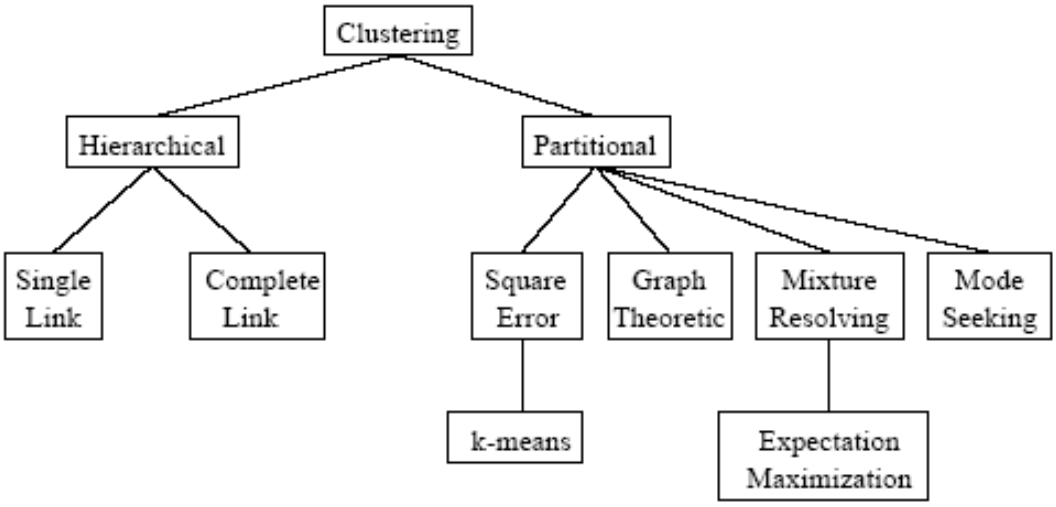


圖 4.1 分群法分類(一)【Jain & Dubes, 1988】

表 4.1 分群法分類表(一)【Jain et al., 1999】

| | 分群方法 | 簡　述 |
|---|---|---|
| 1 | 階層式分群法<br>Hierarchical Clustering Algorithms | 依次將較小群集合併或將較大群集分割，使得分群結果為反映資料結構的樹狀群集。藉由在樹狀圖之特定高度水準切割，可將資料分成若干數目的群集。 |
| 2 | 分割式分群法<br>Partitional Algorithms | 以反覆程序直接將資料分解成若干分離群集，使某能量函數最佳化。 |
| 3 | 混合求解與模式尋找分群法<br>Mixture-Resolving and Mode-Seeking Algorithms | 目的在確認從分配中產生的資料樣式，藉由估計元素密度的參數向量之最大可能性，來確定各分配所屬參數或甚至其數量。經參數評估而置於同一元素的樣式則可視為同一群集。 |

| | | |
|---|---|---|
| 4 | 最近鄰近者分群法<br>Nearest Neighbor Clustering | 將輸入樣式分配至與最鄰近已被指派的樣式之同一群集中，設定距離門檻值加以控制分群過程，直到所有樣式皆被分配完畢為止。 |
| 5 | 模糊分群法<br>Fuzzy Clustering | 使用模糊技術進行資料分群，單一資料可以分配於超過一個群集中，並給予一個介於 0 到 1 的隸屬度(degree of membership)，使得日常生活所遭遇到具有不確定性的實際資料均可以適用。 |
| 6 | 群集表現分群法<br>Representation of Clusters | 透過群心點、分類數節點或是邏輯符號的聯集等方式將來表現分群結果所產生的資料群集之分佈及其結構，以達到資料萃取之目的。 |
| 7 | 類神經網路分群法<br>Artificial Neural Networks for Clustering | 基本概念源自生物神經網路，因具備處理數量化資料，並擁有平行分散式處理架構，且透過神經元間的權重修正可正確學習樣式特徵。 |
| 8 | 演化式分群法<br>Evolutionary Approaches for Clustering | 由自然界演化啟發動機而來，並將所有可行解進行染色體編碼，被合使用演化操作因子如選擇、重組與突變等，來尋找全體母體中最佳的資料分割。 |
| 9 | 搜尋式分群法<br>Search-Based Approaches | 可分成固定與隨機式搜尋法以獲得準則函數之最佳值，固定式搜尋藉著徹底列舉方式保證達到最佳資料分割，而隨機搜尋則先產生一合理而接近最佳解的分割方式，且以漸進方式保證收斂至最佳分割的結果。其中常用的為模擬退火法[Aarts & Korst, 1989]。 |
| 10 | 限制式分群法<br>Incorporating Domain Constraints in Clustering | 分群具有主觀的本質，因此需要根據時空環境或使用者需求而制宜，每種演算法在執行分群時皆牽涉到使用某種直接知識，如專業意見；或是間接知識，如選擇相似度的衡量方法或分群演算法，甚至在特徵選擇與編碼時是否採用經驗值等。 |
| 11 | 大型資料集合分群法<br>Clustering Large Data Sets | 針對數以百萬計高維度的樣式資料集合進行分群以達到資料萃取，尤其是多媒體影音型態，目前最佳求解法如以基因演算法、塔布搜尋法與模擬退火法皆僅適用小規模資料集合。因此，收斂性的 K-mean 演算法與 Kohonen 的 SOM 受到矚目[Mao & Jain, 1996]。隨著資料探勘學門的興起，遂刺激新的分群法產生，如 CLARANS [Ng & Han, 1994]與 BIRCH [Zhang et al., 1996] |

若以分群資料所允許的變數種類作為分類依據，可將分群法作成

如表 4.2 所示的五大項分類【Guha et al., 1999; Huang et al., 1997; Rezaee et al., 1998】。而在 Halkidi(2001)對分群評估技術研究中參考 Jain et al.(1999)的分類並配合群集的定義方式，進一步將分群法歸納為分割式、階層式、密度式與網格式分群等四大類別，如表 4.3 所示【Halkidi, 2001】。其後 Pavel Berkhin(2002)針對分群資料探勘技術所進行調查研究中提出將分群法歸納成共計八大類十七細項【Pavel Berkhin, 2002】，如圖 4.2 所示。

綜合以上分類結果可知，分群法仍以階層式與分割式兩類為主軸，但是配合應用領域之特定需求，如人工智慧或機器學習，因而演化出特殊具有該領域演算精神的分類法形態。另外亦因應資訊爆炸時代的大型資料庫與高維度資訊的分析需求，特別是在資料探勘與知識擷取方面，同樣發展出嶄新的分群演算技術以便進行複雜的資料分析。

表 4.2 分群法分類表(二)【Guha et al., 1999; Huang et al., 1997; Rezaee et al., 1998】

| | 分群方法 | 簡　述 |
|---|---|---|
| 1 | 統計型分群(Statistical) | 以統計分析概念為基礎，並使用相似度來分割目標，限數值性資料。 |
| 2 | 概念型分群(Conceptual) | 用來進行類別資料的分群，並根據其概念資訊設定分群目標。 |
| 3 | 模糊分群(Fuzzy clustering) | 使用模糊技術進行資料分群，單一資料可以分配於超過一個群集中，使得日常生活所遭遇到實際資料的不確定性均可適用。其中最重要的演算法為 Fuzzy C-Means (Bezdeck et al.,1984) |
| 4 | 直觀分群(Crisp clustering) | 僅考慮非重疊分割，以大多數分群結果決定。 |
| 5 | SOM 分群(Kohonen net clustering) | 就是自組織映射神經網路，其演算法訓練結果在輸出層將形成群集。 |

表 4.3 分群法列表分類表(三)【Halkidi, 2001】

| | 分群方法 | 簡　述 |
|---|---|---|
| 1 | 分割式分群法(Partitional clustering) | 以反覆程序直接將資料分解成若干分離群集，使某能量函數最佳化。 |
| 2 | 階層式分群法(Hierarchical clustering) | 依次將較小群集合併或將較大群集分割，使得分群結果為反映資料結構的樹狀群集。藉由在樹狀圖之特定高度水準切割，可將資料分成若干數目的群集。 |
| 3 | 密度型分群法 (Density-based clustering) | 將資料集合中相鄰近的目標在特定密度條件下歸為一群。 |
| 4 | 網格型分群法(Grid-based clustering) | 針對空間性的資料探勘而設計，主要特性為將空間量化為有限數量之單元並進行所有相關之程序。 |

- Hierarchical Methods
  - Agglomerative Algorithms
  - Divisive Algorithms
- Partitioning Methods
  - Relocation Algorithms
  - Probabilistic Clustering
  - *K*-medoids Methods
  - *K*-means Methods
  - Density-Based Algorithms
    - Density-Based Connectivity Clustering
    - Density Functions Clustering
- Grid-Based Methods
- Methods Based on Co-Occurrence of Categorical Data
- Constraint-Based Clustering
- Clustering Algorithms Used in Machine Learning
  - Gradient Descent and Artificial Neural Networks
  - Evolutionary Methods
- Scalable Clustering Algorithms
- Algorithms For High Dimensional Data
  - Subspace Clustering
  - Projection Techniques
  - Co-Clustering Techniques

圖 4.2 分群法分類(二)【Pavel Berkhin, 2002】

### 4.1.5 分群階段

若以資料探勘的觀點來觀察分群程序，如圖 4.3 所示，可分成四大基本步驟(Fayyad et al., 1996)，以下分別介紹各個階段之主要任務：

1. 特徵選擇(feature selection)：本階段的目標是選擇合適的特徵以作為執行分群的依據，讓使用者感興趣的相關資訊得以儘可能完整地包含在輸入訓練資料中。換句話說，執行分群任務前必須先進行資料的前處理。
2. 選擇分群法(clustering algorithm)：此階段所選擇的分群演算法將產生適合資料集合的分群策略之定義。一個分群法的特性與效能可由特徵值之間的近似量(proximity measure)與一組分群準則(clustering criterion)加以描述。
3. 確認並評估分群結果(validation of the results)：在大部分的應用實例中，儘管資料集合中的集群資訊皆為未知，但是不論使用何種分群法，最後的分群結果均必須使用適當的準則或方法來驗證其正確性。
4. 解釋分群結果(interpretation of the results)：這是分群工作最後也是最困難的階段。在許多應用案例中，該領域的專家負起整合分群結果與相關實驗驗證，並在分析後做出正確結論與判斷。

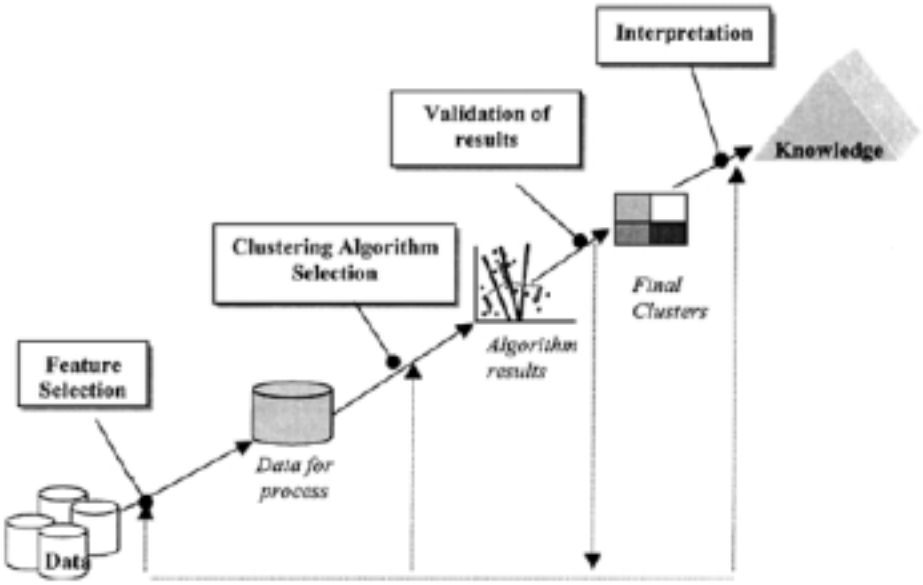


圖 4.3 分群程序之基本步驟【Fayyad et al., 1996】

4.1.6 分群結果評估：分群效標(validity indices)

群集分析中最重要的課題之一就是評估分群結果以找出最符合訓練資料的分割方式，這便是群集是否有效的主要衡量依據。一般而言，分群法應該在訓練資料集合中找到重要群集，使其滿足群內相似性與群間相異性。然而實際上最常遭遇的問題是難以決定符合資料集合的最佳分群數目，而此課題至今仍是相關學者所持續探討之焦點【Dave, 1996; Rezaee et al., 1998; Smyth, 1996; Theodoridis & Koutroubas, 1999; Xie & Beni, 1991】。

通常分群演算法的實驗數據常引用二維的輸入資料集合，其目的就是讓人們能視覺化方便確認分群結果的有效性，例如某演算法可以將資料集合分割地多麼好。然而，顯然地，將資料集合視覺化的做法對於分群結果相當困難驗證，尤其是遇到高維度，例如超過三維的訓練資料集合時，有效的資料視覺化就顯得困難許多，更遑論試圖以視

覺化工具從中找出最佳的分割方式。就因為如此，自組織映射神經網路便展現適用於高維度資料，不僅如此，更具有將其間關係保留並映射一維或二維的網路拓樸圖形上，以方便使用者進行分群結果有效性之評估。這也是自組織映射神經網路相較於其他分群演算法之最主要特性與優點。

以下針對分群有效性(clustering validity)的基礎概念加以說明，並舉出評估分群結果時最基本的兩種分群效標，至於其他為數眾多且各有其用途與特色的分群效標，請參考相關文獻之介紹【Bezdek & Pal, 1998; Davies & Bouldin, 1979; Halkidi et al., 2000; Rezaee et al., 1998; Sharma, 1996; Theodoridis & Koutroubas, 1999; Xie & Beni, 1991】。簡單而言，群集有效性(cluster validity)是探討並評估分群結果的程序，基本上有外部準則(external criteria)、內部準則(internal criteria)與相對準則(relative criteria)等三種方式來進行評估【Theodoridis & Koutroubas, 1999】。外部準則是以一個預先設定好的結構為基礎，再套用在欲評估的資料集合上，以反映分析者對該資料之直觀；內部準則是將包含資料集合本身之向量，例如近似度矩陣(proximity matrix)，加以量化後評估分群結果；而第三種相對準則的基本概念是將同一個分群演算法輔以不同參數值而形成分群策略，並將所得到的分群結構加以比較，再透過緊密性(compactness)與分離性(separation)兩種準則，以進行分群評估與選擇最佳的分群策略【Berry & Linoff, 1996】。前者通常採用群集內個體間最小變異數的統計量；後者大多採用單一連結、完全連結與群心法來計算群集間之距離。

## 4.2 SOM 與傳統分群法

本節中先針對傳統分群法中較具代表性的階層式分群法與 K-means 分群法之原理與特性加以簡述，並進行概要性的文獻回顧、分群應用介紹與簡單的效能評估；其次進一步介紹 SOM 與此兩種分群法

相互結合之分群應用。

### 4.2.1 傳統分群法

傳統分群執行上最常採用多變量統計方法【周文賢, 1998】，其中最具代表性的就是階層式分群法(hierarchical)與分割式分群法(partitioning)中的 K-means 分群法【MacQueen, 1967】，其後才興起如類神經網路或遺傳演算法等以人工智慧為基礎之分群技術。階層式分群法絕大部分是以單一連結法【Sneath & Sokal, 1973】、完全連結法【King, 1967】與最小變異法【Ward, 1963; Murtagh, 1984】作為基礎，而其中 又以前兩者最為常用。惟兩者相異處在於衡量群集間相似度的方法不同，但皆以距離最小準則作為演算基礎，逐漸將資料凝聚成一個較大的群集。完全連結法之分群結果傾向產生較密集的群集邊界【Baeza-Yates, 1992】；反觀單一連結則因連鎖效應而傾向產生細長蔓延狀的分群結果【Nagy, 1968】，同時具有鑑別出如同心圓狀之特殊資料分布。然而，若以務實觀點而言，完全連結法在許多應用實例中較能產生有用的資料層級【Jain & Dubes, 1988】。有別於分割式分群法，階層式分群法可以不需設定分群數目或相關群集屬性的情形下將輸入資料分解成之樹枝狀的群集結構。近年來也應用於基因表現之資料分群與視覺化分析【Szeto, Liew, Yan, & Tang, 2003】。

分割式分群法所獲得的分群結果為單一的資料分割方式，有別於階層式分群法所產生的資料樹狀結構，且大致可分為兩大類，一是強調各個資料點的重要性，而由鄰近的點聚類而成，如 Possibilistic C-means【Krishnapuram & Keller, 1993；1996】；另一種以群心為主要考慮，再將其餘資料點分配到各個群集內，如 K-Means、Fuzzy C-means(FCM)。Possibilistic C-means 比 FCM 多了一個控制參數，主要是降低雜訊的影響，但其計算複雜度也相對的比 FCM 高。

上述方法中皆以平方誤準則(squared error criteria)為最常採用的準

則函數，而其中又以 k 均值法(k-means)【McQueen, 1967】為最簡單常用, 其廣受歡迎的原因尚包括方便執行與時間複雜度僅有 $O(n \cdot k \cdot l)$，$n$ 為樣式數目，$k$ 是群集數，而 $l$ 則是達到收斂所需之循環次數，比起階層式分群法的 $O(n^2 \cdot \log n)$ 低許多，因此相當適合大型資料集合之分群應用(如表 4.4 所示)。然而，k-means 相當敏感於起始群心的選擇，也就是分群數目 $k$ 的決定。若是採用不當的起始群心種子進行演算，則結果便會落入局部最佳解【Selim & Ismail, 1984】。反之，若是能以其他方法獲得良好的資料起始分割，則 k-means 可在即便有瑕疵的大型資料集合中發揮極佳的分群效能並獲得令人滿意的分群結果，尤其是當輸入資料屬於球型分布的時候【Vesanto, 2000】。

表 4.4 分群演算法複雜度比較表

| Clustering Algorithm | Time Complexity | Space Complexity |
|---|---|---|
| k-means | $O(k \cdot k \cdot l)$ | $O(k)$ |
| ISODATA | $O(k \cdot k \cdot l)$ | $O(k)$ |
| Single-linkage | $O(n^2 \cdot \log n)$ | $O(n^2)$ |
| Complete-linkage | $O(n^2 \cdot \log n)$ | $O(n^2)$ |

就 k-means 的改善而言, Peña(1999)針對 random、Forgy、MacQueen 與 Kaufman 等四種 K-means 初始化方法進行實證研究，結果顯示 random 及 Kaufman 兩者較為獨立有效，而進一步比較其收斂速度，以 Kaufman 具有較佳演算效能【Peña, 1999】。張金華(2000)利用熵理論(Entropy theory)進行資料挖掘中的特徵挑選(feature selection)，先刪除解釋力較弱的資料屬性，再以該研究提出的基植於單位區塊的改良式快速 K-means 進行分群, 並証明能較傳統 K-means 提升 15 倍效能之多【張金華, 2000】。Cheung(2003)提出 K-means 分群法的廣義化版本，

不僅可適用於橢圓形資料集合，更能在不需事先決定群數而正確地分群，亦不會造成失效單元的問題(dead unit)【Cheung, 2003】。另外，亦有研究將 K-means 延伸應用於多準則決策(Multiple Criteria Decision Making; MCDM)【Smet & Guzmán, 2004】，由決策者選擇偏好結構(preferences structure)作為定義多準則距離之基礎，並根據此距離將替代方案所構成之集合加以分類，幫助決策者在數目有限的可行方案中做一優劣排序、評估並選擇最符合決策者理想的方案。

### 4.2.2 階層式及分割式分群法之比較

因此，以分群結果的輸出過程而言，階層式分群法以樹狀結構為主，分群後即不再改變；非階層式分群法是在分群的過程中，將原始集群重複的打散並重新分群。若以輸入資料的適用性而言，階層式分群法比分割式分群法來得更具功能性與多用途。例如，單一連結法不僅可適用於包括良好獨立分離之資料群集，同時在鏈狀與同心圓狀的資料集群也可正常運作；反觀 K-means 卻僅能在具有良好分離特性的球形資料分布表現良好【Nagy, 1968】。另外，若以運算時間與空間複雜度進行評估，則分割式分群法一般而言皆遠低於階層式分群法【Day, 1992】，因此相當適合大規模的資料分析應用。有鑑於兩者特性及優缺點，便有學者進一步提出結合型分群法【Murty & Krishna, 1980; Sharma, 1996】。有鑑於這兩種傳統分群法已廣被介紹與討論，因此僅針對兩者演算法加以比較如表 4.5 所示。

綜合以上所述可歸納得知，分群問題所面臨的兩大問題分別是決定最佳群數及評估分群結果。唯有將兩者皆加以妥適解決，方能建立正確的資料模型，並將其應用於現實世界，尤其是解決工程應用上的問題。然而就此觀點來評判傳統分群工具，不論是階層式的單一連結法或是分割式的 K-means 法，兩者皆需事先設定起始群數方能進行演算，但在許多實務應用顯然無法達到。

表 4.5 階層式與分階層式分群法比較表 【周文賢, 1998】

| | 階層式分群法 | 非階層式分群法 |
|---|---|---|
| 集群方法 | 單一連結、完全連結、平均連結、中心法、中位數法與華德法 | K 均值法(K-means)、NCS 法(Nearest Central Sorting) |
| 輸入資料 | 樣本點間或群集間的距離矩陣 | 樣本點到各群或各種子點(群心)距離 |
| 集群法則 | 距離矩陣中最小數值對應之群集則視為同一群 | 將樣本點分派至距離最近的群別 |
| 分群過程 | 樣本一但分至某群後便不再脫離並且可繪成樹狀圖 | 每次分群即重新分派所有樣本到距離最近的群別 |
| 分群次數 | 最多 N 次，直到所有 N 個樣本凝聚成一群為止 | 次數不限直到分群結果與前次完全相同為止 |
| 集群數目 | 可以事後決定，取決於 CCC 準則與分群顯著性之檢定 | 必須事先決定，取決於 CCC 準則與分群顯著性之檢定 |
| 適用情況 | 小樣本，即樣本數不超過 30；偏重理論瞭解但實務上不常使用 | 大樣本，即樣本數超過 30；實務上為普遍使用之分群法 |

同樣就評估分群結果而言，儘管先前介紹許多提出分群效標之研究，但是絕大部分都是立基於假設輸入資料為均一密度分佈，這也與現實狀況大相逕庭。因此，如何先能正確地描述輸入資料空間，找出一適當的資料起始分割，接著進行有效率地分群演算，並且最後能有一套通用的效標來評估分群結果，甚至視狀況彈性修正演算過程以符合使用者需要，這應該是一個優良的分群演算法所應具備的特性。

### 4.2.3 SOM 與傳統分群法之結合

有鑑於現今資訊爆炸時代中資料的儲存成本大幅降低與計算硬體設備的效能大幅提升，再加上個人電腦普及與網路發達等因素，使得資訊型態與規模呈現多元化且爆炸性的發展。最顯著的現象便是大型資料庫的普遍興起，使得如資料倉儲與資料探勘等資料儲存及分析之技術不得不朝向因應大型資料集合之需求而發展。如先前 4.1 節所介紹，隨著資料探勘學門的興起，K-means 與 Kohonen 的 SOM 分群法於

大型資料集合之分群應用日益受到矚目【Mao & Jain, 1996】。

關於 SOM 與傳統分群之結合以兩階段法為主，Sharma 認為兩階段法不僅能克服傳統階層式及分割式分群法兩者之缺點，更能使得分群結果更為精確【Sharma, 1996】。簡單地說，兩階段法結合傳統階層式及分割式分群法，先利用階層式分群法決定分群數目，再以 K-Means 進行分群。基本上，將 SOM 應用於兩階段分群法有兩種做法。第一種是採用 SOM 作為第一階段分群法以決定分群數目【胡承民，1997；鍾文杰，2001】，換句話說，就是從 SOM 所映射的資料拓樸上找出群集的群數及重心，第二階段再將此群數當作 K-means 的設定群數，而找出的重心則用來取代在 K-means 中隨機的起始重心點。在 Sugiyama & Kotani(2002)的研究中亦提出利用 K-means 將完成訓練的 SOM 權重向量進一步分群，如此可改善傳統 SOM 而獲得清楚的群集邊界以利於分群結果之判斷【Sugiyama & Kotani, 2002】。第二種做法就是先將所有資料集合進行 SOM 分群後，再以階層式凝聚分群法(hierarchical agglomerative clustering)與分割式 k-means 分群法(partitive k-means clustering)分別作為第二階段分群法來找出最後的資料群集【Vesanto & Alhoniemi, 2000】。這種做法最大的好處就是可節省大量計算資源，因此相當適用於大型資料集合的叢集分析，並且能夠在有限的時間內考慮數種不同的資料前處理(preprocessing)策略，以達到資料探索與知識萃取的目的。

另外，胡承民(1997)以蒙地卡羅模擬法在資料結構已知的情況下，進行三種分群法的評估，其中包括(1)以往市場區隔常用的兩階段式分群法；(2)非監督式自組織映射圖神經網路(SOM)；以及(3)整合 SOM 及 K-means 分群法的改良式兩階段式。根據評比結果，研究認為整合 SOM 及 K-means 的分群法可以成為實行市場區隔及其它必須使用集群分析時的一項新選擇，且其效能應較使用多變量統計集群為佳。

其後，Krishna 及 Murty 利用 K-Means 找出操作因子，來替換遺傳

演算法中的交配因子，以找出整體最佳化的分群方法(Genetic K-means Algorithm, GKA)【Krishna, 1999】。而鍾文杰(2001)便接著提出了 SOM+GKA 的聚類法，先以自組織映射網路將資料預先分群，以求出的初始群數代入遺傳演算法，將其突變率更動，再透過 K-means 重新安排重心，來縮減其演化的時間。在實際做法上同樣是利用蒙地卡羅模擬法來評估三種集群工具：(1)利用自組織映射圖網路找出初始群數，再將其代入 K-Means 方法(SOM+K-Means)；(2)由 Kuo 所提出的"整合自組織映射圖網路與 K-means 演算法“ (SOM+K);(3)利用自組織映射圖網路找出初始群數, 再代入改良式 Genetic K-Means 演算法找出最佳分群(SOM+GKA)。結果針對國內四家著名物流業者進行配送實例之區隔分析，結果以 SOM+GKA 所分出五個區隔的組內變異為最小，即為該實例中的最佳分群工具。

然而，上述研究所提出以蒙地卡羅模擬法產生訓練用的隨機數據來進行分群工具之效能評比，或許可解決缺乏適當的統計檢定來判斷分群結果之缺憾，然而卻也缺乏真實生活資料之代表性，同時也無法得知在不同或特定的輸入資料分配下，或進行這些分群工具之效能測試與評估。以上是 SOM 與 K-Means 結合應用的兩階段分群法之概述。

若將 GHSOM 與階層式分群法兩者相互比較不難發現，儘管兩者演算過程不盡相同，但其分群結果卻同樣具有以相似度為衡量基準的階層性架構。而 GHSOM 的提出【Dittenbach, Merkl, & Rauber, 2000, 2001, 2002】不僅解決傳統 SOM 的分群結果缺乏獲取輸入資料階層性結構的能力，並且透過一個可配合資料而生長的彈性結構，同時保存網路拓樸在視覺上的直觀性。此外，並提出一組質性的門檻參數來決定合適的粒化性(granularity)作為控制網路增長行為之判斷基準。GHSOM 提出迄今已發展出成熟而且可在 Matlab 環境下執行的工具庫程式【Chan & Pampalk, 2002】並廣被探討及應用。在 Pampalk, Widmer & Chan(2004)的研究中指出 GHSOM 在網路增長決策上的缺點，並提

出延伸模型 GHSOM-TMR (Tension and Mapping Ratio)【Pampalk, Widmer & Chan, 2004】。該改善方法基本上是以 GHSOM 為基礎，並且使用兩個分別用來分析 SOM 拓樸與群集穩定性(stability)之統計量為基礎，不需使用者調整參數而可以自動產生足以表現資料內部階層性結構所需的粒化程度，使得演算結果更令人滿意。總而言之，SOM 與傳統階層式分群之綜合應用可由 GHSOM 相關研究著眼，並由大型資料集合的階層性探索進一步了解目前 SOM 在資料探勘上的實際應用【Vesanto, 2000】。

## 4.3 SOM 與模糊集理論

現實生活中處理分類問題時普遍存在的資料型態大多具有不確定性、非常態分配、不穩定的統計特性，甚至包括語意型態的變數【Zaremba, 2000】，此時模糊集理論【Zadeh, 1965】的引入便產生相當大的幫助。傳統分群法產生的資料分割均明確地與單一群集相對應，因此對這種硬式分群結果而言，群集之間並無關連性；反觀具有模糊處理能力的分群法則進一步延伸原有的分群概念，利用隸屬度函數(membership function)將每筆訓練樣式與每個群集之間均建立程度不一的關聯性，以致於模糊分群之輸出結果為一叢集(clustering)而非一切割結果(partitioning)，亦即單一樣式通常不會完全屬於某一群，而很可能介於兩群之間【Jain, 1999】，如此較符合現實世界普遍存在的模糊現象。

### 4.3.1 模糊分群法(Fuzzy Clustering)

模糊分群法(Fuzzy Clustering)，顧名思義，乃是將傳統分群法結合模糊集合理論而成，其主要目的在有效率地針對模糊非監督的樣式建構合適的模型【Baraldi, 1999a】，首先由 Dunn 提出目標函數概念【Dunn, 1974】，其後再由 Bezdek【Bezdek, 1981】提出通用模式。簡單的說，

就是將一組樣式根據一些準則分配至已知數目的資料群別中，使得每筆樣式對應到一個以上的群別，並且各自擁有不同程度的隸屬度。資料發掘之目的是希望從大量資料中尋找內含的規律性與隱藏的知識，然而處理現實的資料分割問題最大的困難是當兩個不同類別資料產生重疊時無法獲得明確的分群結果。在這方面，模糊分類認為輸入樣式可能會受到誤差或未知原因的影響，因而對所有類別皆具有可能歸屬性，即隸屬度(membership function)，需依據資料本身的特性來決定。基本上，模糊分類顯示輸入樣式的所有可能機率(probabilistic)與可能性(possibility)，透過模糊隸屬函數的定義來衡量上述隸屬度，以作為分群衡量標準。

在解決模糊分群問題中最受歡迎的就是 Fuzzy k-Means 法【Bezdek, 1974】/ Fuzzy C-Means 法【Bezdek, 1981】，其演算型態相當類似由 McQueen 所提出的 k-Means 硬分群法(crisp/hard clustering)，即根據加權最小平方和準則(weighted sums-of-squares criterion)持續尋找隸屬度與新群心，直到目標函數無法獲得更佳解為止。近年來有大量啟發式演算法用來尋找最佳解，如基因搜尋(genetic search)、模擬退火法(simulated annealing)、塔布搜尋法(Tabu search)以及倒傳遞類神經網路(back-propagation neural network)等等。後續 Belacel 提出的 Fuzzy J-Means 保留局部搜尋特性，並將所有可能成為群心的樣式集合定義為鄰近區域，再配合變動型鄰近區域搜尋法(variable neighborhood search)，將目標函數的整數解轉換成連續值的模糊解，結果證明可獲得更佳的品質【Belacel, 2002】。

在眾多的模糊分群研究貢獻中，Rousseeuw 針對模糊群集分析提出綜合性探討【Rousseeuw, 1995】，而 Baraldi 則以不同構面的功能特徵(functional attributes)，針對模糊分類法的自組織策略進行文獻回顧與整合性比較研究，並闡述模糊分群與軟性競爭式學習(soft competitive learning)之概念【Baraldi, 1999a】，接著於後續研究中進而挑選五種分

群法，針對其輸入參數、演算法則、限制條件、優缺點與結構特徵等構面加以驗證比較【Baraldi, 1999b】。模糊分群的應用在樣式識別、機械學習、資料探勘與知識擷取等學門均廣受探討，並且已成功應用於影像分析、化學、能源、生物科學、地質學、醫療科學與資料探勘等廣泛領域【Belacel et al., 2002】。例如將 Fuzzy ART 與 SOM 等非監督式類神經網路視為以分群為基礎的資料分割方法，並應用在影像處理與分割問題【Cinque et al., 2004】；或是結合決策樹方法來發展模糊分類樹並且應用於健檢資料庫，實驗證明模糊分類之效能確實優於傳統分類法【蔣以仁, 1997】。

關於模糊分群的評估，或稱適切度分析(validation analysis)，【譚嘉慧, 2000】將分類適切性指標，即分群效標(cluster validity index)分為兩大類型：第一類型是以隸屬函數計算作為判斷準則；第二類型為結合隸屬函數與資料點結構兩者加以綜合判斷。前者主要缺點為當 Fuzzy C-Means 所得結果並非最佳分割時，僅用隸屬函數的計算並無法提供正確的適切度判斷。而後者加入資料結構的幾何原則，可用以修正上述演算法誤差。該研究提出一個 WB 分群效標，以傳統 DB 效標(Davies & Bouldin index)概念為基礎，接著引入模糊觀念，以隸屬度代表群內變異程度(緊密性)，再以資料點的群間距離來代表群間變異程度(分離性)，該 WB 效標經實驗證實優於上述其他分群效標。另一種模糊分群效標同樣是評估 Fuzzy C-Means 所得的模糊分群結果，但將目標函數定義為群間接近性(inter-cluster proximity)，而最佳 Fuzzy C-partition 產生於某特定 C 值，使群間接近性最小化【Kim, 2003】。

### 4.3.2 模糊自組織映射圖類神經網路(Fuzzy Self-Organizing Map ; FSOM)

關於 SOM 與模糊集理論的理論整合，也就是模糊自組織映射網路，大致上可分為兩大類來探討。第一類是以 SOM 網路架構為基礎，

輔以輸入資料的模糊化與隸屬度最佳化，以下用 FSOM (Fuzzy Self-Organizing Map)表示；而第二類則是以模糊類神經網路(Fuzzy Neural Network)的結構為主，輔以 SOM 來進行各層的權重值訓練，以下用 SOM-based FNN 表示。以下分別針對兩者進行論述與介紹：

傳統的 Fuzzy C-Means 是一種反覆尋找隸屬度與群心，使得目標函數最佳化的求解過程。然而該演算法之主要缺點為當分群數量很大時所產生龐大耗時的計算量。但是當 SOM 與 Fuzzy C-Means 結合應用時，Zaremba 指出可有效地解決兩者演算法的缺點並同時提升分群效能【Zaremba, 2000】。該研究將隸屬函數值取代 SOM 的學習速率，使得愈接近群心的訓練樣本獲得愈大的隸屬度，並且為簡化運算，僅針對 SOM 鄰近區域內的神經元進行隸屬度的計算。換言之，此法利用 SOM 在分群時的資料縮減、非線性的降維映射與鄰近區域內資料關係的保存特性，結合 Fuzzy C-Means 的模糊處理能力與最佳化機制，提供更有力的資料處理策略，如視覺化技術、資料探勘或知識發現等，以因應來自衛星或全球定位系統的空間性高維資料之分析需求。

另外，上述第一類模糊自組織映射網路(FSOM)模式主要以 SOM 為其網路架構，並將輸入向量與權重向量加以模糊化後再進行運算與非線性映射。因此不僅擁有類似人腦『物以類聚』特性，對未知型態的輸入資料可更具有模糊化處理能力與抗雜訊的容錯性。根據本研究蒐集的 FSOM 相關文獻歸納得知，大多是應用在影像辨識，特別是處理群組技術(group technology)工件分群(part family formation)中的工件辨識問題。而此類問題過去主要是採用監督式的倒傳遞(backpropogation)神經網路，但僅適用於間斷型特徵資料；或是非監督式的自適應共振神經網路(ART-1)來進行工件分類，不過仍無法處理模糊型態的特徵值，且應用範圍較為狹隘【Pai, 2001】。

在【鄧博文，1998】、【郭人介、紀勝財，2000】與【Kuo, 2001】研究即探討如何將 FSOM 應用在彈性製造系統中智慧型群組技術的分

群任務上，其作法便是利用上述 SOM 與模糊理論的結合方式，以視覺感應器 CCD 所擷取的工件影像資料為基礎，將 SOM 的輸入值與權重值模糊化後進行訓練，最後將工件歸類到不同群別。該研究與 Fuzzy C-Means 法比較，結果顯示在不同角度的偏斜率下，FSOM 擁有比 Fuzzy C-Means 更高的辨識率，並以人工判讀 SOM 的輸出結果來改善群數須先設定的缺點；此外當雜訊比不大時，其抗雜訊能力也比較強。而相較於 k-Means 而言，FSOM 更能針對工件影像的輪廓邊緣進行模糊化處理，達到更佳的擷取物件特徵與抗雜訊能力。亦有學者從另一個不同觀點切入，即利用 SOM 神經網路的學習特性來改善描述工件特徵的模糊隸屬函數本身所固有的趨近特性(approximate)，因此使得 FSOM 可同時依據間斷型(crisp)、區間型(interval)與模糊型等不同型態的物件屬性進行分群。該研究並提出一個新的 K-winner-take-all 權重更新機制來處理模糊輸入與模糊權重，並衡量模糊數之間的距離作為鑑別兩工件相似性之判斷依據【Pai, 2001】。而陳德華(2003)同樣改善傳統 SOM 只能處理實數型態的輸入資料，而進一步提出將混合特徵資料的距離定義方式建構在 SOM 網路架構上，可處理符號型與模糊型的資料，並根據模糊軟性學習法 Wu(2003)另外提出改善版本加以比較。

其他相關研究如許維宸(2001)以 SOM 為主體提出三種不同組合的分群法，並將自動聚類結果所得的各個聚類群心分別建構三角模糊隸屬函數，再將屬性內所有值代入所屬的模糊隸屬函數，以求得對應的隸屬度，並以該研究提出的模糊資料萃取法則，結合模糊集合概念與 Apriori 萃取法，以獲得最後的模糊關聯性法則。

至於第二類以 SOM 為基礎的模糊類神經網路(SOM-based FNN)主要與第一類 FSOM 之差異在於前者為非監督式學習，用在聚類分群；而後者則屬於監督式競爭式學習，目的在建立模糊規則庫，以作為預測用途。趙志運(2000)提出利用 SOM 演算法的向量量化特性找出最具代表性的群心，再利用拓樸保存特性挑選出一組最具影響力的模糊規

則。如此可在兼顧模糊系統的計算效能與系統表現以計算輸出值。該研究並提出一個迅速搜尋優勝者的啟發式演算法，以提升傳統 SOM 的演算效能。吳育奇(2000)則利用 SOM 結合模糊控制，加入 Grossberg 層作為網路的輸出階段，並經過樣式比對與權重均分的遞迴訓練進行監督式學習，以建立規則庫及產生輸出推估值，並且進一步應用於水庫上游流量推估預測。

Tung(2002) 便提出一般化的自組織模糊類神經網路 GenSoFNN(Generic Self-Organizing Fuzzy Neural Network)，該模式的網路架構有五層並具備兩大特色：其一是使用離散增值分群法 DIC(Discrete Incremental Clustering)來增加其雜訊容忍能力，並且針對關聯性較差的雜訊新增一個新叢集，故不需要事先設定分群數目；另一特色就是以資料導向自動產生模糊規則，並根據訓練迴圈的定義來刪除多餘老化的模糊規則。而網路模式中的參數訓練則是使用以最陡坡降法為基礎的倒傳遞學習演算法。

## 4.4 SOM 與遺傳演算法

演化式計算技術(evolutionary approaches)取法於自然界的演化機制，其作法是將分群問題的候選解進行染色體編碼(chromosomes)，並利用演化操作因子在可能解的母群體中演算，以獲得訓練資料的最佳分割結果，並透過適應性函數來評估代表母代的染色體，其演化到下一子代的存活性【Jain, 1999】。若以數值分析的角度來看，演化式計算可視為類似隨機搜尋(random search)的最佳化演算過程(optimization)；而若以自然演化的角度來看，演化是在動態變化的環境下的一種適應(adaptation)過程，並非在靜態環境下的最佳化過程。因此演化可視為適應性複雜系統(adaptive complex system)的變化過程；若由物理觀點出發，複雜是在非線性動力學中介於混沌(chaos)與秩序(order)間的現象，

並且會帶來資訊(information)的產生與變化。物理學中描述運動的動力學(dynamics)是決定性的，無法描述演化；但在熱力學(thermodynamics)中處理非平衡狀態的消散性結構(dissipative structure)卻可以描述複雜動態系統中的演化現象，亦即系統隨時間愈趨於均衡時，則熵度便會朝向最大化發展。因為演化式計算的本質屬於非決定性的(nondeterministic)，所以能處理現實世界中大量未知型態的資料集合所構成的複雜系統，並可隨系統環境改變而在問題空間中針對所有可能的解答進行全域搜尋，以尋求適應環境的最佳分群。

若由性質來區分，現有演化式計算之相關模型可大略分類如下：【http://cindy.cis.nctu.edu.tw/EC/ACS/frame_theory】

- 最具代表性的遺傳演算法【Holland, 1975；Goldberg, 1989】
- 較偏數值分析的演化策略(evolution strategy)【Schwefel, 1981】
- 介於數值分析和人工智慧的演化式規劃(evolutionary programming)【Fogel et al., 1965】
- 偏向以程式表現人工智慧行為的遺傳規劃(genetic programming)
- 適應動態環境學習的分類元系統(classifier system)
- 用以觀察複雜系統互動的各種生態模擬系統(echo system and etc.)
- 以離散動力系統行為研究人工生命(artificial life)的格構自動機(Cellular Automata)
- 模擬螞蟻群體行為的蟻元系統(ant system)

在上述演化式計算方法中，以前三者較具代表性且都以最小平方誤準則來處理分群問題；而其中又以遺傳演算法最常用，並證實在許多複雜問題的最佳化求解上有不錯的成效。另外，本章所介紹的所有分群法中，唯有遺傳演算法擁有跳離局部最佳解的機制並進行全域搜尋，其餘僅能執行局部搜尋。這便代表以遺傳演算法為基礎的分群法將更有機會找到最佳分群。

基本上，遺傳演算法的分群演算程序可以分成下列三大步驟：

1、 從解母體中隨機抽取一個可能解。每個可能解都會對應至一種資料分割，同時並計算適應性數值。一般而言，適應性數值會與平方誤成反比；也就是說，當可能解的平方誤愈小時，則獲得的適應性函數數值就愈大。

2、 利用演化操作因子，如選擇(selection)、重組(recombination)、突變(mutation)、交配(crossover)與複製(propagation)等，進而產生下一個子代。

3、 重複步驟 2 直到終止條件滿足為止。

### 4.4.1 基因分群法(Genetic clustering)

以遺傳演算法為基礎的分群研究中，Tseng & Yang (2001)提出 CLUSTERING 基因分群法，可適用於結實緊密的球型分佈資料，並且可透過兩種方式進行聚類。首先是由使用者控制之分群，藉由調整參數值來控制分群結果，當參數值小時將產生較多的緊密群集，反之較大的參數值則會導致較少的鬆散群集；另一種是自動化分群，可分為兩階段進行，第一階段是以最近鄰居法(單一連結法)獲得初始分群並降低資料集合規模，以節省下一階段的計算時間；接著第二階段則進行基因演算法，並利用一個啟發式策略尋找最佳的分群結果。最後評估該分群法與 k-means、單一連結法與完全連結法之分群效能。另外，就基因編碼方式而言， Chiou & Lan (2001)提出 SICM(simultaneously clustering method)、STCM(stepwise clustering method)與 CSPM(cluster seed point method)等三種基因分群法，並評估這些分群法與傳統階層式的凝聚分群法在處理不同規模分群問題的效率與正確度。結果顯示在研究設計各種規模的分群問題中，以 CSPM 為最有效，卻也是最缺乏效率的分群法。

Garai & Chaudhuri (2004)所提出的兩階段分群法，其特色是利用分

合機制(split-and-merge)來尋找資料中的群集。第一階段先透過群集分解法(Cluster Decomposition Algorithm, CDA)演算法將原始的資料集合分解成較大數量的破碎群集；第二階段乃以遺傳演算法為基礎，並利用階層式群集合併法(Hierarchical Clustering Merging Algorithm, HCMA)反覆地在整體空間進行最適基因搜尋，並將接近的破碎群集合併成完整的 K-cluster。此外，進行群集合併過程中採用鄰近群集檢核法(Adjacent Cluster Checking Algorithm, ACCA)，目的為衡量兩相鄰破碎群集之間的鄰近性(adjacency)，以作為集群合併之判斷準則。最後以數個含有多個群集的資料集合來驗證此演算法的效能，並同時與其他具有分合機制的分群法加以比較。綜合上述研究方法，可進一步了解基因演算法在處理分群問題時，會在可能解空間中持續搜尋，直到獲得最適化的分群數量與結果為止。不過，訓練資料型態及分布之適用性，目標函數如何適當定義群集間的鄰近性，以及分群效能與正確性的兩難問題等，皆為值得繼續探討的議題方向。

### 4.4.2 SOM 與基因分群法之結合(Genetic SOM)

因此接著將介紹如何應用基因演算法的最佳化求解特性，進一步改善 SOM 演算法或發展結合模式的分群研究。McInerney & Dhawan (1993, 1994)指出傳統的 Kohonen 演算法對於參數設定相當敏感，並且很有可能因而造成網路拓樸的不當對應。為改善此問題，該研究提出一個混合型演算法 GAKN，並引用一成本函數來衡量拓樸映射的正確性，當函數值最小化時便代表正確映射的網路拓樸。此外，GAKN 將 Kohonen 的訓練參數集合視為基因搜尋空間，先利用 Kohonen 學習達到局部最佳化後，再輔以基因演算法來尋找最佳參數組合。根據該實驗結果指出，在較複雜的資料集合中，成本函數最小化的機制將使得 GAKN 較傳統 SOM 擁有更快速的學習收斂性。然而就本研究調查結果

並未發現該作者在其他國際期刊發表相關研究。蔡坤洋(1998)根據基因資訊建構生物神經系統之概念，提出一種建構類神經網路的方法。首先設計建構類神經網路結構的規則，並將其進行基因編碼，再透過遺傳演算法找到最佳法則。

Nissinen & Hyötyniemi (1998a, 1998b)提出行為取向(behavior-based)的演化式 SOM 模式，其演算法在由母群體所構成的二維網格中，以樣本點訓練進行最佳化搜尋。而其中鄰近區域的概念適合用演化計算的方式來產生機率詮釋(probability interpretation)。經過競爭式學習與局部導向的演化機制，使網路可依據適應度在演化個體之間產生組織化結構，以取代過去的參數化表現(parameterized presentation)。該研究並指出所得到的網路結構具備行為模式基礎，可應用於隨時間變化程序的建模，亦即根據時間序列資料的行為不斷即時學習修正。因此可將不同的操作條件儲存於特徵圖中，類似簡易的大腦記憶機制，並且可應用在錯誤診斷與評估上。

Chang & Heh (1998)提出一動態生長的演化式 SOM(Evolutionary SOM)，首先將 SOM 輸出層建構一個鄰近區域圖形，以輸出神經元的權重作為節點(nodes)，鄰近區域關係作為邊(edges)，並透過兩者的分割操作(differentiation operation)來進行圖形演化(graph evolution)。訓練結果使得 ESOM 的每一個世代將產生數種神經網路，其中競爭優勝者將可被區分出來並進入下一世代。實驗結果指出該 ESOM 模式能正確地追蹤輸入資料，並且利用圖形演化而動態修正鄰近區域關係。然而該演算法目前僅有增生機制，尚缺乏即時評量與刪除機制，以達到更完整的演化模擬機制。Kim, Ahn, & Kang(2000)同樣以 Kohonen 學習法則混合基因演算法的概念，提出兩種共適應策略(co-adaptation scheme)，並透過並行的演化(evolution)與學習(learning)機制，來找出最佳向量量化的權重值。而演化機制乃藉由不同權重向量組合(codebook)之間的結構性調整來進行全域搜尋，而學習機制便是將單一權重組合

內部的失真最小化，以完成局部搜尋。根據該研究的模擬結果顯示，以區域學習所主導的演化機制將使演算法得以快速收斂，且經過共適應學習的權重組合亦較未經學習者產生更佳的影像再現品質。

簡順源(2001)利用蒙地卡羅法(monte carlo method)產生模擬資料以用來評估以下三種集群工具：(1)K-means，(2) S+K，(3) S+G。其中前兩者已於前面 4.2 節中介紹，而第三種為整合 SOM 與 GA 的分群法，作法是先使用 SOM 決定適當群數，再以遺傳演算法為基礎的集群方法進行聚群步驟。模擬結果顯示 S+G 為最佳分群方法，因而該研究以電信服務業為研究對象，針對行動電話客戶所重視之不同利益變數進行問卷調查，利用 S+G 分群法，進行消費者市場區隔，以提供業界作為訂定市場行銷策略之參考。

Villmann, B. Villmann & Slowik(2004)屬於回顧性研究，調查結合 SOM 與演化式計算系統的實際發展，並以兩個不同的演算方向進行探討。一者是藉由結合演化環境來改善神經網路的運算，另外就是採用神經網路的學習機制來規劃演化系統。該研究並根據不同的神經圖形提出數種方法，試圖將鄰近區域的協同合作特性納入至演化式計算系統中。

儘管基因演算法在工業工程領域展現極可觀的最佳求解能力，然而在現實世界中仍有些問題卻受限於基因的線性排列型態而無法順利進行編碼。另一方面，有學者提出演化型自組織神經網路(Evolutionary SOM, ESOM)試圖改良傳統 SOM 演算法，使自組織圖(即網路拓樸)的鄰接關係突破傳統固定的限制，而更具備演化可能性【游鴻志，1997】。其作法是以圖學的角度去分析 Kohonen 自組織圖的鄰近關係，並發展出兩套可以互相配合的動態式自組織圖演算法：『演化型自組織圖』與『交配型自組織圖』。其中演化型自組織圖演算法將神經元的權重向量與鄰接關係的組合視為圖形，並發展出圖形演化生長的操作機制，使演化型自組織圖可以依據輸入資料的特性，進而動態調整其權重向量

數目與鄰接關係，以改善 SOM 固定的起始架構之缺點。而交配型自組織圖演算法則是將兩個不同神經網路結構的自組織圖進行基因演算法的改良式交配操作，目的是產生與父代不同結構的子代自組織圖。此外，為要跳離局部最佳解，可將演化型自組織圖設定為區域最佳解的結構，再透過交配型自組織圖的交配與重組，使其有機會離開。從生物學的觀點來看，這兩種改良式演算法便像是生物的無性生殖與有性生殖。因此適合應用在具有動態環境特性的應用領域。而 Chang & Heh(1999)同樣提出圖形演化(graph domain)概念，突破傳統基因的線性編碼，不僅涵蓋原有基因演算法之特性，更將問題求解範圍拓展至圖形領域。該研究並舉例証明 ESOM 較傳統 SOM 擁有較低的分類失誤率與更佳的學習曲線。

## 4.5 其他分群方法

以上各節介紹 SOM 與傳統分群法，或與模糊理論及演化式計算等軟式計算技術的結合應用。除此之外，實際上仍有許多不同特性或適用特定情形的分群技術，本節僅列舉一些較具代表並與研究主題相關之研究分述如后。

另外，Si, Lin & Vuong (2000)所提出 DTRN(Dynamically Topology Representing Networks)分群法，可同時進行拓樸學習與資料分群，來建構具適應性的網路模式。前者利用警戒檢驗機制而適應性增加輸出單元數量，而後者採用贏者取配額(winner-take-quota)學習法則，並輔以退火機制而達到均方誤(mean square error)最小化。同時並利用競爭式 Hebbian 法則學習整體拓樸的資訊，以便動態刪除單元與使用在退火機制中。根據實證模擬結果顯示，該演算法不論在拓樸保存、學習速度與分群有效性都有令人滿意的結果。

傳統的獨立成份分析(independent component analysis)使用線性模

型來描述整體資料空間。儘管在許多情形下可產生有意義的結果，但對普遍的非線性資料分布卻僅能粗略估計而已。為改善此情形，Karhunen, Mălăroiu & Ilmoniemi (2000)提出結合 ICA 與分群法的構想。前者將資料群集以線性 ICA 模式來描述局部性的資料特徵；而後者則負責產生代表整體資料的非線性描述與表現。如此便可避免非線性 ICA 的計算困難度但卻能得到優於線性 ICA 之資料表現結果。該研究採用的分群法包括 k-means、SOM 與 NG(請見第 3.3.2 節)等。

若從相似度衡量的角度來改善分群法時，Wong, Chen & Su (2001)針對已知資料集合提出一高效能的非監督式分群法。該分群法以相似度的衡量指標為基礎，將所有資料點先後進行權重向量的相似度評估與取代同化，如此便可根據分類效度來決定適當的群集，並且不需要事先決定分群群數，而是在最初階段選擇適當群心，再根據資料分布形狀選擇合適的相似性衡量基準(similarity measure)。如此可有效地決定群心位置並進行正確分類。而 Bandyopadhyay (2004)則提出一個稱為 CLUSTER 的自動化分群法，不同大多數其他方法需要事先假設群集數目或結構，該分群法可在分離狀況良好的資料集合中自動偵測任何形狀的群集數目，包括凸面與凹面的群集分佈；同時並能偵測出離散值(outliers)與群集結構的固有階層特性，甚至能鑑別出資料集合沒有自然群集存在的情況。該演算法基本上是以相對鄰近區域圖形(relative neighborhood graph)的反覆切割程序為基礎，並結合群集合併機制的後處理程序，如此來達到更佳的分群效能。

另一方面，隨著現今大量電子化格式的全文資料集合興起，需要創造有效的分析工具或機制來協助使用者組織這些龐大的資料庫，因此使得文件分群的相關探討與應用快速崛起。Hussin & Kamel (2003)便提出結合 SOM 與 ART 的階層式神經網路模式，對路透社新聞資料集合進行文件分群，結果顯示可有效提升分群效能。至於 Deng & Kasabov (2003)則提出一種演化型 SOM(evolving SOM)，其特色是具有

可演化增長的網路架構與快速的線上學習機制，以符合現實世界中資料處理技術所須具備的高效率與適應性。ESOM 與 GNG 演算法擁有類似的網路架構，也就是從沒有任何節點與連結的空網路架構開始演算，根據輸入空間隨機產生的訓練資料，伴隨網路中代表樣式的節點不斷地相互競爭，相連的優勝者會更新其權重，而不相連的優勝者則建立一新連結，以便對未知的資料輸入空間得以呈現真實的群集分布，因而具備隨時修正更新的線上學習能力。值得一提的是，在該演算法最後階段中，設定每隔固定次數的訓練循環之後，便刪除最弱的連結。如此兼具結構演化與鄰近區域內權重修正之樣式區分，確實更強化 SOM 的自組織效能以適應環境變化。

## 4.6 分群方法之比較

在眾多具有不同目的與特性之分群法中，彼此之間相關性與分群效能為何？本研究特別關注各種相關分群法與 SOM 的比較評估。因此特別整理提供一些相關研究作為後續研究者之參考。

近年來關於分群法的比較研究中，Halkidi, Batistakis & Vazirgiannis (2001)曾作過綜合性調查與分類，並進行實證研究與評估。而 Everitt, Landau & Leese (2001)所編著的「Cluster Analysis」中有針對傳統分群法進行整合性評估。其他傳統分群法的比較性研究如【Peña, Lozano & Larrañaga, 1999】,【Hruschka & Natter, 1999】,【Maulik & Bandyopadhyay, 2002】,【Salazar G. et al., 2002】,【Park & Suresh, 2003】，主要以 k-means 為主。Kiang 從群集分析的角度出發，評估 SOM 與傳統分群法在群組問題及兩個機器學習範例之效能表現【Kiang, 2001】。

Mangiameli, Chen & West (1996)提出當今階層式分群法之主要缺點在於當實際資料分布與理想群集分布呈現緊實且分離良好的情形背離時，將會造成分類誤差的產生。該研究採用 252 筆資料集合，並搭配不同程度的資料缺陷，如資料散布(dispersion)、極端值(outlier)、無

關變數(irrelevant variables)與非均勻(nonuniform)的群集密度分佈等，將 SOM 與單一連結、完全連結、平均連結、中心法、華德法、兩階段密度法與 kth 鄰舍法等七種不同的階層式分群法進行評估比較。各項實證結果顯示，其中四分之三的資料集合中，SOM 均達到最高的準確性與穩健性，並且可以根據輸入資料動態調整分群結果。該研究並另外在高度分散的資料下，針對學習速率 α 進行 SOM 網路學習之敏感性分析。結果顯示，當學習速率 $\alpha$ 介於 0.05~0.5 之間，SOM 的分群結果並無顯著差異，且所有實驗的分群結果皆具有可重複性，也就是說權重向量都透過自組織學習而自動緊密地趨近群心。這代表 SOM 相當適用於雜亂的資料並有效改善分群決策的品質，並且應用在市場區隔、信用卡分析問題上。

表 4.6 為 Vesento & Alhoniemi(2000)提出兩階段法，利用三個不同的資料集合之執行效能比較情形。尤表中可看出兩階段最大的特點就是節省大量的計算資源。

表 4.6 不同分群方法的執行效能比較表【Vesento & Alhoniemi, 2000】

| Time Saving Comparison | | Clustering Method | | | | |
|---|---|---|---|---|---|---|
| | | S.L. | C.L. | A.L. | K-means | |
| Tradition SOM | I | 27 | 27 | 114 | 84 | min |
| | II | 2.3 | 2.3 | 9.4 | 32 | |
| | III | 192 | 192 | 780 | 540 | |
| Two-Phase | SOM of I (45s) | 1.9 | 1.9 | 5.5 | 402 | sec |
| | SOM of II (20s) | 1.2 | 1.2 | 3.1 | 228 | |
| | SOM of III (570s) | 11 | 11 | 19 | 600 | |
| Time Saving | I | 97.10% | 97.10% | 99.26% | 91.13% | % |
| | II | 84.64% | 84.64% | 95.90% | 87.08% | |
| | III | 94.96% | 94.96% | 98.76% | 98.21% | |

由上述研究結果而言，SOM 整體說來確實為一有效且便於觀察的分群工具，特別是針對大量高維度資料分析，且使用上不需先備知識。另外又因為對於初始值不甚敏感且分群結果皆具有可重複性之雙重優點，使得 SOM 比起所有的階層式分群法來得方便有效且實用。然而若就分群所呈現結果的使用性與解釋性而言，Rauber, Paralic, & Pampalk 將 SOM 與四種著名的非監督式分群法，包括階層式凝聚分群法(完全連結)、貝氏分群(Bayesian Clustering)、GHSOM 與 GTM(generative topographic mapping)，進行實證評估研究。SOM 可針對大量高維度資料進行視覺化，而關於一些更完整的群集結構資訊則可透過 GHSOM 獲得。而 GTM 為類似 SOM 的一種良好統計視覺化方法，AutoClass 雖無法提供視覺化，但卻可針對每一樣式產生獨立的機率密度函數。至於階層式凝聚分群法則是一種簡單直觀的分群工具，可協助使用者瞭解資料集合階層式架構，特別是與其它沒有此功能之分群法結合使用。

若從與模糊分群法比較的觀點切入，則 Guerrero-Bote et al.(2003), 則針對文件分群問題來評估 SOM、Fuzzy C-Mneas、Fuzzy ART & Fuzzy Max-Min 等分群法。根據該研究針對文件資料庫所進行的實驗結果顯示，SOM 獲得最佳的分群結果並且將群集拓樸化。而其他關於 SOM 的比較文獻，如 Giraudel & Lek (2001)將 SOM 應用至生態社群結構的排列與分類，相較於以往處理此問題所使用的主成份分析(Principal Component Analysis, PCA)與相似性分析(COrrespondence Analysis, CoA)等傳統統計方法，該研究實證結果顯示，SOM 不僅可以完全適用於生態學分析，而且是更加方便可供選擇的工具。其原因主要是透過 SOM 特有的拓樸保存性的映射機制，可將大量複雜的生態資料映射輸出到二維拓樸空間，以達到資料探索目的，並獲得其中隱含的社群次序(community ordination)。

基於前面各節對於 SOM 與軟性計算方法之綜合分群研究介紹，可

勾勒出基於資料探勘與群集分析目的，軟性計算領域中 SOM 綜合分群應用之研究外貌，整體而言，本章收錄研究文獻共計 132 篇，其中經由本研究方法所得者有 60 篇，另外再以文獻滾雪球法將其中具重要代表性之參考文獻共 72 篇摘出整理，以供索引參考。各子節詳細的統計數字如表 4.6 所示。而本研究進一步將各領域重要方法的文獻來源，以及本研究調查與 SOM 結合之綜合分群研究整理列出如表 4.7 所示。表中左側部分代表該領域之重要分群研究，大多由文獻滾雪球法蒐集而得；右側部分則代表與 SOM 結合之分群應用研究，由本研究方法獲得。

表 4.7 第四章各節分群研究文獻統計表

| 編號 | 章節名稱 | 本研究調查 | 文獻滾雪球法 | 小計 |
|---|---|---|---|---|
| 4.1 | 分群理論 | 13 | 22 | 35 |
| 4.2 | 傳統分群 | 6 | 24 | 30 |
| 4.3 | 模糊理論 | 14 | 17 | 31 |
| 4.4 | 遺傳演算法 | 11 | 8 | 19 |
| 4.5 | 其他分群 | 6 | 1 | 7 |
| 4.6 | 方法比較 | 10 | 0 | 10 |
| 小計 | | 60 | 72 | 132 |

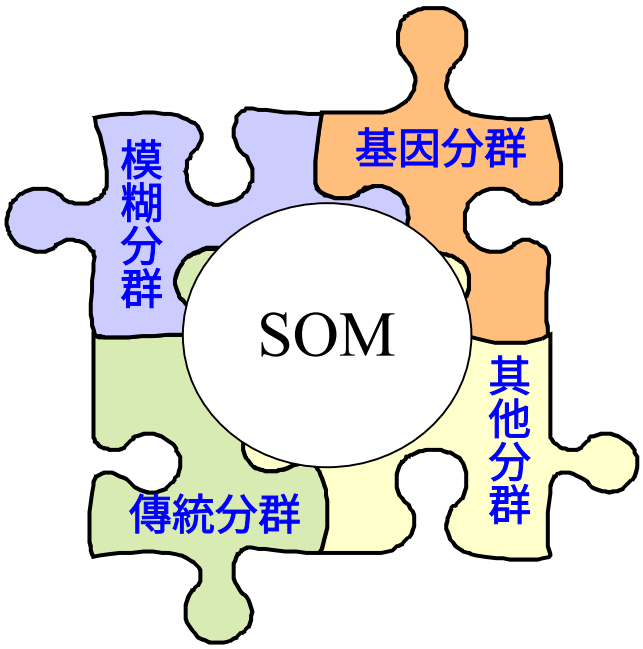


圖 4.4 本研究 SOM 分群整合應用示意圖

表 4.8 第四章分群重點文獻分類表

<table>
<tr><th rowspan="2">編號</th><th rowspan="2">章節名稱</th><th rowspan="2">分類</th><th>文獻滾雪球法</th><th>本研究調查</th></tr>
<tr><th>一般分群</th><th>與 SOM 結合</th></tr>
<tr><td rowspan="2">4.2</td><td rowspan="2">傳統分群</td><td>階層式</td><td>單一連結【Sneath & Sokal, 1973】<br>完全連結【King, 1967】<br>最小變異【Ward, 1963; Murtagh, 1984】</td><td>【Dittenbach, Merkl, & Rauber, 2000, 2001, 2002】<br>【Pampalk, Widmer, & Chan, 2004】<br>【Vesanto, 2000】</td></tr>
<tr><td>分割式</td><td>K-means【MacQueen, 1967】<br>Possibilistic C-means【Krishnapuram & Keller, 1993；1996】</td><td>【Mao & Jain, 1996】【胡承民，1997】【Vesanto, 2000】【鍾文杰，2001】<br>【Sugiyama & Kotani, 2002】</td></tr>
<tr><td rowspan="3">4.3</td><td rowspan="3">模糊理論</td><td rowspan="2">FSOM</td><td rowspan="3">目標函數概念【Dunn, 1974】<br>Fuzzy k-Means【Bezdek, 1974】<br>Fuzzy C-Means【Bezdek, 1981】<br>【Belacel et al., 2002】【Rousseeuw, 1995】【Baraldi, 1999a,b】【Cinque et al., 2004】<br>【蔣以仁, 1997】【呂奇傑, 2001】【譚嘉慧, 2000】【Kim, 2003】</td><td>【Zaremba, 2000】,【Pascual-Marqui et al., 2001】,【許維宸, 2001】,【周建興, 2003】,<br>【Wang et al., 2003】</td></tr>
<tr><td>【鄧博文，1998】【Venugopal, 1999】【郭人介、紀勝財，2000】,【Kuo, 2001】,<br>【Lee & Chen, 2001】【Pai, 2001】【陳德華, 2003】【Wu, 2003】</td></tr>
<tr><td>SOM-FNN</td><td>【趙志運，2000】【吳育奇，2000】【Tung, 2002】</td></tr>
<tr><td rowspan="2">4.4</td><td rowspan="2">遺傳演算法</td><td>基因分群</td><td>Genetic Algorithm【Holland, 1975；Goldberg, 1989】<br>Evolutionary Strategy【Schwefel, 1981】<br>Evolutionary Programming【Fogel et al., 1965】</td><td>【McInerney & Dhawan, 1993, 1994】【蔡坤洋, 1998】【Nissinen & Hyötyniemi 1998a, 1998b】【Chang & Heh, 1998】【Kim, Ahn, & Kang, 2000】【簡順源，2001】【鍾文杰，2001】【Villmann et al., 2004】</td></tr>
<tr><td>圖論</td><td>【Jain, 1999】【Maulik & Bandyopadhyay, 2000】【Tseng & Yang , 2001】【Chiou & Lan, 2001】【Garai & Chaudhuri , 2004】</td><td>【游鴻志，1997】【Chang & Heh, 1999】</td></tr>
<tr><td>4.5</td><td>其他分群</td><td>無</td><td colspan="2">DTRN【Si, Lin, & Vuong (2000)】, ICA【Karhunen et al., 2000】, 相似度【Wong et al, 2001】<br>SOM + ART【Hussin & Kamel, 2003】, Evolving SOM【Deng & Kasabov, 2003】, CLUSTER【Bandyopadhyay, 2004】</td></tr>
<tr><td>4.6</td><td>方法比較</td><td>無</td><td>【Erika Johana Salazar G. et al., 2002】【Peña et al., 1999】【Hruschka & Natter, 1999】【Halkidi et al., 2001】【Maulik & Bandyopadhyay, 2002】【Park & Suresh,</td><td>【Giraudel & Lek, 2001】【Mangiameli et al., 1996】【Guerrero-Bote et al., 2003】<br>【Rauber, et al., 2000】</td></tr>
</table>

| | | | | |
|---|---|---|---|---|
| | | | 2003】 | |