第一章 緒論

1.1 研究背景與動機

集群分析或群聚分析(cluster analysis)與分群或聚類(clustering)為傳統統計學中由來已久的資料分析技術，主要探討如何將眾多個案縮減成少數具有代表性的群別，以達到了解集群結構與進行特定分配的資料分析與推論之主要目的【周文賢，1998】。集群分析同時在不同學科領域中受到廣泛地研究討論，並發展出許多成熟具效能的演算法。

然而因應知識經濟與網路時代的來臨及快速變遷, 傳統產業之資訊更新率平均為五年，而高科技產業更只有十八個月。有效地將龐大資料歸納分析出有用的資訊，不僅可作為管理者的決策參考，更能在競爭激烈的微利時代，產生企業營運管理不可或缺的企業智能(business intelligence)。因此，資料探勘(或資料挖礦)(data mining)理論與應用技術儼然同為九 O 年代研究學者與實務經營者的關切議題。

Data Mining 利用資料來建立一些模擬真實世界的模式，利用這些模式來描述資料中的特徵以及關係，以提供作為決策所需之資訊。隨著資訊科技的快速進展，許多新的電腦分析工具相繼問世，諸如關聯式資料庫、模糊理論、基因演算法以及類神經網路等，使得看似無關的雜亂資料得以經由一種系統性且可實行的程序進而發掘其中所隱含有價值的知識寶藏。一般而言，Data Mining 的理論技術可分為傳統技術與改良技術兩支。傳統技術以統計分析為代表，舉凡統計學內所含之敘述統計、機率論、迴歸分析、類別資料分析等皆屬之，尤其 Data Mining 的對象常為多變數且大規模的資料，因此需要多變量分析中用來精簡變數的因素分析(factor analysis)、用來分類的鑑別分析(discriminant analysis)以及用來區隔群體的群集分析等分析技術。另外在改良技術方面，應用較普遍的有決策樹理論(decision trees)、類神經網路(neural networks)以及規則歸納法(rules induction)等。運用上述各式理論技術，Data Mining 可建立六種模式：分類(classification)、迴歸

(regression)、時間序列(time series)、聚類(clustering)、關連(association)與次序(sequence)。分類、迴歸與時間序列模式主要是預測應用，而關連與次序模式則是用來描述行為(例如消費行為)，而分群模式則是兩者都適用。

分類是先將變數進行數值計算，再依其結果分派歸類，或針對已分類的資料來研究它們的特徵並建立模型後，再根據這些特徵對其他未經分類或是新資料進行預測。分類通常會牽涉到兩種統計方法：邏輯迴歸(logistic regression)與鑑別分析。然而隨著 Data Mining 逐漸普遍，因此也常採用類神經網路與決策樹等方法進行分類。

然而，為達到商業快速回應(quick response)之目的，因此最好能夠在大型資料庫中將多維度資料按其相似度先行聚類，以作為後續分類工作的前處理，如此不僅能解決資料特徵值不明顯所產生的問題，更將有助於提升整體資料探勘或知識發掘的效率與精確度。所謂「聚類」或稱「分群」是將資料分為數群，其目的是要發掘群間差異與群內相似性。聚類與分類不同的是，不曉得會以何種方式或根據什麼特徵來聚類，這也就是所謂的非監督式學習(unsupervised learning)。因此必須要有一套良好機制來解讀分群的意義。但也正因如此，一個優良的分群演算法將可適度地忽略嚴謹的統計假設前提，而大幅提昇資料的分析效率。

在眾多分群理論中，相對於傳統統計多變量分析中的群集分析，近十年來「類神經網路」當中的「自組織映射圖神經網路」相當受到關注。類神經網路是人工智慧學門中屬於較新的「計算智慧」，有別於傳統人工智慧的「符號智慧」，計算智慧是以資料為基礎，通過訓練建立聯繫，進行問題求解，並具有如人腦的學習容錯功能，主要研究以人工的方法和技術，模仿、延伸和擴展人的智慧，透過機器學習實現機器智慧。人工神經網路、遺傳演算法、模糊系統、進化程式設計、人工生命等都可以包括在計算智慧當中。換言之，類神經網路是一種模

擬人腦思考結構的資料分析模式，由輸入變數之數值中自我學習，並根據學習經驗所得知識，不斷調整參數以期建構資料的樣式(patterns)。類神經網路為非線性設計，與傳統迴歸分析相比，好處是在進行分析時無須限定模式，特別當資料變數間存有交互效應時可自動偵測出；缺點則在於其分析過程為一黑盒子，故常無法以可讀之模型格式展現，每階段的加權與轉換亦不明確，是故類神經網路多利用於資料屬於高度非線性且帶有相當程度的變數交感效應時。

「自組織映射圖神經網路」(Self-Organizing Map Neural Network，以下簡稱 SOM)是由芬蘭赫爾辛基科技大學的 Teuvo Kohonen 教授於 1982 年提出【Kohonen, 1982】，且為最具代表性的非監督式類神經網路模式。該模式最主要的特色就是將事先無定義的輸入資料集合按其相似性進行網路訓練，最後將輸出資料映射到二維網路拓樸的格點上，除了實踐資料視覺化外，更能藉由分析者判斷網路拓樸上的資料分佈狀態來進行資料分群。SOM 演算原理簡單且跳脫傳統統計分群較為嚴謹的資料假設條件，因此至今仍受到各領域學門研究學者與實務者廣泛的研究與應用。

有鑑於 SOM 為最具代表性的非監督式類神經網路，且該模式不僅可將複雜的高維資料表示成二維圖像，達到簡化及區分資料的目的；更成為解決分群問題中最受歡迎的競爭式學習法之一。芬蘭赫爾辛基科技大學的類神經網路研究中心(Neural Networks Research Center, NNRC)以其累積的豐富經驗所彙整的文獻資料庫，詳實收錄歷年來 SOM 相關文獻五千餘篇，並整理刊登於國際期刊中【Kaski et al., 1998; Oja et al., 2002】；甚至提出整合數百萬筆網頁的資料自組織應用(WebSOM)，並發展諸如生化資訊、學習矩陣、語音辨識與自然語言建模等以 SOM 為基礎的多元化研究路線。儘管如此，就方法論而言，仍欠缺諸多改善模式之整理歸納；其次對於整合性的分群應用也無多加著墨。

反觀國內與 SOM 相關的研究迄今雖有五十多篇，但不僅尚無以分群方法為出發的回顧性研究，特別是與傳統分群法(hard clustering)或軟性計算技術(Soft-computing)之分群整合與應用。綜合考量研究時間限制、資料蒐集便利性與類神經網路所帶動 SOM 研究風氣等因素，本研究遂透過當今著名資料庫的搜尋引擎，並以條件式的資料檢索策略，回顧自 1990 迄今 SOM 相關的國外學術期刊，探討 SOM 模式改善與分群應用之相關研究，從演算原理探討到與其它分群法的結合應用作一歷史脈絡之整理分析，希冀能提供後續有興趣的研究者對於 SOM 模式改善或分群應用一個知識入口與資源平台，亦可將本研究當作一 SOM 入門的綜合參考目錄。

1.2 研究目的

基於上述背景與動機，本研究將朝向以 SOM 為研究主體的整合型回顧研究進行，並將研究核心問題確立為「如何改善 SOM 模式以提升資料分群效能」，進一步分解為兩大子題「SOM 模式改善的研究調查」與「SOM 的分群應用研究」。透過條件式的文獻檢索策略進行資料的蒐集、篩選、整理歸納與結果分析，預計將完成以下的分析工作：

1. 了解 SOM 模式改善的歷史脈絡，包括研究方法、最新研究方向及趨勢等，並利用 SOM 演算法的重點特徵構面進行蒐集文獻之分析論述。
2. 了解 SOM 與其它分群法之綜合分群應用，包括與傳統分群法、模糊集理論、遺傳演算法結合的分群綜合應用之歸納整理與比較分析。
3. 建立一綜合介紹性之參考書目，因為文獻檢索策略是以上述兩大研究核心問題為中心而擬定，並且針對五個國際性電子資料庫作為檢索途徑，因此檢索結果將廣泛地包括 SOM 改善模式

與分群綜合應用之相關研究，換句話說，相當適合作為後續 SOM 或分群相關研究者參考。

4. 提供一 SOM 資源整合入口，包含有學術期刊、國際研討會、書籍、國內外研究機構與開發工具等介紹。

1.3 研究方法及流程

若以研究性質而言，本研究屬於初探性研究；而就研究內容而言，則屬於整合型研究回顧，所探討的包括以下兩大問題：

1. SOM 相關的研究脈絡與現況，即「SOM 方法改善調查」。
2. SOM 與其他分群法之結合應用，即「SOM 分群整合應用」。

根據上述兩大問題，本研究引用 Cooper 對整合型回顧研究所提出之五階段研究模型【Cooper, 1982】作為主要架構與流程，並結合「主題層面配對檢索策略」【Harter, 1986】與本研究提出的「檢索策略矩陣」作為資料蒐集與文獻檢索的主要研究方法。而所得到的結果亦屬於階段性，仍有賴後續進一步針對各改善模式或綜合分群方法進行特定資料的分群測試與評估，以更清楚了解 SOM 分群特性與效能，進而發展出可適應資料特性且創新具效能的改善模式。

Cooper 之五階段研究模型包括有(1)問題陳述；(2)資料蒐集；(3)資料評估；(4)分析和解釋；以及(5)結果的發表。其中資料蒐集階段採用「主題層面配對檢索策略」，透過三大主題層面的界定來定義研究範圍，並作為文獻檢索之關鍵詞；此外並結合本研究提出的「檢索策略矩陣」，其中每項策略均包含檢索主題、途徑與符合其研究目的之篩選條件組。換句話說，按照此檢索矩陣的文獻搜尋法進行系統化的文獻檢索，再將所得結果根據所屬策略之篩選條件組進行篩選後，最後進行構面分析，以釐清 SOM 演算原理特性。

本研究的研究流程，如圖 1-1 所示。

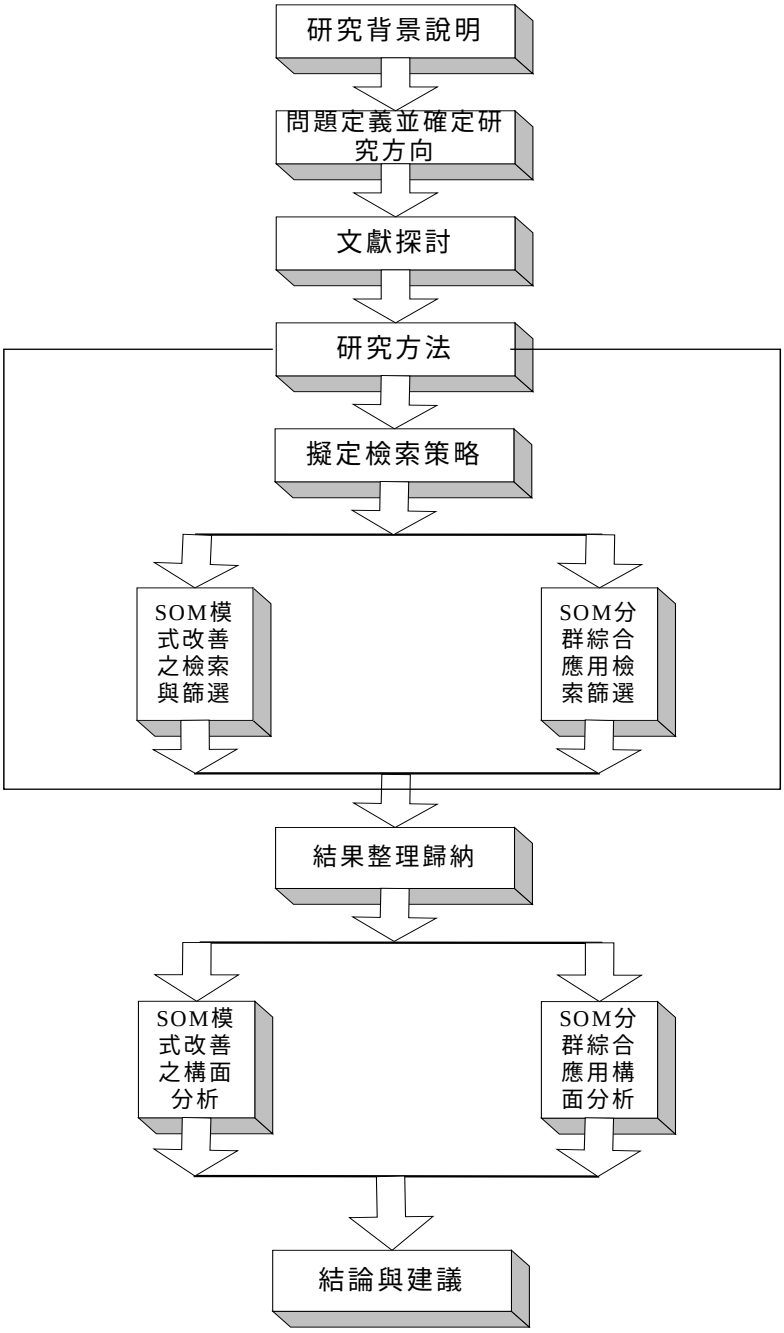
研究背景說明
問題定義並確定研究方向
文獻探討
研究方法
擬定檢索策略
SOM模式改善之檢索與篩選
SOM分群綜合應用檢索篩選
結果整理歸納
SOM模式改善之構面分析
SOM分群綜合應用構面分析
結論與建議

圖 1.1 研究流程圖

1.4 研究範圍與限制

有鑑於 SOM 特有的自組織特性以及對高維資料的視覺化能力，使得相關的應用研究在多個學門領域如人工智慧、機器學習、樣式偵測/擷取與資料探勘等均受到廣泛探討與研究，特別是在計算科學中用以探討分群問題。

另外，針對本研究法獲得的目標文獻所進行之構面分析，主要考量研究屬性為調查性質，且目標文獻之內容涵蓋甚廣。因此採用文字論述方式進行研究主題之探討。

然而本研究為達到研究目的，並避免研究失焦以及考量有限的時間與資源等因素，遂將研究範圍分別針對以下的構面加以定義與限制。其中若是檢索所得之文獻無法滿足允收條件時亦不納入研究範圍，以維持研究主題之整體精神。

1. 時間

SOM 自 1982 年提出以來已在許多學門以及各界學者造成相當廣泛的後續討論與多元化的應用，本研究考量整個理論發展歷程中的重要性與代表性以及文獻資源蒐集的可行性，擬將時程設定自 1990 到 2004，以利研究工作之進行。

2. 研究主題

本回顧研究將核心問題為「如何改善 SOM 模式以提升資料分群效能？」，進一步分解為「SOM 模式改善的研究調查」與「SOM 與分群應用研究」兩大子題，並作為本研究主要研究主題。正因如此，本研究根據資料檢索技術中的「主題層面配對檢索策略」設定「SOM」、「Learning Algorithm」與「Clustering」等三大主題層面，並取 SOM 與其他兩者之交集後再取其聯集作為研究範圍。

換句話說，本研究假設論文關鍵字若同時包含「SOM」與「Learning Algorithm」者即代表該研究為探討 SOM 演算法之相關論文；同理，若關鍵字同時包含「SOM」與「Clustering」者，則

代表該論文為探討 SOM 在分群上的相關研究。關於主題層面配對檢索策略之進一步介紹請參閱第 2.2.2 節。

3. 檢索途徑

本研究於資料蒐集階段所採用的「主題層面配對檢索策略」提供了文獻檢索所需的關鍵詞，至於實際檢索對象與途徑，本研究選擇資料庫及期刊兩大方向進行文獻的檢索與蒐集，並提出「檢索策略矩陣」來定義研究主題與檢索途徑之間的關係，詳細說明請參見第 2.2.2 節。

在資料庫方面，本研究採用 SDOS、IEL、EBSCO、SCIE 與 ACM 等五個當今著名且與本研究相關的國際性資料庫，其中 SDOS、IEL 與 ACM 為全文電子資料庫，而 EBSCO 與 ACM 為部分全文，SCIE 則是文獻目錄資料庫。

在學術期刊方面，本研究遂以 INNS(International Neural Networks Society)的 Neural Networks 與 IEEE transactions on Neural Networks 兩種與 SOM 最為相關的學術期刊作為搜尋中心起點，除了以主題層面作為檢索關鍵詞外，並結合資訊檢索技術中「引用文獻滾雪球法」，針對符合的文獻進行參考文獻延伸查詢，若所引用之文獻符合篩選條件便予以收錄至檢索結果中。

另外，本研究假設所有高品質具代表性的文獻最終都會刊登於正式學術期刊(journal papers)中，因此將「SOM 模式改善的研究調查」的文獻蒐集範圍鎖定於此（本論文第三章)。至於「SOM 與分群應用研究」將需要廣泛涉獵最新結合應用之研究，因此將範圍放大至各種國際研討會(conference proceedings)或是博士論文與技術報告(本論文第四章)。

4. 篩選條件組

「檢索策略矩陣」除了上述以研究主題與檢索途徑來定義研究範圍之外，還包括「篩選條件組」針對初步檢索所得之文獻加以

進一步篩選。在此根據本研究兩大研究主題：「SOM 模式改善的研究調查」與「SOM 與分群應用研究」，分別設計兩組篩選條件來過濾文獻，詳細說明請參見第 2.2.2 節。

1.5 各章節介紹

本研究共分為五章，各章節內容概述如下：

第一章為緒論，主要包括研究背景與動機，研究目的，研究方法，研究範圍與限制以及研究流程等部分。

第二章主要是將研究回顧作一概述，並介紹其扮演的角色與重要性，接著將過去相關的回顧性研究整理歸納其回顧方法，最後提出本研究之研究回顧法，其中包含兩大文獻檢索策略與相對應之篩選條件組，並解釋如何有效協助相關文獻的回顧與蒐集，以達成研究目的。

第三章共分為三大部分。首先從 SOM 基本原理介紹作為開端，並說明其演算法特性與優缺點，並闡述在資料探勘中所扮演的角色；第二部分對 SOM 的發展沿革作一簡要綜覽，並介紹相關重要期刊、國際研討會、Kohonen 所屬芬蘭赫爾辛基大學的類神經網路研究中心與相關最新消息等。第三部份則是將檢索策略一「SOM 模式改善的研究調查」所得之文獻經篩選分類後進行構面分析, 而構面的選擇是依據 SOM 演算法之特色來決定。

第四章則是先針對分群理論加以簡介，包括分群定義、分類、階段與評估以及 SOM 與其它分群法之比較，再將檢索策略二「SOM 與分群應用研究」所得之文獻經篩選分類後，依據不同的演算法結合對象分別介紹其結合模式之特色與應用。

第五章為結論及建議，將本研究所獲得的結果與貢獻加以整理歸納，並建議後續可能之研究方向。