第三章 SOM 變型模式與比較分析

大自然中不論是生命或非生命系統都普遍存在「自我組織」(self-organization)的共同現象，然而這類現象雖容易辨別但卻不易定義。Kohonen 提出的 SOM，其根源來自哺乳類動物大腦中所發現的同質性結構，結合知覺中樞處理資料的自組織特性加以簡化而來【Allinson, Yin, & Obermayer, 2002】。根據生理學與解剖學之證據顯示，通常在許多生物的腦部組織中，大量神經元之間會存在側向連結，並形成二維的層狀結構，而側向連結的強度與神經元間的距離有關【陳慶翰，2002】。而這裡所指的自組織，換句話說，也就是非監督式學習(unsupervised learning)的降维現象(dimension reduction)。

在神經計算科學的專有名詞定義中，自組織是一種經由資料驅動(data-driven)而針對神經突觸強度(synaptic strength)進行調適修正(adaptive adjustment)，這不僅構成樣式辨認(pattern recognition)中動態修正(dynamic adjustment)的基礎，同時也成為高階資料的表現架構。

本章將從 SOM 基本原理與歷史沿革作為開端，前者還包括演算法特性與優缺點之闡述，以及在資料探勘中所扮演的角色；而後者是以時間為主軸進一步介紹 SOM 發展沿革、相關期刊與國際研討會、T. Kohonen 教授所屬芬蘭赫爾辛基科技大學的類神經網路研究中心與近期發展等資訊。

在具備 SOM 基礎知識後，本章將重心集中探討 SOM 相關之改善模式，其作法是利用先前檢索策略所得(121 篇)，再經條件篩選而得目標文獻(31 篇)，並分別以數個 SOM 特徵構面進行闡述。

3.1 SOM 概述

在群聚分析(或集群分析)中，自組織映射圖神經網路為 1982 年由 T. Kohonen 所提出最具代表性的非監督式類神經網路模式。從提出【Kohonen, 1982】到完整描述【Kohonen, 1989, 1995】，SOM 基本上不僅可處理二元值，更可處理連續值之輸入資料，其主要特色是可將

分佈型態尚未明確定義的高維度資料，透過從問題域中取得適當的特徵變數，並經由網路訓練便可學習內在的聚類規則，以進一步應用於新範例。另外，透過將高維度的輸入資料 x, $x \in \Re^n$ 加以視覺化，並在通常是二維的輸出平面產生映射點，所獲得的網路拓樸關係可忠實地保留輸入資料之間的非線性關係。

SOM 的成功應用範圍由財務資料分析，經醫療資料分析到時間序列預測，工程控制以及許多其他方面等【Dittenbach, Rauber & Merkl, 2002】，使其成為兼具穩定性與可塑性的模型架構。此外，有兩篇近期的參考目錄詳細歸納出 3,000 篇以上 SOM 相關研究與應用【Kaski, Kangas & Kohonen, 1998】、【Oja, Kaski & Kohonen, 2002】。

3.1.1 基本原理

SOM 屬於競爭式學習的啟發式演算法，基本精神是在未經標示的樣本群尋找某些相似的特徵、規則或關係，然後再將具有共同特徵的樣本聚集成群(類)。競爭式學習有時又稱為「贏者全拿」(winner-take-all)學習法。正如名稱所述，當輸入資料產生時，網路上的神經元會彼此競爭來爭取被活化的機會。最後在網路輸出層僅有一個神經元會被激發成活化狀態(active state)，即輸出為 1；而其他神經元便會被抑制成休止狀態(inactive state)，即輸出為 0；當競爭完成後，只有獲勝神經元才會進行調整(學習)，其他神經元則保持不變。至於 SOM 與一般競爭式學習不同點是競爭方式採用「有福同享」，也就是不只獲勝的神經元有資格學習，連鄰近區域內的神經元亦可利用墨西哥帽(斗笠帽)函數作為側向抑制(lateral inhibition)作用函數進行學習。

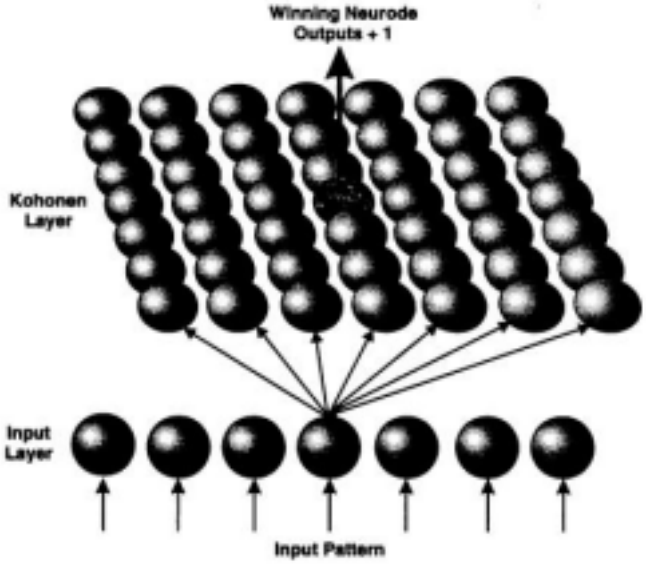


圖 3.1(a) SOM 網路拓樸架構圖【陳慶翰，2002】

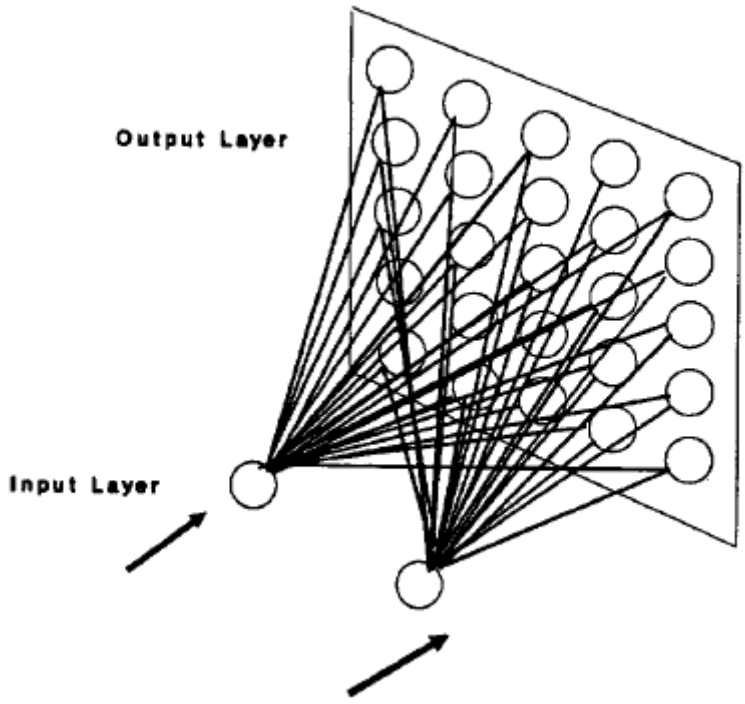


圖 3.1(b) SOM 網路拓樸架構圖【Nour & Madey, 1996】

在自組織特徵學習過程中，輸出層的神經元是以矩陣方式排列於一維或二維的空間中，並且根據目前的輸入向量彼此競爭以爭取到調整鏈結值向量的機會，最後輸出層的神經元會根據輸入向量的「特徵」以有意義且具次序性的「拓樸結構」(topological structure)展現在輸出

空間中。而這種陣列的拓樸關係也就是神經元之間的鄰近關係(neighboring relationship)。由於所產生的拓樸結構圖可以反應出輸入向量本身的特徵，因而被稱為「自組織特徵映射網路」(self-organizing feature map)。

SOM 學習演算法相當簡單，這也是其受到各學科領域高度研究與應用之主要原因。整體而言，演算法可分為三大部分，首先定義由神經元所構成的一維或二維矩陣；其次尋找優勝神經元，其輸出值可反映出對目前輸入最有反應的神經元；最後是調整優勝神經元及其鄰近區域內神經元的鏈結值(權重向量)，使其更接近輸入向量。以下將 SOM 學習演算法步驟分述如下：

【步驟一】：初始化

在進行 SOM 網路訓練前必須針對網路結構與權重向量進行初始化。在網路架構定義方面，基本上 SOM 的網路拓樸是由 i 個單元的集合所構成，並根據事先定義的固定拓樸型態來排列，最常使用的就是 $N*M$ 的二維網格。其次，每一個單元 i 都會被賦予與輸入資料相同維度的權重向量 m_i，$m_i \in \Re^n$，而權重向量(鏈結強度)可能有兩種方式決定，一是以隨機亂數指定，或以隨機策略如主成份分析(principle component analysis, PCA)來設定。值得注意的是，所有 i 個權重向量的初始值均應相異，且通常會加以正規化(normalize)成為長度為 1 的單位向量。

【步驟二】：輸入範例特徵向量

針對時間/訓練循環，輸入向量 $\underline{x} = (x_1, x_2, \cdots, x_i)^T$，在此使用了一離散時間標記 t，代表目前訓練重複次數。在每次訓練循環，$x(t)$ 均由輸入向量集合 $\Re^n$ 中隨機抽出。

【步驟三】：尋找優勝神經元

以最小歐基里德距離的方式尋找時間/訓練循環 t 之優勝單元 c (winner)，顯示其含有最高活動力。而對於隨機選擇的輸入向量 $x(t)$ 而言，$c(t)$ 則將顯出更高的活動力。換句話說，$c(t)$ 在特定輸入向量的未來表現皆會因為呈現較高的活動力而較為合適成為優勝單元。

一般而言，一個單元的活動力（作用值）是根據輸入向量(input pattern)與該單元的權重向量(鏈結強度)之間的歐幾里得距離(Euclidian distance)來衡量。換句話說，若有一單元其權重向量m_i與目前輸入向量 x 之間的歐幾里德距離為最小時，該單元便稱為優勝單元。因此，優勝單元c的選擇方式可用下列算式(3-1)加以表示：

$$c(t) = \| x(t) - m_c(t) \| = \min\{ \| x(t) - m_i(t) \| \}. \qquad (3\text{-}1)$$

$c(t)$　時間點t的優勝單元

$x(t)$　時間點t隨機抽取的輸入向量

$m_i(t)$ 時間點t，i單元的權重向量

$m_c(t)$ 時間點t，優勝單元c的權重向量

$\|x(t) - m_c(t)\|$：在時間點t,優勝單元c的活性(activity),用來計算Δm_i(鏈結強度修正量)

【步驟四】：調整權重向量

適應行為(adaptation)發生在每次學習重複過程中，其執行方式是根據各別輸入向量與權重向量之差異量，朝向坡降（gradual reduction）方式進行學習。而至於適應調整的總次數是由學習速率α主導，其亦會隨著時間過程而逐漸遞減。這種適應(調整)強度的漸減特性使得在學習初期有大量的適應步驟，並且權重向量亦需從隨機初始值逐漸調整轉向至輸入向量的實際需求(群心)。而在學習末期，微幅的適應讓權重向量在輸入空間宛如進行微調動作(如圖 3.2 所示)。

若將標準競爭學習(winner takes all)加以延伸應用，針對隨時間變動且處於優勝單元周圍坡降鄰近區域的單元而言也同樣加以調整。實際上，在 SOM 學習步驟過程中，一組位於優勝單元周圍的單元將會漸漸修正為目前所呈現的輸入樣式。這使得輸入向量能夠有一空間安排，也就是讓相似的輸入向量得以在輸出單元的網格中映射在相互鄰近的區域。於是 SOM 學習過程將產生輸入向量的網路拓樸次序。

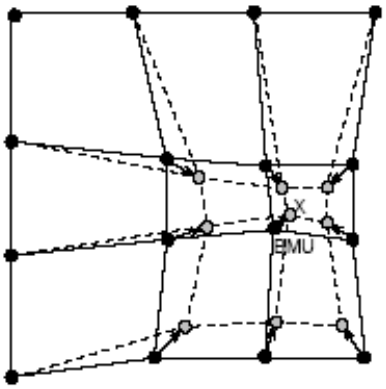


圖 3.2 優勝單元與鄰近區域內權重修正示意圖【Vesanto, 2000】

對於位在優勝單元周圍鄰近區域的單元，就輸出空間而言，可用與鄰近中心的距離h_{ci}，也就是該單元i與該次學習循環的優勝單元c之間距離，加以間接地表示。換句話說，我們以該次學習循環的優勝單元c當作是輸出空間中的鄰近中心，並指派由 0 到 1 的調整量，以確保距離優勝單元愈近者，其適應調整程度也就愈大。

在結合上述 SOM 的原理原則後，我們可以寫出一如公式(3-2)的學習法則來表示鄰近區域內單元之權重修正。在此使用離散時間標示 t 以代表目前學習循環次數。

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t)\left[x(t) - m_i(t) \right] \qquad (3\text{-}2)$$

t　表示目前學習循環次數

α　則代表隨時間變化的學習速率

h_{ci} 則表示隨時間變化的鄰近函數值

x　表示目前的輸入樣式(向量)

m_i 則代表分配給單元i的權重向量

圖 3.3 的簡單圖形可呈現 SOM 的結構與學習流程。在該圖中其輸出空間為 6*6 共 36 個單元的網格所構成。然後一個隨機選取的輸入向量$x(t)$映射至輸出單元的網路格點上。在下一階段的學習過程中，顯示出最高活性(activation)的優勝單元c則被選擇。圖 3.3 中標繪成黑色的單元視為優勝單元，而其權重向量$m_c(t)$朝向目前的輸入向量$x(t)$移動，這樣的移動表現在圖 3.3 左側的輸入空間中。由修正

(adaptation)的結果看出，單元c在下一次的學習循環(t = t+1)就輸入樣式而言，將會產生更高的活性(作用值)，因為單元的權重向量$m_i(t+1)$在輸入空間中更加接近輸入樣式x。除優勝單元外，對於鄰近單元亦修正其權重向量，受到修正調整的單元由圖中輸出空間標繪有陰影的單元可看出，其陰影的深淺即對應於該單元的權重修正量，亦可代表其鄰近函數的空間寬度。一般而言，在優勝單元的鄰近區域中愈靠近中心者所受到的修正愈強，這可由圖中單元標繪的深淺顏色說明。

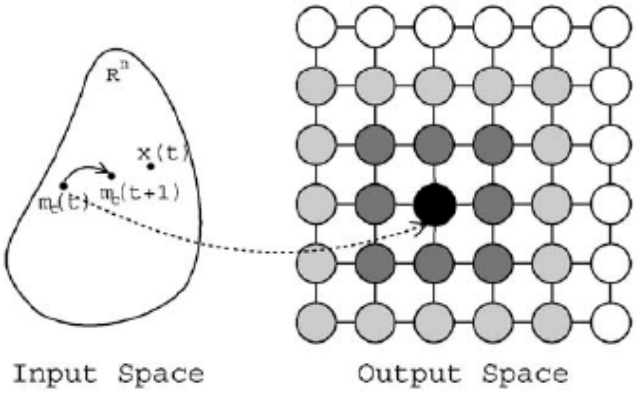


圖 3.3 由 SOM 輸入空間與輸出空間觀察權重修正情形

【Dittenbach, Rauber, & Merkl, 2002】

【步驟五】：返回步驟二，直到特徵映射圖形成後終止。

【參數設定】

1. 學習速率參數$\alpha(t)$：本參數為用來調整權重向量且應隨著時間而調整，一般而言將隨時間逐漸變小。至於參數遞減的形式可以是線性遞減、指數遞減或是與時間成反比等，此時的學習循環可以視為演算法的「排列階段」(ordering phase)。之後學習循環的主要目的在於進行特徵映射圖的細部調整，因此可視為演算法的「收斂階段」(convergence phase)，此時學習循環的參數值應保持相當小的數值。
2. 鄰近區域參數$h_{ci}(t)$：

   鄰近區域函數通常定義包圍著優勝神經元c的正方形區域，其形式可以是矩形、六邊形或八邊形等(如圖 3.4 所示) 。不管是何者形式都應遵守一開始先包含全部或較大範圍的神經

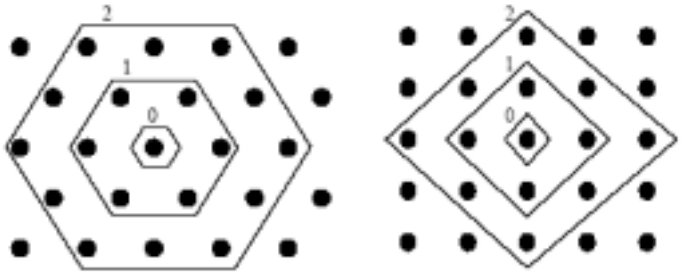
元，然後隨時間增加而慢慢縮減鄰近區域的大小。舉例來說，在「排列階段」可以隨時間而縮減至較小範圍；而在「收斂階段」則應僅包含一個或兩個神經元，甚至只針對優勝神經元進行權重向量的調整。

高斯函數(Gaussian function)可以用來定義鄰近區域與核心之關係，及鄰近區域函數$h_{ci}(t)$，如公式（3-3）所示：

$$h_{ci}(t) = \exp(-\frac{\|r_c - r_i\|^2}{2 \cdot \delta(t)^2}) \qquad (3\text{-}3)$$

其中$\|r_c - r_i\|$代表在輸出空間中單元c與單元i之間的距離(也就是鏈結強度的向量差)；換句話說，r_i表示在輸出網格內的指向單元i的二維向量(也就是鏈結強度)，δ為隨著時間變化的鄰近半徑縮小因子。

常見的實例就是在學習過程初期，通常會選擇較大的鄰近區域-中心，以便涵蓋較大的輸出空間範圍。然而，鄰近區域-中心的寬度會隨著學習過程逐漸縮小，到學習末期只剩下優勝單元本身進行適應調整。這樣的縮減是由一個隨著時間變化的鄰近半徑縮小因子δ來表現，如公式（3-3）所示。而如此策略將使學習初期可形成大型的聚落（集群），而在學習末期則使輸入向量具有良好的鑑別力。



(a)六角形鄰近區域　　(b)矩形鄰近區域

圖 3.4 常見鄰近區域類型【Vesanto, 2000】

3.1.2 演算法特性與優缺點

自組織映射圖神經網路模式(SOM)【Kohonn, 1982; Kohonen, 1990; Kohonen, 1995; Kohonen et al., 1996】是目前最受歡迎的非監督式類神經網路模式，其應用範圍相當廣泛並橫跨許多領域，原本主要用來處理工程問題，但日漸普遍應用在資料分析上。Kokonen 亦曾以專書針對其理論基礎提出詳盡論述，並介紹廣泛且多元化的應用領域(Kohonen, 1995, 1997, 2001)。簡言之，SOM 具備以下兩大特色：【Vesanto, 2000】

A. 向量量化(vector quantization):將訓練資料構成之輸入空間加以量化。

B. 向量映射(vector projection)：同時執行具拓樸保留特性之映射，將訓練結果呈現在一規則低維度的網格上。

在向量量化方面，SOM 將原本擁有 N 個訓練樣本的原始資料集合縮減成 M 個仍具代表性的原型(prototype)。後續進一步如分群(clustering)與視覺化(visualization)便將特徵向量(prototype vectors)取代原有的資料集合而作為分析對象。

另外在向量映射方面，雖然 SOM 的特徵向量在低維度的拓樸格點已有定義明確位置，而有助於優勝單元產生時輸入資料映射之定義。然而這種映射仍相當粗造，一方面異質性高的特徵向量有可能會映射到同一神經元，而且映射的資料來源僅能夠由事先決定形狀大小的 SOM 網格單元獲得而非完整的原始資料，有別於其他映射方法如 Sammon’s mapping 擁有連續性的輸出結果。儘管如此，SOM 仍具有一項重要優點，就是單元的拓樸次序(topological ordering)主要乃依據網格上定義的局部鄰近區域。因為當資料密度高時會有較多數量的圖形單元，而鄰近區域便如輸入空間所量測會有變小情形，如此說明 SOM 映射有朝向局部的資料密度調整之特性。

SOM 特別適合應用於資料理解(data understanding)，但在資料預備

(data preparation)與建模(modeling)同樣也是一項健全的工具。SOM 提供一便利的工作平台，不僅可協助獲得欲分析資料的初步理解，也可作為建立資料初始模型之用。SOM 在資料探勘與資料庫知識發掘(Knowledge Discovery in Database, KDD)領域中應用於全文檢索或財務資料分析也被証明為相當有價值的工具。高維資料的視覺化正是 SOM 主要應用領域之一。SOM 擁有下列幾項優點，使其成為資料探勘之利器：【Vesanto, 2000】

A. 將訓練資料進行有規則的降維與映射

B. 產生的映射圖是依據資料的機率密度函數而產生

C. 映射圖具有高容錯性

D. 具有容易解釋，簡單與直接視覺化的特性

就演算過程來說，一旦演算法達成收斂狀態，輸出層的特徵圖形展現了輸入空間重要的統計特性。換句話說，對於一筆輸入向量，特徵圖將於輸入空間顯示一優勝單元，而權重向量將提供該優勝單元於輸入空間中的圖形座標，以下為 SOM 演算法的四個重要特性【陳慶翰，2002】：

A. 輸入空間的趨近性(approximation of the input space)

特徵圖形經由權重向量集合展現在輸出空間，並且針對輸入空間提供一個極佳的趨近特性。換句話說，SOM 的目的為經由找尋較小的樣式集合來儲存大量的輸入向量，以便針對原始的輸入空間提供一個較佳的趨近性。這理論基礎為「向量量化」，也是資料降維或壓縮的主要原理。

B. 網路拓樸的次序性(topological ordering)

經由 SOM 演算法計算所得到的特徵圖具有一拓樸次序性，就是在輸出網絡中神經元的空間位置與特定輸入樣式的特徵相對應。網路拓樸的次序性來自於將優勝單元的迫使權重向量朝向輸入向量調整；請注意，此時權重修正亦將使得最靠近優勝單元的鄰近神經元產生向中央權重修正的效果。如此便使得整個拓樸特徵圖形所形成的輸出空間產生適切的次序性，並

以虛擬的拓樸網格來呈現。網格上的每個輸出單元均可以其相對應的權重向量作為輸入空間的座標值。因此，若是輸出空間中的相鄰單元，其在輸入空間相對應資料點亦為相連時，便可直接觀察到網路拓樸的次序性。

C. 機率密度的對應性(density matching)

SOM 特徵圖同時也反映出輸入分配的統計量變化情形，也就是說，樣本訓練向量中發生的機率密度較高者在輸入空間所佔的區域將映射至輸出空間的較大部分區域，因此，相較於輸入空間中機率密度較低者，擁有較佳的解釋能力(resolution)。

D. 特徵的選擇性(feature selection)

從輸入空間中給定一非線性分配，SOM 擁有掘取一組最佳特徵集合來詮釋(或趨近)資料分配之能力。這項特徵同時也是前三項特性之集合。儘管主成份分析法可藉由關聯矩陣中具有最大特徵值之向量計算，求得訓練資料中擁有最大變異的輸入維度(向度)，但仍僅限於線性或平面的輸入空間；至於曲線或曲面(surface)時，主成份分析的表現便不如 SOM 的拓樸次序特性。

綜合以上所述，SOM 具有下列優點：【Vesanto, 2000】

- 穩健性(robustness)：假設鄰近區域函數延伸至足夠遠如高斯函數，則 SOM 會擁有相當穩健特性，這是因為競爭式學習所產生的原型會受到所有資料樣本之影響。
- 局部調整(local tuning)：拓樸的次序性將在每個優勝單元的鄰近區域發揮作用，因而形成朝向資料密度特性作局部調整。
- 易現性(ease of visualization)：SOM 有規則的網路格點可使建立一個有效率且視覺化的使用者介面容易許多。

以上優點大多歸因於 SOM 鄰近區域間的關係，這也是構成 SOM 架構的基礎。然而鄰近區域卻也有以下缺點：【Vesanto, 2000】

- 邊際效應(border effect)：鄰近區域的定義在 SOM 拓樸圖形的邊

緣為非對稱性。因此中央區域的單元，其鄰近區域函數(密度估計)必不等同於邊界單元。

- 收縮效應(contraction)：在向量量化過程中的平均分配使得變數數值的範圍縮小，並且受到鄰近函數的增強，極端值將因此去除，這在某些情況如分析者關切離散值時是不樂見的。
- 內插單元(interpolating units)：當資料群的分佈為不連續時，在資料群之間插入單元可便於資料分配的推估；然而，對於某些分析工具如單一連結分群法的例子顯示如此作法可能會提供錯誤的資料形狀線索。

有鑑於上述特性與優缺點，本研究遂採用以下四個分析構面作為本回顧研究第四階段資料分析與解釋之依據。

1. 競爭式學習(competitive learning)
2. 資料/集群視覺化(data / cluster visualization)
3. 網路拓樸映射與次序保存(topological mapping and order preservation)
4. 網路拓樸之動態網格(topological dynamic grid structure)

3.1.3 SOM 在資料探勘之應用

SOM 在資料探勘與資料庫知識發現領域中為一相當有價值的分析工具，例如應用於全文資料或財務資料的分析，或許多相關工業問題應用，如樣式辨認(pattern recognition)、影像分析(image analysis)、製程控制(process monitoring)與故障診斷(fault diagnosis)等方面，皆獲得相當成功的結果。【Vensento, 2000】

在大量高維度的資料集合中，若欲透過觀察量測值、統計數據或是文字型態的文件等資料來找尋其資料結構是相當困難及耗時的。然而資料項目間常常隱含某些特定關係需要加以調查，這也正是資料探勘的主要精神。SOM 基本上是一視覺化，分群與映射之工具，特別適合應用在資料探索(data exploration)或資料洞悉(data understanding)

領域中，透過特定圖形呈現出資料集合之結構狀態【Kaski, 1997】。不過也可應用在資料準備(data preparation)與針對資料局部模型進行建模或機率密度估計(probability density estimation)【Vensento, 2000】。

以資料分析(data analysis)角度來看，SOM 同時具有向量量化與向量映射演算法之特性，並能根據不同資料之局部資料密度(local data density)提供低維度的特徵圖，使得資料視覺化更有效率。除此之外，SOM 亦能依照訓練資料的機率密度產生健全穩健的特徵向量集合，以利資料分群與分析後續應用。

資料探勘(或資料挖掘)，為運用人工智慧或統計分析等相關技術，從大量雜亂無序且未結構化的資料中找出其間的規則，並整理出有用的資訊。Michael & Gordon(1997)曾做以下定義：「資料挖掘是利用自動或半自動方式從大量資料中找出樣式(patterns)及規則(rules)的探勘或分析過程」。此外，資料探勘可同時視為資料庫知識發現之核心程序【Vensento, 2000】，由圖 3.5 可知，知識發現的過程是由原始資料經選擇目標後，分別去除異常值與轉換為適當格式等前處理程序，便進入資料探勘階段。在此階段資料亦經過準備(preparation)、洞悉(understanding)與建立模型(modeling)等步驟，便產生具有代表性的資料樣式(patterns)，以供使用者自動或半自動地分析、評估與解釋真正感興趣的部份，所得知識可作為整體分析系統採取行動之依據。

圖 3.5 資料庫知識發掘流程圖【Kaski, 1997】

由圖3.6之循環圖可進一步說明如何將SOM應用至資料探勘各階段中，同時表 3.1 列出 SOM 於各階段之主要任務。SOM 於 Data mining 之應用可由探索性資料分析(exploratory data analysis)開始。Kaski(1997)於其博士論文中介紹 SOM 精簡特性並與其他分群法(階

層式/分割式)與映射法(PCA/MDS)作一概念性評析，指出 SOM 同時具有分群與非線性映射之特性，可分別達到降低資料量與低維空間之資料映射目的。

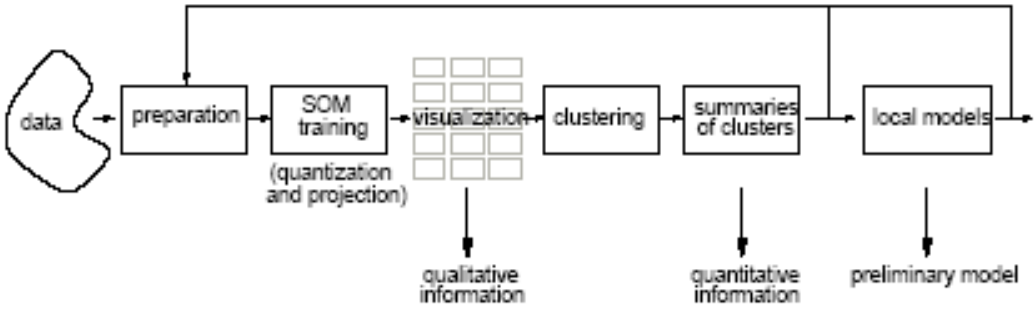


圖 3.6 在資料準備-調查循環中使用 SOM【Vensento, 2000】

表 3.1 Data Mining 各階段 SOM 之主要任務

階段名稱	SOM 主要任務/功能
資料準備(data preparation)	資料選擇與縮減(data selection & reduction) 斟酌(discretization) 累計(aggregation) 符號資料數量化(symbolical to numerical) 遺落資料置換(missing value replacement) 正規化(normalization)
資料洞悉(data understanding)	利用網格次序性作為一方便的視覺化平台來展現不同的資料特徵，並根據使用者需要提供資料的質化或量化資訊。
➢ SOM 訓練(training)	執行 SOM 訓練，包括輸入資料量化與映射
➢ SOM 的視覺化 (visualization)	產生代表輸入資料之質化資訊(輸出層的網路拓樸)
分群與摘要 (clustering and summarization)	經過分群後，使用者可由上述資訊選擇有興趣者進一步分析來獲得代表集群特性的量化摘要資訊
資料建模(Data Modeling)	先對輸入變數進行非監督式量化將輸入資料空間加以分割，每個分割區域再根據輸出變數執行監督式學習，以建立資料的局部模型