南　　華　　大　　學

資訊管理學系

碩士論文

應用類神經網路法於遺漏值問題之研究

The Study of Employing Artificial Neural Networks into the Problem of Missing Value

研 究 生：林俊男

指導教授：謝昆霖 博士

中華民國　九十四 年 六 月 十六日

南　華　大　學

資訊管理學系

碩 士 學 位 論 文

應用類神經網路法於遺漏值問題之研究

研究生：林俊男

經考試合格特此證明

口試委員：邱宏彬

陳彥旭

謝邦昌

指導教授：謝邦昌

系主任(所長)：資訊管理學系 系主任吳光閔

口試日期：中華民國 九十四 年 六 月 十六 日

誌　　謝

碩士兩年，時光匆匆，還未熟悉南華的氣息，就準備要離去。感謝謝昆霖老師的細心指導，使得駑鈍的我，也能成果豐碩；感謝邱宏彬老師時常的鼓勵，使得本無信心的我，也漸漸肯定了自己；感謝陳彥匡老師深入淺出的意見，使得我的碩士論文更趨完整；感謝學長姊們的幫助，使得剛入學懵懂的我，能順利的開始了研究所生涯；感謝同學們的關懷，使得在研究路途上，我不孤單；感謝學弟妹們的激勵，使得身為學長的我，能不斷的充實自己；感謝系助理伊汝和淑媛，兩年麻煩您們不少的事物；感謝前女友的陪伴，使得辛苦的碩士生涯中，能努力不懈；最後要感謝我的家人，由於你們的支持，才能讓我無後顧之憂的繼續求學；總之，感謝大家。

林俊男 謹誌於 南華大學資訊管理研究所

民國九十四年六月二十九日

應用類神經網路法於遺漏值問題之研究

學生：林俊男　　指導教授：謝昆霖博士

南　華　大　學　資訊管理學系碩士班

摘　　要

幾年來，資料挖掘(Data mining)技術受到許多企業的重視，更被廣泛地應用到顧客關係管理中，尤其是具智慧型的挖掘技術，例如類神經網路(Artificial neural networks, ANNs)中的非監督學習模式(unsupervised learning)，透過類神經網路的學習特性從資料中來挖掘出有用的資訊是它和傳統的演算法的不同；而在挖掘的過程中，由於種種可能的原因，都會有所機會發生資料遺漏或是不完整的情況，以往使用者只能捨棄該筆資料，此情形越多其挖掘出來的資訊偏差率就越大。因此，陸續有研究學者開始研究討論此相關領域的問題，最常見的方法為直接刪除、平均值或是眾數來取代等，但這種近乎直覺式的取代法對於最後的決策行為無法提供有意義的參考資訊。

本研究改以群組的「物以類聚」特性原理來思考，以此觀點來尋求

獲得遺漏值較適推估值。而類神經網路法中自組織映射圖網路模式，正是一個常見且發展多年的神經網路模式，其特點則是對於輸入值的資料型態並無任何的限制，所以廣受各領域學者的歡迎，其成效也令人滿意；因此，本研究嘗試利用自組織映射圖網路來建立一個針對遺漏值問題的推估模式架構，透過自組織映射圖網路的群聚方法來找出遺漏值最適推估值，讓使用者可以在引用資料挖掘法時仍能保有最大的資訊量，期使挖掘出的結果更有意義。並且以資料庫（RFM 資料庫資料）以及工業製程（半導體銅製程實驗資料）兩種類型的應用問題來展現本研究方法的可行性與合理性。

關鍵詞：資料挖掘(data mining)、遺漏值(missing value)、類神經網路(ANNs)、自組織映射圖網路(SOM)

The Study of Employing Artificial Neural Networks into the Problem of Missing Value

Student：Chun-Nan Lin. Advisors：Dr. Kun-Lin Hsieh.

Department of Information Management
The M.B.A. Program
Nan-Hua University

ABSTRACT

Recently, the technique of data mining had been applied into the issue of customer relationship management (CRM) by most enterprises. Among those techniques, the artificial neural networks (ANNs) had been mentioned as an intelligent approach. Especially, the unsupervised learning mode during ANNs can mine the available information by learning the clustering characteristic from data. However, the data may be lost or incomplete (i.e. the missing value) as for some particular reasons. To delete the missing value since making decision analysis is frequently employed to do it. Besides, the replaced estimate, e.g. the average value or the mode value, will be another approach to manipulate this problem. No matter what approach we take, the information we got will be limited for our decision analysis. In this study, we take the logistic concept of "clustering" to deal with such problem. The self-organize mapping (SOM) model of ANNs, which had been mentioned well for many applications, with the unsupervised learning will be taken to construct our solution herein. Finally, two illustrative examples including the missing value for database and for the experimental design are employed to demonstrate the rationality and availability for our proposed approach.

Keywords: data mining, missing value, artificial neural networks (ANNs), self-organize mapping (SOM)

目　錄

表　　目　　錄

圖　　目　　錄

第一章　緒 論

本章將描述本研究的背景、研究動機與目的、研究的方法和簡介本文的整體架構。

第一節　研究背景

隨著網際網路興起和科技的進步使得資料和資訊的取得更為便捷，而在這個知識爆炸的時代，如何快速的取得自身所需的資料和資訊更是各界重視的焦點。

資料的型態從早期的人工手寫紙本到現今電腦處理的電子資料其格式、內容皆為不同，早期人工手寫的紙本資料由於無電腦的存檔而造成保存不易、容易散失的情形；現今的電子資料又因為作業平台的不同、軟硬體設備的不同、通訊網路環境的不同也有所差異，透過各種方式所收集的資料，由於上述的種種因素而可能進一步的造成資料缺漏或遺失的情形，我們稱之為遺漏值問題(missing value problem) [1]。

遺漏值問題[25,27]，在資料分析領域裡已經不算是新穎的問題了。不論是學術界或是實務界所應用的資料分析方式，皆面臨了一個重要的課題，就是當今的資料庫技術，雖具有大量儲存資料的處理能力，但由於不同的技術平台及環境而造成了異質性資料庫的問題。在兩個以上的異質性資料庫間的資料處理過程，可能因交易(transation)、程式(programming)和操作(operation)等因素，發生了隨機性遺漏值問題[33]。傳統填補遺漏值問題的方法有許多，例如：直接去除遺漏值、人工比對、使用固定值填補、使用眾數填補、使用

平均數填補等[27]，雖然這些已近乎直覺式的方法估計出的值經填補原遺漏值後，對於後續的決策分析有某些程度上的意義，但其仍未能適切地反映出該遺漏值在實際的資料結構上的意義[1]。

第二節　研究動機與目的

在遺漏值的研究中，雖然過去有許多學者相繼提出相當多的方法。然而過去這些回填遺漏值的方法中皆有一些缺點，無法將所估計出的值能適切的反映在實際的資料結構上[7]：

1.直接去除遺漏值

在處理含有遺漏值資料的方法中最直接的方式，便是將含有遺漏值的資料整筆去除。若是在金融業等資料量非常龐大的環境下，直接去除該筆資料或許對整體資料的分析影響甚微，但是如果在資料量不大的產業裡直接刪除該筆資料的話，將造成資料量的縮減，使得可以分析、挖掘出的資訊變的更少。

2.人工比對

人本身是個複雜的動物，以人工比對的方法來做回填值的動作，將造成很多個人主觀或不客觀的因素。不同人在不同的時間和地點，受到不同的外在因素、環境所影響，其想法也會有所差異，因此所估計出的回填值其客觀性是值得商確的。

3.使用固定值填補

從不含遺漏值的資料中，將相同屬性欄位以機器式學習(Machine learning)方法，找出一個固定值填補。通常此方法是利用資料庫本身所含的資訊來預測推估一個固定的值，並以回填的方法來處理，雖

然較省時省力，但是推估出來的回填值卻不能保證他的正確性，可能會有某種程度的偏差(bias)。

4.使用眾數填補

找出資料庫裡所有資料出現次數最多的值來當作遺漏值的回填，但是當考量的屬性資料其出現不是具有高度重複性的話，在分析應用上將會有其限制。

5.使用平均數填補

使用平均數填補是目前最為常見的方法，利用資料庫中不含遺漏值的所有資料計算出其平均值加以填補。雖然此種方式所估計出的回填值基本上並不會有太大的落差，但是也容易受到極偏值或是資料型態分佈的影響，而造成求出的平均值有所偏差。

綜觀前述各項傳統作法均有其利弊，因此本研究擬透過資料群聚特質的思維來解析這類的問題，應用類神經網路之自組織映射圖網路模式來發展出一套遺漏值推估的解析流程，該程序能適切的提供一個有意義的遺漏推估值，讓使用者可以進行資料挖掘法或是解析該資料庫中的資料時，能保有最大的資訊量以期許挖掘出或是分析出的結果更具意義。

第三節　研究方法

在本研究論文中，我們首先整理、歸納與分析有關資料探勘的各種參考文獻，尤其是相關的資料分群演算法應用在遺漏值問題上之探討，期望能設計出一套適切的演算流程，研究流程如圖 1-1。

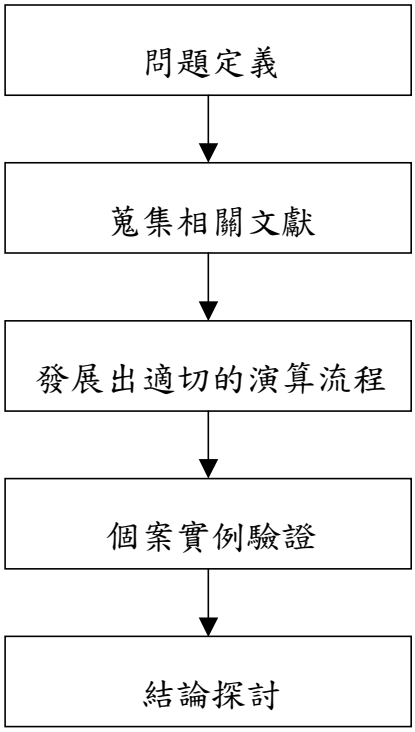

圖 1-1　本研究流程圖

根據我們所蒐集的文獻中，傳統估計遺漏值的方法幾乎是像前述的近乎以直覺式的方法或是傳統的統計方式來處理此相關問題，而效果往往不甚理想。因此本研究利用類神經網路中的自組織映射圖網路(Self-Organizing Map, SOM)為主體，來建構本研究所提出的估計遺漏資料值演算程序。主要的概念是利用自組織映射圖網路「物

以類聚」的特質來進行資料群聚並加以解析，接著我們也將其應用於半導體製程實驗參數與 RFM(購買時間(**R**ecency)、購買頻率(**F**requency)與購買金額(**M**onetary)三個參數)模式資料庫行銷等案例上以驗證其可行性。

第四節　論文架構

這篇論文的其餘章節組織如下，第二章我們將整理與回顧本研究相關的背景知識。第三章將介紹我們主要的研究方法—以類神經網路為基礎之遺漏值估計演算程序(Neural Network-based Estimate Missing value Solution, NBEMS)，詳述我們所整理發展出的演算程序。第四章將使用我們所方法進行個案的實驗，分別應用於工業製程參數最佳化問題與資料庫行銷問題，透過個案資料來展現我們所提方法的可行性和合理性。最後，相關的建議、結論與未來研究方向則在第五章說明。

第二章　文獻回顧

本章將介紹在此研究中所需了解的相關背景知識。

第一節　資料探勘

80 年代，由於資料庫技術開始發達，使得許多企業積極地投入大量資源來建置資訊系統，也使得企業所保存的資料變得龐大且複雜，但由於傳統統計侷限在小樣本的問題，使得傳統統計無法處理大量且複雜的資料。因此為了能獲得潛在資訊來幫助企業解決問題，遂發展出獨立之新領域－資料探勘[15]。M.S. Chen et al.[22]整理得資料探勘的演化步驟如表 2-1：

表 2-1：資料探勘的演化步驟 [22]

時間軸	→			
	1960s	1980s	1990s	迄今
演進	資料收集	資料處理	資料倉儲、決策技援	資料挖掘
企業問題	過去五年中公司的總收益多少?	在加州的分部去年三月的銷售額是多少?	在加州的分部去年三月的銷售額是多?舊金山據此可得出什麼結論?	下個月舊金山的銷售會如何?為什麼?
促成技術	電腦、磁帶、磁片	關聯式資料庫、SQL、ODBC	OLAP、資料倉儲、多維式資料庫	新演算法、多處理器電腦、大型資料庫
特性	提供歷史性的，靜態的資料	以記錄來提供歷史性的、動態資料	在各種層次上提供回溯的、動態的資料	提供預測性的資訊

一、資料探勘的定義與流程

資料探勘是近年相當熱門的一門新興技術，它結合了人工智慧和資料庫技術。它可從大量的資料中，萃取出潛在的、隱藏的、過去不為人所知道且可信與有效的知識。也可以說是依照使用者所設定的參數條件下，在一群大量且未經處理的資料中找到使用者感興趣的或有意義的資訊，經過特殊的處理後，作為使用者決策判斷的參考依據[15]。以下是常被引用的定義：

(一) Fayyad[24]的定義則嚴格區分資料探勘與資料庫中之知識發掘(KDD)。其定義資料庫中知識發現為自資料中選取合適資料，進行資料處理、轉換、資料探勘至結果評估之一系列過程。而資料探勘為其中一步驟(圖 2-1)。

(二) Berry[21]則認為資料探勘是為挖掘有意義的特徵或法則，而必須從大量資料之中以自動或是半自動的方式來探索與分析資料。

(三) Kleissner[29]認為資料探勘是一種新的且不斷循環的決策支援分析過程，它能夠從資料中，發現出隱藏價值的知識，以提供給企業專業人員參考。

在本研究，我們還是遵照 Fayyad *et al.*的定義，將知識發掘與資料探勘分開看待，以免觀念上有所混亂。

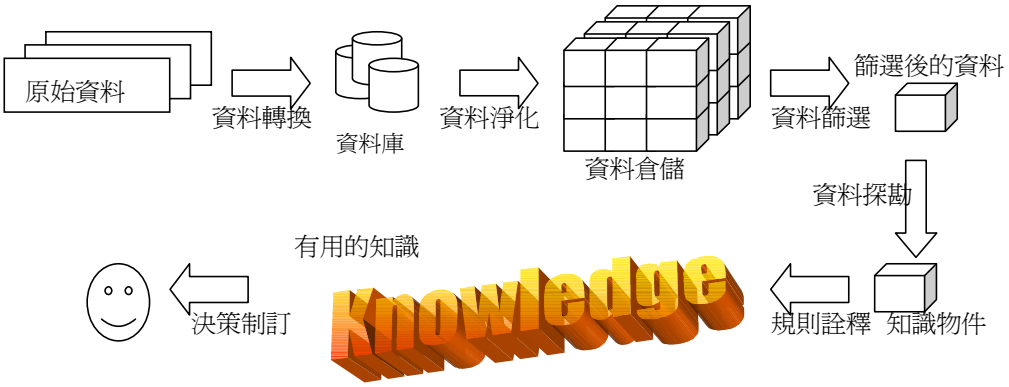

圖 2-1：知識挖掘過程示意圖[15]

二、資料探勘模式

為了適用在不同領域中資料探勘的問題，資料探勘區分為許多種方法，例如關聯法則(Association Rule)、資料分類(Data Classification)、資料分群(Data Clustering)、路徑尋訪樣式(Path Traversal Pattern) [23]……等不同的資料探勘模式。詳細介紹分別如下：

(一)關聯法則

由資料庫交易中找出項目(items)之間的關聯性，常使用的參數為信賴度(confidence) 和支持度(support)，來評量一條關聯法則的發生強度和頻率。例如某大賣場發現購買麵包的顧客同時會購買牛奶，則該大賣場在行銷規劃可將牛奶與麵包放置同一架上，以增加連帶銷售目的。而關聯法則依層次來區分，可分為單層次關聯法則與多階層關聯法則；若把交易時間順序的因素來進行關聯法則的考量，則為稱為序列型樣[20]。關聯法則模式最常見的演算法如 Apriori 演算法[19]、Sampling 演算法[34]等演算法[15]。

(二)資料分類

根據一些資料的屬性來進行計算，從歷史資料中進行特徵及規則擷取，根據這些特徵來建立模式，最後針對其他未分類或是新的資料進行預測。例如，從以往信用卡核卡歷史資料中找出核準與否的特徵，建立分類模式，此分類模式便可依據新的客戶資料(年齡、職業、收入、教育程度、婚姻狀況…)推論是否核準此新客戶的開卡申請。其模式最常見的技術如決策樹、倒傳遞網路等[15]。

(三)資料分群

是將資料分群，目的是找出同群集中資料的相似性，及各群集之間的差異性，使得同群中資料相似度最大，而各群之資料差異度最大。例如，銷售業者將客戶依其年齡、收入、居住地點、興趣、…等的屬性進行分群，這樣市場區隔能讓行銷人員了解最適合行銷的客戶群，並提供最合適的產品及服務適當的顧客。其模式最常見的技術如 K-means、Fuzzy C-means、SOM、ART 等[15]。

由於本研究所採取的邏輯思維是以「群聚」的觀念，因此在實驗時為了印証我們方法的可行性並與傳統的方法作個比較，我們選取了 k-means 演算法來進行實驗。以下我們便介紹 k-means 演算法。

1.k-means 簡介

集群分析係根據樣本的某些特性之相似程度，將樣本劃分成幾個集群，使同一個集群內的樣本具有高度之同質性，而不同集群間之樣本則具有較高度的異質性。而集群分析依照分類的方式不同可分為階層式集群分析(Hierarchical Cluster Analysis)及非階層式集群

分析(Nonhierarchical Cluster Analysis)，k-means 便是屬於非階層式集群分析裡最常被使用的一種方法[10]。k-means 演算法的演算過程如下：

步驟一、決定初始的分群數中心點。

步驟二、將中心點鄰近的資料分為該群。

步驟三、計算各群內的中心點，並更改為該群的新中心點。

步驟四、同步驟二，直到分群結果已穩定沒變化為止。

以下舉個小例子來說明 k-means 演算法的演算過程：

步驟一、先任意選擇三點作為初始的分群中心點

步驟二、利用這三點將所有的資料作分群，如圖 2-2 所示。

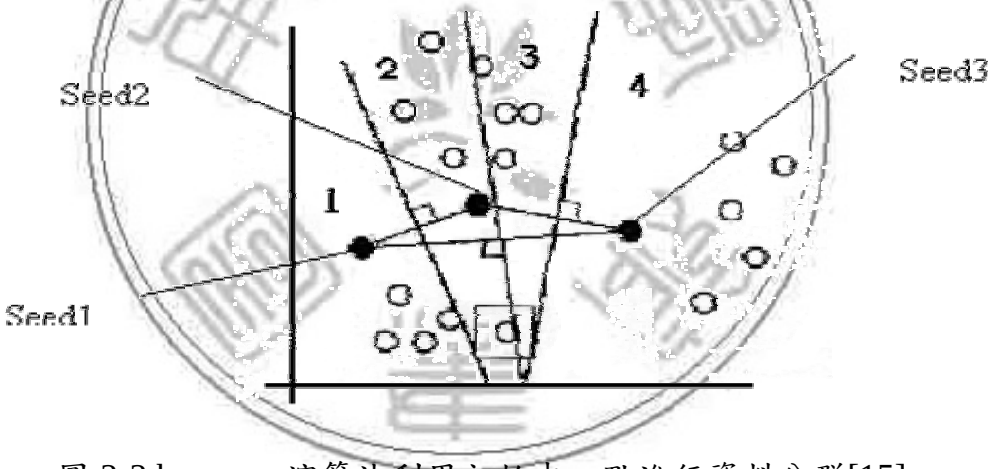

圖 2-2 k-means 演算法利用初始中心點進行資料分群[15]

步驟三、計算各群內的中心點，並更改為該群的新中心點，如圖 2-3 所示。

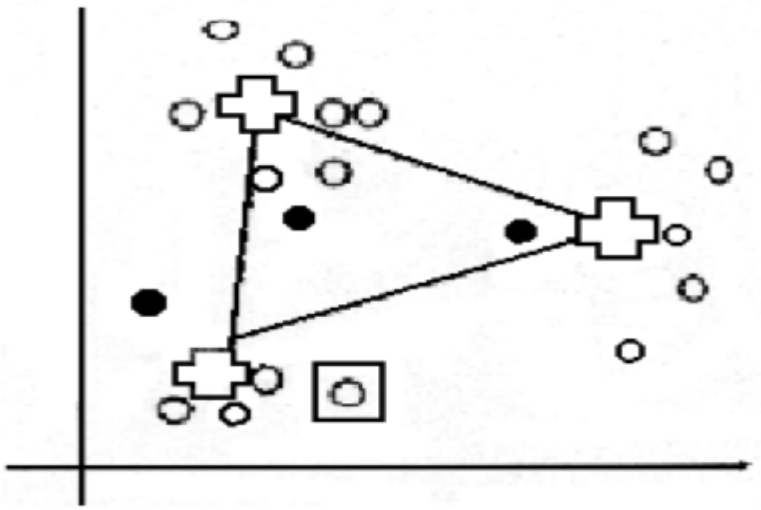

圖 2-3 k-means 演算法計算新的中心點[15]

步驟四、利用新的中心點再次進行資料分群，如圖 2-4 所示。

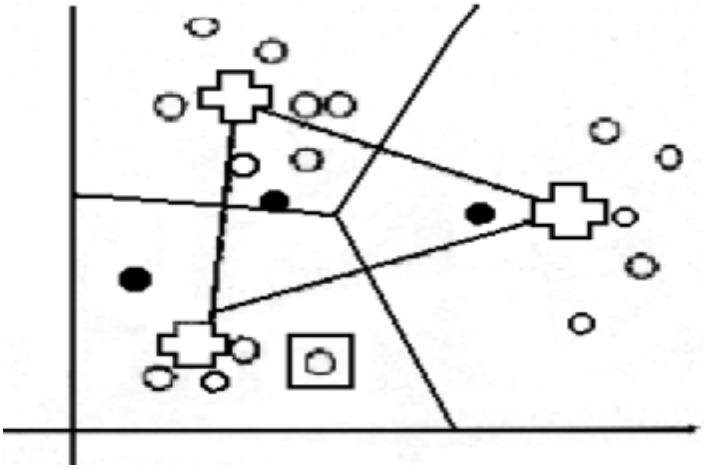

圖 2-4 k-means 演算法利用新的中心點再次進行資料分群[15]

(四)路徑尋訪樣式

在網際網路的環境中，此模式主要是擷取使用者瀏覽路徑存取特徵來瞭解使用者瀏覽網站之行為進而訂定出適合該使用者之個人

化行銷環境。大部分的方法都是將網站使用的記錄檔轉換成路徑順序的格式，然後判斷特徵出現之頻率[15]。

本研究在資料探勘模式的部分，主要是利用類神經網路中之自組織映射圖網路模式來進行資料的分群。所以接下來第二節將介紹類神經網路的背景知識及介紹其中的自組織映射圖網路模式。

第二節　類神經網路之相關研究

在第一節，我們介紹了資料探勘的演化步驟、定義與流程以及模式。由於我們主要是利用類神經網路進行資料分群的動作，因此本節我們將針對資料探勘中的類神經網路技術作為探討對象，介紹其基本的概念及我們所應用的自組織映射圖網路模式。

一、類神經網路

本部分將簡單介紹類神經網路的基本定義、類型和類神經網路的學習模式等相關的知識。

(一)類神經網路基本定義

雖然現今的電腦有著強大的計算能力，但在許多方面的表現仍無法超越人腦或稱生物腦，例如辨識或學習之類的能力，因此電腦嚴格定義只能稱為「計算機」(computer)而非「電子腦」(electric brain)[4,5]。

生物腦是由巨量的神經細胞所組成，形成一個高度連結網狀的神經網路(neural network)，生物或人的資訊處理工作即透過這些連結來完成。據估計共有 1000 億個(10^{11})，每個神經細胞又有近 1000 個連結與其他神經細胞相連，所以就有近 100 萬億(10^{14})個連結，比一般電腦僅一個中央處理器來的強大許多，在面對分類與決策之類的

複雜問題上，人腦的表現比電腦出色許多。為了讓電腦也能有人類面對複雜問題的良好處理能力，人工智慧、模糊理論、基因演算法、類神經網路等技術因此產生。其中類神經網路便是以模仿生物神經網路而產生的資訊處理系統[4,5]。

類神經網路(artificial neural network)，或譯為人工神經網路，是指模仿生物神經網路的資訊處理系統其基本架構如圖 2-5。Aleksander 與 Morton(1990)對類神經網路所下的定義為[18]「類神經網路是一種以自然特性儲存並運用經驗知識的平行分散處理器」；葉怡成(1993)對類神經網路有著較為精確的定義[4]：「類神經網路是一種計算系統，包含軟體與硬體，它使用大量簡單的相連人工神經元來模仿生物神經網路的能力。人工神經元是生物神經元的簡單模擬，它從外界環境或者其他人工神經元取得資訊，並加以非常簡單的運算，並輸出其結果到外界環境或者其他人工神經元。」

近年來由於類神經網路自身的理論上突破，加上從現代生物學、認知學、心理學等各學門對於生物神經網路的瞭解，皆增加了類神經網路的研究風氣。在應用上若依其輸出變數的特性，可分為兩大類[4,5]：

1.函數型問題：

網路的輸出為一個連續值的變數，例如：

(1)物理化學變量(濃度、溫度、PH 值、強度、流量、座標、尺寸)

(2)經濟社會變量(股價涨跌百分比、匯率、利率、成本、銷售量)

2.分類型問題：

(1)決策(醫藥處方、替代方案、買賣決策)

(2)診斷(疾病種類、故障原因、訊號識別)

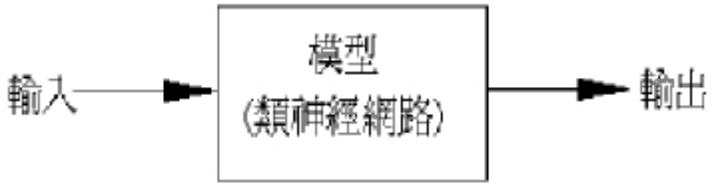

圖 2-5 類神經網路基本架構[1]

(二)類神經網路的類型

介紹完類神經網路的定義後，接下來針對類神經網路的類型做介紹。類神經網路依照學習模式可以大概分為四種類型[4,5]：

1.監督式學習網路(Supervised learning network)

從問題領域中取得訓練範例(有輸入變數值、也有輸出變數)，並從中學習輸入變數與輸出變數的內在對映規則，以應用於新的案例(只有輸入變數值，而需推論輸出變數值的應用)。

2.非監督式學習網路(Unsupervised learning network)

從問題領域中取得訓練範例(只有輸入變數值)，並從中學習範例的內在聚類規則，以應用於新的案例(有輸入變數值，而需推論它與那些訓練範例屬同一聚類的應用)。

3.聯想式學習網路(Associate learning network)

從問題領域中取得訓練範例(狀態變數值)，並從中學習範例的內在記憶規則，以應用於新的案例(只有不完整的狀態變數值，而需推論其完整的狀態變數值的應用)。

4.最適化應用網路(Optimization application network)

類神經網路除了「學習」應用外，還有一類特殊應用—最適化應用：對一問題決定其設計變數值，使其在滿足設計限制下，使設計目標達最佳狀態的應用。設計應用與排程應用屬之。此類應用的網路架構大都與聯想式學習網路的架構相似。

(三)類神經網路的學習規則

從人類所接受的分類方式來看，學習的策略可以大略分成下列幾種[17]：

1.機械式的背誦學習(rote learning)

屬於層次最低的學習法，僅單純背誦而沒有推廣性(generalization)。這樣的學習效果，就像是只建立一個輸入及輸出的對照表而已，雖然簡單，但是並無法達到推廣、學習變化的目的，當有新的資料不斷產生，就無法一一存載的。

2.指令式的學習(learning by instruction)

學習者扮演的角色只是將外界的知識(例如：書本、老師、操作手冊等)轉化為本身可懂的表示法或語言。這就像現今電腦執行的程式，成果好壞取決於程式的撰寫者，而非執行者或學習者，因此這種方法也無法模擬出像人一樣的機器。

3.類推式的學習(learning by analogy)

學習者從與目的狀況最相似的過去經驗中得到新技術。像是原本從事排球運動的人，能夠很快的掌握到打好排球的技巧。此方法

目前不斷有研究者投入，但尚未有十分突出的研究，若類神經網路採用此學習法，會有極為明顯的學習效果，所以此方面的研究仍待大家努力的。

4.歸納式的學習(learning by induction)

此即為類神經網路學習過程中的學習方式，可分為下列兩種：

(1)從範例中學習(learning from examples)

此法相當於於監督式學習，學習者從一組含有正例(positive examples)與反例(negative examples)的範例中，歸納出一個能解釋範例的整體觀念(concept)，較簡單的說法就是，環境中會有所謂的「老師」，告訴學習者，什麼樣的刺激，該有什麼樣的反應，以數學語言來說就是存在有輸入/輸出對映的資料。

(2)從觀察及發現中學習(learning from observation and discovery)

又稱為非監督式學習，這種學習是讓學習者自己發現資料本身的重要特徵或結構，沒有老師會提供任何其他資訊給學習者。

從以上的文獻中可得知，各類的類神經網路應用在不同類型的問題，像監督式學習網路適用在分類(診斷、決策)、預測(函數合成)問題，非監督式學習網路適用在聚類問題，聯想式學習網路適合應用在雜訊過濾、資料擷取上，最佳化問題網路則應用於設計和排程上。而且也可得知類神經網路對於分類、聚類型問題，可以提供一個不錯的解決方法。除此之外，類神經網路的非監督式學習，則可依照問題的特徵，分隔出群內差異小，群間差異大的不同群組，進

而達到分群的目的，所以我們可以說類神經網路是一種可以應用在多種學科及功能的方法。

二、自組織映射圖網路模式介紹

本部份將介紹自組織映射圖網路模式，包含基本的觀念與架構等。

(一)自組織映射圖網路模式基本觀念

自組織映射圖網路(Self-Organizing Map, SOM)是一種非監督式學習網路模式，早在 1980 年 Kohonen 即提出此模式，至今仍是非監督式學習網路模式的典範[30]。非監督式學習網路模式應用可再依其輸入值特性分成兩類[12,30,32]：

1.輸入值為二元值者(例如：ART1)

2.輸入值為連續值者(例如：SOM)

(二)自組織映射圖網路模式架構

自組織映射圖網路模式的架構如圖2-6，其組成為[2,5]：

1.輸入層

用以表現網路的輸入變數，即訓練範例的輸入向量，或稱特徵向量，其處理單元數目依問題而定。使用線性轉換函數，即f(x)=x。

2.輸出層

用以表現網路輸出變數，即訓練範例的聚類，其處理單元數目依問題而定。

3.網路連結

每個輸出層單元與輸入層處理相連連結的加權值所構成的向量，表示一個訓練範例對映樣本點聚類之形心座標。當網路學習完畢後，其輸出處理單元相鄰近者會具有相似的連結加權值。

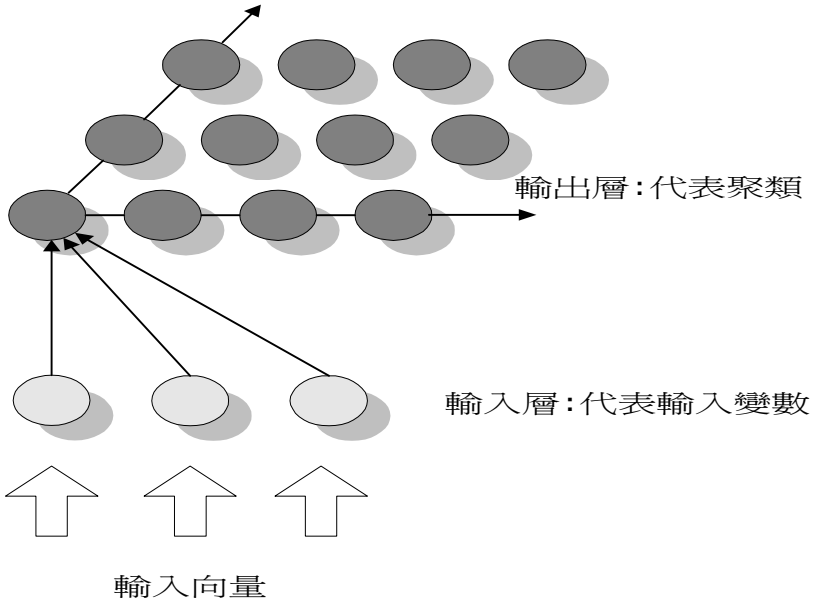

圖 2-6 自組織映射圖網路模式基本架構圖[5]

4.網路拓樸

自組織映射圖網路模式與其他類神經網路模式有一點重要的差異，它的輸出層處理單元的相對位置具有意義，而一般的網路模式則否。通常它的輸出層處理單元以二維的型態排列居多，形狀以矩形居多。但實際上可以用三角形、圓形、甚至任意形狀，而且一維、三維排列亦可。圖2-7顯示一個二維矩形網路拓樸的輸出層。

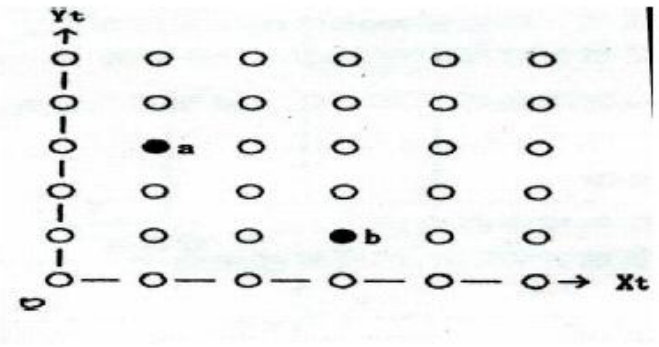

圖 2-7 二維矩形網路拓樸[5]

5.拓樸座標

拓樸座標是指標定一輸出層處理單元在網路拓樸中位置的座標。對於一個二維型態排列的網路拓樸，每一個輸出層處理單元具有一個二維的拓樸座標；如採用一維或三維排列，則每一個輸出單元將具一維或三維的拓樸座標。拓樸座標與樣本空間座標必須釐清，樣本空間座標的維次由輸入層處理單元的數目決定，通常從數維到數十維都有可能，是用來標示一訓練範例的輸入向量，或稱特徵向量，在樣本空間中的位置，即訓練範例所對映的樣本顯示的其三維點之位置。拓樸座標的取法有很多，例如圖2-7顯示的具二維矩形網路拓樸的輸出層，可取左下方的單元為座標原點，每向上一橫列與每向右一直行其座標值增一單位。例如圖2-7的a表示輸出處理單元的拓樸座標為(1，3)，而b表示輸出處理單元的拓樸座標為(3，1)。

6.鄰近區域

鄰近區域是指在網路拓樸中，以某一輸出處理單元為中心的區域，稱此單元之鄰近區域。參見圖2-8，鄰近區域內的輸出處理單元會相互影響。鄰近區域會因網路學習過程而逐漸縮小。

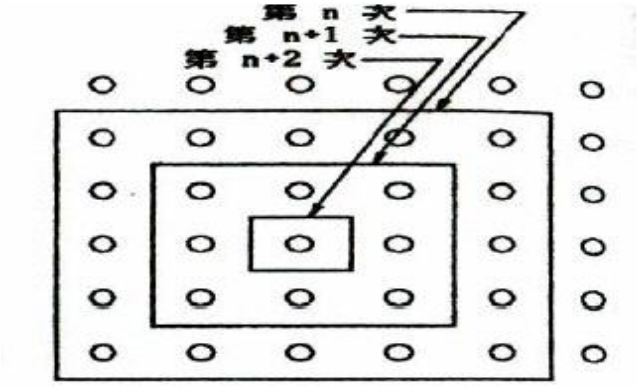

圖 2-8 自組織映射圖網路「鄰近區域」觀念[5]

7.鄰近中心

控制鄰近區域中心位置的參數，即一輸出處理單元在網路拓樸的拓樸座標。

8.鄰近半徑

控制鄰近區域大小的參數。

9.鄰近距離

一輸出處理單元在網路拓樸中距鄰近中心的距離，由此一輸出處理單元的拓樸座標以及鄰近中心的拓樸座標來決定，以公式表示如下：

$$r_j = \sqrt{(X_j - C_x)^2 + (Y_j - C_y)^2} \qquad (2\text{-}1)$$

其中 (X_j, Y_j) = 第j個輸出處理單元拓樸座標。

其中 (C_x, C_y) = 鄰近中心拓樸座標。

10.鄰近係數

控制鄰近區域內輸出處理單元相互影響程度的參數。鄰近係數是「鄰近半徑」與「鄰近距離」的函數。

11.鄰近函數

控制鄰近係數和「鄰近半徑」與「鄰近距離」關係式的函數：

$$R\_\ factor_{\ j} = f(r_j, R) \tag{2-2-a}$$

圖2-9顯示幾種常見的鄰近函數，本文將採用斗笠帽鄰近函數：

$$R\_\ factor_{\ j} = \exp(-r_j / R) \tag{2-2-b}$$

此式當 $r_j = 0$ 時 $R\_\ factor_j = 1$ ； $r_j = \infty$ 時 $R\_\ factor_j = 0$ ； $r_j = R$ 時 $R\_\ factor_j = 0.368$ 。

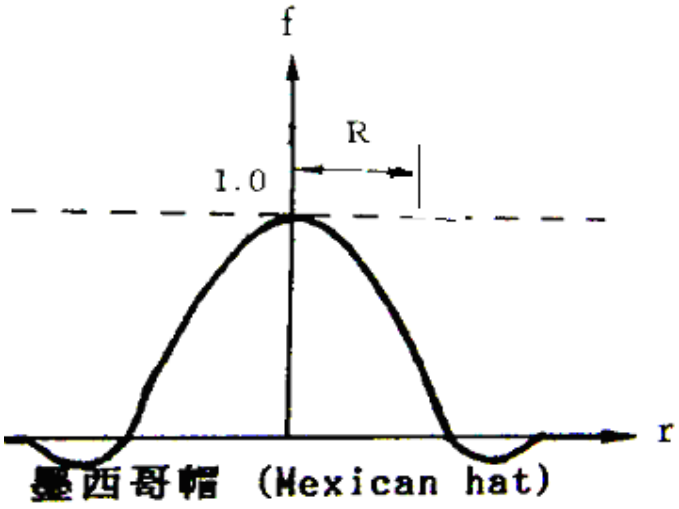

圖 2-9 自組織映射圖網路模式「鄰近函數」的觀念[5]

12.鄰近區域收縮

鄰近區域會在網路學習過程中逐漸縮小，實際上即鄰近半徑逐漸縮小，以公式表示如下：

$$R^n = R\_rate * R^{n-1} \tag{2-3}$$

其中 R_rate = 鄰近半徑縮小因子(<1.0)。

在本文中採以學習循環為單位，網路每執行一個學習循環鄰近半徑收縮一次。

(三)自組織映射圖網路模式演算法

自組織映射圖網路模式的基本原理相當簡單，可說是非常直覺。其網路演算法和反傳遞網路模式隱藏層的非監督式學習網路演算法相似，兩者的差異在於自組織映射圖網路模式多出「鄰近區域」(neighborhood)的觀念。鄰近區域內的輸出處理單元會相互影響，鄰近區域會在網路學習過程中逐漸縮小。其步驟如下[5]：

1.計算訓練範例與各輸出層單元的距離

每次載入一個訓練範例便計算各輸出層單元與其輸入向量的距離，以公式表示如下：

$$\begin{aligned}\left\|X(C)-X(C_j)\right\| &= X(C)\text{與}X(C_j)\text{間歐氏距離平方}\\ &=\left[X(C)-X(C_j)\right]^t * \left[X(C)-X(C_j)\right]\\ &=\sum_i \left[X_i(C)-X_i(C_j)\right]^2 \end{aligned} \qquad (2\text{-}4)$$

其中 $X(C)$ = 訓練範例C的特徵向量。

$X(C_j)$ = 第j個輸出層處理單元對映之特徵向量。

= 第j個輸出層單元與輸入層單元間的加權。

$X_i(C)$ = 訓練範例C的特徵向量的第I個元素。

$X_i(C_j)$ = 第j個輸出層單元對映之特徵向量第I個元素。

$=W_{ij}$

2.找出優勝單元

距離最短的輸出層單元稱為**優勝單元(winner)**。以公式表示如下：

$$\text{如果}\left\|X(C)-X(C_{j^*})\right\| = \min_j \left\|X(C)-X(C_j)\right\| \qquad (2\text{-}5)$$

則第j*個輸出層處理單元為優勝單元。

3.調整輸入層與輸出層間的連結加權值

網路連結加權值需修正，以公式表示如下：

$$\Delta W_{ij} = +h * (X - W_{ij}) * R\_factor_j \qquad (2\text{-}6)$$

其中 h = **學習速率**。

R_factor_j = 第j個輸出處理單元的鄰近係數

= f(R, r_j) 。

注意不只有優勝單元的連結加權值需修正，而是與連結所連輸出層單元距優勝單元(鄰近中心)的鄰近距離有關。鄰近距離越大，鄰近係數越小，連結加權值修正也越小。

4.對所有訓練範例重複步驟1到3稱為一學習循環,每執行一個學習循環，將鄰近半徑收縮一次，且學習速率折減一次。

以上演算法規則我們可以用一個簡單的例題來說明，則更容易瞭解。

(1)假設有一聚類問題的聚類規則是由二維的特徵向量所決定，其真實聚類規則如圖2-10所示，共有五個聚類。

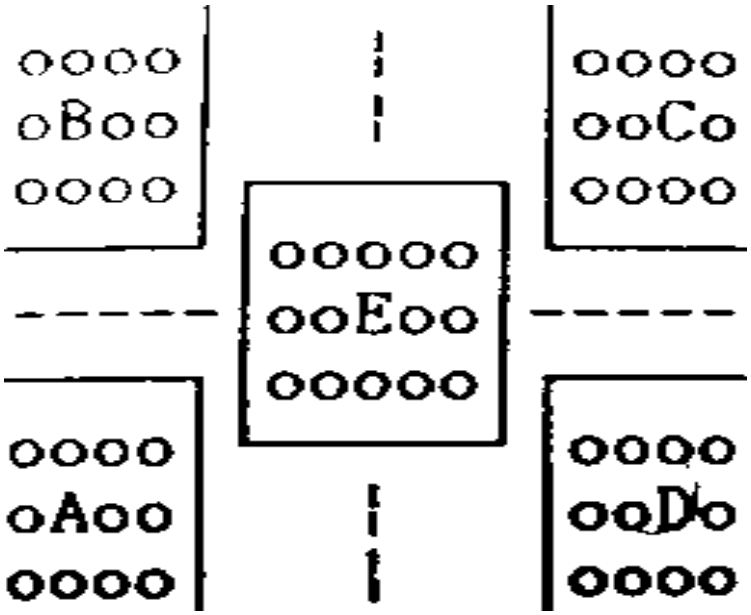

圖 2-10 真實聚類規則[5]

(2)這5種聚類各有2個樣本(訓練範例)，如圖2-11所示。

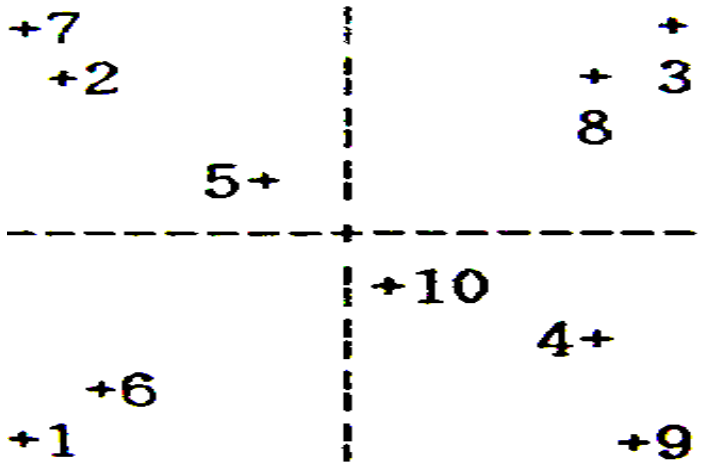

圖 2-11 訓練範例採樣點[5]

(3)假設以一自組織映射圖網路模式解此問題，採3*3=9個輸出層單元，如圖2-12所示。其對映的初始聚類群形心點，如圖2-13所示。

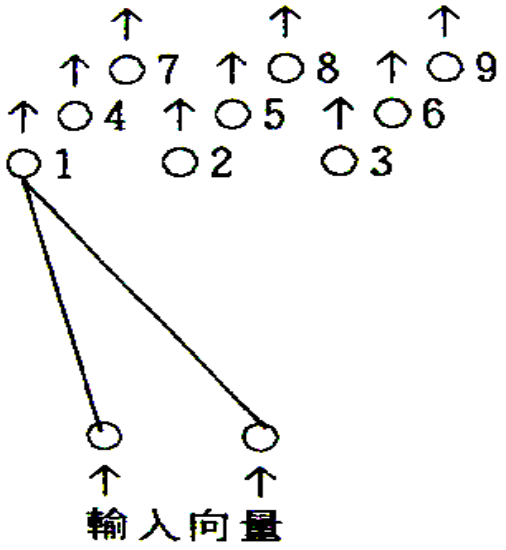

圖 2-12 自組織映射圖網路模式架構圖[5]

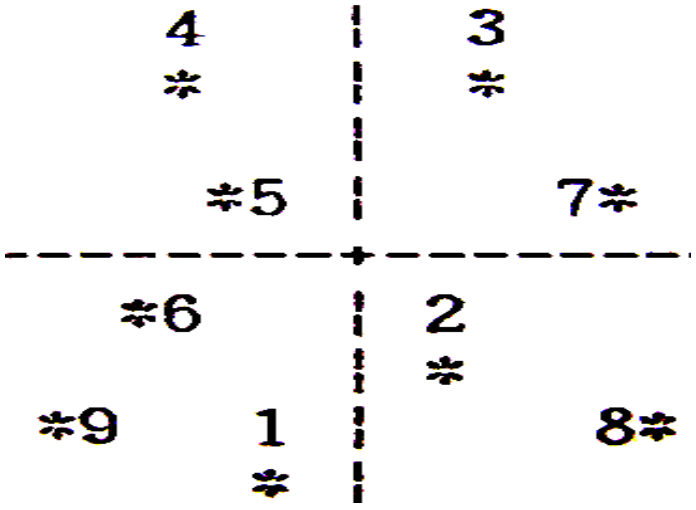

圖 2-13 初始樣本點群形心[5]

(4)假設鄰近中心與鄰近區域的關係如圖2-14所示。連結加權值調整規則為優勝單元相連者大幅調整，其鄰近區域單元相連者中幅調整，其餘單元則不調整。

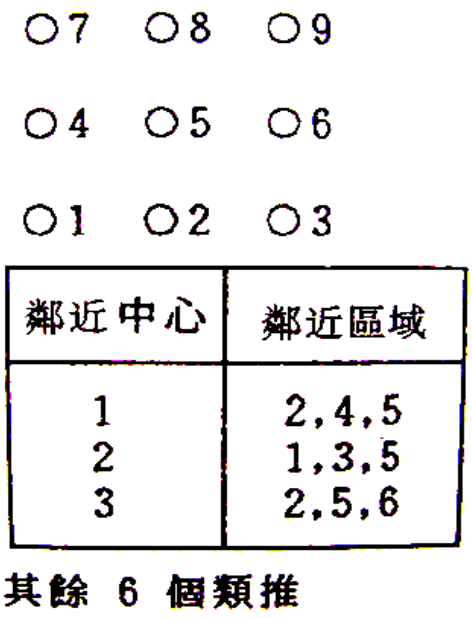

圖 2-14 自組織映射圖網路模式鄰近區域[5]

(5)首先載入第一個訓練範例，其優勝單元為輸出層單元9，調整優勝單元以及其鄰近區域內輸出層單元之連結加權值，即將輸出層單元9向訓練範例大幅移動，輸出層單元5，6，8向訓練範例中幅移動，如圖2-15所示。

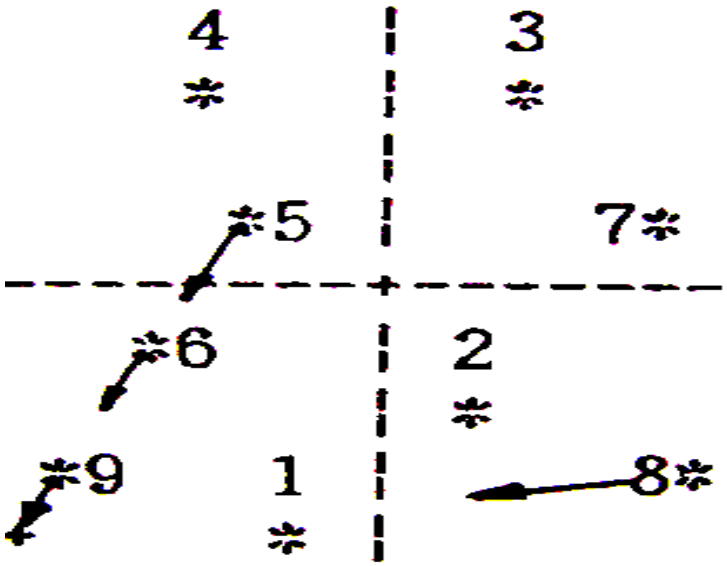

圖 2-15 第1個訓練範例(優勝單元：9；鄰近區域：5,6,8)[5]

(6)接著載入第二個訓練範例，其優勝單元為輸出層單元4，調整優勝單元以及其鄰近區域內輸出層單元之連結加權值，即將輸出層單元4向訓練範例大幅移動，輸出層單元1，5，7向訓練範例中幅移動，如圖2-16所示。

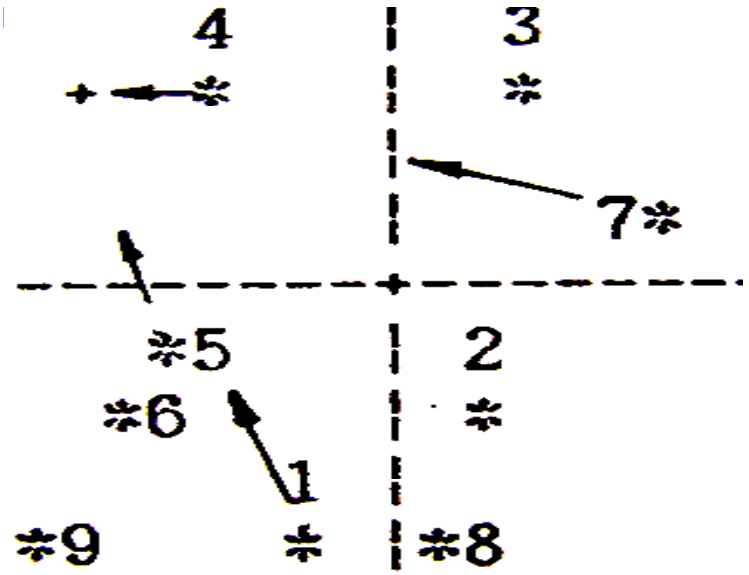

圖 2-16 第2個訓練範例(優勝單元：4；鄰近區域：1,5,7)[5]

(7)接著依序載入其餘八個訓練範例，即完成第一個循環(第三、四個訓練範例的學習過程如圖2-17、2-18所示)。第一個學習循環結束時，9個輸出層單元對映的訓練範例樣本點的聚類形心座標如圖2-19所示。

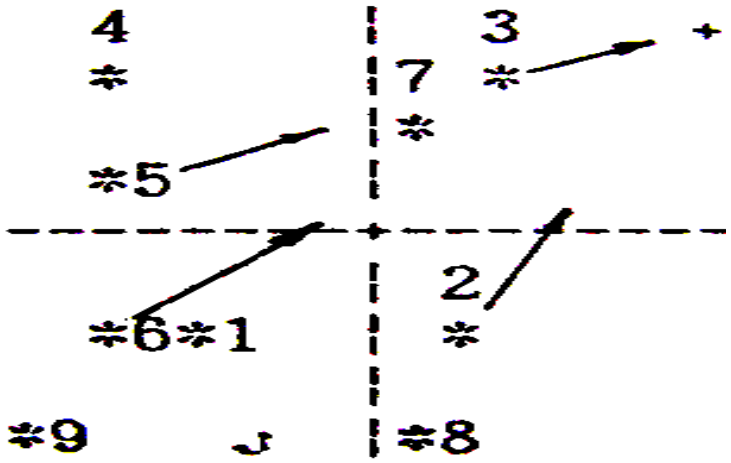

圖 2-17 第3個訓練範例(優勝單元：3；鄰近區域：2,5,6)[5]

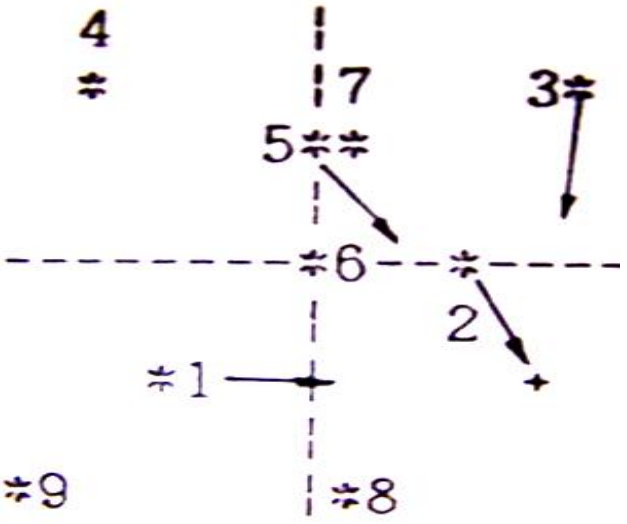

圖 2-18 第4個訓練範例(優勝單元：2；鄰近區域：1,3,5)[5]

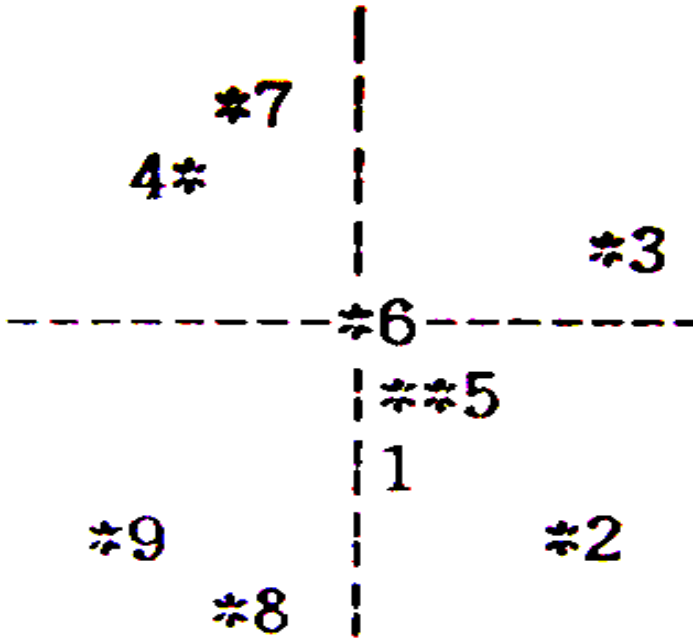

圖 2-19 第一個學習循環結束時的樣本點群形心[5]

(8)反覆載入上述十個範例，即可得到樣本點的聚類形心的精確位置，如此即完成非監督式學習，圖2-20表示第一次學習循環結束結果與真實聚類規則之比較。

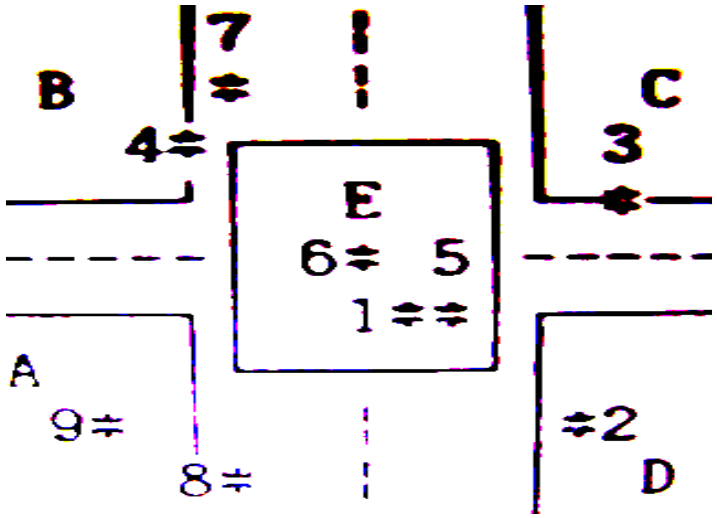

圖 2-20 第一個學習循環結束時結果與真實聚類規則之比較[5]

第三節　遺漏資料

資料探勘的過程中，不論是準備資料或是建立模型時都會面臨到大量資料的處理程序，尤其是進行資料探勘時，大量資料的使用則是更為重要。從相關的實務面來看，其常會發現具有遺漏或含有雜訊的資料[7]。

一、 遺漏資料的型態

遺漏資料(Missing data)屬於雜訊資料其中一種。在關聯式資料庫中，這些遺漏值皆被以 NULL 來表示。雖然所有遺漏值的表示方式都一樣，但其發生的原因可能不盡相同：

(一)空數值(empty value)

會發生空數值的情形，可能是和較私密的資料相關資訊，例如顧客可能因為不想被打擾而未填入聯絡電話。

(二)不存在的數值(nonexistent value)

不存在的數值之所以會發生，是由於欲解決的問題之特性所導致。例如一個模型需要使用 12 個月份的歷史資料來預測某些未來事件，但最近新增的顧客資料並無過去的歷史資料而產生遺漏值。

(三)不完整的資料(incomplete data)

不完整的資料會發生在資料來源無法產生所有的相關資料時。例如同時從銀行的活期存款部門、信用卡部門以及貸款部門所收集到的資料中，許多顧客可能只有和某一個部門有關係，而在其他的部門並無該顧客的資料。

(四)未被收集到的資料(uncollected data)

未被收集到的資料是指從未被收集到的遺漏值。例如大部分的電話交換機，在顧客關閉話中插接服務時，並不會紀錄通話情形。

二、 遺漏資料的處理方式

發現遺漏資料時，大都採用統計觀點的方式，以離群值(outlier)的方式來處理：

(一)不處理

某些演算法(例如：多種建立決策樹的演算法)可以容許遺漏值的存在，如果遺漏值的情形不嚴重的話，就不會影響到此模型的結果。

(二)將包含遺漏值的資料列過濾掉

這種方式主要是將資料維度予以縮減，但是如此一味地資料過濾排除掉，相對地也可能會造成資料抽樣上的偏差。

(三)忽略該資料欄位

將重點放在完整的資料上，而忽略這些有遺漏值的資料欄位。此方法和上個方法一樣，可能會造成資料抽樣上的偏差。

(四)預測新數值

先求算出不含遺漏資料筆的平均數或是眾數後，再將其值插入該遺漏的資料欄位中是最常使用的作法。

(五)建立各自獨立的模型

一般情形下，通常可根據顧客資料的不同，將顧客分群，然後針對不同群的顧客，建立個別模型。

(六)修改營運系統

直到所有的資料皆可以被收集完整後為止，但顯見的是此方式不適用於大部分的短期專案計畫。

資料探勘發展至今，並無一種方式可以完全通用於任何一種情形，只能依據實際的情形加以判斷嘗試，所以本研究嘗試結合 SOM 網路模式發展出一套演算流程以便處理遺漏資料值。

第三章　以類神經網路為基礎之估計遺漏值演算程序

本章我們將詳細介紹如何應用類神經網路為基礎，建構一套估計遺漏值的演算程序(Neural Network-Based Estimate Missing value Solution, NBEMS)。本研究所整理發展出的概念程序如圖 3-1，步驟如下所示：

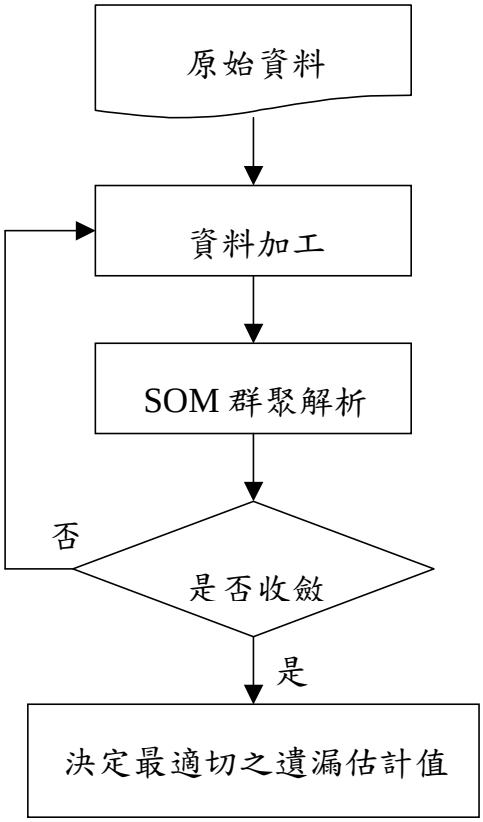

圖 3-1 本研究概念程序圖

步驟一、找出遺漏值。

首先，檢視所有的原始資料，並且將所有屬於遺漏值的欄位先行予以標記。

步驟二、去除極偏值。

這個步驟主要是用於篩選我們將用來估計遺漏值的資料集。由於一個資料集中，可能會出現和平均水準落差甚大的值(我們稱之為極偏值)，如果極偏值是因為某些人為上的疏失或是無特殊性意義的情況所造成的，倘若不加以去除，將造成資料展現時的偏差，因此在這裡我們經過資料的標準化用來檢視各筆資料是否為極偏值以增加後續在估計遺漏值上的精準度。而經過資料標準化後(我們採用 Z 分配)，當其標準化值大於 3 或小於-3 時且已確認並無特殊理由所造成的，我們便視該值為極偏值並予以刪除。

標準化公式：$Z = \frac{(x_i - \overline{x})}{S}$ （3-1）

X_i = 第i筆資料。

$\overline{X}$ = 各屬性扣除遺漏值後的平均值。

S = 標準差。

步驟三、給予遺漏值初始值。

在整個演算流程裡，我們必須給予我們已標記的遺漏值欄位一個初始的估計值，由於平均值(Average)有著一個資料群裡共同表現的趨勢[13]，因此我們選定使用平均值來作為我們已標記的遺漏資料值欄位之初始估計值。在這裡平均值的計算是將所有該屬性欄位過

濾掉已標記的遺漏資料欄位外的加總平均。平均值有三個重要的性質(或作用) [13]，即

1.簡化作用

指平均值能簡化一群體(或分配)的所有數值而為一數值。

2.代表作用

指平均值能代表一群體(或分配)的平均水準。

3.比較作用

指平均值簡化所有數值為一數值後，以該數值代表群體(或分配)的平均水準，而便於兩個或兩個以上群體(或分配)間作比較。

平均值計算的公式如下：

$$\overline{X} = \frac{1}{n}(x_1 + x_2 + x_3 + \mathbf{L}\,\mathbf{L} + x_n) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3-2}$$

n　為資料數

步驟四、利用 SOM 進行學習分群，檢視分群後結果，判斷條件：「各分群內，除了含遺漏值的資料外，必須含有兩筆(含)以上的資料。」

在這裡我們利用試誤法(try-and-error)，經過多次的實驗來決定我們所感到適切的網路模式結構。我們經由一個限制條件來限定分群的結果，因為倘若一個分群內只有兩筆資料，且其中一筆還是我們欲估計的遺漏值，這樣會造成我們所估計出的遺漏值較不為客

觀，因此在本實驗中，增加了該限制式的存在，並且以這個限制式來輔助判斷是否停止群聚的解析過程。

步驟五、檢視 SOM 的輸出值，過濾原先含遺漏值的資料，並計算該遺漏值所屬的群集之平均值，如果求出的估計值和輸入前的估計值相等時，則結束演算程序，決定最適切之遺漏估計值；反之，將前一步驟學習分群後所求得之估計值回填原遺漏值，並回到步驟四。

在本步驟中，我們將檢視透過自組織映射圖網路模式所學習訓練出的分群，並且計算原先我們已標記的遺漏值所屬的分群，扣除掉該筆資料後的該分群該屬性欄位的平均值，該值計算出後，便與輸入至 SOM 學習分群前的值進行比較，如果兩值相等便直接結束演算程序，如果不相等則將前一步驟學習分群後所求得之估計值回填原遺漏值反覆執行步驟四。

因為會有數組群聚的結果，為能適切地決定出最適的群聚數，我們利用歐氏距離為考量的依據，選擇具有最小歐氏距離者為最適網路架構(歐氏距離公式請參閱公式 2-4)。因為一個好的分群結果應具有兩個特性：內聚力強和耦合力弱；內聚力是代表群內各筆資料的相似程度或關聯程度，因此如果該值越高則代表該群內的資料相似度或是關聯程度越高；而耦合力剛好相反，它是表示群與群間的相似或關聯程度，因此該值越小則表示群間的差異度大。而歐氏距離則表示群內各筆資料與群中心點的距離，因此該值越小表示各筆資料越接近該分群的群中心點，也間接表示其內聚力越大，這也就是我們為什麼要採用歐氏距離值為考量的依據。

第四章　個案實驗

本章將應用我們的方法來實現於兩個案例中，並且與傳統的估計遺漏值的方法所估計出的結果進行比較。第一節先使用傳統的方法來估計工業製程實驗參數之遺漏值，其中包含了直覺式的推估方式—平均數和 k-means 演算法，接著應用我們的方法來進行工業製程實驗參數之遺漏值的推估，以及這些方法估計出的遺漏值結果的比較。第二節則針對行銷資料庫的研究個案，首先也是先使用傳統方法來進行估計，接著引用我們的方法進行推估行銷資料庫模式的遺漏值，最後則是傳統方法和我們的方法在行銷資料庫模式的資料下，所推估出的遺漏值結果比較分析。

第一節　估計工業製程實驗參數之遺漏回填值

由文獻的探討後我們可以得知，最常使用於估計遺漏值的方法便是平均數的取代。而本研究所發展出的估計遺漏值思維，主要為資料分群的原理，所以在本節中，我們嘗試利用計算整體平均值和統計方法裡的 k-means 資料分群方法以及我們的方法來估計在工業製程實驗的環境下，其實驗參數的遺漏回填值。

一、 個案描述

本實驗個案採取使用唐麗英和王春和[6]論文中的半導體銅製程的資料進行實驗(資料型態如表 4-1)。本資料一共有 18 筆資料，並隨

機假設 5%的資料遺漏情況，每筆資料有三個屬性變數，分別是研磨速、均勻度及 Selectivit，反白部份為我們隨機假設的遺漏值。

表 4-1 工業製程實驗參數資料原始型態

第 n 筆資料	研磨速(RR)	均勻度(NU;%)	Selectivit(Tan/Cu)
1	294	14.3	4
2	289	15.7	4.3
3	314	23.2	5.6
4	375	12.1	3.7
5	437	8.7	4.9
6	498	6.5	6.1
7	481	8.99	4.2
8	588	11.8	4.3
9	660	12.4	5.3
10	242	16.2	4.6
11	268	26.9	4.1
12	340	10.5	5.3
13	377	16.9	3.9
14	434	5.06	4.7
15	494	7.08	5.4
16	483	8.76	5.2
17	580	15.1	4.6
18	651	5	5.8

(一)利用平均值來估計遺漏值

利用公式 3-2 計算各屬性的平均值(扣除我們假設的遺漏值)，其求得之結果如表 4-2。

表 4-2 利用平均數所估計的工業製程實驗參數遺漏值

研磨速(RR)	均勻度(NU;%)	Selectivit(Tan/Cu)
430.7059	12.7347	4.8176

(二)利用 k-means 來估計遺漏值

除了根據第二章所介紹的 k-means 演算法，我們一樣利用我們前面所提到的限制式「每組分群中，扣除遺漏值後，必須有兩筆以上(含)的資料。」加以檢視各分結果。其結果見表 4-3。

表 4-3 經 k-means 演算法所求的工業製程實驗參數資料分群結果

第 n 群	群內各筆資料
1	8、9、17、18
2	1、2、3、10、11
3	4、5、6、7、12、13、14、15、16

檢視 k-means 分群出的結果，找出遺漏值所屬的分群，並且計算該群扣除遺漏值後的平均值，其求得的值便是利用 k-means 演算法所找出的估計值，如表 4-4 所示。

表 4-4 k-means 估計出的工業製程實驗參數遺漏值

RR	NU	Tan/Cu
429.50	9.49	4.63

(三)利用 NBEMS 來估計遺漏值

在本節中，主要介紹我們所發展整理出來的一套估計遺漏值的演算程序並利用本演算程序來進行遺漏值的推估。

步驟一、找出遺漏值。

檢視所有的原始資料，並且將所有屬於遺漏值的欄位先行予以標記。表 4-5 表示工業製程實驗參數部分原始資料型態。

表 4-5　工業製程實驗參數部分原始資料型態

第 n 筆	研磨速(RR)	均勻度(NU;%)	Selectivit(Tan/Cu)
1	294	14.3	4
…	…	…	…
5	437	*missing*	4.9
…	…	…	…
16	*missing*	8.76	5.2

步驟二、去除極偏值。

標準化公式請參考公式 3-1，在這裡我們以第一筆資料為例：

標準化值=(294-430.7058824)/417.8460815

= -0.327168037 (因為-3<-0.327168037<3，所以該筆資料保留不去除)

步驟三、給予遺漏值初始值。

平均值公式請參閱公式 3-2，求得之平均值如表 4-6。

表 4-6　工業製程實驗參數之各屬性平均值

	RR	NU	TAN/Cu
mean	430.7059	12.7347	4.8176

步驟四、利用 SOM 進行學習分群，檢視分群後結果，判斷條件：「各分群內，除了含遺漏值的資料外，必須含有兩筆(含)以上的資料。」

由於現在是第一次進行學習，因此利用步驟三所求得之值當作遺漏值的初始值，經回填原遺漏值後，輸入至 SOM 進行學習分群。(操作視窗如圖 4-1、4-2、4-3)。

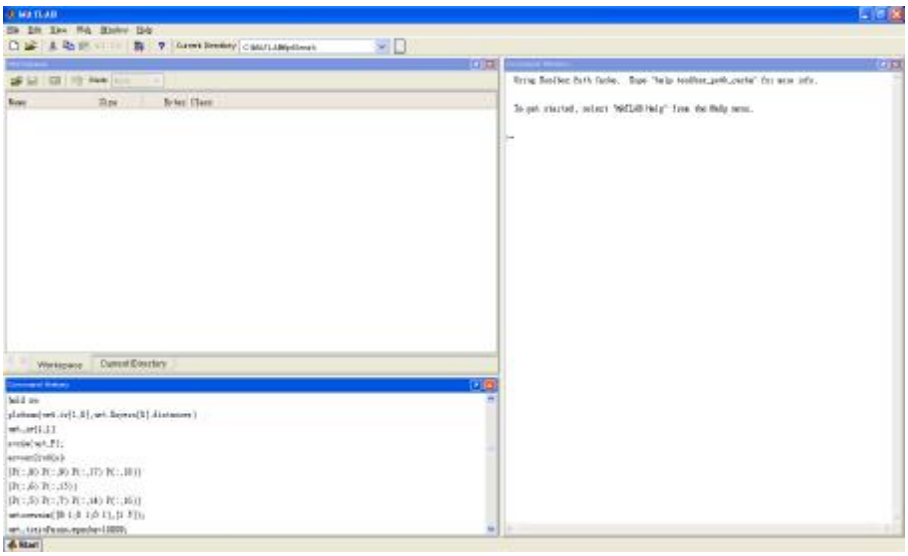

圖 4-1　Matlab設定SOM起始視窗[3,14]

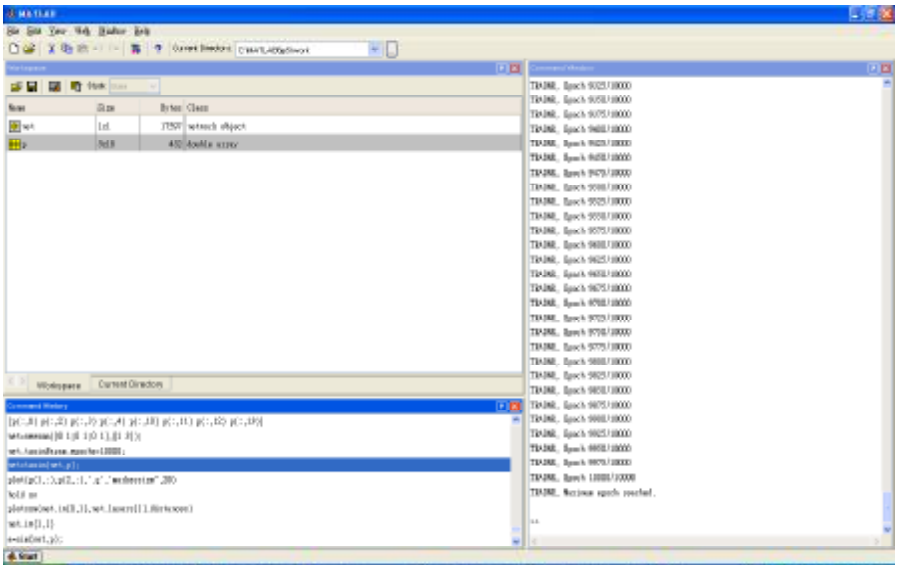

圖 4-2　SOM分群過程[3,14]

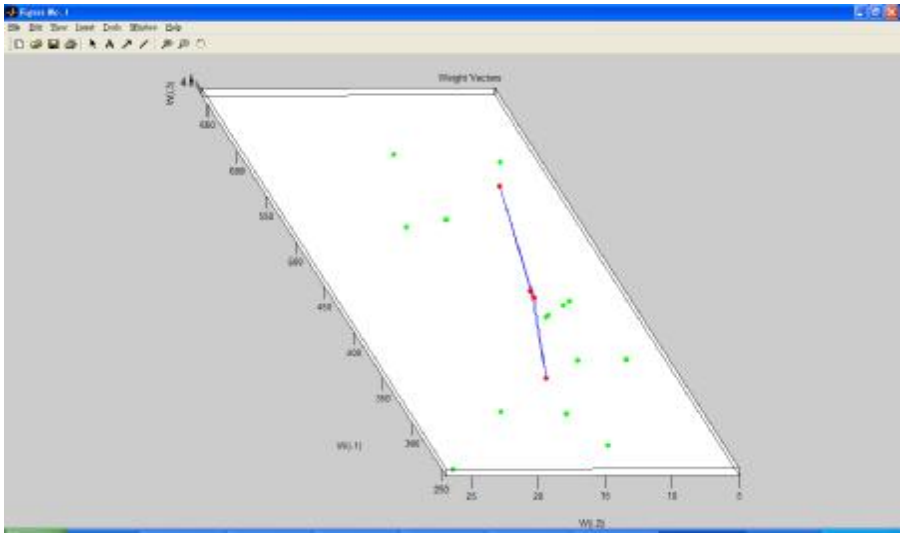

圖 4-3　SOM分群後的工業製程實驗參數之向量端點圖[3,14]

經檢視後，發現當 SOM 劃分成四群的情形時，第一個分群並不符合該限制條件，因此將不考慮劃分成四群的情形。

步驟五、檢視 SOM 的輸出值，過濾原先含遺漏值的資料，並計算該遺漏值所屬的群集之平均值，如果求出的估計值和輸入前的估計值相等時，則結束演算程序，決定最適切之遺漏估計值；反之，將前一步驟學習分群後所求得之估計值回填原遺漏值，並回到步驟四。

例如:第五筆資料經 SOM 分於第二群，第二群扣除原遺漏資料後的平均數為 7.603 並且取代第五筆 NU 欄位的值，由於和輸入至 SOM 進行學習分群前的值不相同(學習分群前：12.7347)，因此將繼續輸入至 SOM 進行學期分群，直到學習分群前後的值相同為止。

決定最後的估計值結果，我們計算各分群內的歐氏距離值，在此採用其值為最小的分群(歐氏距離公式請參閱公式 2-4，各群歐氏距離值見表 4-7)。

表 4-7　SOM應用於工業製程實驗參數資料時各群之歐氏距離值

3c	4c	5c	6c	7c	8c	9c
40.79	31.23	52.96	42.35	59.06	46.73	210.07

由於分成四群時，有筆遺漏值單獨自成一群，所以分成四群時是屬於不適切的分群數，因此不納入考量，所以我們取歐氏距離值最小的三群(如表4-7所示)，在確定最好的結構後，接著將遺漏值的最適推估值予以計算出來，結果如表4-8所示。

表 4-8　經NBEMS所估計出的工業製程實驗參數遺漏值

RR	NU	Tan/Cu
450.6667	7.6033	4.4857

二、各種方法於工業製程實驗參數之實驗比較

本段將以前三部分的實驗結果為基礎，進行相互比較及討論。

1.評估指標

本研究主要是提出處理遺漏值的方法，經文獻整理後找出可適切用來衡量本方法和傳統方法的優缺。

(1)相關係數

相關係數是一個無單位的數值，因此可以衡量任何兩個不同單位及性質的隨機變數間之線性相關方向及程度大小[9]，其公式定義如下：

設隨機變數 X、Y 之聯合機率分配為 f(x,y)，若 mx、my、s^2x 、 s^2y 及 Cov(X,Y)均存在，則 X 與 Y 之相關係數為

$$rx,y = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{sx,y}{sxsy} \tag{4-1}$$

(2)平均絕對誤差(mean absolute error, MAE)

由 MAE 值之大小，可瞭解預測值與實際值之離散程度。其值越小代表模式之離散程度越小，其效果亦較佳[8,28]。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|P_i - O_i\right| \tag{4-2}$$

其中 N 為多少筆資料，P_i 為實際值，O_i 為預測值。

2.推估值的比較

利用平均值、k-means 演算法和我們的方法在工業製程實驗參數的推估值之相關係數比較如表 4-9 和圖 4-4、MAE 的比較如表 4-10、真實實驗數據如表 4-11。整體來說以平均值直接取代的效果最差，而 k-means 演算法所估計出的回填值其次，最接近原始資料的回填值則是由我們的方法所估計出來的。

表 4-9 各方法與工業製程實驗參數原始資料結構之相關係數比較表

	RR	Nu	Tan/cu
對平均值	0.9955	0.9871	0.9706
對 k-means	0.9953	0.9995	0.9845
對 NBEMS	0.9983	0.9991	0.9918

註：陰影部分為最接近 real data

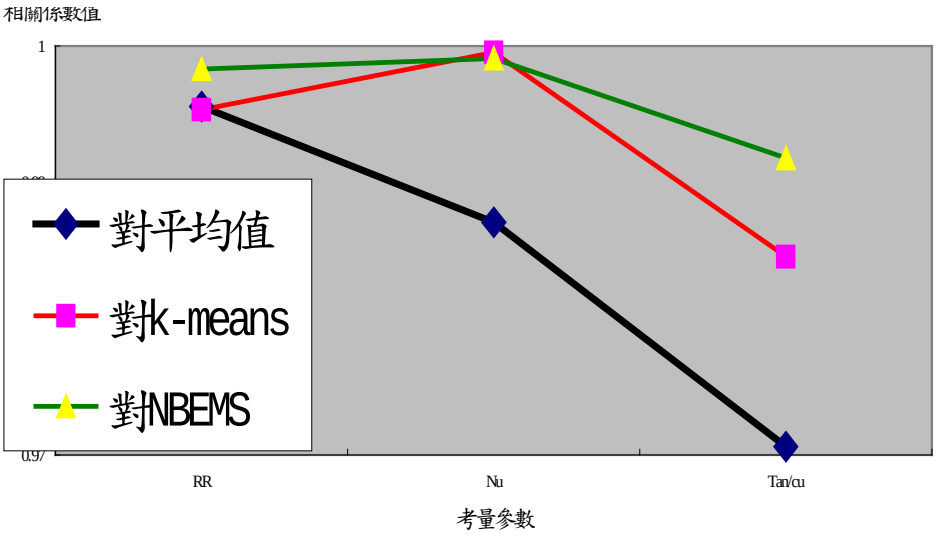

圖 4-4 各方法與工業製程實驗參數原始資料結構之相關係數比較圖

表 4-10 各方法與工業製程實驗參數原始資料結構之 MAE 值比較

	RR	NU	Tan/Cu
平均值	52.2941	4.0347	0.7176
k-means	53.5000	0.7863	0.5250
NBEMS	32.3333	1.0967	0.3857

註：陰影部分為最接近 real data

表 4-11 各方法與工業製程實驗參數原始資料之比較

	RR	NU	Tan/Cu
Real data	483.0000	8.7000	4.1000
平均值	430.7059	12.7347	4.8176
k-means	429.5000	9.4863	4.6250
NBEMS	450.6667	7.6033	4.4857

註：陰影部分為最接近 real data

第二節　估計行銷資料庫模式之遺漏回填值

本節將利用行銷資料庫模式的資料來進行遺漏值回填的實驗，利用傳統和我們的方法分別進行遺漏值的估計，並在最後進行比較。

一、個案描述

在進行實驗之前，我們先簡單介紹什麼是 RFM Model？在行銷管理或顧客關係管理的許多研究中，發展出相當多的模式或衡量標準來提供企業有效的工具，來幫助決策者瞭解其顧客，進而制訂適

切的行銷策略與方法。而在這許多模式方法中，RFM Model 是其中相當著名的方法。由於顧客的消費行為是相當抽象且難以衡量的，因此必須透過一些可衡量的指標性參數來進行分析與衡量。而 RFM Model 最大的特徵就是量化顧客的消費行為，並分析且衡量這些數據，而其中最近購買時間(**R**ecency)、購買頻率(**F**requency)與購買金額(**M**onetary)三個參數，就是一個具參考性且指標性的參數，簡稱為 RFM。

分析顧客過去的購買行為以判斷那些顧客值得進一步的接觸是發展行銷策略上重要的議題[11,16,26,31]。不同特性的顧客適合不同的行銷策略，因此，顧客分群是客製化行銷策略決策上首要的工作。透過 RFM 的分析可以量化顧客消費行為並且衡量顧客忠誠度和貢獻度，以利顧客分群及目標客戶的鎖定。R 值愈小，則隱含著該顧客再次選購此產品的購買程度愈高；反之，R 值愈大，隨著時間之拉長該顧客之持續購買慾隨之降低，則表示著此顧客的購買行為可能改變或是變節至他處消費。F 值主要是在測量顧客在此時間內與公司之互動程度，若 F 值愈高則代表此顧客與公司互動程度愈高；該顧客對此產品有愈高的熱衷程度，持續購買之動機亦較強。M 值主要是代表顧客對此產品之興趣指標，也是對企業之實質金錢貢獻。某顧客的 M 值愈多，代表該顧客大量購買此產品，對此商品具有大量之需求。

本部分所使用的個案資料為 Microsoft SQL Server 2000 中所提供的範例資料庫—Food Mart 在 1997 和 1998 年的交易資料，我們隨

機選擇其中 500 筆資料並計算出其 RFM 分數，且假設三個屬性分別有 5%的 missing data(如表 4-12)來做進一步的實驗分析。

表 4-12 原始 RFM Model 資料型態(以前 6 筆資料為例)

第 n 筆	R	F	M
1	365	18	231.22
2	365	2	21.45
3	327	33	649.73
4	365	2	21.45
5	365	3	***Missing data***
6	253	13	448.87
…	…	…	…

(一)利用平均值來估計遺漏值

將遺漏資料欄位給標記出後，利用公式 3-2 計算各變數的平均值(扣除我們假設的遺漏值)，其結果如表 4-13

表 4-13 利用平均值方法所估計的行銷資料庫模式之遺漏回填值

	R	F	M
推估值	226.2372	18.3504	387.4545

(二)利用 k-means 來估計遺漏值

由於前部分我們已簡單介紹 k-means 演算法，因此在這裡我們不再重複介紹，我們直接利用相同的個案資料進行實驗，實驗出的分群結果見表 4-14，估計出的遺漏回填值見表 4-15。

表 4-14 經 k-means 演算法所求的行銷資料庫模式資料分群結果

第 n 群	群內各筆資料
1	157 258 295 337 462 479
2	2 4 7 9 11 12 13 14 15 16 18 19 22 27 29 31 32 33 34 35 37 38 40 46 47 51 53 54 56 58 61 62 63 66 67 68 69 70 72 73 78 82 84 86 88 90 92 94 95 98 99 101 102 106 107 108 110 112 113 115 121 122 123 128 129 130 131 133 137 141 145 146 148 149 150 152 154 155 165 166 167 171 172 173 174 175 182 183 185 188 193 194 196 197 198 200 201 204 205 207 210 212 213 218 221 225 228 229 230 231 234 235 237 245 247 249 251 252 254 257 260 261 262 264 267 268 269 273 274 276 280 281 286 287 288 289 292 296 299 300 304 305 306 309 310 313 315 320 322 323 327 328 330 334 335 338 339 343 344 345 347 352 353 354 355 358 362 363 364 367 368 370 371 374 375 378 379 381 382 383 385 388 389 391 398 399 405 406 407 408 409 410 411 413 414 415 416 421 422 423 424 426 428 429 432 436 437 438 439 446 447 450 453 454 455 461 463 464 465 468 471 477 478 482 486 489 490
3	1 3 5 6 10 17 21 23 24 25 26 28 30 36 41 42 43 44 45 49 50 52 55 57 60 64 65 74 75 77 79 80 81 83 85 87 89 91 93 96 97 100 104 109 111 114 118 119 120 124 125 126 127 132 135 136 138 139 140 142 143 144 147 156 159 160 161 162 163 164 169 170 178 179 181 184 190 191 192 195 202 203 208 211 216 217 219 220 222 223 226 227 232 233 236 238 239 241 242 243 244 248 250 253 255 256 263 265 266 270 271 272 275 278 279 282 284 285 290 291 294 297 298 301 302 303 307 308 311 314 316 317 319 321 325 326 329 336 341 346 349 356 357 359 361 365 366 369 373 377 380 384 390 392 394 396 397 400 401 402 403 404 412 417 418 419 420 427 431 433 435 440 442 445 448 451 452 457 459 460 466 470 473 474 475 480 483 484 485 487 488 492 493
4	39 76 103 134 168 180 189 206 215 277 283 393 395 443 449 467
5	8 20 48 59 71 105 116 117 151 153 158 176 177 186 187 199 209 214 224 240 246 259 293 312 318 324 331 332 333 340 342 348 350 351 360 372 376 386 387 430 434 441 444 456 458 469
6	425 472 476 481 491

表 4-15 k-means 估計出的行銷資料庫模式資料之遺漏值平均

R	F	M
218.7755	15.42157	452.5211

(三)以 NBEMS 估計行銷資料庫模式之遺漏回填值

本節中，主要應用我們所發展出來的一套估計遺漏值的演算程序來進行行銷資料庫模式之遺漏值的推估。

步驟一、找出遺漏值。

檢視原始資料並標記出遺漏值，如表 4-16 所示。

表4-16 行銷資料庫模式部分原始資料型態

第 n 筆	R	F	M
1	365	18	231.22
2	365	2	21.45
3	327	33	649.73
4	365	2	21.45
5	365	3	***Missing data***
6	253	13	448.87
…	…	…	…

步驟二、去除極偏值。

計算各筆資料標準化值，檢視並去除偏值。標準化公式如公式 3-1。

例如: 第一筆資料標準化值= (365-224.73)/122.38=1.1461

因為-3<1.1461<3，所以該筆資料保留不去除。

步驟三、給予遺漏值初始值。

計算原始資料各屬性欄位的平均值(不含已標記含遺漏值的資料)，平均值公式請參閱公式 3-2，求得結果如表 4-17 所示，取代後如表 4-18。

表 4-17 行銷資料庫模式之各屬性平均值

	R	F	M
mean	226.2372	18.3504	387.4545

表 4-18 行銷資料庫模式資料求得之平均值取代原遺漏資料欄位

第 n 筆	R	F	M
1	365	18	231.22
2	365	2	21.45
3	327	33	649.73
4	365	2	21.45
5	365	3	387.4545
6	253	13	448.87
7	365	2	21.45
8	102	43	796.45
…	…	…	…

註 1：以前 8 筆資料為例 註 2：灰色部分為原遺漏資料欄位

步驟四、利用 SOM 進行學習分群，檢視分群後結果，判斷條件：「各分群內，除了含遺漏值的資料外，必須含有兩筆(含)以上的資料。」

將含有遺漏值的資料(將步驟三所求得之值取代原已標記的遺漏值欄位，如表 4-18 所示)的原始資料輸入至 SOM 進行學習分群，分群過程如圖 4-5 所示。

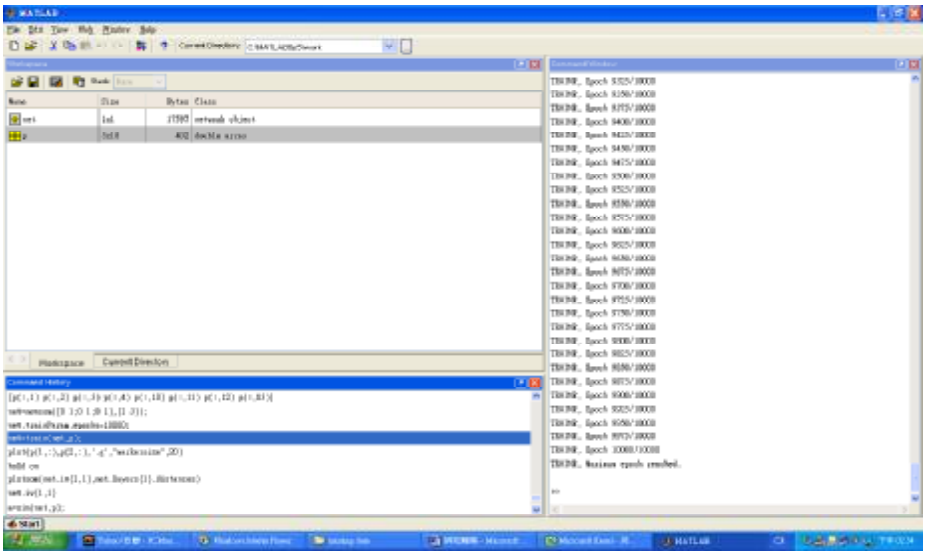

圖 4-5 SOM 學習訓練行銷資料庫模式資料的分群過程[3,14]

經檢視後，發現當 SOM 劃分成十群以上的情形時，皆有一個分群並不符合該限制條件，因此將不考慮劃分成十群以上的情形。

步驟五、檢視 SOM 的輸出值，過濾原先含遺漏值的資料，並計算該遺漏值所屬的群集之平均值，如果求出的估計值和輸入前的估計值相等時，則結束演算程序，決定最適切之遺漏估計值；反之，將前一步驟學習分群後所求得之估計值回填原遺漏值，並回到步驟四。

例如:第五筆資料經 SOM 分於第三群，第三群扣除原遺漏值後的平均數為 400.5012 並且取代第五筆 M 欄位的值，由於和輸入至 SOM 進行學習分群前的值不相同(學習分群前：387.4545)，因此將繼

續輸入至 SOM 進行學期分群，直到學習分群前後的值相同為止。圖 4-6 為經 SOM 學習分群後的端點向量圖。

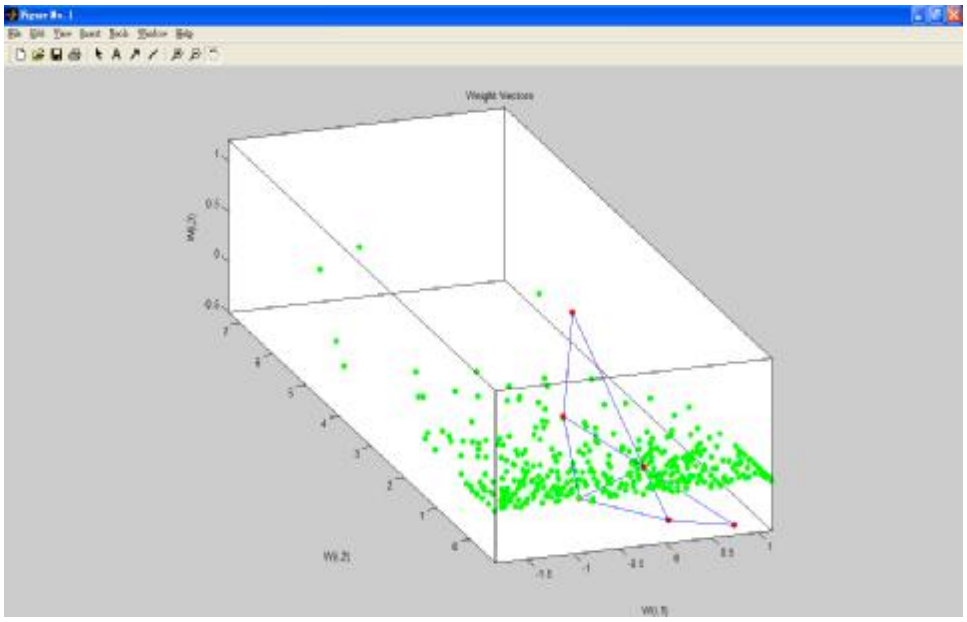

圖 4-6 SOM分群後的行銷資料庫模式資料之端點向量圖[3,14]

決定最後的估計值結果，我們計算各分群內的歐氏距離值，在此採用其值為最小的分群(歐氏距離公式請參閱公式2-4，各群歐氏距離值見表4-19)。

表 4-19 行銷資料庫模式資料各分群之歐氏距離值

分群數	3c	4c	5c	6c	7c	8c	9c	10c
歐氏距離值	1.01	1.55	1.47	0.80	1.26	1.02	1.98	1.16

由表4-19我們可以得知分成六群時，其歐氏距離值為最小(陰影部分)，因此本實驗選擇使用分成六群的結果。在確定選擇的群數後，接著我們再將各群所獲得的遺漏推估值計算出來，本個案的遺漏推估值如表4-20所示。

表4-20 經NBEMS所估計出的行銷資料庫模式資料之遺漏回填值

	R	F	M
1c	343.16	3.39	*
2c	138.19	9.17	*
3c	286.8	16.79	399.84
4c	84.17	18.29	348.15
5c	172.95	52.25	*
6c	191.73	27.74	*

註：標記*號者為該分群內並無屬於該屬性的遺漏資料值

二、各種方法於行銷資料庫模式之實驗比較

本部分和前面一樣，將比較傳統方法與我們的方法所估計出的遺漏資料回填值之實驗結果。

1.評估指標

和前面一樣，我們使用二個指標(相關係數，請參閱公式 4-1、MAE，請參閱公式 4-2)來評估各方法的效果。

2.推估值的比較

利用平均值、k-means 演算法和我們的方法在工業製程實驗參數的推估值之相關係數比較如表 4-21 和圖 4-7、MAE 的比較如表 4-22、真實實驗數據如表 4-23；整體來說以我們的方法所估計出的回填值透過相關係數與 MAE 指標的比較後，其效果皆優於平均值直接取代法和 k-means 演算法。

表 4-21 各方法與 RFM Model 資料結構之相關係數比較表

	R	F	M
平均值	0.9896	0.9757	0.9995
k-means	0.9880	0.9816	0.9998
NBEMS	0.9913	0.9931	0.9994

註：陰影部分為最接近 real data

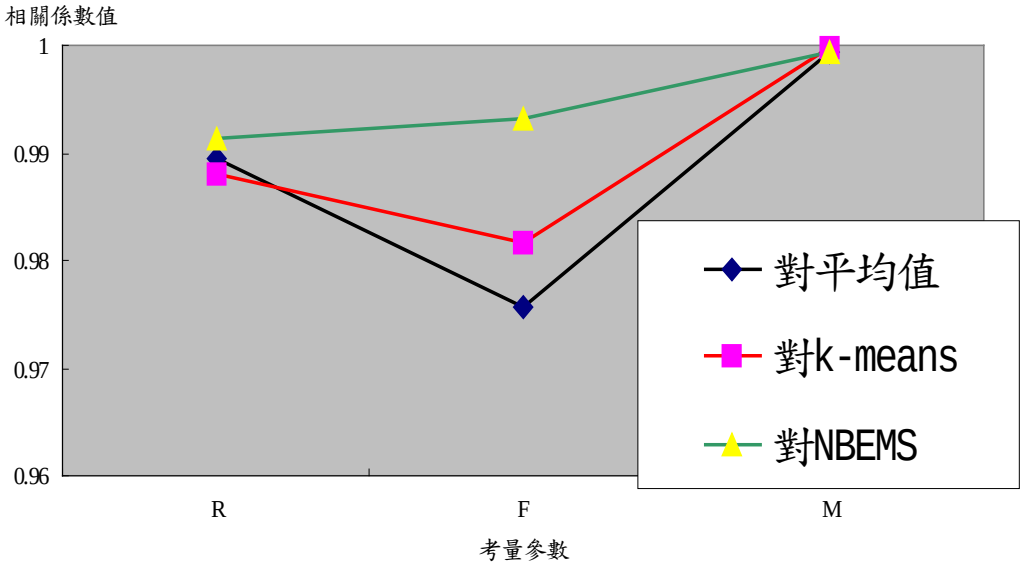

圖 4-7 各方法與 RFM Model 資料結構之相關係數比較圖

表 4-22 各方法與 RFM Model 資料結構之 MAE 值比較

	R	F	M
平均值	32.7228	6.0496	36.5409
k-means	40.1845	8.9784	101.6075
NBEMS	12.0548	6.0270	26.1828

註：陰影部分為最接近 real data

表 4-23　各方法與 RFM Model 原始資料之比較(平均值)

	R	F	M
Real data	258.96	24.40	350.91
平均值	226.24	18.35	387.45
k-means	218.78	15.42	452.52
NBEMS	246.91	18.37	377.10

註：陰影部分為最接近 real data 值

第五章　結論與未來展望

隨著資訊科技的快速發展，資料處理速度與效能也越來越佳，如何應用現有的資料挖掘出有價值的資訊來提升企業本身的競爭優勢，更是日前所有企業相當重視的課題之一，有鑑於此，智慧型資料探勘技術便逐漸被人們所注意起來。

在資料探勘的過程中，由於種種的因素可能會造成資料或資訊遺漏的情形，直接使用該資料進行資料挖掘動作時，容易挖掘出不正確的知識，倘若刪除不用，又會造成資料量的縮減，而導致挖掘出的知識也有所偏差，因此在現今資料探勘技術逐漸重視的時刻，如何處理、估計遺漏值便顯得越來越重要。

本研究以「物以類聚」群組特性原理為出發點來思考，嘗試發展出一套以類神經網路為基礎的估計遺漏值演算程序(NBEMS)，並且分別以兩個個案實例來驗證本演算程序的可行性與合理性，透過我們的方法所找出的遺漏值其精準度皆優於較直覺式的平均值取代方法與 k-means 演算法方式，而除此之外，資料經回填後，整體資料的型態與結構皆和真實資料相異不遠，這結果可使使用者在進行資料挖掘的動作時，能保有最大的資訊量，以期許挖掘出資訊能更為適切、更有意義。

由於受到人力、時間、資料來源等種種因素的限制下，使得本研究初步僅針對 5%的資料遺漏情形下進行實驗，未來可嘗試不同的資料種類、型態、格式和含遺漏資料的程度的不同來做進一步的實驗與後續的討論。再者，也可以嘗試其他不同的技術的分群方法，例如 Fuzzy C-means，或是不同的資料處理程序、不同的類神經網路

模式等來進行本研究的延展及後續的探討，或許能有更多意外的發現。

參考文獻

一、中文部分：

[1] 林俊男、謝昆霖，「應用通用型類神經網路於遺漏值問題之解析」，2004 商業現代化學術研討會，高雄，民國 93 年 12 月 18 日。

[2] 侯瑞芳，「知識管理整合性研析於永續發展之應用」，南華大學資訊管理研究所碩士論文，民國 93 年 6 月。

[3] 周鵬程，類神經網路入門，台北，全華科技圖書，民國九十三年。

[4] 葉怡成，應用類神經網路，台北，儒林圖書，民國九十年。

[5] 葉怡成，類神經網路模式應用與實作，台北，儒林圖書，民國九十年。

[6] 唐麗英、王春和，「應用灰色關聯分析法於動態系統具多品質特性製程最佳化之研究」，工業工程學刊，第十七卷第二期，147~156 頁，89 年。

[7] 陳銘宗，「組合式關聯法則應用於缺值問題之研究」，朝陽科技大學資訊管理研究所碩士論文，民國 80 年 6 月。

[8] 陳莉、江柏寬、蔡家盛，「進化演算法應用於流量預測之研究」，第七屆人工智慧與應用研討會論文集，746~751 頁，台中，民國 91 年 11 月 15 日。

[9] 程大器，統計學理論與應用上，台北，智聖文化，民國九十年。

[10] 張紹勳、張紹評、林秀娟，SPSS For Windows 多變量統計分析，台北，松崗電腦，民國八十九年。

[11] 張晉赫、謝文天、蔡佳偉、邱宏彬，「運用資料探勘技術進行一對一行銷之電子商務推薦系統」，2003 企業管理學術研討會暨 2003 電子商務經營管理研討會，台中，民國 92 年 12 月 13 日。

[12] 薛芳宜、蔣志堅、謝昆霖，「環境永續知識管理系統應用：以資料探勘技術為探討」，大葉大學研討會，彰化，民國 93 年。

[13] 顏月珠，統計學，台北，三民書局，民國八十七年。

[14] 羅華強，類神經網路—Matlab 的應用，新竹，清蔚科技，民國九十年。

[15] 蘇建源，「模糊邏輯與資料探勘技術為基礎在顧客關係管理上之研究與應用」，南華大學資訊管理研究所碩士論文，民國 93 年 6 月。

[16] 蘇建源、邱美倫、邱宏彬、吳光閔，「應用資料探勘技術支援顧客導向影片檢索及推薦之智慧型人機介面」，2003 企業管理學術研討會暨 2003 電子商務經營管理研討會，台中，民國 92 年 12 月 13 日。

[17] 蘇木春、張孝德，機器學習類神經網路、模糊系統以及基因演算法則，台北，全華科技圖書，民國八十六年。

二、英文部分：

[18] Aleksander, I., H. Morton, An Introduction to Neural Computing, Chapman and Hall, New York, NY, 1990.

[19] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *Proc. of the 20th Conf.* on Very Large Databases, pp.487-499,1994.

[20] R. Agrawal and R. Srikant, "Mining Sequential Patterns," *Proc. of the 11th Conf. on Data Engineering,* pp.3-14, 1995.

[21] M. Berry and G. Linoff, Data Mining Techniques for marketing, sales, and Customer Support, New York. Wiley Computer Publishing,1997.

[22] M.S. Chen, J. Han, and P.S. Yu, "Data Mining: An overview from a database perspective ," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 8, No. 6, pp.866-883, 1996.

[23] M. S. Chen, J. S. Park, and P. S. Yu, "Efficient data mining for path traversal patterns," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 10, No. 2, pp.209-221,1998.

[24] U. M. Fayyad, "Data mining and knowledge discovery: making sense out of data,"*IEEE Expert,* Vol. 11, No. 5, pp.20-25,1996.

[25] F. Hair, E. Anderson, L. Tatham and C. Black, *Multivariate Data Analysis*, Fifth Edition, Prentice Hall, pp.56-64, 1998.

[26] Yi-Chung Hu , Ruey-Shun Chen, and Gwo-HshiungTzeng, "Discovering fuzzy association rules using fuzzy partition methods", *Journal of Knowledge-Based Systems,* Vol.16,pp.137-147,2003.

[27] J. Han and M. Kamber, *Data Mining : Concepts and Techniques*, Morgan Kaufmann Pub, pp.105-251, 2000.

[28] Heikki Junninen, Harri Niska, Kari Tupporainen, Juhani Ruuskanen and Mikko Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmospheric Environment*,38,2895-2907, 2004

[29] C. Kleissner, "Data mining for the enterprise," *Proc. of the Thirty-First Hawaii International Conference,* Vol. 7, pp. 295-304, 1998.

[30] Lu H., R. Setiono, and H. Liu, "Effective Data Mining Using Neural Network," *IEEE Transactions on Knowledge and Data Engineering,* Vol.8, No.6, Dec. 1996.

[31] Shaw Michael J , Chandrasekar Subramaniam ,Gek Woo Tan, and Michael E. Welge, "Knowledge Management and Data Mining for Marketing ,"*Journal of Decision Support Systems,*Vol.31,pp.127-137, 2001.

[32] Markey M. K., J. Y. Lo, G. D. Tourassi, and C. E. Floyd, Jr., "Self-organizing map for cluster analysis of a breast cancer database," *Artificial Intelligence in Medicine*, Vol.27, pp.113-127. 2003.

[33] A. Ragel and B. Cremilleux,"Treatment of Missing Values for Association Rules," *Proceeding of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-98),* Melbourne, Australia, pp.258-270, 1998.

[34] H. Toivonen, "Sampling large databases for association rules," *Proc. of the 22nd Conf.* on VLDB, pp.134-145,1996.