朝陽科技大學
資訊管理系

碩士論文

部落格探勘—以網路電話產品為例

Blog mining：An example of VoIP phone products

指導教授：陳隆昇 博士

研 究 生：杜逸普

中華民國九十七年六月六日

朝陽科技大學資訊管理系

Department of Information Management

Chaoyang University of Technology

碩士論文

Thesis for the Degree of Master

部落格探勘—以網路電話產品為例

Blog mining：An example of VoIP phone products

指導教授：陳 隆 昇 博士 (Long-Sheng Chen)

研 究 生：杜 逸 普 (Yih-Puu Duh)

中華民國九十七年六月六日

June 6, 2008

摘要

部落格（Blog）是一種由使用者編輯、更新以及維護的網站，近年來變成一種新奇且受歡迎的媒體，特別是新世代的網際網路使用者而言。在 2007 年 1 月份利用部落格搜尋引擎 Technorati 搜尋，我們可以搜尋到全球約有六千萬個部落格。因此，部落格可以被視為一種新的行銷平台，近期也有越來越多的企業嘗試在部落格上，挖掘有利於商業目的的資訊以及知識。所以，本研究最主要的目的，是提出一個創新的部落格探勘模組（Blog Mining Model）來擷取部落格上的知識。

本研究提出的部落格模組有兩個目的：1. 我們利用自我組織特徵映射圖（Self-organizing Map, SOM），來區隔市場。2. 以隱含語意索引（Latent Semantic Indexing, LSI）方法為基礎，利用倒傳遞類神經網路（Back-propagation Neural Networks, BPN）、支撐向量機（Suppot Vector Machines, SVM）以及決策樹（Decision Tree）三種分類方法，來發展稀疏性資料分類器，並且預測新進客戶所屬的分類。最後，我們實際在部落格上找尋討論網路電話產品的文章，來說明我們所提的方法，並證明其優越性。

關鍵字：部落格、資料探勘、自我組織特徵映射圖、分類、隱含語意索引

Abstract

A blog is a user-generated website and becomes a new and popular media, especially for new generation of Internet users. Depending on the blog search engine, Technorati, we can track about 60 million blogs in January 2007. Therefore, blogs are viewd as new marketing channels today. Recently, more and more companies attempt to discover useful information and knowledge for business purposes. Therefore, the major objective of this study is to present a new structure which is called “Blog Mining Model” for extracting knowledge from blogs.

There are two purposes in our proposed model. Firstly, we segment market segmentation by using Self-organizing Map (SOM) neural network. Secondly, for sparse text data, we propose a Latent Semantic Indexing (LSI) based classifier whose learning methods include Back-Propagation Neural Networks (BPN), Support Vector Machines (SVM), and Decision Trees, have been developed for classifying sparse data. These built classifiers can be used to predict category of segmentation of new customers. Finally, a real case regarding VoIP (Voice over Internet Protocol) phone product will be provided to illustrate the superiority of our proposed methodology.

Keywords：Blog, Data Mining, Self-organizing Map, Classification, Latent Semantic Indexing.

誌謝

首先，感謝我的指導教授陳隆昇博士，在我研究所兩年之中，亦師亦友的照顧我。當我研究遇到瓶頸，他總是給予我最大的協助與指導，讓我的研究進度得以繼續向前。同時也特別感謝陳穆臻教授及許俊欽教授，在百忙之中撥冗擔任學生口試委員，給予諸多寶貴意見，讓本論文更完整充實。

其次我要感謝台中市育仁小學校長施麗蘭修女。2006 年初，我適逢親人離世，在這兩年多的日子，施修女每星期五以天主的話語勉勵我，讓我走出悲傷，用更樂觀的態度去看待這個美好的世界。

感謝 M315 的阿旺學長、丸子學姊於我剛入學時，給予我最大的關心及協助；感謝 M317 的學長姊、同學、及學妹，給予我快樂的兩年研究生涯；感謝麗陽學院王大哥、王大嫂以及柴犬 Lucky 給予我居住上的協助；感謝我女友這些年對我的包容、諒解及照顧。

最後要感謝的是我父母，是他們窮自己也不能窮小孩的態度，辛苦栽培十個小孩，才能讓我拿到全家的第七個碩士學位，我的一點一滴，都要感謝我的父母及家人。最後，僅以此文獻給在天國的父親及五哥。願天主的平安與喜樂，常與大家同在！

杜逸普謹識

中華民國九十七年六月六日

目錄

表 目 錄

圖 目 錄

第一章 緒論

1.1 研究背景

部落格（Blog）一詞，是由 Web log 所衍生而來，原本指的是網路伺服器所記錄的訪客記錄檔。後來 John Barger（1997）提出「Weblog」一詞，代表使用網頁作為媒介，呈現個人記錄。最後 Peter Merholz（1999)將 Weblog 分開唸為「We Blog」，意思是讓我們一起來寫部落格，Blog 變成了動詞，指的是一種在網路上發表感想或是撰寫日誌的行為，每個人都可以在網路上，自由的發表自己的言論，而撰寫或經營部落格的人，便稱為部落客（Blogger）或博客。由於部落格具備了資料的真實性、資訊的即時性，以及使用操作上的便利性，使得部落格在近幾年內快速成長，成為一種新興的傳播媒體。

在台灣，部落格的使用情形從 2005 年約有 61.2%的網友擁有自己的部落格（蕃薯藤[1]，2005），到 2006 年有 8 成的網友已使用過部落格，其中閱讀他人部落格有 69%，略多於自己編輯的 55%（資策會[2]，2006），2007 年則有 93.8%的網友瀏覽過部落格，有 66.6%的網友擁有自己的部落格（波特仕[3]，2007）。在全球，部落格的使用數量於 2007 年 4 月也已突破 7200 萬個，是 2005 年 3 月的 9 倍（Technorati[4]，2007）。從這些數據，我們可以發現，

[1] 2005 年蕃薯藤網路調查報告：http://survey.yam.com/survey2005/index2.html
[2] 資策會資訊市場情報中心：http://mic.iii.org.tw/intelligence/
[3] 波特仕線上市調網：http://www.pollster.com.tw/report/20070720/index.htm
[4] Blog 搜尋引擎：http://technorati.com

部落格在台灣已經是新興的一股科技熱潮，在全球更是不容忽視的力量。

部落格不僅是個人抒發想法及表達意見的一種新主流，上百萬人可以透過鏈結其他人的部落格，進而公開或者交換知識及資訊，並且在部落格的世界建立網路或關聯，稱為部落格圈（Blogosphere）（Rosenbloom, 2004）。微軟（Microsoft）董事長比爾・蓋茲（Bill Gates）曾說，部落格是繼電子郵件（E-Mail）、電子佈告欄（Bulletin Board System, BBS）與即時訊息（Instant Messaging, IM）之後，第四個改變世界的網路應用（陳穆臻，2005）。

此外，部落格在眾多網路媒體中脫穎而出，是基於他的四零優點：「零技術、零成本、零編輯、零形式」，與六大特性：「強調個人主義、具有時間順序、形成小眾市場、作者與讀者的互動、相互連結性、分眾化的閱讀」（傅大煜，2005）。因為四零優點，無論是個人使用者抑或企業經營者，都能以極低的成本，輕易的將部落格經營得非常出色；因為六大特性，瀏覽部落格的讀者，能夠即時的閱讀或利用 RSS[5]訂閱讀者所需的資訊，克服網路世代中資訊爆炸的窘境，而讀者於閱讀後亦能對部落格經營者寫下回應，或者是透過部落格的相互連結性，做更多的延伸閱讀。

然而部落格的使用，除了受到一般的網友的喜愛之外，更有許多企業利用部落格來從事商業活動，如通用汽車（General Motors）成立一個官方部

[5] RSS：RSS 是一種用於網上新聞頻道、網誌和其他 Web 內容的數據交換規範，起源於網景通訊公司（Netscape）的推送技術（push technology），將訂戶訂閱的內容傳送給他們的通訊協同格式（Protocol）。資料來源：維基網路百科全書 http://zh.wikipedia.org/wiki/RSS

落格— Fastlane[6]，讓消費者與企業主管在部落格上交流各種汽車的心得與意見。通用汽車可以在第一時間聆聽消費者心聲，一方面降低錯誤策略所帶來的損失，另一方面助於評估未來策略發展的正確方向（陳品均，2006）。

由此可見，部落格從早期以網頁形式呈現個人記錄，或者是網友抒發心情的網路日誌，逐漸的受到企業的重視，成為企業宣傳、廣告及行銷的強力媒體，在資訊經濟的時代中，為企業帶來了新的商機與經營模式。而以往利用個人網頁作為宣傳、廣告及行銷工具等企業，也因為更新或維護網頁耗時費工，以及必須投入較高的資金及技術（戈立秀，2007），逐漸的捨個人網頁不用，而改用部落格作為行銷媒體。

[6] Fastlane Blog：http://fastlane.gmblogs.com/

1.2 研究動機

由於部落格操作及使用上的便利性，近年來廣被網友用來抒發心情或者撰寫日誌，甚至會有瀏覽者透過回應欄與部落客進行對話與交流。這些看起來平凡的對話，往往有許多不平凡的資訊與知識潛藏著。例如[7]：部落客會在自己的部落格上分享買手機的心情，然後說明了手機的價位是多少，功能有相機、MP3、來電鈴聲等，之後對購買的手機寫下自己的評論。這樣的日誌乍看之下平凡無奇，實際上卻有顧客對手機產品的意見，如果手機廠商能夠取得並瞭解顧客在部落格上發表的的消費想法、資訊及知識，對於廠商下一款手機的推出，是有極大的參考價值及幫助。

目前企業最常應用部落格行銷的方式，是將試用品或者產品活動等相關資訊，放在網路部落格上，並告知消費者部落格網址，讓消費者前往該部落格討論（郭芷婷，2005）。而應用部落格行銷最成功的案例則是 Nike 運動用品廠商，於 2004 年設立「速度的藝術」（Art of Speed）部落格，由 15 組散佈全球各地的廣告創意團隊，製作一系列以「速度的藝術」為主題的短片，放置在部落格上，並提供下載。為期 20 天的時間裡，有數萬人下載並將該部落格網址加入自己部落格裡的連結，讓「速度的藝術」在部落格圈裡廣為討論的話題，一改「Just do it！」的品牌形象，並給與消費者感性

[7] 範例資料來源：
http://tw.myblog.yahoo.com/jw!eoXY3BuFREVOmRK0sitT/article?mid=3801&pk=%E7%85%A7%E7%9B%B8%E6%89%8B%E6%A9%9F

與創意的印象（林昭妘，2006）。

由此可見，部落格不論是在個人的使用上，或者是企業的應用上，都富涵著潛在的商業資訊及知識，而透過資料探勘的技術，可以將這些潛在的商業資訊與知識擷取而出，並加以應用。目前資料探勘（Data Mining）的技術應用在網路上，多半是針對網站內容探勘（Web Content Mining）、網站結構探勘（Web Structure Mining）及網站使用探勘（Web Usage Mining）（Facca & Lanzi, 2005）。自從部落格成為新的網路行銷媒體後，部落格資料探勘（Blog Data Mining）也逐漸的受到重視：Kumar et al. （2005）研究在部落格圈裡形成的社群以及社群所帶來的爆炸性改革；Tseng et al. （2005）利用群聚法來對部落格圈裡的社群做分群並將其形象化。不難想見，『部落格探勘（Blog Mining）』將成為未來新的趨勢，並受到重視。

因此，本研究希望建立一個『部落格探勘』的機制，透過這個機制，我們可以將部落格上的文章進行資料探勘，並擷取出隱含在文章裡的商業資訊及知識，作為企業改良產品或推出產品的參考方針。

1.3 研究目的

基於擷取潛在於部落格文章商業資訊與知識的立場，本研究將提出一個『部落格探勘模組（Blog Mining Model）』，來針對部落格上的文章做資料探勘，並且以網路電話（Voice over Internet Protocol, VoIP）產品為研究對象進行探討，擷取出隱含於部落格文章中，關於網路電話的商業資訊以及知識。

本研究所提之『部落格探勘模組』，其擷取知識方法可分為兩階段：第一階段是針對所蒐集的資料，以自我組織特徵映射圖（Self-organization Map, SOM）的群聚技術（Clustering），將撰寫這些部落客所撰寫探討網路電話文章，依照其文章內容不同而區隔成數群，進而找出適當的顧客區隔(Customer Segmentation)。第二階段則是以倒傳遞類神經網路（Back-propagation Neural Network, BPN）、決策樹（Decision Tree）以及支援向量機（Support Vector Machines, SVM）等分類方法，能對部落格的網路電話文章進行顧客分類預測。

最後，透過本研究之部落格探勘模組，對關於網路電話的部落格文章進行資料探勘後，我們預期可以達到下列研究目的：

1. 顧客區隔：透過 SOM 的群聚技術，我們得以將部落格上的文章得以分為數群，然而每一篇部落格的文章代表了每一位發表心得的顧客，所以我們可以依照顧客發表的文章內容，將顧客區隔成數群的網路電話使用

者，供行銷決策使用。

2. 顧客分類：我們可以從顧客的部落格文章，攫取知識，建立分類器，除了幫助企業瞭解顧客對網路電話的需求外，更可將未知的新進顧客分類在適當的消費群，一方面協助顧客選擇適合的網路電話產品，另一方面也協助企業針對不同消費群的客戶，推銷合適的網路電話產品，以滿足顧客所需。
3. 發展處理稀疏資料（Sparse Data）分類器：在本研究提出的部落格探勘模組第二階段中，針對部落格文章的稀疏性(Sparsity)可能會造成分類績效的影響，我們導入『隱含語意索引』（Latent Semantic Indexing, LSI）方法，作為屬性擷取（Feature Extraction）工具，希望能有效的改善空間浪費、影響分類預測準確率等稀疏資料的缺點，進而提高分類預測的正確率。

1.4 研究架構與流程

在確定了以部落格為研究背景，希望建立並提出一個部落格探勘模組為研究動機、目的之後，本研究將在第二章進行隱含語意索引、非監督式學習（Unsupervised）以及監督式學習（Supervised）相關研究之文獻探討，深入的瞭解研究背景的知識，以發展研究架構。經過文獻探討後，第三章我們將設計研究架構與方法，並提出確切的部落格探勘模組，能擷取出潛在部落格文章中的商業資訊與知識。在第四章中，透過蒐集部落格中探討網路電話的文章，實際的操作我們提出的部落格探勘模組。最後我們將在第五章討論整個部落格探勘模組應用於部落格文章中，討論網路電話產品的研究結果。研究流程如下圖 1-1：

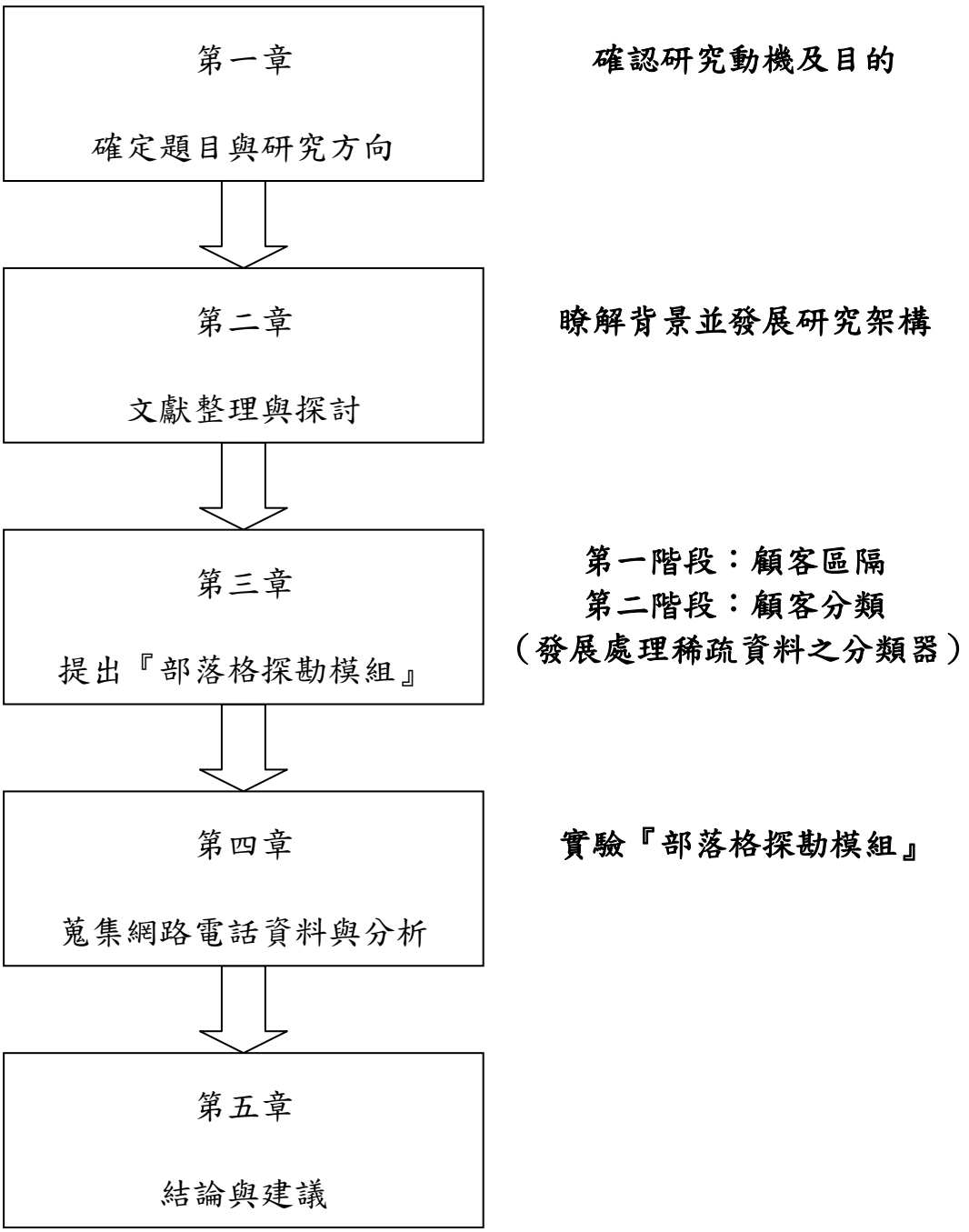


圖 1-1 本研究之研究流程

第二章　文獻探討

本章將針對與研究主題相關之文獻進行探討，如分群方法、分類方法以及屬性擷取等研究文獻，以瞭解相關研究議題之發展現況。

2.1 分群方法簡介

分群，簡單來說，就是將類似的抽象物件或實體聚集，形成數個類別類別叢集的過程，即稱為分群。在同一叢集中，物件與物件彼此的相似度非常高；在不同叢集中，物件與物件彼此的相異性非常高。分群與分類（Classification）的差異，是分群要劃分的類別是未知的，而分類所要劃分的類別是已知的（曾韋榮，2006）。

目前資料分群的應用面相當廣泛，包括資料探勘、統計學、機器學習、空間資料庫技術及市場行銷等，亦可作為資料探勘技術裡的資料前處理步驟（Han & Kamber, 2001）。資料分群的技術有相當多種，本文列出幾種常見的分群法，並整理為表，如表 2-1 所示。

在眾多分群方法中，楊東昌（2004）曾經回顧文獻並於整理後提出，SOM 為一個有效並且方便於觀察的分群工具，特別是針對大量且高維度的資料分析，使用上也不需要先備知識。又 SOM 可針對大量高維度的資料進行視覺化，且 SOM 演算法與網路結構簡單並易於延伸應用，同時又具備資料降維與拓樸保存的特性，自 Kohonen（1982）提出至今，已經成功應用於不

表 2-1 常見的分群技術

方法	簡介
切割法 Partitioning Method	假設全部資料集合有 n 個物件，被分割為 k 個子集，一個子集代表一群，每群至少一個物件，每個物件只能被劃分於一群。分群的規則是依各物件的特徵去劃分到相似的集合，再藉由修正各子集的特徵屬性來調整子集範圍。著名的切割演算法有 K-means 以及 K-medoids。
階層法 Hierarchical Method	資料集合以階層分解的方式，由上而下或者由下而上建立樹狀結構的分群。可分為凝聚（Agglomerative）法或分裂（Divisive）法。
密度基礎法 Density-based Method	當某群集的鄰近區域（Neighborhood）密度高於某門檻值，該群集則會持續擴大。換言之，在所給予的群集中，其鄰近區域半徑的設定必須包含最低限制的資料點。
格子基礎法 Grid-based Method	將物件空間（Object Space）量化為有限區間，如格子般的結構，所有的分群過程，都是在格子結構上進行。
模型基礎法 Model-based Method	假設每個分群都有一個模型，且會找出適合放入該模型的資料。此類方法可分為統計方式（Statistical Approach）以及類神經網路方式（Neural Network Approach）。
類神經網路法 Neural Network Method	將每個群集視為一個樣本（Exemplar），依據某些距離的量測，新物件將會被分配到與新物件樣本最相似的群集中。此類的方法有競爭學習（Competitive Learning）以及 SOM 兩種方法。

（資料來源：整理自 Han & Kamber, 2001；李韋承，2005；曾韋榮，2006）

同領域中，SOM 是一種非監督式學習及競爭式學習的類神經網路，透過將高維度的資料，映射到一個二維的格子拓樸圖上，能夠將這些複雜的高維度資料有效的視覺化（Kohonen, 1998）。在映射的過程中，是利用輸入資料的訓練，並以資料的特徵向量進行訓練，將高維度的資料映射到二維的地圖上，以達到資料維度的縮減。此時，映射於二維拓樸圖上的資料，相似性高的資料將會鄰近在一起；而相異性高的資料則不會鄰近，甚至在二維地圖的對角上，利用 SOM 這樣的特性，可以應用於市場區隔上（Kiang et al., 2006）。圖 2-1 為 SOM 應用在美國電信電報市場的區隔上，依照使用情況將顧客區隔為六大類，可以做為行銷決策支援。也因此，SOM 的映射過程也是一種群聚的過程，亦可視為群聚演算法的一種（張斐章，2003）。

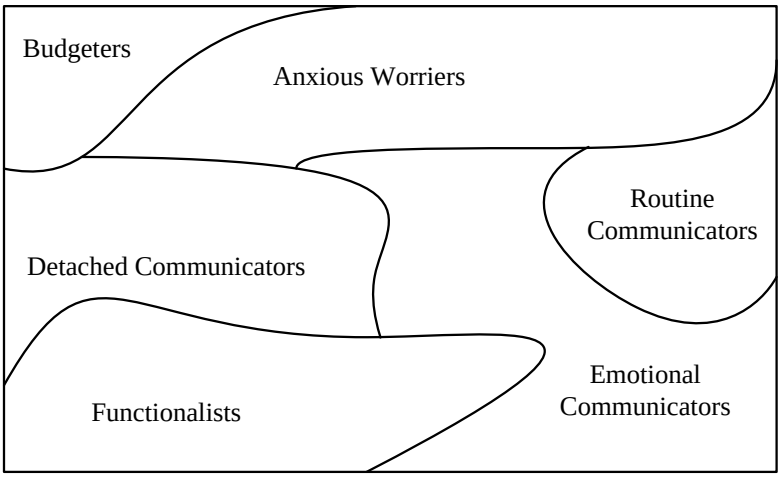


圖 2-1 應用 SOM 區隔美國電信電報市場

（資料來源：修改自 Kiang, 2006）

2.1.1 SOM

SOM，是一種被廣泛應用於分群的類神經網路，由 Kohonen 所提出的非監督式學習的網路模式，亦是一種競爭式學習的神經網路。SOM 演算是以特徵映射的方式，將任意維度的輸入向量，映射到較低維度的特徵映射圖上，形成拓樸層架構的映射圖，如圖 2-2 所示，並且依據目前的輸入向量在神經元之間彼此相互競爭，優勝的神經元可以獲得調整連結權重向量的機會，如圖 2-3 所示。最後輸出層的神經元會依據輸入向量的特徵，以有意義的拓樸架構展現在輸出空間中，並由拓樸結構圖來反應出所有輸入值之間的分佈關係。因此將此網路稱為自我組織特徵映射圖，而輸出的映射圖也稱之為拓樸圖（張斐章，2003）。

SOM 的神經網路，初始化階段先隨機給予權重於輸出層神經元。接下來的訓練階段，將每一筆輸入資料利用方程式（1）的公式來計算，找出歐式距離最小的最適配神經元，亦即擁有調整權重的勝利神經元。

$$\|x_v - w_i\| = \min\|x_v - w_i\| \quad , \quad i = 1,\ldots,N, \qquad (1)$$

接著勝利神經元擁有調整權重的權力以及更新鄰近距離裡的其他神經元，調整的公式是依照 Kohonen 的規則，如公式（2）。其中，α 是神經網路的學習速率，x_v 是輸入向量，N_r 是鄰近距離半徑 r 內的所有神經元集合。

$$\Delta w_i = \alpha\left(x_v - w_i^{old}\right) \quad , \quad \text{for} \quad i \in N_r \quad , \qquad (2)$$

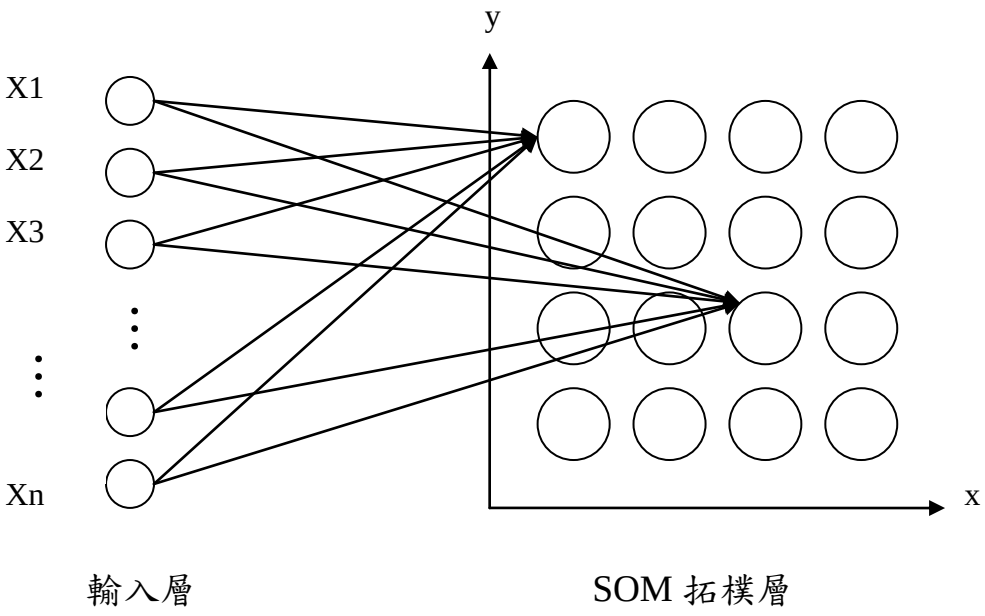


圖 2-2 映射至二維拓樸圖之 SOM 架構圖

（資料來源：修改自張斐章，2003）

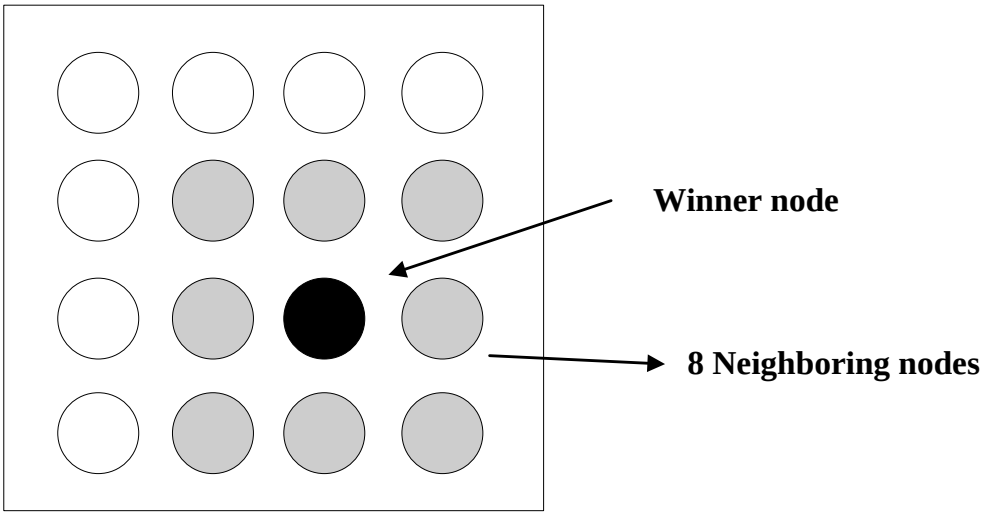


圖 2-3 優勝神經元及鄰近神經元關係圖

（資料來源：修改自 Kinag, 2006）

最後當達到迭代次數，或者是特徵映射圖形成時，終止訓練。整體的 SOM 演算流程如圖 2-4 所示。

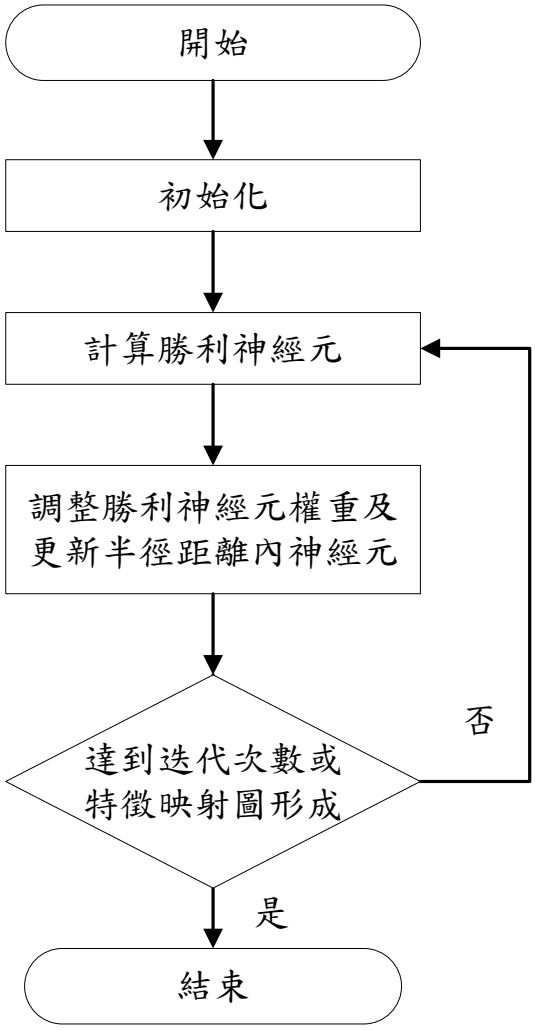


圖 2-4 SOM 演算流程圖

目前 SOM 被廣泛的應用在各個研究領域上，包括了聲音辨識（speech recognition）、影像資料壓縮（image data compression）、影像或特性辨識（image or character recognition）、機器人控制（robot control）及醫學診斷（medical diagnosis）等等（Kiang et al., 2006）。其他相關文獻詳列如表 2-2。

2.2 分類方法簡介

分類，是一種「先判斷明確的類別標籤（Class Labels），再利用訓練資料（Training Data）以及在分類屬性中的類別標籤，學習出相關分類規則，之後將這些規則用來分類新進的資料」（Han & Kamber, 2001）。而分類的模式主要有兩個步驟，第一步驟先建立描述資料的類別集合，然後透過分類演算法分析訓練資料，以產生分類相關規則，第二個步驟是利用測試資料（Test Data）來評估先前產生的分類規則之正確性，如果正確性高，則這些規則即可被應用於新進資料的分類上（Han & Kamber, 2001）。目前常見的分類技術有許多種，本研究將其整理如表 2-3。

本研究利用的分類方法是以 BPN、SVM 以及決策樹來做為分類的技術，主要是因為 BPN 具有能處理雜訊資料或處理有遺漏值的資料之能力。而 SVM 近年來是分類的熱門工具，亦被與其他分類工具做比較預測的正確性。決策樹則是普遍，並且能夠產生簡單易懂的分類規則的工具。本文將以這三種方法進行分類預測的研究，並比較其結果。

表 2-2 國內外應用 SOM 之文獻整理表

國內文獻			
研究者	時間	研究名稱	研究簡介
王彥翔	2003	自組特徵映射與學習向量量化神經網路於河川流量之預測	以 SOM 為基礎提出強制型自組特徵映射，以期能改善洪流量預測精準度，再以學習向量量化神經網路來調整群聚的結果。
何鴻聖	2004	自我組織神經網路在選股策略的應用	由於 SOM 是非線性的群聚方法，能捕捉股票報酬的非線性現象，因此，以 SOM 建構非線性選股策略與投資組合。
林長富	2006	自組織映射圖網路於碎形影像壓縮之研究	以 SOM 將定義域區塊分群，並將定義域區塊轉換為頻率域訊號，有效的加快編碼的素度，並維持影像的品質。
國外文獻			
研究者	時間	研究名稱	研究簡介
Garcia & Gonzalez	2004	Self-organizing map and clustering for wastewater treatment monitoring.	以 SOM 為基礎，發展工廠裡之廢棄污水處置的監督管理技術。
Kiang et al.	2006	An extended self-organizing map network for market segmentation — a telecommunication example.	利用一種使用者能指定拓樸圖上群聚群數之延伸的 SOM，並以 AT&T 電信公司資料進行實驗，發現延伸 SOM 比因式分析以及 K-means 二階段步驟的群聚分析，更能發現市場區隔。
Kiang et al.	2007	The effect of sample size on the extended self-organizing map network — A market segmentation application.	利用隨機樣本與規律樣本兩種不同的輸入資料，再次驗證延伸 SOM 的確比因式分析及 K-means 二階段步驟的群聚分析，更能精確的區隔市場。

表 2-3 常見的分類技術

方法	簡介
案例式推理方法 Case-Based Reasoning	假設相似性的問題有相似的解決方法，主要是利用過去類似問題的個案解決經驗的累積，推論出相關的知識，以重複應用於新的相似問題上。
K-Nearest Neighbor	以類比式學習（Analogy Learning）為基礎。給定一個未知樣本 y，藉由 k 個最近鄰居法搜尋模式空間，並找出最接近未知變數的 k 個訓練樣本。而這 k 個訓練樣本即是未知樣本 y 的 k 個近鄰。
基因演算法 Genetic Algorithms	先產生初始母體（Initial Population）染色體，染色體內的基因皆是布林函數，以表示分類的法則。接著透過交配、突變方式的產生，直到染色體適合度（Fitness）達到門檻為止。
粗略集合 Rough Set	能發現不準確資料或雜訊資料的內在結構關係，主要應用於離散型資料。實務上的資料，通常有些許類別無法以可用的屬性來區分，而粗略集合可近似或粗略的定義這樣的類別。
模糊集合 Fuzzy Set	分類規則允許使用模糊的界限及門檻，以範圍介於 0～1 之間的值表示在某類別或某集合內物件的相似程度。
類神經網路 Neural Network	一連串的輸入與輸出單位的集合，單位與單位之間的連結皆有權重，學習時期透過不斷的修改連結權重，已找到輸入單位的正確類別屬性。
決策樹 Decision Tree	普遍的分類或預測演算法，以樹狀結構由上而下表現。最上面的點代表跟節點（root），樹葉節點（Leaf Node）代表類別，非葉節點（Nonleaf）代表某屬性的測試，樹的分支（Branch）代表測試的結果。
支撐向量機 Support Vector Machines	由統計學習理論衍生而來的演算法，利用區分超平面（Separating Hyperplane）來區隔兩個或是多個不同的類別。自 Vapnik（1995）提出後，已經廣泛的應用於各個領域。

（資料來源：整理自 Shin et al., 2005；曾韋榮，2006）

2.2.1 BPN

BPN 是目前類神經網路學習模式中，最具有代表性，應用也最普遍的模式（葉怡成，2002）。BPN 是一種監督式學習網路，並具有學習（Learning）精度高、回想（Recall）速度快、輸出值可為連續值、能處理複雜的樣本識別以及高度非線性的函數合成問題，例如：樣本識別（Sample Recognition）、分類、預測（Prediction）、雜訊過濾（Noise Filter）、資料壓縮（Data Compression）等問題（張斐章，2003）。BPN 主要的架構為輸入層、隱藏層及輸出層；演算過程則是利用最陡坡降法（The Gradient Steepest Descent Method），將 BPN 輸出值與實際輸出值之誤差（Error）降低至最小（Lee, 2008）。本文整理了一些國內外應用 BPN 的相關研究文獻，如表 2-4 所示。

BPN 原理是使用最佳化中的最陡坡降法來迭代運算，將誤差值予以最小化求得最佳值，並加入隱藏層處理神經元，使最佳化問題藉著自行調整參數，進而得到更精確的最佳解。BPN 基本的架構包含了三個部分：

1. 輸入層：代表著網路的輸入變數，其神經元數量視研究問題而定，轉換函數使用線性函數，即 $f(x)=x$。
2. 隱藏層：用以表現輸入神經元間交互影響，隱藏層的神經元數量沒有一定規則，通常都是以試誤法（Trial & Error）決定。隱藏層的層數一般而言都是一層，但可視問題的複雜度而增加層數。隱藏層的轉換函數是採用非線性函數。

表 2-4 國內外應用 BPN 之文獻整理表

國內文獻			
研究者	時間	研究名稱	研究簡介
楊啟洲	2005	以倒傳遞類神經網路作為授信風險預測之研究	利用 BPN 作為授信風險之預測工具，預先掌握重要資料，降低營運風險，謀取更多利潤。
黃鈁嫕	2005	運用類神經網路預測新進顧客產品喜好之個人化商品推薦技術	利用大型賣場交易資料顧為實作樣本，以 RMF 分析法對顧客作分群，搭配 BPN 來預測新進顧客喜好的個人化商品，最後以 Top-N 做推薦。
蔡正修	2007	台灣上市電子類股價指數走勢預測之研究	運用迴歸分析、時間數列、BPN 及適應性網路模糊推論系統等方法建立模式，並以各模式預測台灣上市電子類股價指數隔月收盤指數，並比較各模式預測結果優劣。
國外文獻			
研究者	時間	研究名稱	研究簡介
Chang & Chao	2006	Application of back-propagation networks in debris flow prediction.	以 BPN 來分析預測 1983 至 1993 年間的土石流，準確率高達 93.82%。因此可利用 BPN 來預測土石流發生，以降低土石流的威脅。
Huang et al.	2007	Residual life predictions for ball bearings based on self-organizing map and back propagation neural network methods.	利用 SOM 來擷取軸承退化的訊號，並以 BPN 來預測軸承殘餘的有用生命價值。
Lee	2008	Back-propagation neural network for the prediction of the short-term storm surge in Taichung harbor, Taiwan.	以台灣的台中港研究資料顯示，利用 BPN 進行瞬間浪潮的預測，可以提早 1-6 小時預測到颱風帶來的瞬間浪潮。

3. 輸出層：代表著網路的輸出變數，其神經元數量亦視研究問題而定，轉換函數為非線性函數。BPN 基本架構如圖 2-5 所示。

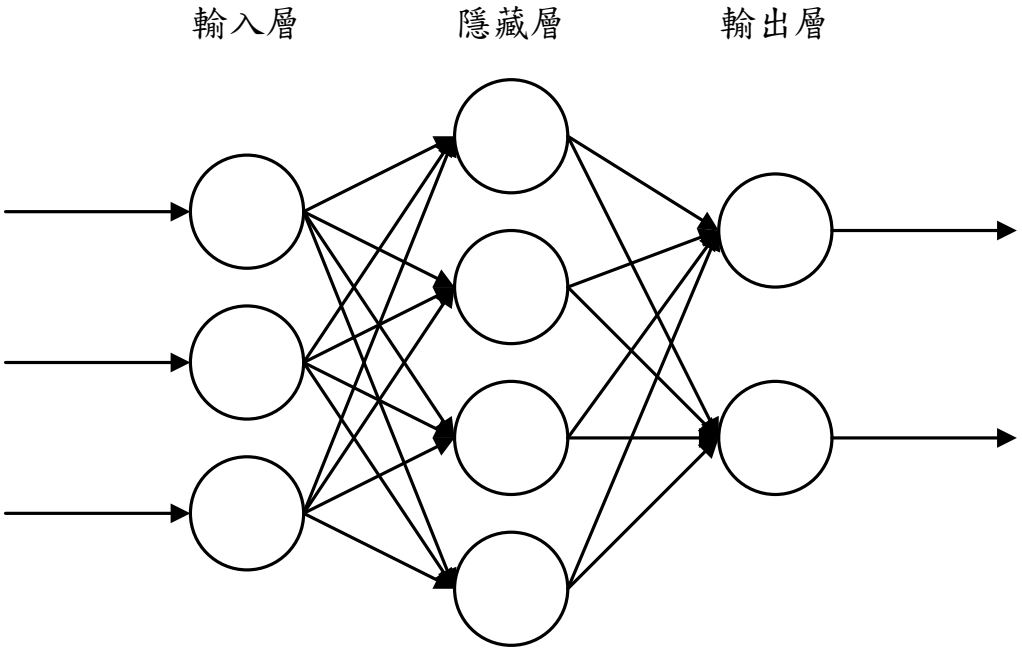


圖 2-5 BPN 網路架構圖

（資料來源：修改自張斐章，2003）

BPN 的演算法步驟如下：

1. 計算輸入層到隱藏層的輸出值，以及隱藏層到輸出層的輸出值，利用公式（3）及活化函數轉換公式（4）進行計算。

$$Y_j = f(\sum W_{ij} X_i - \theta_i) \qquad (3)$$

$$f(x)=\frac{1}{1+e^{-x}} \tag{4}$$

其中，Y_j：輸出變數，類似生物神經元模型的輸出訊號；

f：轉換函數，為一個非線性函數；

X_i：輸入變數，類似生物神經元模型的輸入訊號；

W_{ij}：連結權重值，類似生物神經元模型的突觸強度；

θ_i：閥值，類似生物神經元模型的閥值。

2. 計算網路輸出值與目標輸出值的差距，並利用誤差函數作為目標函數公式（5），降低網路輸出值與目標輸出值的差距，達到網路學習的目的。

$$E=\frac{1}{2}\sum(T_j-Y_j)^2 \tag{5}$$

其中，T_j 為輸出層第 j 個輸出神經元之目標輸出值；

Y_j 為輸出層第 j 個輸出神經元之網路輸出值。

3. 修正調整連結權重值如公式（6），以及閥值（7）。

$$\Delta W=-\eta\frac{\partial E}{\partial W} \tag{6}$$

$$\Delta\theta=-\eta\frac{\partial E}{\partial \theta} \tag{7}$$

其中，ΔW 為連結權重值修正量；

$\Delta\theta$ 為閥值修正量；

η 為學習速率，決定權值修正量大小。

4. 重複步驟 1 至步驟 3，直至網路訓練達到收斂狀態，亦即網路輸出值與目標輸出值的誤差最小。

BPN 執行的步驟如下：

步驟 1　設定網路參數

步驟 2　亂數產生初始連結權重值及閥值

步驟 3　計算隱藏層與輸出層之輸出值

步驟 4　計算目標函數

步驟 5　計算權重修正量

步驟 6　調整各層之連結權重及閥值

步驟 7　判斷是否仍有訓練樣本，若是則回到步驟 3，若否則到步驟 8

步驟 8　判斷是否達到網路停止原則，若是則停止網路訓練，若否則回到步驟 3

BPN 執行的流程如下圖 2-6 所示。

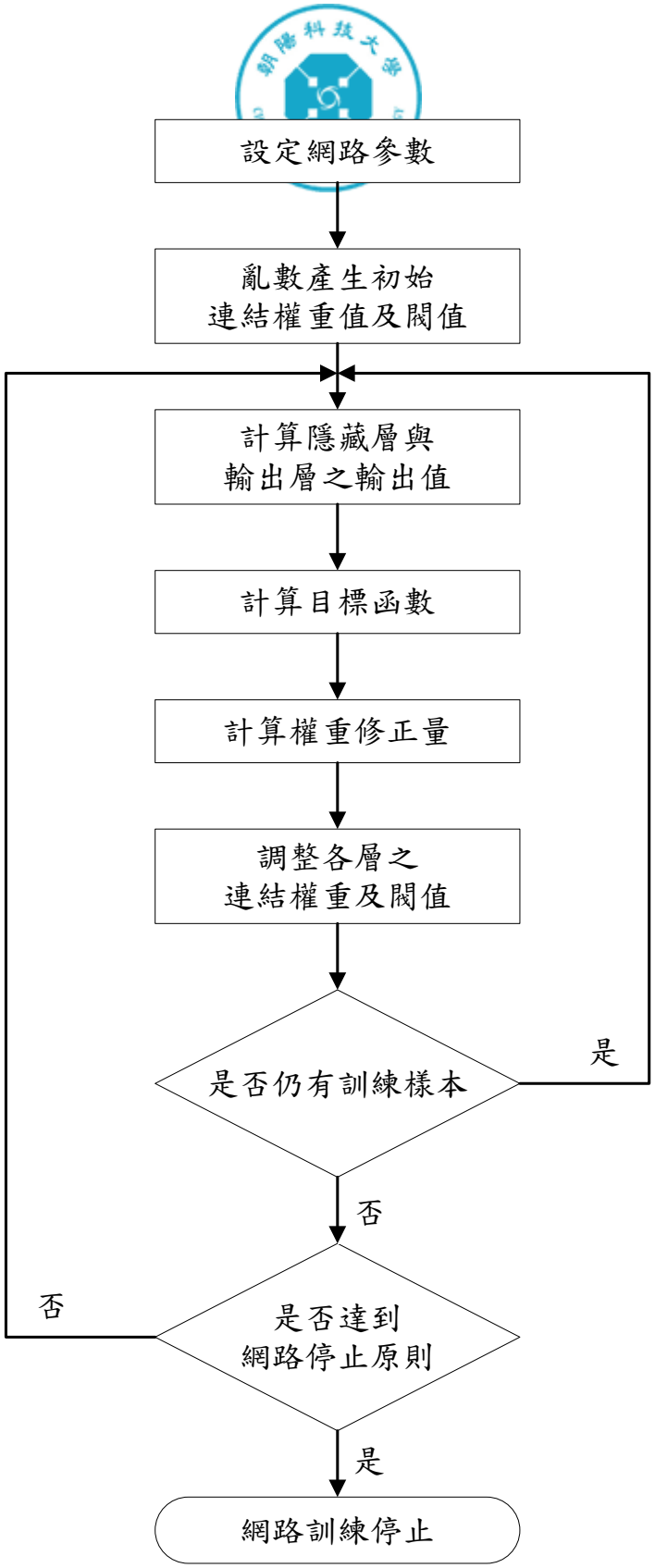


圖 2-6 BPN 執行流程圖

（資料來源：修改自張斐章，2003）

2.2.2 SVM

SVM 是一種被廣泛應用的分類演算法，特別是像文字這樣高維度的資料（Shima et al., 2004）。SVM 是由統計學習理論（Statistical Learning Theory）衍生而來的學習演算法，從簡易向量分類器（Simple Vector Classifiers），逐漸發展為超平面分類器（Hyperplane Classifiers），最後才衍生為支撐向量分類器（Support Vector Classifiers）（黃承龍 等人，2004）。自 Vapnik（1995）提出後，就逐漸成為機器學習（Machine Learning）研究領域中，極為熱門的一種方法（粘志鵬，2006）。

目前 SVM 應用的範圍非常廣泛，Shin et al.（2005）大致上整理了 SVM 應用面如：財金時間序列預測（financial time-series forecasting）、行銷方面（marketing）、製造業產量評估（estimating manufacturing yields）、文件分類（text categorization）、臉部影像偵測（face detection using image）、手寫辨識（hand written digit recognition）以及醫學診斷（medical diagnosis）。本文亦整理一些國內外 SVM 應用的相關文獻，如表 2-5 所示。

SVM 是一種基於統計學習理論的機器學習方法，最主要是利用一個區分超平面（separating hyperplane）來區隔兩種或多種不同類別的資料，來解決資料分類的問題。自 V. Vapnik（1995）提出後，短短十幾年間，已成為被廣泛應用的機器學習方法。首先，SVM 分類方法的基礎定義如下：

x_i：是一個向量，用以表示某筆資料的屬性，$x_i \in R^N$，$i=1,2,3,..m$。

表 2-5 國內外應用 SVM 之文獻整理表

國內文獻			
研究者	時間	研究名稱	研究簡介
徐豐智	2005	Support Vector Machines 分類技術應用於文件相關性量測之探討	提出一種新的「語意相關性測量」的方法與運算平台，利用 SVM 技術在文件分類上高效能表現的優點，來支援文件相關性量測的計算。
黃敏菁	2005	支援向量機在財務時間序列預測之應用	使用技術性指標搭配 SVM 與 SVR，預測台灣加權股價指數下一交易日之變動方向與變動幅度探討 SVM 與 SVR 應用在財務時間序列預測之可行性。
粘志鵬	2006	基於支援向量機之中文自動作文評分系統	提出一種「特徵義原空間」的方法，並使用 SVM 理論模型來自動建立一套中文作文評分系統。
國外文獻			
研究者	時間	研究名稱	研究簡介
Shima et al.	2004	SVM-based feature selection of latent semantic features.	提出一種基於 SVM 的特徵排序方法，使得選取的 LSI 特徵更適合拿來做分類。
Shin et al.	2005	An application of support vector machines in bankruptcy prediction model.	研究 SVM 應用於破產預測問題的效能，並且和 BPN 方法做比較。
Lee	2007	Application of support vector machines to corporate credit rating prediction.	應用 SVM 於法人信譽評價的問題上，並嘗試提出更有說明力及穩定的新模組，最後和 MDA、CBR 及 BPN 做比較。

y_i：稱為標註（Label），通常用$\{\pm 1\}$表示+1 和-1 兩個不同的類別，

$y_i \in \{\pm 1\}$，$i = 1,2,3,..m$。

f：決定函數，$f: R^N \to \{\pm 1\}$

當給定一筆資料x_i時，支撐向量機會判斷該資料是屬於哪一個類別（+1 或-1）。而分類的原理，是在已給予的訓練資料群中，找到一個超平面，將這兩類的資料區分開來，該超平面則可稱為區分超平面，如下圖 2-7 所示。落在區分超平面的所有x必須滿足公式（8）。

$$w \cdot x + b = 0 \qquad (8)$$

其中，w為超平面之法向量；

b為閥值。

此時我們的決定函數f則為公式（9），當輸入一筆測試資料時，即可依

$$f(x) = w \cdot x + b \qquad (9)$$

照決定函數來決定類別。若$f(x) > 0$，則該筆資料歸類為+1；若$f(x) < 0$，則該筆資料歸類為-1。

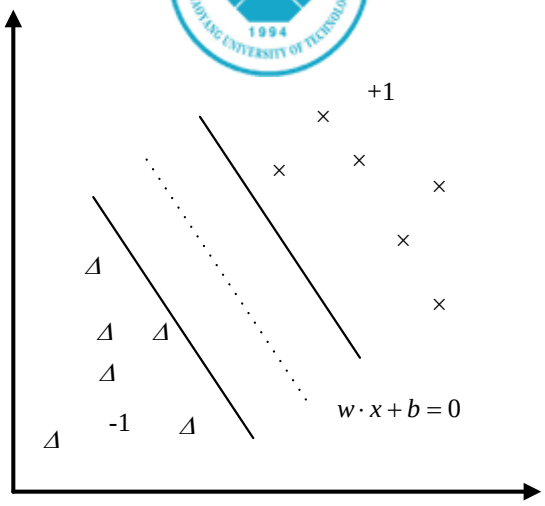


圖 2-7 超平面示意圖

（資料來源：修改自粘志鵬，2006）

由於w和b的可能性之組合有上千種或更多，支撐向量機的目的則是為了解決這樣的問題，找到一個最適合的區分平面，使得兩種不同類別之間的超平面能夠擁有最大邊界（Margin）。而支撐向量機的演算法如下：

1. 先定義區分平面的邊界為d^+和d^-，分別表示+1 類別和-1 類別距離區分平面的最短距離，且此類的資料需符合下列條件：

$$x^i \cdot w + b \geq +1 \quad \text{for} \quad y_i = +1 \qquad (10)$$

$$x^i \cdot w + b \leq -1 \quad \text{for} \quad y_i = -1 \qquad (11)$$

組合公式（10）與公式（11）可得到下式（12）：

$$y_i(x^i \cdot w + b) - 1 \geq 0 \qquad \forall i \qquad (12)$$

由（10）和（11）可知，$d^+ = d^- = \frac{1}{\|w\|}$，因此邊界為$\frac{2}{\|w\|}$。若我們欲尋求

最大邊界值，必須在符合限制式（12）下，求得$\|w\|^2$之最小值。而在限制式（12）中，若有任一個向量x_i可以讓等號成立，則稱該向量x_i為支撐向量（support vector）。如圖 2-8，即為最大邊界以及支撐向量圖。

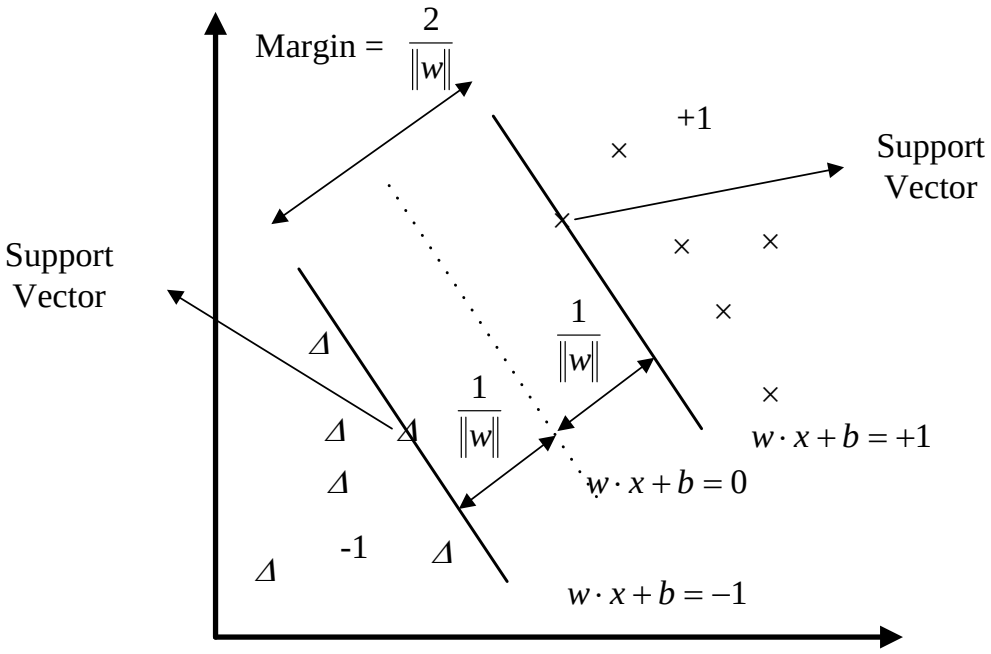


圖 2-8 最大邊界及支撐向量示意圖

（資料來源：修改自黃承龍等人，2004）

本論文的研究資料屬於高維度，因此我們採用放射型（Radial Basis Function, RBF）核心函數（Kernel Function），如公式（13）。

$$k(x_i, x_j) = \exp(-\gamma \|x_i - y_j\|^2) \qquad (13)$$

2.2.3 決策樹

決策樹是一種眾所皆知的法則歸納演算法，可以用來挖掘知識，常被應用在分類或者是預測的問題上（Lee et al., 2007），和 BPN 及 SVM 皆屬於監督式學習。決策樹是以樹狀資料結構呈現，包含了一個樹根（Root）、數個節點（Node）以及數個樹葉（Leaf），每個節點都代表著一個屬性（Attribute）（Kumar et al., 2008）。每一個從樹根到樹葉的分支，就代表著一個分類的規則，而決策樹的優點正是能產生容易被人類所瞭解並應用的決策法則（尤春惠，2004）。本文整理了一些國內外應用決策樹的相關文獻，如表 2-6 所示。

決策樹是利用樹狀資料結構的為基礎的分類方法，優點是能產生讓人類容易瞭解的決策規則。決策樹的建構是以監督式學習方法建立，每個內部節點（Internal Node）為某屬性的測試，而分支（Branch）代表對某屬性測試的結果，最後的樹葉節點（Leaf Node）代表類別。當一筆資料進入根部後，依照演算法則開始向子節點（Child Node）歸納，最後分類到適合的樹葉節點，亦為找到適合的類別。從根部進入，最後找到適合的類別，這樣的路徑稱為分類規則。如圖 2-9 所示。

決策樹的演算法規則很多，常見的有 ID3、C4.5、CART（Classification and Regression Tree）及 CHAID（Chi-square Automatic Interaction Detector）等。在本研究中，我們所使用的演算法規則為 C4.5 演算法。C4.5 是由 Quinlan

表 2-6 國內外應用決策樹之文獻整理表

國內文獻			
研究者	時間	研究名稱	研究簡介
林盈源	2003	決策樹在資料庫行銷決策之應用	利用決策樹之建構，做為行銷決策之參考，進而建立顧客的市場區隔及找出目標客戶群，以作為企業支援行銷策略之用。
劉麗蘭	2006	以決策樹分析台灣上市櫃紡織業公司的財務危機	建構一般財務指標對企業財務危機的預測，並對企業可能產生的財務危機分類，藉以提供詳細的說明，讓企業有簡單的判斷規則可以參考。
鄭志強	2006	以決策樹演算法建構台灣企業財務危機預警模式	以決策樹建立最佳化預測模型，並透過財務比率分析構面，解釋公司營運狀況，以期能區別出正常公司與財務危機公司，並找出發生重大弊案之公司的特性。
國外文獻			
研究者	時間	研究名稱	研究簡介
Wu et al.	2006	An effective application of decision tree to stock trading.	介紹一種以濾嘴法則結合決策樹的股票投資方法。
Tso & Yau	2007	Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks.	介紹迴歸分析、決策樹及類神經網路三種電能消耗預測的方法，並比較其結果。
Lee et al.	2007	A prediction model for success of services in e-commerce using decision tree: E-customer's attitude towards online service.	嘗試找出促使顧客線上購物的服務特性，並利用決策樹，在電子商務中發展以顧客識別提供服務為基礎的預測模組。

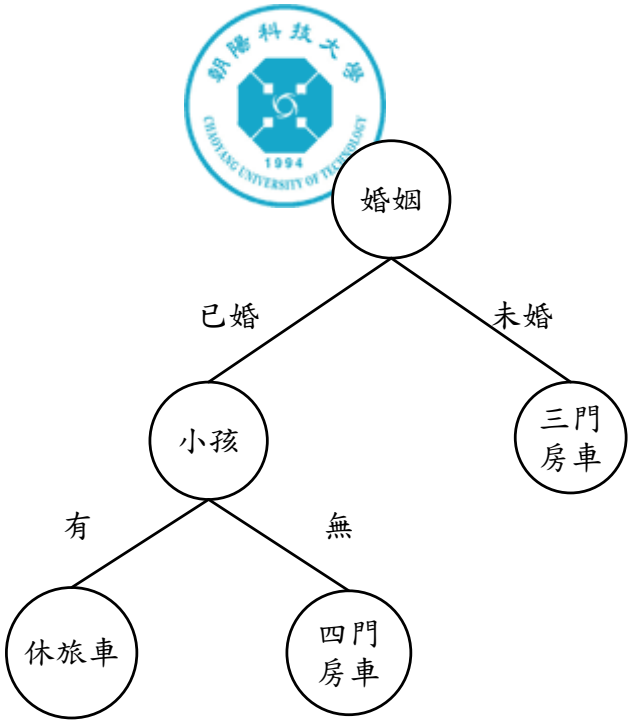


圖 2-9 決策樹狀圖

（資料來源：修改自林盈源，2003）

（1993）改進 ID3 演算法而來的，利用連續性數值分割法，改善 ID3 無法處理數值屬性分類的缺陷。改善的地方包括：1. 遺漏值的預測子仍可以被使用；2. 含有連續值的預測子可以被使用；3. 加入修剪樹的功能；4. 規則衍生。

C4.5 利用的連續性數值分割法，先將資料集依照數值屬性的屬性大小排列，接著計算兩個屬性數值可能的中點，稱之分割點（Cutpoint），將資料集分割為兩個部分，再以決策屬性評估函數計算該分割點之決策屬性值，最後在眾多分割點中，挑選最佳決策屬性值分割點，作為該數值屬性的資料樣本區分點。

基本決策樹建構規則如下：

1. 設定根節點（Root Node）C，此時所有訓練樣本皆屬於 C 物件集合。若 C 物件集合都屬於同一個類別，則此類別為 C 的決策結果，然後停止建構決策樹。如果不是，則進行下一步。

2. 針對 C 物件集合，計算其熵值$E(C)$，計算公式如公式（14）。

$$E(C)=-\sum_{j} p_j \times \log_2 p_j \qquad (14)$$

其中，p_j=(屬於類別 j 的物件數) / (集合 C 的總物件數)；

$\log_2$ 預設為 0。

3. 對於所有尚未出現於根節點與目前節點路徑上之候選屬性A_i，以A_i對 C 物件集合分割，並計算以A_i產生子決策樹的熵值$E(A_i)$，與資訊獲利$G(A_i)$。計算公式如公式（15）與公式（16）。

$$E(A_i)=\sum_{k}(\frac{n_k}{n})\times E(C_k) \qquad (15)$$

$$G(A_i)=E(C)-E(A_i) \qquad (16)$$

其中，C_k為物件集合 C 中，A_i屬性相同的物件子集合k；

$E(C_k)$為物件集合C_k之熵值；

n為物件集合 C 之物件總數；

n_k為物件子集合C_k之物件總數。

4. 選擇資訊獲利最大之候選屬性，將其作為節點 C 之分類屬性。

5. 在 C 節點下，建立子節點C_1、C_2、C_3、…、C_m（假設有m個分類屬性

值），依照分類屬性，將 C 物件集合中的物件，分配到適合的子集合之中。

6. 將子節點C_i，依照步驟 1 繼續執行。

當決策樹達到下列任一條件時，即停止建構：

1. 分割後，每一個子節點的物件都屬於同一個類別。
2. 沒有任何剩餘屬性可以再被分割。
3. 沒有任何物件可以再被測試。

最後，當決策樹完整的建構後，稱為完全成長的樹（Fully Grown Tree）。這樣的樹通常無法直接作為分類規則，主要是礙於訓練資料中，可能包含雜訊（Noise）或游離值（Outlier）容易導致分類規則異常或錯誤，因此需要透過修剪（Prune）的動作，使決策樹的分類規則更貼近實際的狀況，更有利用的價值。

C4.5 演算法的修剪標準是依據預估錯誤率（Predicated Error Rate）的值，當作決策樹修剪的判斷。預估錯誤率是由訓練組資料的錯誤率，來估計非訓練組資料的錯誤率。修剪的方法是從決策樹的底部，也就是從樹葉節點向上測試節點形成的子樹，接著將子樹以葉節點代替，測試其預估錯誤率。若是新的葉節點之預估錯誤率低於原本的子樹，則由新的葉節點取代原本的子樹，並且將原本子樹底下的類別全部歸於新的葉節點；若是新的葉節點之預估錯誤率沒有原本子樹的預估錯誤率低，則保留原本的子樹

不動。至此，整體決策樹的建構，才能提供有價值並貼近現實的分類規則。

2.3 維度縮減

在文件資訊檢索（Information Retrieval）的相關研究中，通常以詞彙來代表整份文件，每一個不同的詞彙都可以視為一個特徵（Feature）。這樣的作法，往往產生上萬個特徵的高維度特徵空間（Feature Space）（Hao et al., 2007）。此外，文件樣本通常具有維度高、雜訊多、樣本稀疏，以及特徵不明顯等四個特性（姚力群 & 陶卿，2005）。為了解決這樣的問題，可以應用維度縮減（Dimensionality Reduction）的方法。維度縮減的方法，大致上可以分為特徵選取（Feature Selection）以及屬性擷取（Feature Extraction）（Jain et al., 2000）。

2.3.1 特徵選取

特徵選取主要是從眾多特徵中，挑取最重要的特徵成為一個新的子集合來代表原始的特徵集合，新子集中的特徵還保留原始特徵的本義，著名的方法有連續向前選取法（Sequential Forward Selection, SFS）、連續向後選取法（Sequential Backward Selection, SBS）、徹底搜尋法（Exhaustive Search）等方法（Jain et al., 2000）。本研究使用的特徵選取方法為連續向前選取法。

連續向前選取法，主要是由空特徵變數集合開始進行，接著由一維的變

數組合測試，透過一次增加一個特徵變數，直到已達到要求特徵變數數目為止（陳昭穎，2006）。目前多用於醫學診斷、圖形辨識、錯誤診斷及分類預測等方面（Su & Yang, 2008；Sugumaran et al., 2007；Liu & Zheng, 2006；Selamat & Omatu, 2004）。本文整理國內外應用特徵選取之相關文獻，如表 2-7 所示。

2.3.2 屬性擷取

屬性擷取是利用壓縮、萃取原始特徵的方式，得到一個更具有辨識能力的新子集合，不過在新子集合中，無法保有原始特徵的本義，著名的方法有主成分分析法（Principal Component Analysis, PCA）、獨立成分分析法（Independent Component Analysis, ICA）、核心主成分分析法（Kernel Principal Component Analysis, KPCA）以及 LSI 等方法（Jain et al., 2000）。本研究採用的屬性擷取方法為 PCA、ICA 以及 LSI 等三種方法。

1. PCA

PCA 為多變量分析中的一種方法，由 Pearson 在 1901 年所創，接著由 Hotelling 根據 Pearson 之理論所衍生的一套統計方法。最初 Pearson 之研究是在 p 維空間點的集合中，找到最佳進似的線段及平面，而 Hotelling 則由原本 p 的變數集合中，找出比原始 p 變數少的基本變數 p'，此 p' 變數足以

表 2-7 國內外應用特徵選取之文獻整理表

國內文獻			
研究者	時間	研究名稱	研究簡介
楊敦翔	2003	以類神經網路與特徵選取技巧處理空氣能見度預測問題之研究	以類神經網路預測隔日高雄市空氣能見度，欲從眾多的氣象參數中，選取與能見度最有相關性之特徵變數組合，以找到此變數組合與能見度間的映射關係。
陳昭穎	2006	資料探勘技術於超音波旋轉肌肌群影像之診斷應用	使用資料探勘技術去擷取有用紋路特徵，並將這些特徵應用於放射狀基礎函數網路，以及支援向量機來把肩部超音波影像分為四類：正常、發炎、鈣化、和斷裂。
國外文獻			
研究者	時間	研究名稱	研究簡介
Su & Yang	2008	Feature selection for the SVM: An application to hypertension diagnosis.	以高血壓診斷為例，先利用特徵選取來決定最重要的特徵，再以 SVM 進行分類預測。
Liu & Zheng	2006	FS_SFS : Anovel feature selection method for support vector machines.	利用一種改良的特徵選取方法—FS_SFS 搭配 SVM 來進行分類預測。
Selamat & Omatu	2004	Web page feature selection and classification using neural networks.	先利用特徵選取擷取最相關的特徵，再以類神經網路來進行新聞網頁分類。

表達原始 *p* 變數的分佈特徵，即為 PCA 的主要概念（許邦輝，2006）。

主成份分析法之主要目的乃是希望以較少數目的變數解釋原本資料中大部份的變異，目前廣泛應用於分類預測、維度縮減、錯誤偵測等問題（Fortuna & Capson, 2004；許邦輝，2006；張結雄，2002）。本文整理了一些關於 PCA 以及 ICA 的國內外相關研究文獻，如表 2-8 所示。

2. ICA

在 ICA 中，每一個特徵在此分析法中，皆稱為成分，而 ICA 是欲讓成分之間的統計相依性（Statistical Dependence）降至最低，亦即要使成分與成分之間，互相為一種獨立情形，這樣的方法由於能擷取資料基本結構的成分，成為非常熱門的線性轉換法（林巧苑，2001）。ICA 目前在許多領域上的應用非常廣泛，本文整理了一些關於 PCA 以及 ICA 的國內外相關研究文獻，如表 2-8 所示。

3. LSI

LSI 具有計算上的優勢，並且已建立穩定的地位，因此在資訊檢索中是廣泛被使用的工具，甚至常被用來作為資料探勘的預先處理步驟（Kaban & GiroLami, 2002）。由於利用詞彙索引為基礎的檢索方法，向來面對著兩個影響索引結果的問題：一字多義（polysemy）以及一義多字（synonymy），

表 2-8 國內外應用 PCA 及 ICA 之文獻整理表

國內文獻			
研究者	時間	研究名稱	研究簡介
許邦輝	2006	以主成分分析法為基礎之文件自動分類模式	提出以 PCA 為基礎的文件類別自動判定方法，建立一套知識文件自動分類系統。
李銘浚	2007	應用獨立成分分析、對數頻譜預估、及頻率成分調整技術作語音增強之研究	目的為發展一個能將語音訊號裡之背景雜訊去除的有效方法。
國外文獻			
研究者	時間	研究名稱	研究簡介
Ture et al.	2007	Comparison of dimension reduction methods using patient satisfaction data	比較傳統 PCA 與 GPCA、PCA-NN 及 NLPCA-NN 四種方法的成效。
Ekenel & Sankur	2004	Feature selection in the independent component subspace for face recognition	探討 ICA 及 PCA 在臉部辨識的應用成效。
Fortuna & Capson	2004	Improved support vector classification using PCA and ICA feature space modification	透過 PCA 及 ICA 來修改特徵變數，改善 SVM 的分類預測率。

主要是因為同樣的概念，可以用不同詞彙來表達，而這樣的問題便會影響檢索的準確性。因此 Deerwester et al.（1990）提出 LSI 方法，來改善上述的問題（吳忻萍，1997）。此外，LSI 也能改善樣本資料稀疏的問題（曾韋榮，2006）。

隱含語意是指同一個概念下，語法上不同的表達方式。LSI 是建構在向量空間模型（Vector Space Model）的基礎上，並以奇異值分解（Singular Value Decomposition, SVD）的方法，將原本的詞彙和文章以簡化的 k 個概念（Concept）表示，縮小空間維度，將相關的詞彙及文章聚集在一起，以解決資訊檢索中，隱含語意的問題（吳忻萍，1997；曾韋榮，2006）。本文亦整理了一些國內外應用 LSI 相關的研究文獻，如表 2-9 所示。

在經過了相關的文獻探討之後，可以得到下列心得：

1. SOM 是利用資料本身的輸入特徵向量，來映射到輸出拓樸層上，因此所形成的群聚，更能表現出原始輸入資料彼此的相關性，所以非常適合用來分析部落格所蒐集而來的文章，對於應用在區隔顧客市場，相信能夠更精準更精確。
2. BPN、SVM 及決策樹，都是適合用來分類以及預測的技術，因此本研究於顧客分類的研究目的上，將採用這三種方法，並且比較其效果與準確性。
3. 由於資訊檢索利用的詞彙-文章向量空間矩陣，往往過於稀疏，因此曾韋

表 2-9 國內外應用 LSI 之文獻整理表

國內文獻			
研究者	時間	研究名稱	研究簡介
林家民	2005	基於潛藏語意分析之多語言文件自動分群技術	現有的文件分群技術大多只處理單語文件，但國際化的趨勢及網路的發展，常讓人需要產生、獲取及儲存不同語言的文件。因此設計一個基於潛藏語意分析的多語言文件自動分群技術。
洪淑芬	2006	潛在語意索引在生醫文件分類之應用	以能自動判別一份文件是否探討蛋白質與蛋白質之間的交互影響為目標，並利用不同機器學習演算法與文件特徵表示法進行實驗。
曾韋榮	2006	結合潛在語意檢索及資訊粒化於資料探勘	結合潛在語意檢索及資訊粒化於資料探勘，希望能達到縮減資料屬性維度、資料筆數以及有效處理非平衡資料。
國外文獻			
研究者	時間	研究名稱	研究簡介
Gao & Zhang	2005	Clustered SVD strategies in latent semantic indexing.	有鑑於 SVD 在大量不同性質的資料集下效果不好，因此提出先將這些大量不同性質的資料分割成數個小群聚，再進行 SVD。
Husbands et al.	2005	Term norm distribution and its effects on Latent Semantic Indexing.	研究文字檢索在一個大規模的文件集中，並著重在文字標準規範的分配上，然後觀察其成效。
Kontostathis & Pottenger	2006	A framework for understanding Latent Semantic Indexing (LSI) performance.	介紹一個理論的模組，用來瞭解 LSI 在搜尋以及檢索應用上的成效。

榮(2006)提到以LSI來解決稀疏性資料的問題。本研究也將嘗試導入Feature Selection、PCA、ICA 及 LSI 等維度縮減方法來結合 BPN、SVM 與決策樹，試圖發展改善稀疏性資料的分類器，並觀察維度縮減方法導入前後的預測精準度差異。

第三章 研究方法

本論文之研究目的，主要是提出一個「部落格探勘模組」，並利用這個模組來擷取出隱含於部落格文章中，富涵商業資訊之知識。我們期望可以透過擷取出的知識，達到：1. 顧客區隔，2. 顧客分類，以供行銷決策使用。同時我們也發展處理稀疏性資料分類器，提高分類預測的正確率。在本章節中，我們將詳細的介紹本論文提出的「部落格探勘模組」所使用的相關研究方法。

首先在 3.1 節裡，介紹我們所提出之「部落格探勘模組」的研究方法與流程。接著我們要先蒐集部落格上，有關網路電話這樣產品的文章，在 3.2 節中詳述部落格文章的蒐集。完成資料蒐集後，在進行資料探勘前，必須先對所蒐集的資訊做前處理，在 3.3 節裡會詳述資料前處理。完成資料前處理後，我們緊接著進行資料探勘，本研究使用自我組織特徵映射圖來做資料分群，達到本論文第一個研究目的—顧客區隔，在 3.4 節中會詳述資料分群的方法。在 3.5 節中，我們利用 3.4 節資料分群的結果，分別使用倒傳遞類神經網路、支撐向量機以及決策樹等分類預測研究方法，對於未知的部落格文章來進行分類預測，以求達到本論文第二個研究目的—顧客分類。然而使用詞彙-文章矩陣進行研究，往往矩陣中的資料具有稀疏性，容易影響分類預測的準確率以及耗費儲存空間，因此我們導入隱含語意索引技巧來發展處理稀疏性資料分類器，在 3.6 節中將詳述之。

3.1 部落格探勘模組

本節我們將提出本研究之「部落格探勘模組」，其研究流程如圖 3-1，詳細步驟說明如下：

步驟 1　部落格文章蒐集：針對欲探討的主題，進行部落格上的文章蒐集，在本論文中，我們以網路電話產品的文章為探討的主題，因此對部落格上討論網路電話產品的文章進行蒐集。

步驟 2　資料前處理：任何欲進行資料探勘的資料，都必須先經過資料前處理，才容易挖掘出有用的資訊及知識。本論文中，我們是利用詞彙-文章向量矩陣作為訓練輸入資料，因此我們資料前處理的流程是定義關鍵字、刪去贅詞關鍵字、合併關鍵字、刪去未出現關鍵字以及產生詞彙-文章向量矩陣，並對詞彙-文章向量矩陣進行正規化。

步驟 3　資料分群：區隔這些討論網路電話產品的部落客，觀察不同群聚裡的部落客有不同的需求面，希望藉此達到顧客區隔。因此在資料分群的技術上，我們使用的是自我組織特徵映射圖，利用部落客所發表的文章，依照關鍵字的特色，能夠自我形成群聚，實現顧客區隔的目的。

步驟 4　資料分類：透過既已蒐集網路電話產品文章的顧客區隔結果，來預測未來發表網路電話產品文章的部落客會是屬於既有群聚中的哪一群，以供決策行銷之用。因此我們使用倒傳遞類神經網路、支撐

向量機及決策樹等監督式學習的分類預測技術，來進行資料分類預測。

步驟 5　維度縮減：由於詞彙-文章向量矩陣屬於稀疏性資料的矩陣，往往耗費儲存空間以及影響分類預測準確率，因此本論文希望能藉由導入維度縮減技術，發展能有效改善稀疏性資料分類器，提高分類預測的準確率。

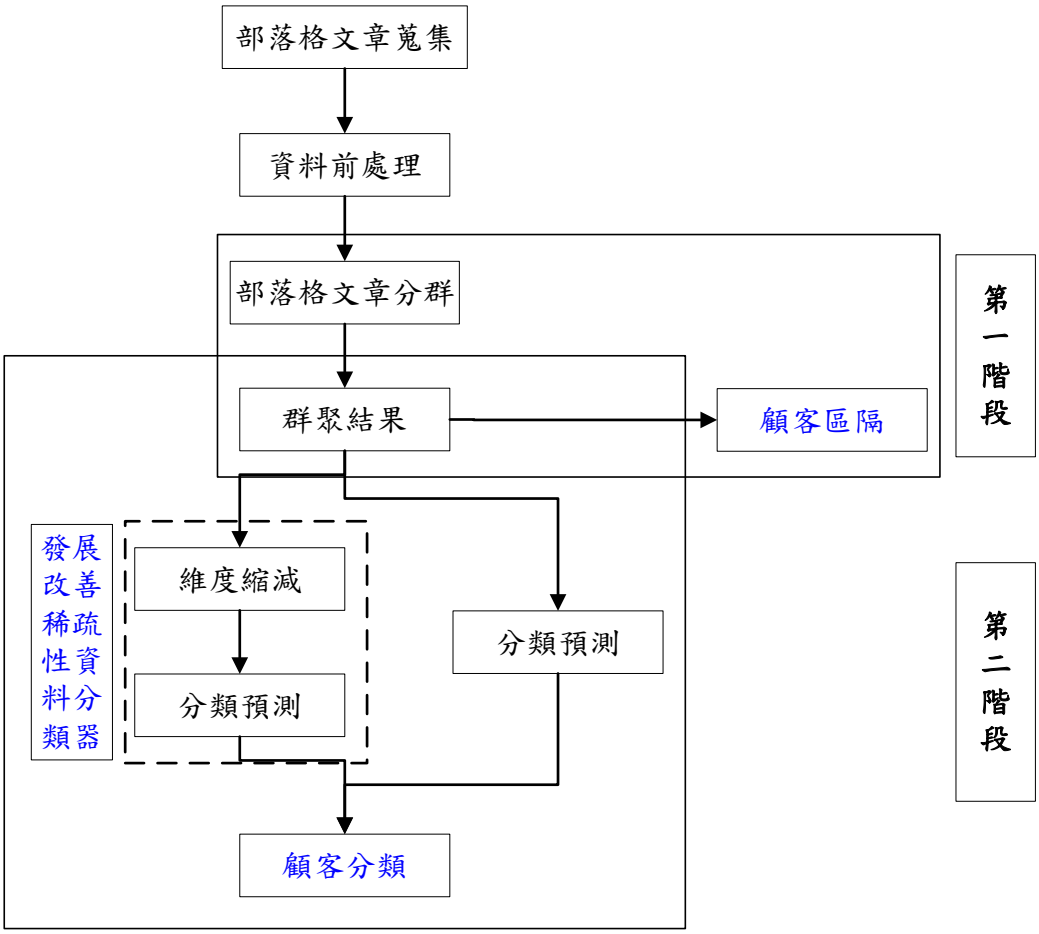


圖 3-1 部落格探勘模組流程圖

3.2 部落格文章蒐集

本研究蒐集了台灣各大部落格裡，談論到網路電話的文章，一共兩百筆，來作為我們的研究資料及對象。例如：「使用網路電話，會比一般電話更便宜，因為他把聲音數位化，並使用封包交換，來代替回路交換所花的費用[8]。」該部落客所發表的文章提到，網路電話會比一般電話便宜等重點；又如：「今天用 skype 的網路電話，通話費是很便宜，但是通話品質不是很好，還有很大的進步空間，剛開始不會斷訊，過了兩三分鐘後就開始會斷訊了，不好的地方在這，傳話的速度也是有點太慢了[9]。」這篇文章則是提到了網路電話便宜，但是通話品質不良以及傳輸速度過慢等重點。這樣平凡無奇的文章，卻隱含了部落客對網路電話的使用心得，對本研究而言是相當重要的研究資料，每一筆都是我們值得蒐集的寶貴資料。

3.3 資料前處理

在文件的資訊擷取上，通常選擇「詞彙-文章」(Term-Document）向量矩陣的方法來解決。利用詞彙-文章向量矩陣的方法，可以將蒐集的文章轉

[8] 資料來源：台灣雅虎奇魔部落格 http://tw.myblog.yahoo.com/jw!_BUauruBBQdqBZq2qfiSH4kR/article?mid=3&pk=%22%E7%B6%B2%E8%B7%AF%E9%9B%BB%E8%A9%B1%22

[9] 資料來源：台灣雅虎奇魔部落格 http://tw.myblog.yahoo.com/jw!Jepn1EeeHwHSLTtAITLq1kSw/article?mid=748&pk=%22%E7%B6%B2%E8%B7%AF%E9%9B%BB%E8%A9%B1%22

為向量矩陣。首先要先定義關鍵字，接下來文章中有出現關鍵字，給予 1 的向量值，沒有出現則給予 0 的向量值。詞彙-文章向量矩陣方法的優點，是可以擷取出能夠代表文章的特徵向量，並且解決兩個中國文字詞彙的困擾—相同意義卻能以不同詞彙表達之「一義多字」，以及相同詞彙卻有不同意義之「一字多義」(曾韋榮，2006)。

因此，在進行我們的研究時，首要的步驟便是針對蒐集來的網路電話文章，進行資料前處理，其流程如圖 3-2 所示，詳細的步驟與目的如下：

步驟 1　定義關鍵字：首先我們先蒐集 50 筆網路電話部落格文章，並針對網路電話的特性，分別以『網路電話服務品質』及『網路電話產品品質』為主要衡量構面來定義出關鍵字，希望能精準的擷取能代表該文章的特徵向量。同時，我們為了瞭解部落客對於網路電話使用的滿意程度，新增『火星文及使用者情緒字眼』構面來定義關鍵字，協助我們瞭解部落客使用網路電話的滿意度。例如：「通話延遲」、「通話品質」屬於網路電話服務品質構面之關鍵字；「網路攝影機」、「視訊」屬於網路電話產品品質構面之關鍵字；「Orz」、「囧」屬於火星文及使用者情緒字眼構面之關鍵字。

步驟 2　刪去贅詞之關鍵字：由於定義的關鍵字，有些是屬於口吻性用語，在文章中不斷的出現，也不具有意義，嚴重的影響研究的效果及意義，因此我們將此類的關鍵字視為雜訊，將其刪去。例如：「有」、

「沒有」等關鍵字即是屬於贅詞之關鍵字。

步驟 3　整合關鍵字：在步驟 1 所定義的關鍵字，有許多關鍵字是具有相似的意義，因此我們選擇將其整合，並給予整合後的關鍵字適當的詞彙。例如：「藍芽」與「免持聽筒」兩個關鍵字皆有無線傳輸的功能及意義，因此我們將其整合為「無線傳輸」關鍵字。

步驟 4　刪去未出現或出現頻率極低之關鍵字：在步驟 1 定義的關鍵字，有些未曾在所蒐集的網路電話文章中，因此在本研究中不具有意義，我們選擇將這些沒有出現的關鍵字刪去。例如：「來電保留」關鍵字即未在本研究出現過，因而刪去。

步驟 5　產生詞彙-文章矩陣：經過了步驟 1 至步驟 4 的階段，我們可以得到最後的關鍵字定義表。接著我們將部落格上所蒐集的網路電話文章，依照最後的關鍵字定義表，產生一個詞彙為 5、文章為 5 的詞彙-文章矩陣，如表 3-1 所示。

步驟 6　詞彙-文章矩陣正規化：將步驟 5 所得之詞彙-文章矩陣值進行正規

表 3-1　詞彙-文章矩陣表

	通話品質	網路攝影機	視訊	Orz	無線傳輸
文章 1	1	0	0	0	0
文章 2	2	0	1	0	0
文章 3	0	0	0	0	0
文章 4	4	0	0	0	0
文章 5	1	1	0	0	0

化，使得矩陣內的向量值介於 0~1 之間，方便我們在往後的研究步驟中進行實驗。例如：將表 3-1 正規化後，得到表 3-2。

表 3-2　正規化後之詞彙-文章矩陣表

	通話品質	網路攝影機	視訊	Orz	無線傳輸
文章 1	0.25	0	0	0	0
文章 2	0.5	0	0.2	0	0
文章 3	0	0	0	0	0
文章 4	1	0	0	0	0
文章 5	0.25	0.33333	0	0	0

當我們從步驟 1 開始，到步驟 6 結束這樣一個處理程序，就完成了資料前處理，並獲得一個正規化後的詞彙-文章向量矩陣，接著我們就可以著手進行下一個研究方法。

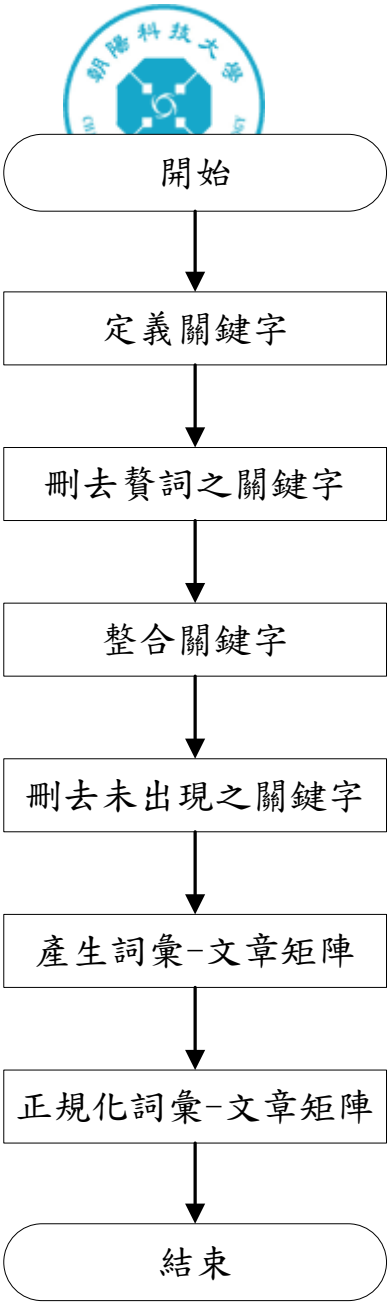


圖 3-2 資料前處理流程圖

3.4 資料分群

在資料分群的研究步驟中，本論文使用的研究方法是自我組織特徵映射圖。因為自我組織特徵映射圖的優點，能夠將複雜且高維度的輸入資料，轉換投射到較低維度的一維或二維的地圖上，因此自我組織特徵映射圖可以用來協助建構市場區隔的決策支援系統（Kiang at al., 2006）。因此在本節中我們將詳述本研究在資料分群上，所使用的自我組織特徵映射圖之研究方法。

本論文提出的部落格探勘模組，第一個研究目的乃是針對部落客發表的文章，應用 SOM 作為顧客區隔之研究方法。本論文應用 SOM 的流程如圖 3-3 所示，詳細的步驟描述如下：

步驟 1　建構神經網路：決定輸入層的神經元個數，以及二維輸出層的神經元個數。

步驟 2　讀取輸入資料：將部落格蒐集而來的詞彙-文章矩陣作為輸入資料，並且隨機給予初始權重。

步驟 3　計算勝利神經元：將輸入資料投射到輸出層，並計算出若干個勝利神經元。

步驟 4　調整權重及更新鄰近距離：　求出勝利神經元後，調整勝利神經元的權重，並且更新鄰近距離內的神經元，使得輸出層的向量資料接近輸入層的向量資料。

步驟 5　訓練終止：當特徵映射圖形成或者是迭代次數終止條件達成後，即停止訓練。如果尚未形成特徵映射圖或達到迭代次數終止條件，則回到步驟 3。

步驟 6　顧客區隔：當完成了步驟 5 之後，將獲得若干群聚，這些群聚代表著部落格文章發表者對網路電話產品需求上的差異，達到顧客區隔的目的。

當完成了顧客區隔後，我們便應用這個顧客區隔的結果，來進行顧客分類。換句話說，針對往後的部落格上，討論網路電話的文章，我們可以預測該發表者屬於哪一群使用網路電話的顧客。

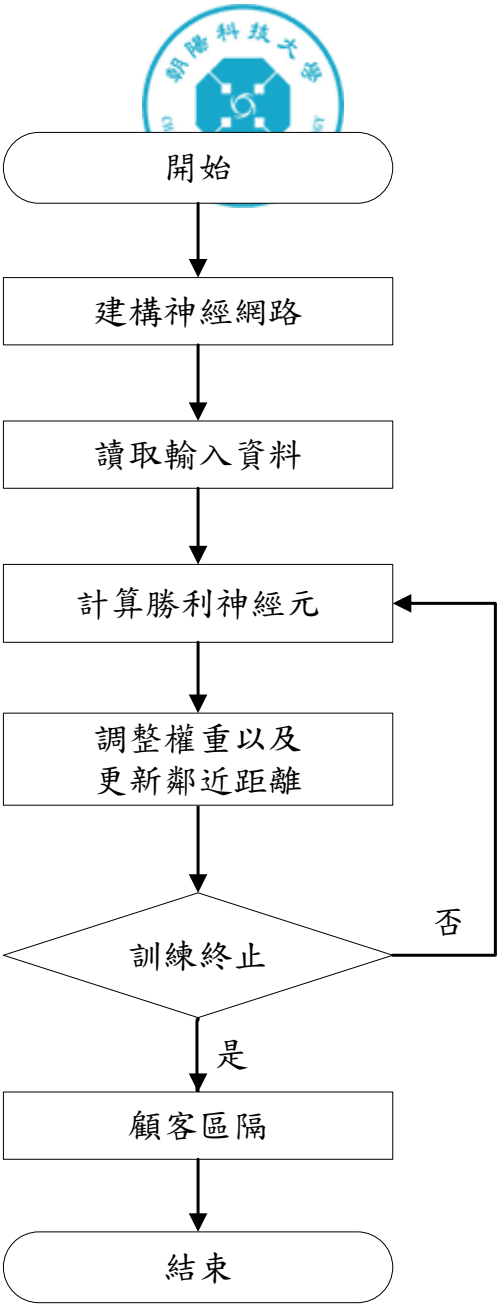


圖 3-3 應用 SOM 於顧客區隔流程圖

3.5 資料分類

在完成了資料分群的步驟之後，我們會獲得數個部落格文章分群的群聚，每一個群聚都將擁有自己特殊的特色，代表著不同部落客對網路電話這項產品，將會有不同的需求面。於是我們利用資料分群步驟所獲得的數個群聚，作為我們分類預測的目標；從部落格上蒐集的網路電話文章則作為我們欲分類的對象，進行分類預測的實驗。

本研究所採用的實驗工具，為 BPN、SVM 以及決策樹等三種知名的分類方法。在所蒐集部落格上的網路電話文章中，我們取 80%比例的文章作為訓練資料，剩下的 20%比例的文章作為測試資料，並且將資料切成 5 個等分資料，以作為交互驗證（Cross-Validate）。資料分類模型建立的流程，如圖 3-4 所示，詳細步驟敘述如下：

步驟 1　確定輸入變數：在這個步驟，我們利用先前 3.2 節得到的詞彙-文章向量矩陣，將其依照 80%與 20%的比例分成訓練資料以及測試資料，並將資料切成 5 個等分資料，以作為交互驗證。

步驟 2　確定訓練輸出目標：我們利用 3.3 節資料分群的群聚結果，將這些群聚作為監督式學習之訓練資料輸出的目標值。

步驟 3　確定各監督式學習方法參數：在 BPN 方面，要確定網路設置的參數、學習速率、動量係數及迭代次數等相關參數；在 SVM 方面，要確定放射型核心函數的C值與γ值等參數；在決策樹方面，則是

要確定預估錯誤率的修剪參數。

步驟 4　執行各監督式學習方法

步驟 5　獲得分類預測數據並結束。

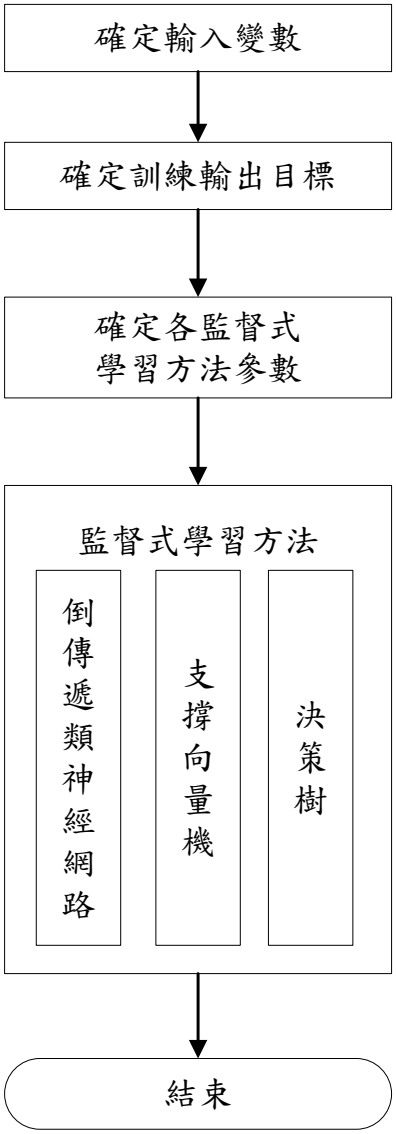


圖 3-4　資料分類模型建立流程圖

3.6 維度縮減

3.6.1 特徵選取

特徵選取通常被視為一種特徵衡量的方法，來決定哪一個特徵要放在第一個位置或最後一個位置。一般而言，特徵選取的演算步驟可以分割為三個平行的步驟：封裝（Wrapper）、過濾（Filter）以及嵌進（Embedded）（Su & Yang）。

封裝是一種新子集合的特徵選取方法，第一階段先利用預測精準度來挑選新子集合的特徵，接著第二階段是學習訓練資料並在測試資料上作測試。過濾是一種預先處理的方法，第一階段是利用距離、資訊或依賴度來衡量所要選取的特徵，接著第二階段一樣是學習訓練資料並在測試資料上作測試。嵌進則是一種機器學習的演算法（Su & Yang）。

本研究採用的特徵選取為連續向前選取法，主要是由空特徵變數集合開始進行，接著由一維的變數組合測試，透過一次增加一個特徵變數，直到已達到要求特徵變數數目為止。我們利用的是倒傳遞類神經網路的權重，作為我們特徵選取的特徵。蕭宇翔（2005）指出，倒傳遞類神經網路的特徵選取可由以下的步驟進行：

步驟 1　在每個輸入神經元，計算所鏈結的「輸入層-隱藏層」權重，以及「隱藏層-輸出層」權重之絕對值乘積，並且相加。

步驟 2　將每個輸入神經元依據步驟 1 所得之值，以遞減的方式排列，排

列越後面的，表示該神經元越不重要。

步驟 3　透過步驟 2 即可篩選出重要的特徵。

因此，我們篩選出重要的特徵後，由最重要的特徵開始進行分類預測，接著增加次重要的特徵進行分類預測，逐步的增加特徵，直到已達要求的特徵數目為止。特徵選取示意圖如下圖 3-5 所示。我們從原始矩陣中，得到 Attribute 5 的權重值為第一優先挑選特徵，而 Attribute 2 之特徵為第二優先選取特徵。

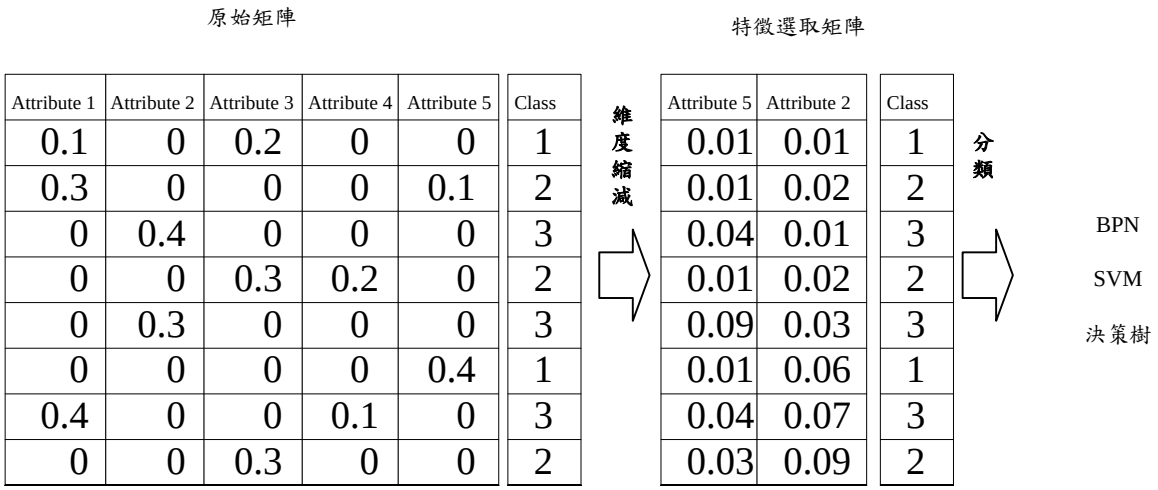


圖 3-5 特徵選取示意圖

3.6.2 屬性擷取

1. PCA 方法及理論

PCA 之主要目的，是希望以較少數目的變數解釋原本資料中大部份的變異，更期望能將既有之高相關性變數轉化為線性組合變數，亦即以較少數目之線性組合變數解釋大部份原始資料中之變異。PCA 的方法及步驟如下（許邦輝，2006）：

步驟 1　利用 Z-score，將原始變數標準化。

步驟 2　求取任兩變數之相關係數，並計算相關係數矩陣。

步驟 3　計算特徵值及特徵向量。

步驟 4　求取新成分，自第一特徵值、第二特徵值逐步取之。

步驟 5　計算各成分之貢獻率及累積貢獻率。

步驟 6　選擇主成分，亦即選擇對原始資料變異解釋能力較高之新成分。

PCA 示意圖如下圖 3-6 所示。原始矩陣中，有五個屬性值，經由 PCA 屬性擷取之後，我們得到新成分 PC 1 代表第一特徵值，PC 2 代表第二特徵值，利用所得的 PCA 矩陣，即能解釋原先五個屬性值所表示的原始矩陣。

原始矩陣

Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Class
0.1	0	0.2	0	0	1
0.3	0	0	0	0.1	2
0	0.4	0	0	0	3
0	0	0.3	0.2	0	2
0	0.3	0	0	0	3
0	0	0	0	0.4	1
0.4	0	0	0.1	0	3
0	0	0.3	0	0	2

維度縮減

PCA矩陣

PC 1	PC 2	Class
0.01	0.01	1
0.01	0.02	2
0.04	0.01	3
0.01	0.02	2
0.09	0.03	3
0.01	0.06	1
0.04	0.07	3
0.03	0.09	2

分類

BPN

SVM

決策樹

圖 3-6　PCA 示意圖

2. ICA 方法及理論

ICA 目的是要讓成分之間的統計相依性降至最低，也就是成分與成分之間，互為獨立情形，根據李銘浚（2007）的研究，ICA 最主要有三個流程如下：

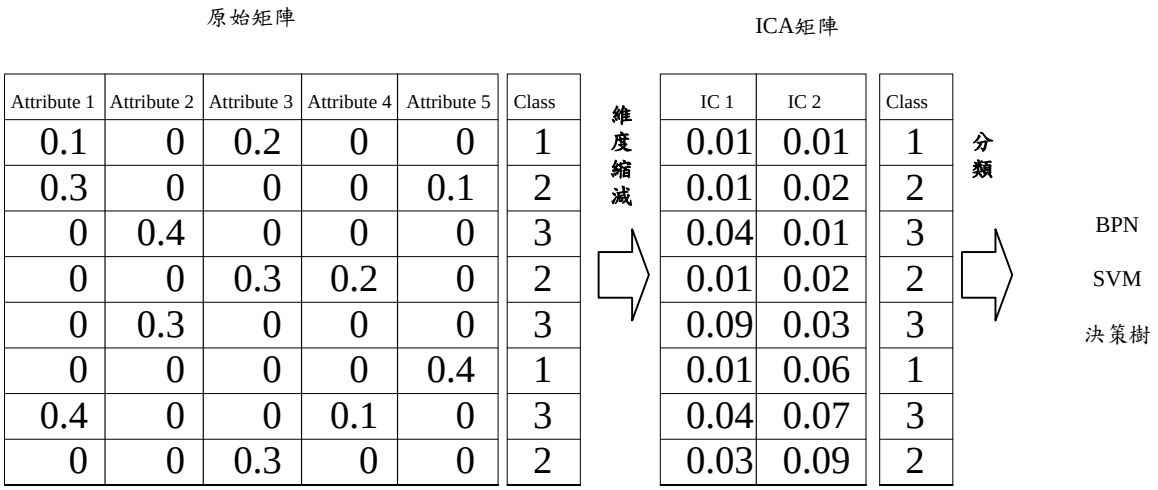
步驟 1　集中變數（Centering），扣除訊號的平均值，使其平均值為零，可以簡化推導的過程。

步驟 2　資料白色化（Whitening），目的是使轉換後的資料，彼此之間具有非相關性。

步驟 3　解混合矩陣（Demixing Matrix），分離出彼此互相獨立的訊號。

ICA 示意圖如下圖 3-7 所示。原始矩陣中，有五個屬性值，經由 ICA 屬

性擷取之後，我們得到新成分 IC 1 代表第一特徵值，IC 2 代表第二特徵值，利用所得的 ICA 矩陣，即能解釋原先五個屬性值所表示的原始矩陣。



Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Class
0.1	0	0.2	0	0	1
0.3	0	0	0	0.1	2
0	0.4	0	0	0	3
0	0	0.3	0.2	0	2
0	0.3	0	0	0	3
0	0	0	0	0.4	1
0.4	0	0	0.1	0	3
0	0	0.3	0	0	2

IC 1	IC 2	Class
0.01	0.01	1
0.01	0.02	2
0.04	0.01	3
0.01	0.02	2
0.09	0.03	3
0.01	0.06	1
0.04	0.07	3
0.03	0.09	2

圖 3-7　ICA 示意圖

3. LSI 方法及理論

資訊檢索的過程中，常常受到「多字一義」、「多義一字」及「詞彙獨立性」等問題的困擾，而影響檢索的精確度。因此 Deerwester et al.（1990）提出 LSI 方法，來解決資訊檢索所面臨的困擾。隱含語意索引主要的方法，是使用 SVD 將原詞彙-文章向量矩陣，分解投射到維度較小的向量空間，而詞彙和文章有較顯著相關的將聚集在一起。維度縮減圖如下圖 3-8，將原始的 A 矩陣分解後取 k 維度，得到一個維度縮減後的 A_K 向量矩陣。在此，本研究將 LSI 視為一種屬性擷取工具，希望能藉由 LSI 能解決在資訊檢索上

的困擾，以及向量空間維度縮減的特性，發展改善稀疏性資料分類器。我們分別採用 Global-LSI 及 Local-LSI 兩種方法進行，接著取 k 維度後，以維度縮減後的 A_K 向量矩陣進行分類預測的實驗，並與未施行 LSI 方法比較，

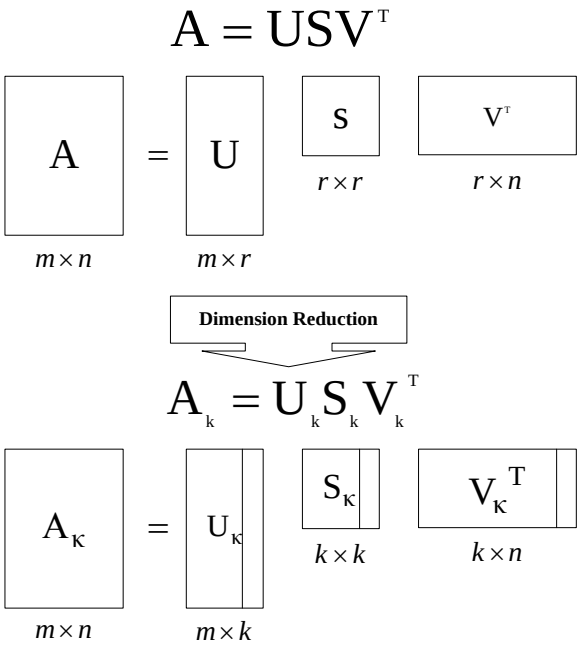


圖 3-8 詞彙-文章向量矩陣縮減為 k 維度過程

（資料來源：修改自曾韋榮，2006）

最後驗證是否有效改善稀疏性資料的分類準確率。不論是哪一種方法，我們一樣取 80%比例的文章作為訓練資料，剩下的 20%比例的文章作為測試資料，並且將資料切成 5 個等分資料，以作為交互驗證。詳細方法與步驟如下：

方法一　Global-LSI

步驟 1　原始矩陣 A 分解：利用 SVD 將 3.2 節所獲得的原始詞彙-文章向量矩陣 A 分解。

步驟 2　縮減為 A_k 矩陣：取 k 維度，縮減為新的 A_k 矩陣。k 值介於 1～30。

步驟 3　資料分類預測：以縮減維度後之 A_k 矩陣，取代原始矩陣 A 作為 3.4 節分類預測之輸入資料，進行分類預測。

步驟 4　獲得分類預測數據並結束。

Global-LSI 配合資料分類流程如下圖 3-9 所示，Global-LSI 示意圖如下圖 3-10 所示。原始矩陣中，有五個屬性值，經由 Global-LSI 屬性擷取之後，我們得到新成分 Atrribute 1'代表第一特徵值，Atrribute 2'代表第二特徵值，利用所得的 A_k 矩陣，即能解釋原先五個屬性值所表示的原始矩陣。

方法二　Local-LSI

步驟 1　群聚矩陣分解：Local-LSI 的作法有別於 Global-LSI 最大的差別，便是在這一步驟。Global-LSI 做 SVD 矩陣分解是全部矩陣下去分解，假設有 100 筆資料，那麼便是將 100 筆資料做 SVD 矩陣分解；而 Local-LSI 的作法，針對不同的類別（class）進行分解。一樣假設 100 筆資料，可是分為兩類，第一類為 60 筆，第二類為 40 筆資料，那麼 Local-LSI 便是將 60 筆與 40 筆的矩陣分別做 SVD 矩陣分解。

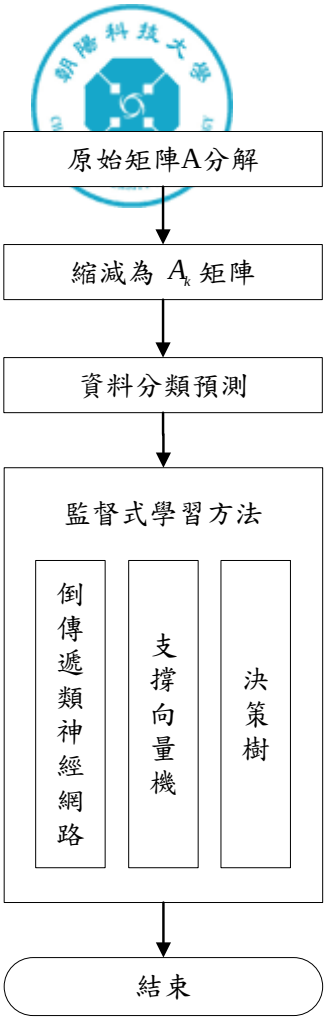


圖 3-9　Global-LSI 配合資料分類流程圖

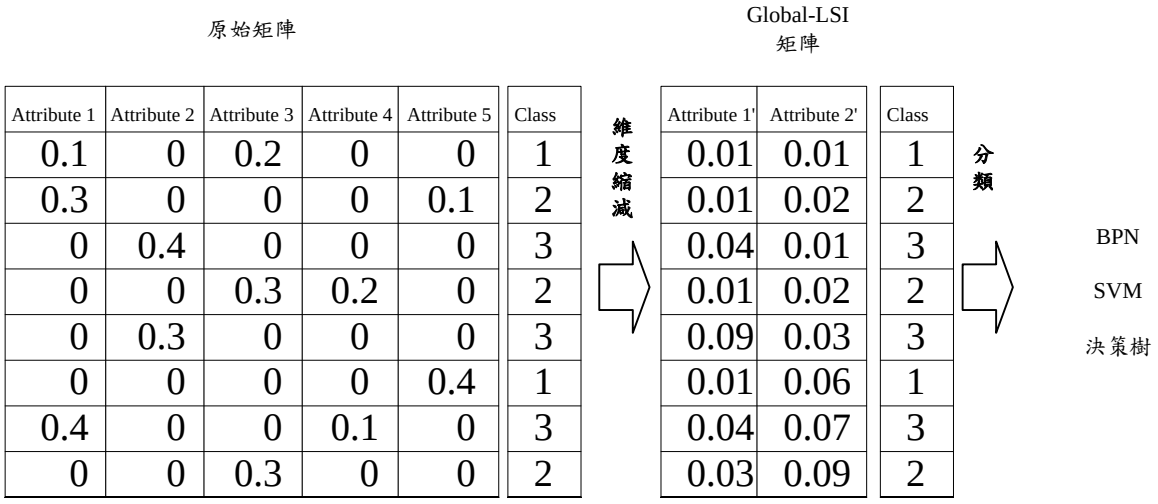


圖 3-10　Global-LSI 示意圖

步驟 2　縮減為 A_{kc} 矩陣：為了區分 Global-LSI 的 A_k 縮減矩陣，我們將 Local-LSI 縮減為 k 維度的矩陣命名為 A_{kc} 縮減矩陣。Global-LSI 的 A_k 矩陣是將原始矩陣分解後，取 k 維度即可得；而 Local-LSI 的 A_{kc} 矩陣則是將各群聚分解的矩陣，按照原始矩陣的資料順序排列回去。承步驟 1 的例子，將 60 筆資料縮減為 A_1 矩陣，將 40 筆資料縮減為 A_2 矩陣，最後將 A_1 及 A_2 按照 A 矩陣的順序排列回去，成為 A_{kc} 縮減矩陣。

步驟 3　資料分類預測：以縮減維度後之 A_{kc} 矩陣，取代原始矩陣 A 作為 3.4 節分類預測之輸入資料，進行分類預測。

步驟 4　獲得分類預測數據並結束。

Local-LSI 配合資料分類流程如下圖 3-11 所示，Local-LSI 示意圖如下圖 3-12 所示。原始矩陣中，有五個屬性值，經由 Local-LSI 屬性擷取之後，我們得到新成分 Atrribute 1'代表第一特徵值，Atrribute 2'代表第二特徵值，利用所得的 A_{kc} 矩陣，即能解釋原先五個屬性值所表示的原始矩陣。

在執行完維度縮減搭配資料分類預測後，我們可以將其分類預測準確率與未經過維度縮減分類預測的原始資料分類預測準確率做個比較，觀察是否有效改善稀疏性資料的分類預測準確率。

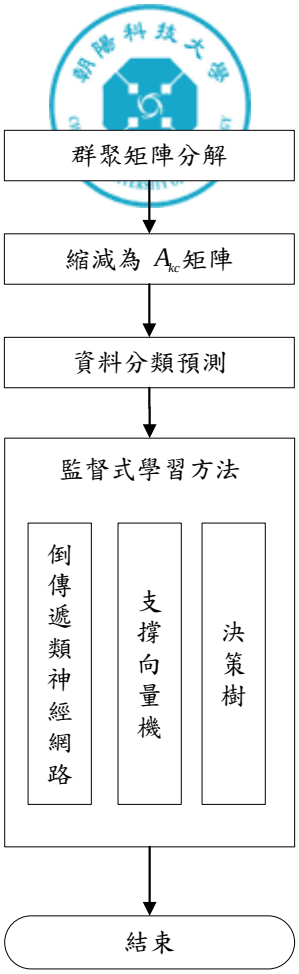


圖 3-11　Local-LSI 配合資料分類流程圖

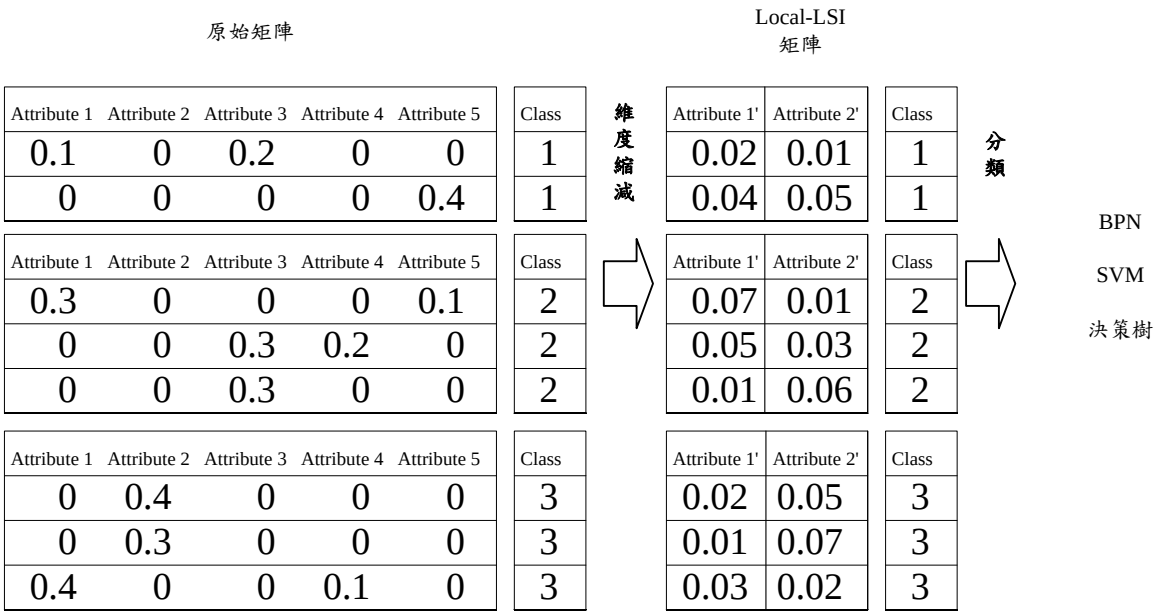


圖 3-12　Local-LSI 示意圖

第四章　實驗結果與分析

本章實際蒐集發表於部落格之網路電話文章，實證本研究提出之部落格探勘模組。實驗資料及環境敘述如下：

1. 實驗資料：以 Yahoo!奇摩部落格、無名小站部落格為主，蒐集兩百筆討論網路電話的文章。文章範例如：

(1) 我下午掛完跟你們的網路電話，馬上跟我同學連線，我跟他的通話就非常非常清楚，跟在台灣講電話一樣，完全不會 LAG 應該要這樣才對，不知道是不是我們家寬頻太慢，我搞清楚再跟你們說，要是是寬頻問題，你們就去升等好了，你們今天跟我對話的軟體叫做 skype，說明書在電腦桌子墊子下面，視訊我今天是跟你用 MSN，所以有兩種軟體，兩種都可以語音跟視訊，不用錢請多利用。

(2) 今天用 skype 的網路電話，通話費是很便宜，但是通話品質不是很好，還有很大的進步空間，剛開始不會斷訊，過了兩三分鐘後就開始會斷訊了，不好的地方在這，傳話的速度也是有點太慢了。

2. 實驗環境：中央處理器為 AMD Athlon 64 Processor 3200+

記憶體為 DDR 2　667　1G

作業系統為 Microsoft Windows XP Professional

實驗軟體有 Matlab 7.0.1、Libsvm-2.83、See5/C5.0 2.05 版

4.1 資料前處理

資料蒐集完畢後，接著即是進行資料前處理，共分為定義關鍵字、刪去贅詞之關鍵字、整合關鍵字、刪去未出現或出現頻率極低之關鍵字、產生詞彙-文章向量矩陣及向量矩陣正規化等 6 個步驟，分敘如次：

步驟 1　定義關鍵字：以「網路電話服務品質」、「網路電話產品品質」以及「火星文集使用者情緒字眼」構面來定義關鍵字，共定義出 55 個關鍵字，如表 4-1 所示。

表 4-1 關鍵字定義

	關鍵字	同義字	定義
網路電話服務品質構	通話監聽		網路電話傳輸的封包遭到竊取
	通話干擾	雜訊、雜音	網路電話傳輸的線路受到雜訊干擾。
	通話延遲	Lag、delay	網路電話收話方，沒有和發話方同步，出現延遲的現象。
	通話品質	音質	整體網路電話在通話上的品質。
	加值服務		網路電話的加值服務一般而言有：來電轉接、語音信箱、簡訊服務、圖鈴下載、交友中心等等，依照各家網路電話服務提供商而有所不同。[林士玄，民 95]
	通話費率		依照每家網路電話服務提供商不同，會有不同的網路電話撥打費用方案。
	通話費	電話費	每個月撥打網路電話所花的費用。
	配號	門號	以 070 開頭的網路電話號碼［黃應欽，民 95］
	帳單		網路電話服務提供商每月寄發的單子，上頭註明應該繳費的金額以及明細。
	多方會談		能夠允許三人或三人以上的多人，同時加入網路電話會談。
	國際電話		跨越國界的電話。例如台灣撥打至美、加、大

面			陸、澳洲等地。
	國內長途		同在一個國界內，分屬不同縣市。例如台灣台中縣撥打至台灣台北市。
	頻寬		在一定時間內，有多少資訊可以從一個地方傳輸到另一個地方的資訊流度量標準；也代表傳輸媒體或設備的傳輸能力。[孫明源, 民 92]
	撥通成功率		撥打網路電話，成功通話的機會
	斷線率		撥打網路電話，成功通話之後，中斷通話的機率。
	網內互打		收發話方，同屬相同的網路電話服務提供商。
	Skype		網路電話軟體
	MSN		網路電話軟體
網路電話產品品質構面	麥克風	耳機麥克風	聲音的輸入裝置
	藍芽		
	免持聽筒		通話時，不需手執話筒的功能。
	擴音		通話時，能夠將聲音放到最大。
	重量		話機的重量
	喇叭		聲音的輸出裝置
	迴音		通話時，能夠聽到話筒傳來自己說話的迴音。
	液晶螢幕	液晶顯示器、LCD	利用液晶材料的特性，加上電路的驅動促使液晶轉向，再以控制外部光源穿透，來達到明暗效果的光電顯示元件，能將電子訊息轉變為光學訊號的光電裝置。[張錦村, 民 94]
	網路攝影機	Webcam	能夠將本地的鏡頭捕捉，並且將影像傳送到網路的另一端。
	視訊		透過攝影機，將彼此的影像傳輸給對方。
	來電鈴聲		當來電時，話機所呈現的鈴聲。
	來電保留		通話至一半時，能夠將通話保留，不會中斷。
	待機時間		話機不通話時，電池能保持待機的時間之電力衡量單位。
	通話時間		話機通話時，電池能保持通話的時間之電力衡量單位。
	電池		話機的電力儲存設備
火	Orz	OTZ、or2	佩服的五體投地。
	囧	囧 rz、冏	無奈之意。

星文及消費者情緒字眼構面	^_^	^^”、^”^	開心之意
	XD		大笑之意
	：)	：-)、：D	微笑之意
	>_<	><、>”<	難受之意
	貴		通話費用價格高
	便宜		通話費用價格低
	:(	：-（	愁眉苦臉之意
	= =	= =”、=.=	
	方便	便利	
	清晰	清楚	對話的內容很清楚
	模糊		對話的內容不清楚
	省錢	節費	可以降低通話的費用
	免費		不必花任何的通話費用
	好		
	不好		
	會		
	不會		
	有		
	沒有		
	讚	棒	

步驟 2　刪去贅詞之關鍵字：將不具意義、口吻性用語之關鍵字刪去，如表 4-2 所示。

表 4-2　刪去贅詞關鍵字

好	不好	有	沒有	會	不會

步驟 3　整合關鍵字：將相似意義之關鍵字合併，如表 4-3 所示。

表 4-3　整合前後之關鍵字

整合前關鍵字	整合後關鍵字
通話干擾 通話延遲 通話品質	通話品質

通話費率 通話費 帳單	通話費用
斷線率 撥通成功率	斷線率
藍芽 免持聽筒	無線傳輸
待機時間 通話時間 電池	電池
0rz 囧	0rz
^_^ ：)	：)
便宜 省錢	省錢
>_< ：(= =	：(

步驟 4　刪去未出現或出現頻率極低之關鍵字：未出現及出現頻率低，在本研究不具任何意義，將其刪去，如表 4-4 所示。

表 4-4　刪去未出現之關鍵字

來電保留	液晶螢幕	重量

步驟 5　產生詞彙-文章向量矩陣：經過了步驟 1 至步驟 4 的階段，我們可以得到最後的關鍵字定義表，如表 4-5。接著我們將部落格上所蒐集的 200 筆網路電話文章，依照表 4-5 所列的 33 個關鍵字，產生一個詞彙為 33、文章為 200 的詞彙-文章向量矩陣。

表 4-5　前處理後之關鍵字定義表

通話監聽	通話品質	加值服務	通話費用	配號	多方會談
國際電話	國內長途	頻寬	斷線率	網內互打	SKYPE
MSN	麥克風	無線傳輸	擴音	喇叭	迴音
網路攝影機	視訊	來電鈴聲	電池	Orz	：）
XD	：(	貴	省錢	方便	清晰
模糊	免費	讚			

步驟 6　向量矩陣正規化：將步驟 5 所得之向量矩陣值進行正規化，使得矩陣內的向量值介於 0~1 之間，方便我們在往後的研究步驟中進行實驗。

4.2 文章分群

以 Matlab 7.0.1 toolbox 執行 SOM，實驗參數設置為：輸入資料維度為 33 個關鍵字詞彙*200 筆文章，輸出拓樸層為 1*3 的二維地圖，粗調學習速率為 0.9，細調學習速率為 0.1，半徑距離為 1，終止條件為迭代次數 100 次。執行完畢後，得到群聚如表 4-6 所示。

表 4-6 自我組織特徵映射分群結果

	關鍵字	比例		關鍵字	比例		關鍵字	比例
第一群	Skype	19.08%	第二群	Skype	27.69%	第三群	Skype	18.92%
	麥克風	9.92%		通話品質	20.92%		免費	11.64%
	通話費用	9.16%		省錢	6.46%		省錢	11.02%
	視訊	8.65%		通話費用	6.15%		通話費用	9.98%

第一群	省錢	7.89%	第二群	MSN	5.54%	第三群	通話品質	8.73%
	通話品質	5.85%		麥克風	4.92%		國際電話	7.28%
	國際電話	5.09%		視訊	4.62%		網內互打	4.57%
	MSN	4.07%		免費	3.38%		視訊	3.74%
	斷線率	3.56%		無限傳輸	3.08%		方便	3.53%
	配號	3.31%		國際電話	2.77%		MSN	3.12%
	無限傳輸	2.80%		網路攝影機	1.85%		貴	2.91%
	頻寬	2.54%		方便	1.85%		國內長途	2.29%
	貴	2.04%		迴音	1.54%		配號	1.87%
	國內長途	1.78%		XD	1.54%		清晰	1.87%
	方便	1.78%		國內長途	1.23%		斷線率	1.66%
	網路攝影機	1.53%		頻寬	1.23%		多方會談	1.25%
	：）	1.53%		斷線率	0.92%		頻寬	1.25%
	免費	1.53%		：）	0.92%		麥克風	1.25%
	通話監聽	1.27%		貴	0.92%		模糊	1.25%
	迴音	1.27%		清晰	0.92%		通話監聽	1.04%
	：（	1.02%		多方會談	0.31%		加值服務	0.42%
	擴音	0.76%		網內互打	0.31%		迴音	0.21%
	來電鈴聲	0.76%		電池	0.31%		讚	0.21%
	清晰	0.76%		：（	0.31%		無限傳輸	0.00%
	多方會談	0.51%		讚	0.31%		擴音	0.00%
	電池	0.51%		通話監聽	0.00%		喇叭	0.00%
	喇叭	0.25%		加值服務	0.00%		網路攝影機	0.00%
	Orz	0.25%		配號	0.00%		來電鈴聲	0.00%
	XD	0.25%		擴音	0.00%		電池	0.00%
	模糊	0.25%		喇叭	0.00%		Orz	0.00%
	加值服務	0.00%		來電鈴聲	0.00%		：）	0.00%
	網內互打	0.00%		Orz	0.00%		XD	0.00%
	讚	0.00%		模糊	0.00%		：（	0.00%

接著我們取前 70%累加比例來分析群聚的結果，如表 4-7、4-8、4-9 所示。

表 4-7 第一類前 70%累加比例表

	關鍵字	比例	累加比例
第一類	Skype	19.08%	19.08%
	麥克風	9.92%	29.01%
	通話費用	9.16%	38.17%
	視訊	8.65%	46.82%
	省錢	7.89%	54.71%
	通話品質	5.85%	60.56%
	國際電話	5.09%	65.65%
	MSN	4.07%	69.72%

表 4-8 第二類前 70%累加比例表

	關鍵字	比例	累加比例
第二類	Skype	27.69%	27.69%
	通話品質	20.92%	48.62%
	省錢	6.46%	55.08%
	通話費用	6.15%	61.23%
	MSN	5.54%	66.77%
	麥克風	4.92%	71.69%

表 4-9 第三類前 70%累加比例表

	關鍵字	比例	累加比例
第三類	Skype	18.92%	18.92%
	免費	11.64%	30.56%
	省錢	11.02%	41.58%
	通話費用	9.98%	51.56%
	通話品質	8.73%	60.29%
	國際電話	7.28%	67.57%
	網內互打	4.57%	72.14%

從表 4-7 來看，可以發現第一類除了基本的通話費用、品質外，他們還著重在「麥克風」及「視訊」上。這說明著，會在部落格發表這些文章的使用者，他們希望的是功能多樣化的網路電話。因此，本文將第一類的用戶定義為「玩家級」的網路電話部落客。

在表 4-8，第二類有相當高的比例是在探討著「通話品質」。那麼發表這些注重網路電話通話品質文章的部落客，可能是企業相關的用者，因為對於企業而言，每一通商業電話，都可能是一筆重要的交易，因此他們會非常要求網路電話的品質，其次才是通話費用。所以，本文將第二類定義為「企業用戶」的網路電話部落客。

從表 4-9 來看，可以發現第三類有百分之五十的比例，著重在「免費」、「省錢」、「通話費用」等項。這意味著，這一類的文章，是非常重視網路電話的通話金額，在部落格上發表這樣的文章，可能是一些對網路電話要求不高的，只希望有便宜，甚至是免費的網路電話可以使用的部落客所發表的。本文將第三類定義為「經濟實惠」的網路電話部落客。

因此，透過資料分群的方法，本文將發表網路電話產品文章之部落客，分為三類：

第一類：玩家級網路電話部落客。

第二類：企業用戶網路電話部落客。

第三類：經濟實惠網路電話部落客。

4.3 分類預測

在完成了文章分群之後，緊接著要實驗的部分是分類預測。將未知的新進部落格文章，分類在適合的群組，也就是 4.2 節所完成之「玩家級」網路電話部落客、「企業用戶」網路電話部落客以及「經濟實惠」網路電話部落客等三個類別。

我們使用了 BPN、SVM 以及決策樹三種分類預測之方法來進行實驗，觀察其分類預測之正確率。同時間，為了發展改善稀疏性資料分類預測的正確率，我們也加入了 Global-LSI、Local-LSI、PCA、ICA 以及 Feature Selection 等五項維度縮減的技術，並且與未經過維度縮減的分類預測結果作比較。

4.3.1 進行分類預測之資料描述

本研究進行分類預測之資料，為部落格上所蒐集之兩百筆有關網路電話之文章。屬於第一群玩家級網路電話部落客之文章有 107 筆，屬於第二群企業用戶網路電話部落客之文章有 48 筆，屬於第三群經濟實惠網路電話部落客之文章有 45 筆。我們將所有文章以亂數排列，並分割為 Fold 1、Fold 2、Fold 3、Fold 4 及 Fold 5 等 5 個資料集，以作為交互驗證。5 個資料集之資料描述如表 4-10 所示。

表 4-10　各分類預測之資料集描述

	資料集	群組	文章筆數	佔有比例
訓練組	Fold 1	第一群	85	53.13%
		第二群	39	24.38%
		第三群	36	22.49%
	Fold 2	第一群	84	52.50%
		第二群	39	24.38%
		第三群	37	23.12%
	Fold 3	第一群	85	53.13%
		第二群	40	25.00%
		第三群	35	21.87%
	Fold 4	第一群	83	51.88%
		第二群	38	23.74%
		第三群	39	24.38%
	Fold 5	第一群	91	56.88%
		第二群	36	22.49%
		第三群	33	20.63%
測試組	Fold 1	第一群	22	55.00%
		第二群	9	22.50%
		第三群	9	22.50%
	Fold 2	第一群	23	57.50%
		第二群	9	22.50%
		第三群	8	20.00%
	Fold 3	第一群	22	55.00%
		第二群	8	20.00%
		第三群	10	25.00%
	Fold 4	第一群	24	60.00%
		第二群	10	25.00%
		第三群	6	15.00%
	Fold 5	第一群	16	40.00%
		第二群	12	30.00%
		第三群	12	30.00%

4.3.2 BPN、SVM 及決策樹之分類預測結果

下列表 4-11~4-13 與圖 4-1~4-3，為各資料集執行 BPN、SVM 及決策樹之分類預測結果。下列表 4-14 與圖 4-4 為 BPN、SVM 及決策樹分類預測結果比較分析。

表 4-11 執行 BPN 之分類預測正確率

	BPN 分類預測正確率
Fold 1	82.50%
Fold 2	85.00%
Fold 3	92.50%
Fold 4	92.50%
Fold 5	85.00%
平均	87.50%
標準差	4.68%

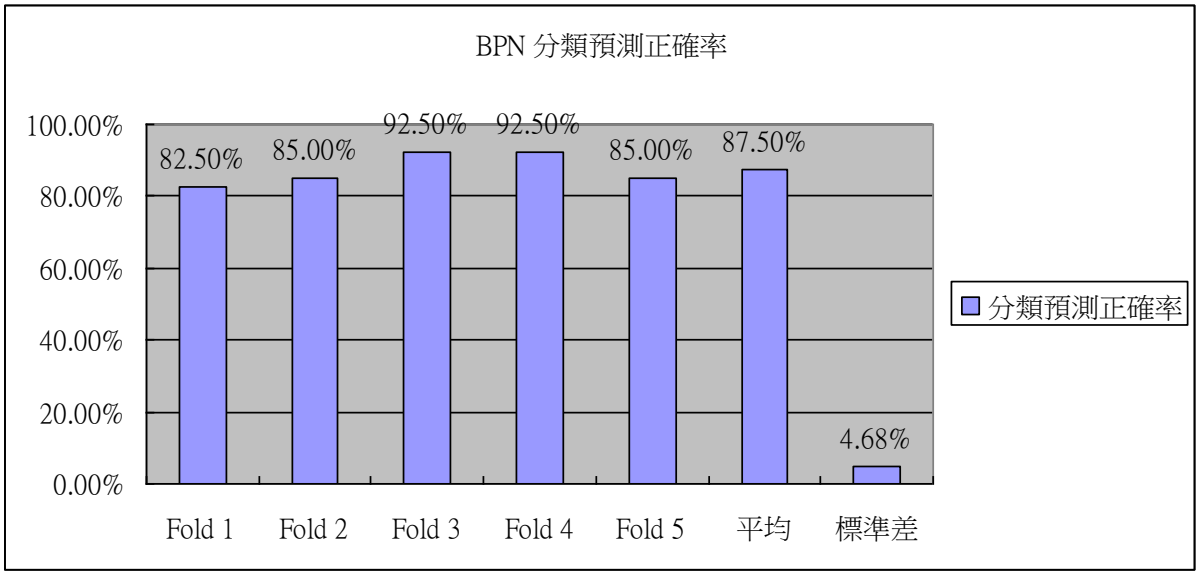


圖 4-1 執行 BPN 之分類預測正確率

表 4-12 執行 SVM 之分類預測正確率

	SVM 分類預測正確率
Fold 1	92.50%
Fold 2	95.00%
Fold 3	87.50%
Fold 4	82.50%
Fold 5	92.50%
平均	90.00%
標準差	5.00%

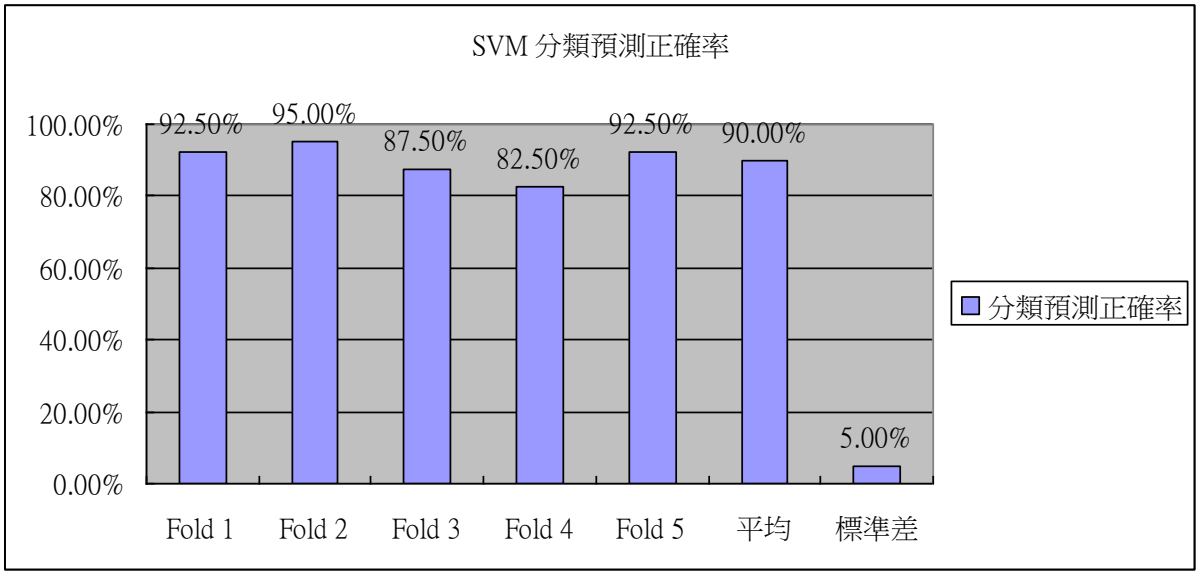


圖 4-2 執行 SVM 之分類預測正確率

表 4-13 執行決策樹之分類預測正確率

	決策樹分類預測正確率
Fold 1	90.00%
Fold 2	92.50%
Fold 3	77.50%
Fold 4	77.50%
Fold 5	75.00%
平均	82.50%
標準差	8.10%

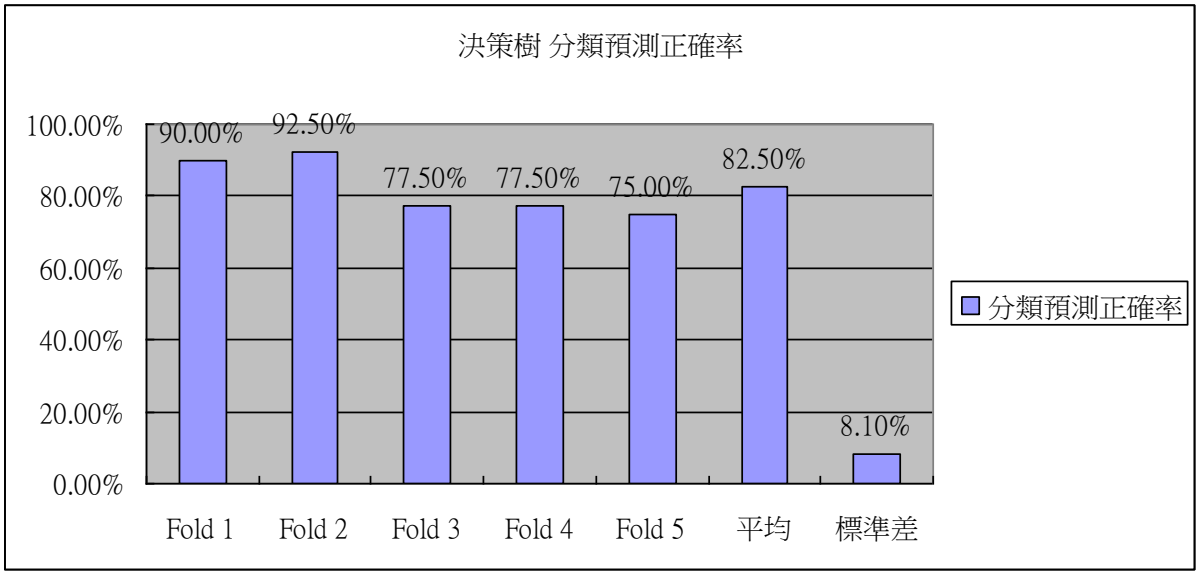


圖 4-3 執行決策樹之分類預測正確率

表 4-14　BPN、SVM 及決策樹之分類預測結果比較

	平均預測正確率	標準差
BPN	87.50%	4.68%
SVM	90.00%	5.00%
決策樹	82.50%	8.10%

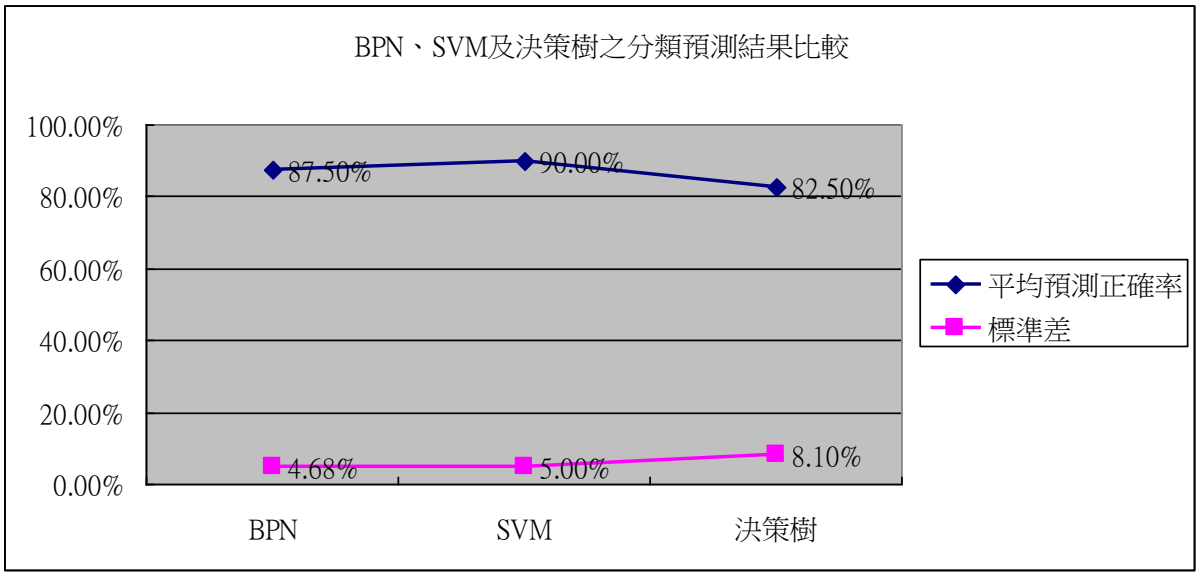


圖 4-4 BPN、SVM 及決策樹之分類預測結果比較

由上表 4-14 及圖 4-4 可以得知，將 5 個資料集之預測正確率平均後，取得平均預測正確率，SVM 以平均正確率 90.00%優於 BPN 之平均正確率 87.50%，並且優於決策樹之最佳正確率 82.50%。以平均預測正確率來看，SVM 為三種方法中之最佳方法，其次為 BPN，最後為決策樹。同時間，我們也觀察其標準差，發現 BPN 及 SVM 標準差皆在 5.00%以下，可以解釋所得到的平均預測正確率之差異不會太大，而決策樹的表現則是稍嫌不

足。因此就平均預測正確率及標準差來看，SVM 之分類預測效果，都優於 BPN 及決策樹，所以利用 SVM 來進行未知新進顧客分類預測到適合的消費群，是最適合的。

4.4 維度縮減

為了發展改善稀疏性資料分類器，我們使用維度縮減的方法來進行，分別是以屬性擷取以及特徵選取來進行。

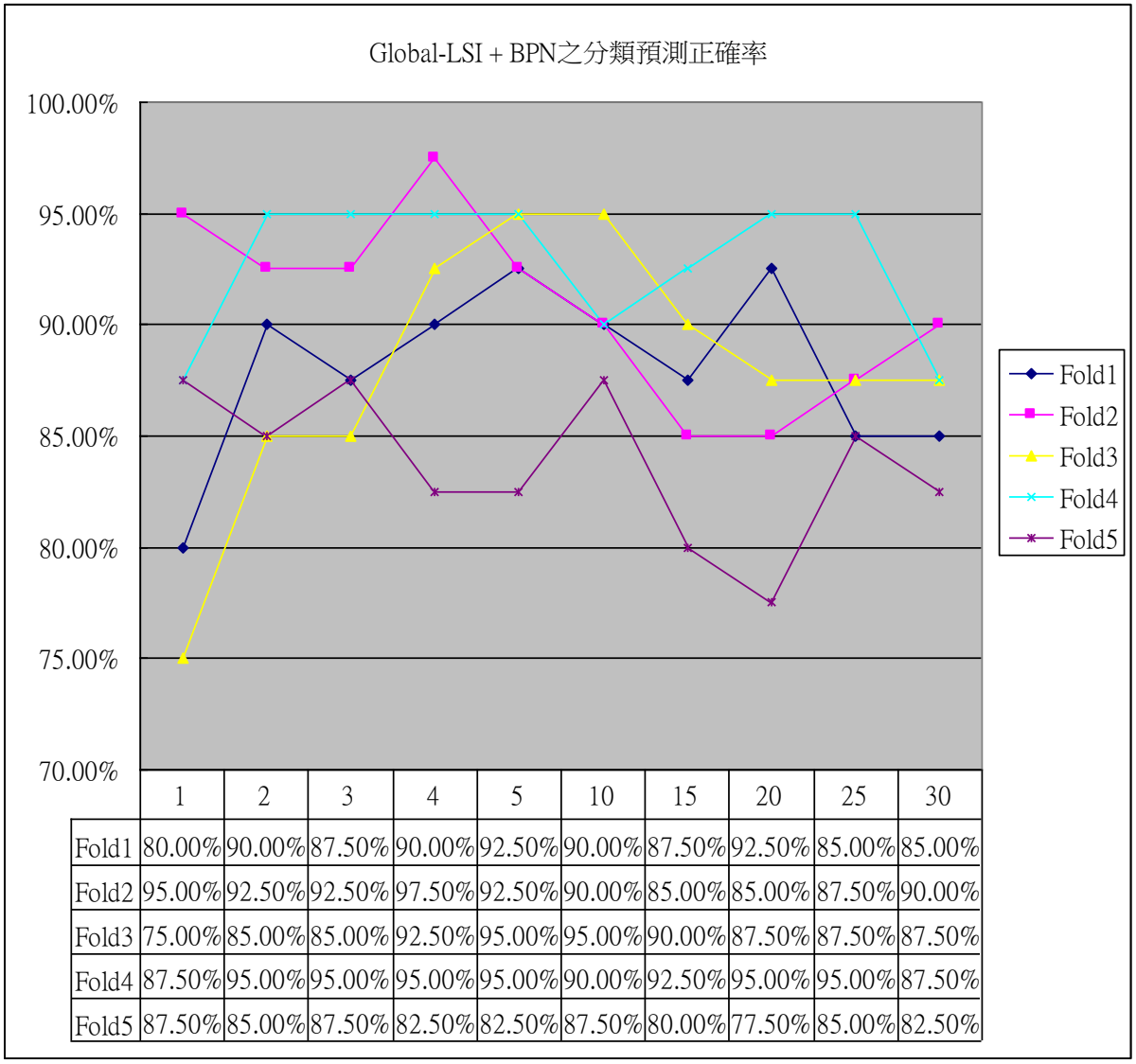
4.4.1 屬性擷取

在屬性擷取的實驗中，我們以 LSI、PCA 以及 ICA 來進行。其中，LSI 又細分為 Global-LSI 與 Local-LSI。在屬性擷取的過程中，擷取的維度我們分別擷取 k =1、2、3、4、5、10、15、20、25 及 30。

下列表 4-15~4-17 與圖 4-5~4-7，為先執行 Global-LSI，再執行分類預測方法，表 4-18 與圖 4-8，為先執行 Global-LSI，再執行分類預測之結果比較。表 4-19~4-21 與圖 4-9~4-11 為先執行 Local-LSI，再執行分類預測方法，表 4-22 與圖 4-12 為 Local-LSI 為先執行 Local-LSI，再執行分類預測之結果比較。表 4-23~4-25 與圖 4-13~4-15 為先執行 PCA，再執行分類預測方法，表 4-26 與圖 4-16 為先執行 PCA，再執行分類預測之結果比較。表 4-27~4-29 與圖 4-17~4-19 為為先執行 ICA，再執行分類預測方法，表 4-30 與圖 4-20 為先執行 ICA，再執行分類預測之結果比較。

表 4-15　Global-LSI + BPN 之分類預測正確率

	Global-LSI + BPN 分類預測正確率	*k*
Fold 1	92.50%	5
Fold 2	97.50%	4
Fold 3	95.00%	5
Fold 4	95.00%	2
Fold 5	87.50%	1
平均	93.50%	3.4
標準差	3.79%	

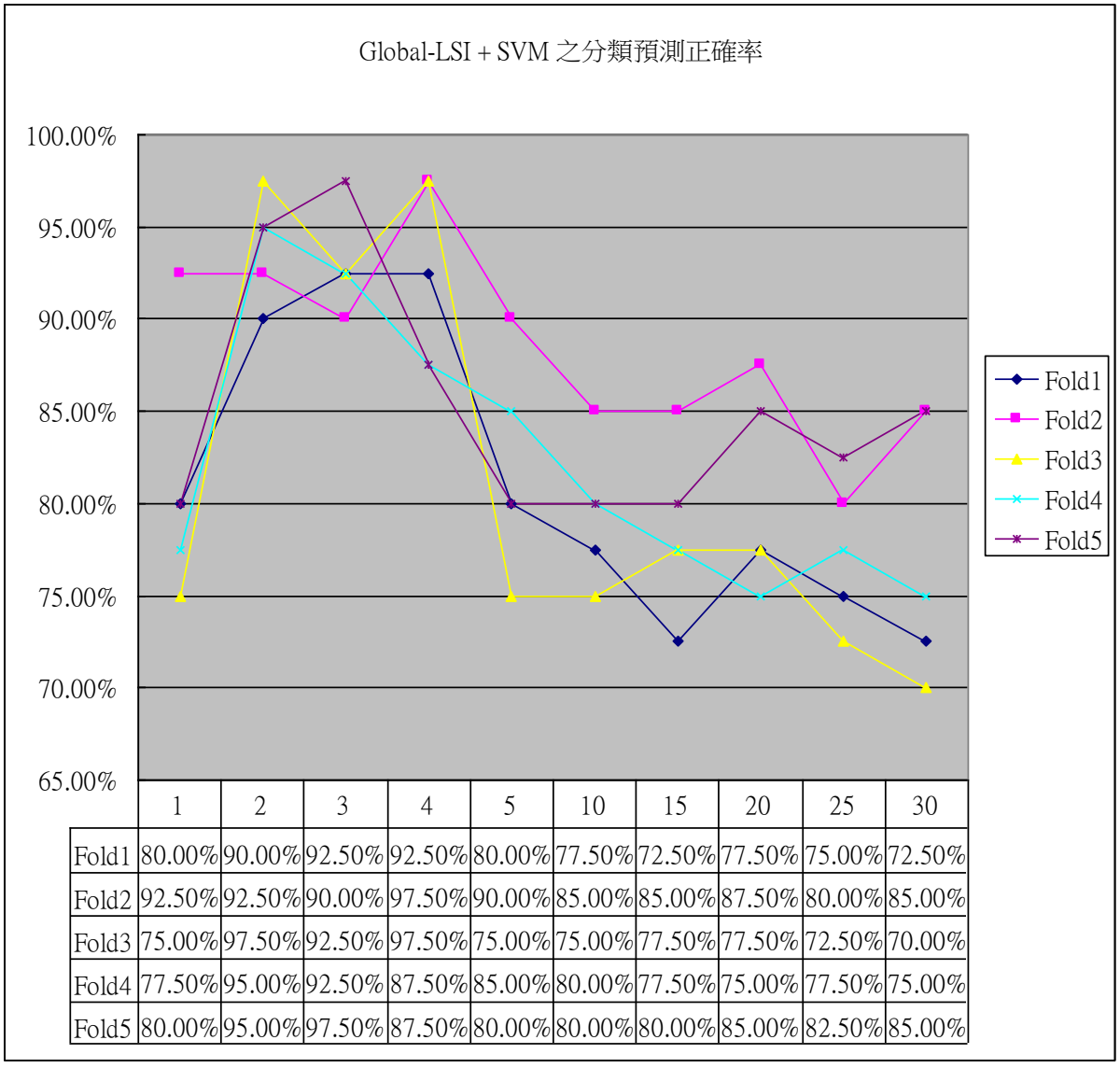


	1	2	3	4	5	10	15	20	25	30
Fold1	80.00%	90.00%	87.50%	90.00%	92.50%	90.00%	87.50%	92.50%	85.00%	85.00%
Fold2	95.00%	92.50%	92.50%	97.50%	92.50%	90.00%	85.00%	85.00%	87.50%	90.00%
Fold3	75.00%	85.00%	85.00%	92.50%	95.00%	95.00%	90.00%	87.50%	87.50%	87.50%
Fold4	87.50%	95.00%	95.00%	95.00%	95.00%	90.00%	92.50%	95.00%	95.00%	87.50%
Fold5	87.50%	85.00%	87.50%	82.50%	82.50%	87.50%	80.00%	77.50%	85.00%	82.50%

圖 4-5　Global-LSI + BPN 之分類預測正確率

表 4-16　Global-LSI + SVM 之分類預測正確率

	Global-LSI + SVM 分類預測正確率	k
Fold 1	92.50%	3
Fold 2	97.50%	4
Fold 3	97.50%	2
Fold 4	95.00%	2
Fold 5	97.50%	3
平均	96.00%	2.8
標準差	2.24%	

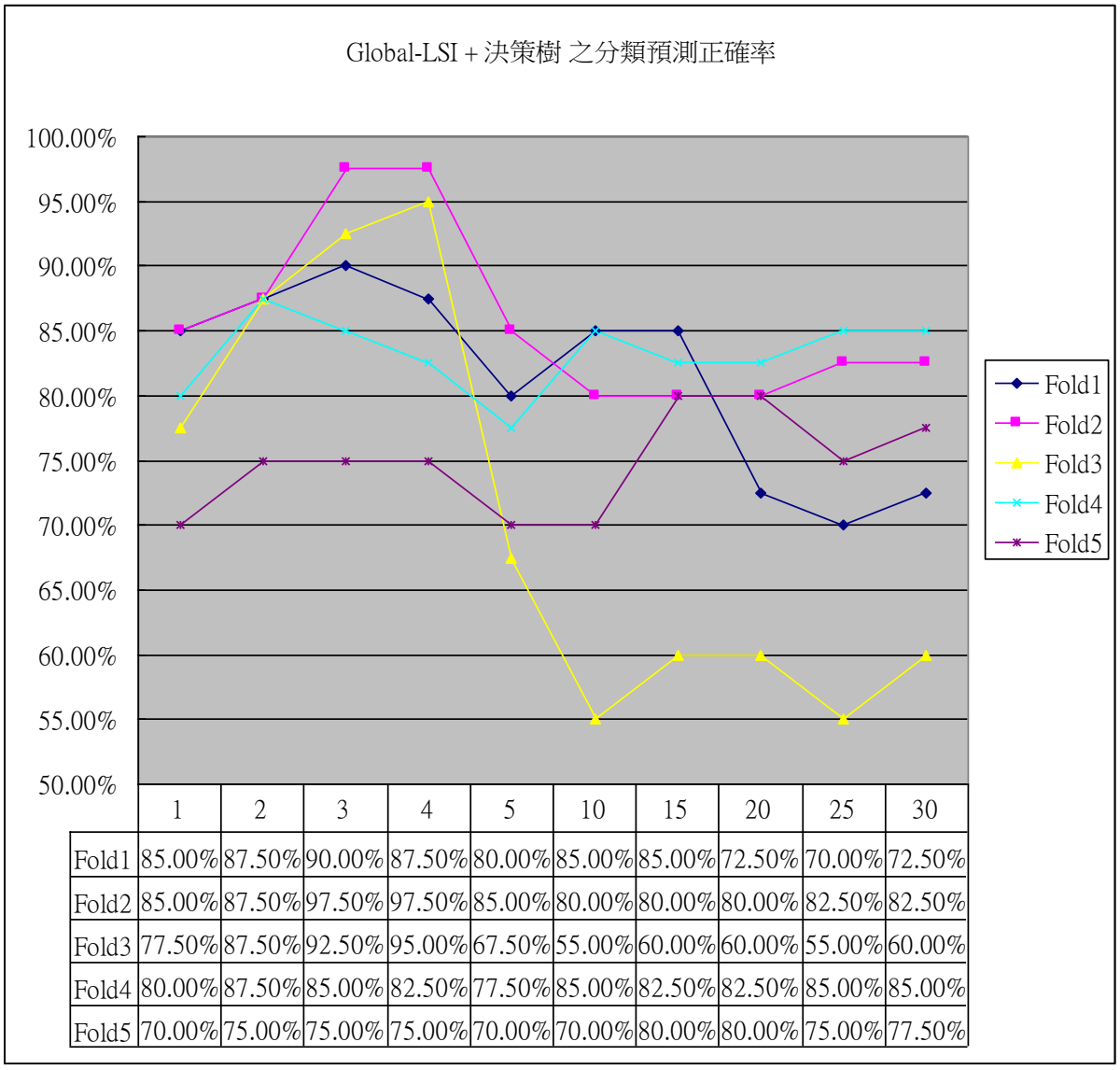


	1	2	3	4	5	10	15	20	25	30
Fold1	80.00%	90.00%	92.50%	92.50%	80.00%	77.50%	72.50%	77.50%	75.00%	72.50%
Fold2	92.50%	92.50%	90.00%	97.50%	90.00%	85.00%	85.00%	87.50%	80.00%	85.00%
Fold3	75.00%	97.50%	92.50%	97.50%	75.00%	75.00%	77.50%	77.50%	72.50%	70.00%
Fold4	77.50%	95.00%	92.50%	87.50%	85.00%	80.00%	77.50%	75.00%	77.50%	75.00%
Fold5	80.00%	95.00%	97.50%	87.50%	80.00%	80.00%	80.00%	85.00%	82.50%	85.00%

圖 4-6　Global-LSI + SVM 之分類預測正確率

表 4-17　Global-LSI + 決策樹 之分類預測正確率

	Global-LSI + 決策樹分類預測正確率	*k*
Fold 1	90.00%	3
Fold 2	97.50%	3
Fold 3	95.00%	4
Fold 4	87.50%	2
Fold 5	80.00%	15
平均	90.00%	5.4
標準差	6.85%	



	1	2	3	4	5	10	15	20	25	30
Fold1	85.00%	87.50%	90.00%	87.50%	80.00%	85.00%	85.00%	72.50%	70.00%	72.50%
Fold2	85.00%	87.50%	97.50%	97.50%	85.00%	80.00%	80.00%	80.00%	82.50%	82.50%
Fold3	77.50%	87.50%	92.50%	95.00%	67.50%	55.00%	60.00%	60.00%	55.00%	60.00%
Fold4	80.00%	87.50%	85.00%	82.50%	77.50%	85.00%	82.50%	82.50%	85.00%	85.00%
Fold5	70.00%	75.00%	75.00%	75.00%	70.00%	70.00%	80.00%	80.00%	75.00%	77.50%

圖 4-7　Global-LSI + 決策樹 之分類預測正確率

表 4-18　先執行 Global-LSI 後執行分類預測之結果比較

	平均預測正確率	標準差	平均縮減維度 K
Global-LSI + BPN	93.50%	3.79%	3.4
Global-LSI + SVM	96.00%	2.24%	2.8
Global-LSI +決策樹	90.00%	6.85%	5.4

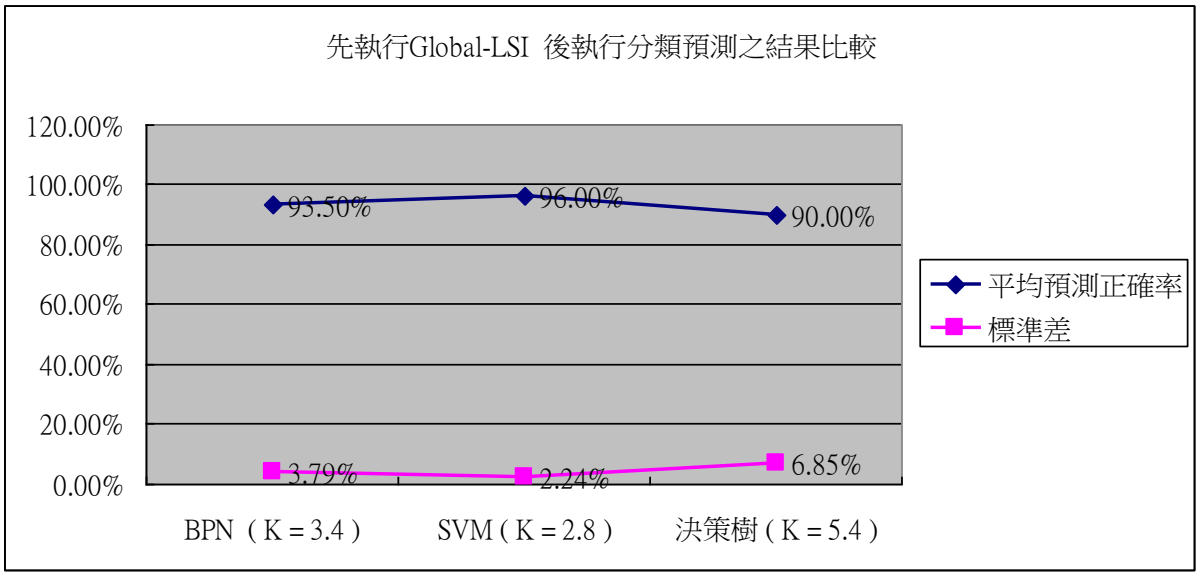
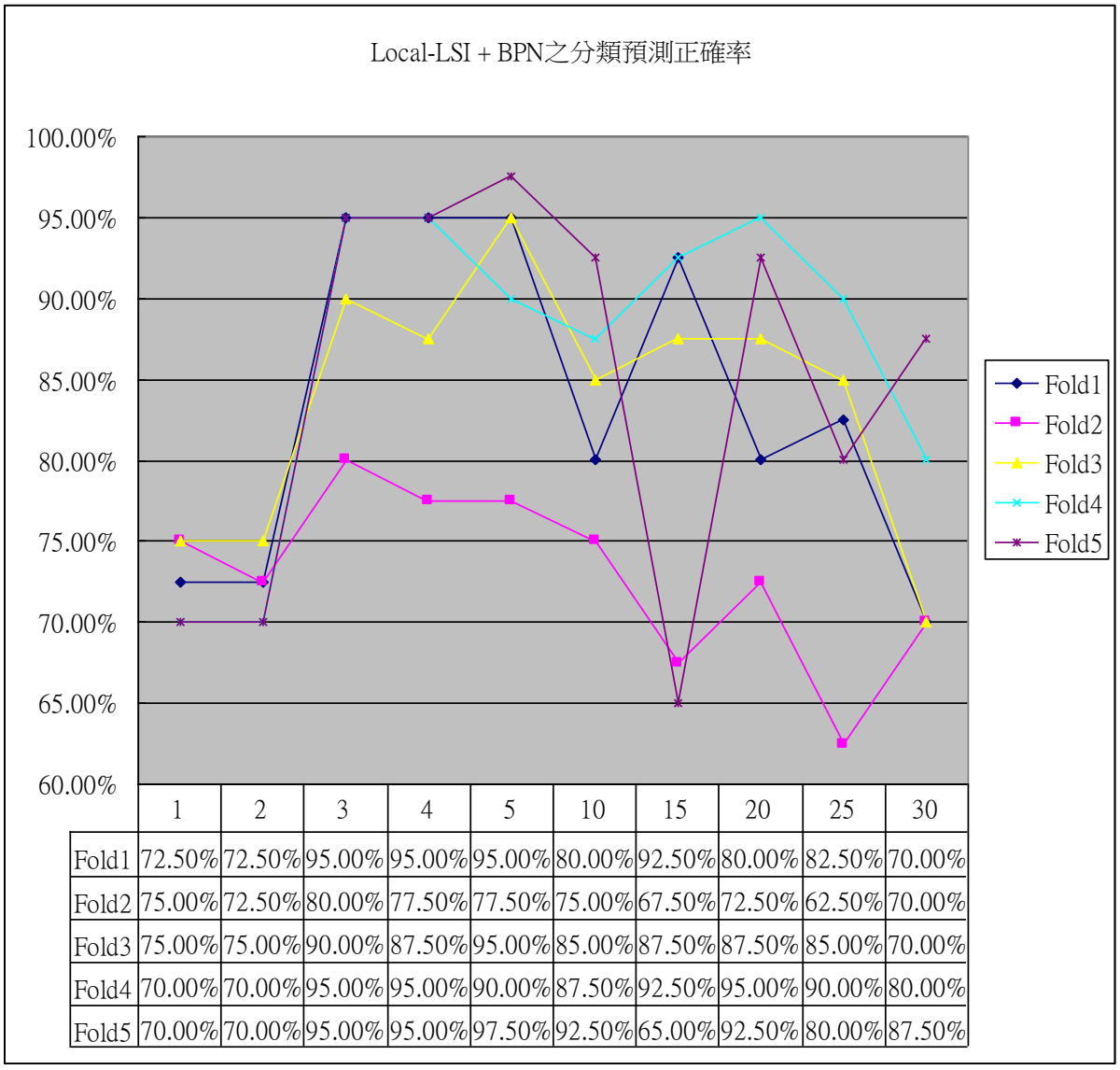


圖 4-8 先執行 Global-LSI　後執行分類預測之結果比較

表 4-19　Local-LSI + BPN 之分類預測正確率

	Local-LSI + BPN 分類預測正確率	k
Fold 1	95.00%	3
Fold 2	80.00%	3
Fold 3	95.00%	5
Fold 4	95.00%	3
Fold 5	97.50%	5
平均	92.50%	3.8
標準差	7.07%	

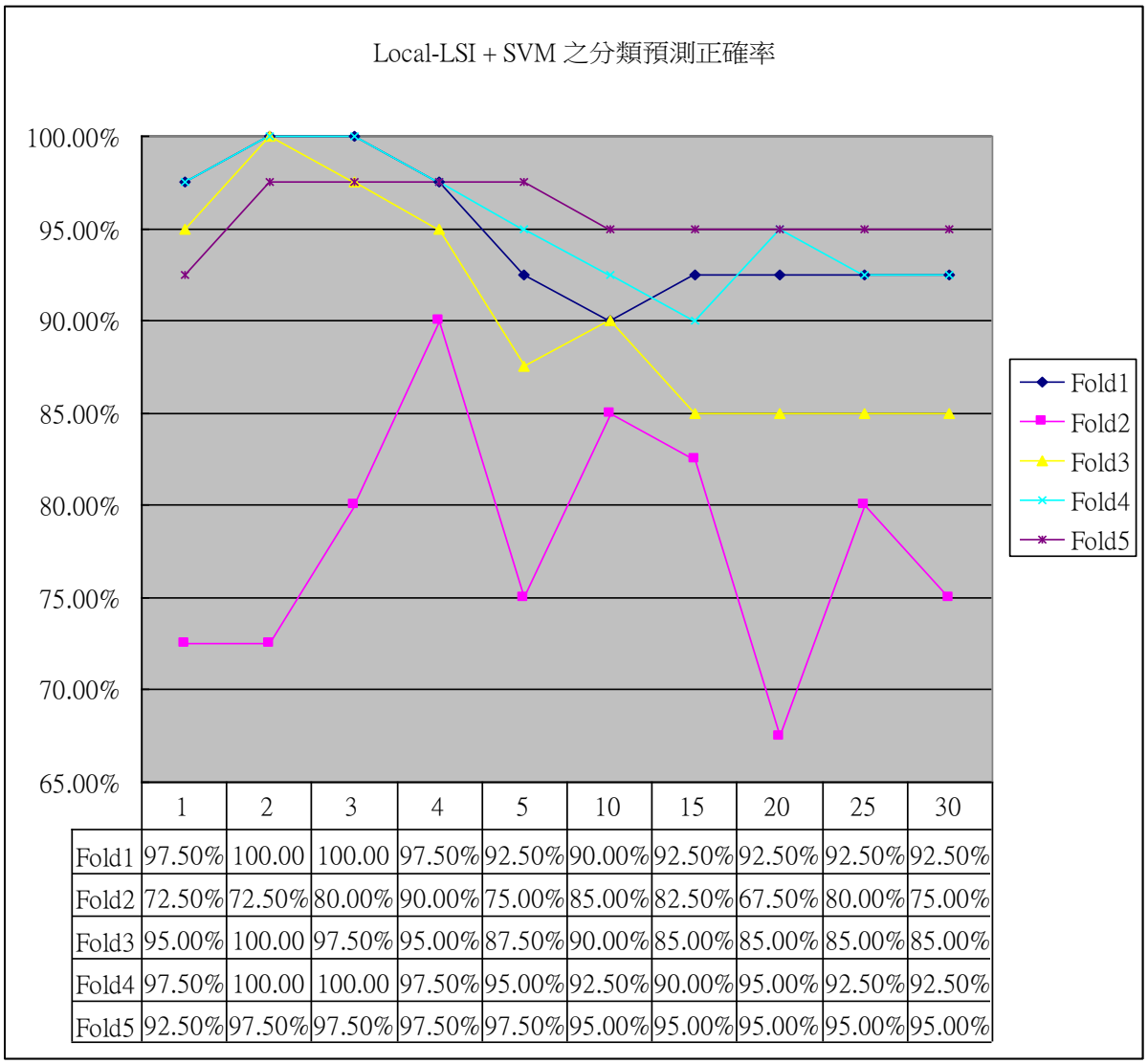


	1	2	3	4	5	10	15	20	25	30
Fold1	72.50%	72.50%	95.00%	95.00%	95.00%	80.00%	92.50%	80.00%	82.50%	70.00%
Fold2	75.00%	72.50%	80.00%	77.50%	77.50%	75.00%	67.50%	72.50%	62.50%	70.00%
Fold3	75.00%	75.00%	90.00%	87.50%	95.00%	85.00%	87.50%	87.50%	85.00%	70.00%
Fold4	70.00%	70.00%	95.00%	95.00%	90.00%	87.50%	92.50%	95.00%	90.00%	80.00%
Fold5	70.00%	70.00%	95.00%	95.00%	97.50%	92.50%	65.00%	92.50%	80.00%	87.50%

圖 4-9　Local-LSI + BPN 之分類預測正確率

表 4-20　Local-LSI + SVM 之分類預測正確率

	Local-LSI + SVM 分類預測正確率	k
Fold 1	100.00%	2
Fold 2	90.00%	4
Fold 3	100.00%	2
Fold 4	100.00%	2
Fold 5	97.50%	2
平均	97.50%	2.4
標準差	4.33%	

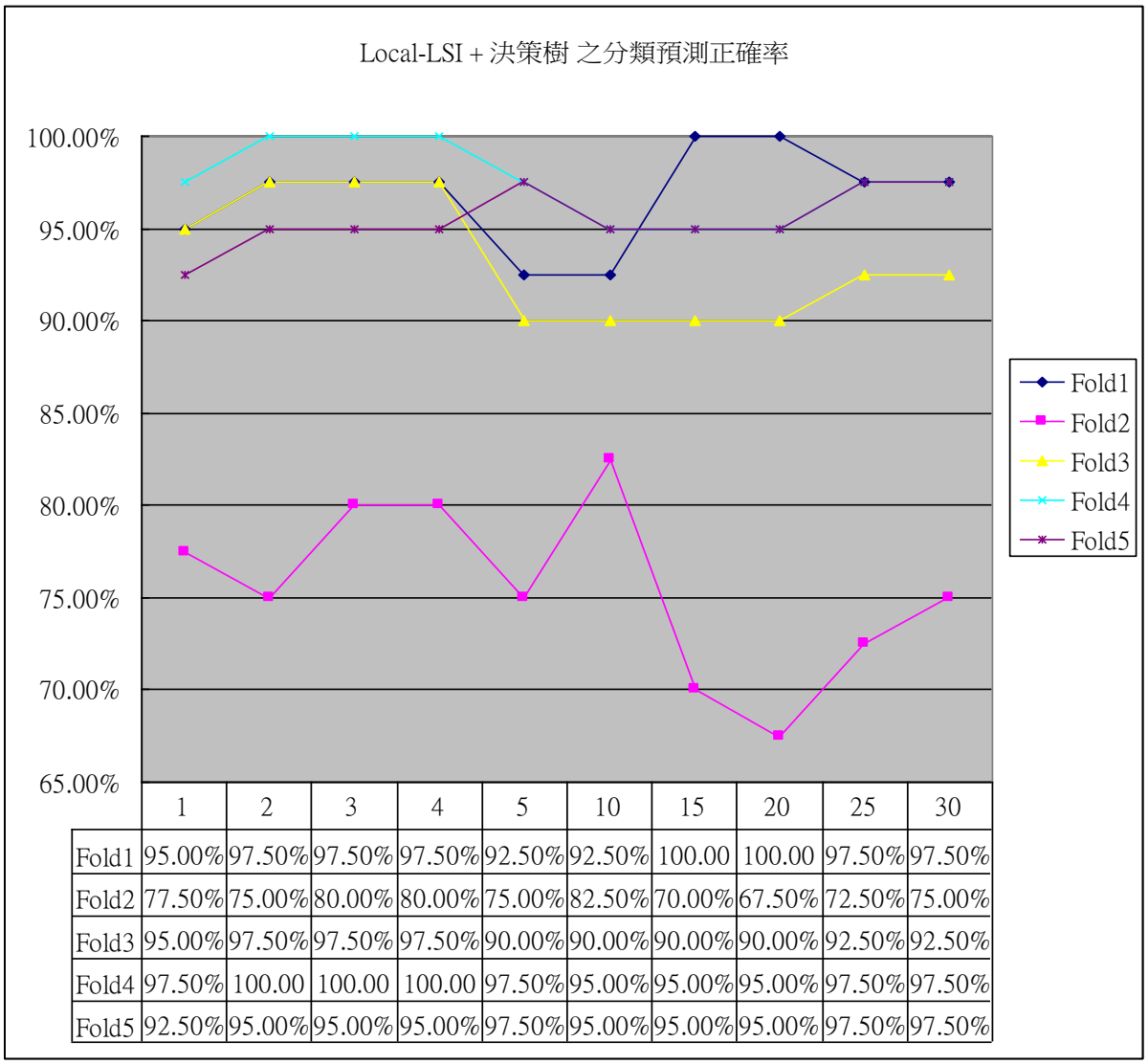


	1	2	3	4	5	10	15	20	25	30
Fold1	97.50%	100.00	100.00	97.50%	92.50%	90.00%	92.50%	92.50%	92.50%	92.50%
Fold2	72.50%	72.50%	80.00%	90.00%	75.00%	85.00%	82.50%	67.50%	80.00%	75.00%
Fold3	95.00%	100.00	97.50%	95.00%	87.50%	90.00%	85.00%	85.00%	85.00%	85.00%
Fold4	97.50%	100.00	100.00	97.50%	95.00%	92.50%	90.00%	95.00%	92.50%	92.50%
Fold5	92.50%	97.50%	97.50%	97.50%	97.50%	95.00%	95.00%	95.00%	95.00%	95.00%

圖 4-10　Local-LSI + SVM 之分類預測正確率

表 4-21　Local-LSI + 決策樹 之分類預測正確率

	Local-LSI + 決策樹分類預測正確率	*k*
Fold 1	100.00%	15
Fold 2	82.50%	10
Fold 3	97.50%	2
Fold 4	100.00%	2
Fold 5	97.50%	5
平均	95.50%	6.8
標準差	7.37%	



	1	2	3	4	5	10	15	20	25	30
Fold1	95.00%	97.50%	97.50%	97.50%	92.50%	92.50%	100.00	100.00	97.50%	97.50%
Fold2	77.50%	75.00%	80.00%	80.00%	75.00%	82.50%	70.00%	67.50%	72.50%	75.00%
Fold3	95.00%	97.50%	97.50%	97.50%	90.00%	90.00%	90.00%	90.00%	92.50%	92.50%
Fold4	97.50%	100.00	100.00	100.00	97.50%	95.00%	95.00%	95.00%	97.50%	97.50%
Fold5	92.50%	95.00%	95.00%	95.00%	97.50%	95.00%	95.00%	95.00%	97.50%	97.50%

圖 4-11　Local-LSI + 決策樹 之分類預測正確率

表 4-22　先執行 Local-LSI　後執行分類預測之結果比較

	平均預測正確率	標準差	平均縮減維度 K
Local-LSI + BPN	92.50%	7.07%	3.8
Local-LSI + SVM	97.50%	4.33%	2.4
Local-LSI +決策樹	95.50%	7.37%	6.8

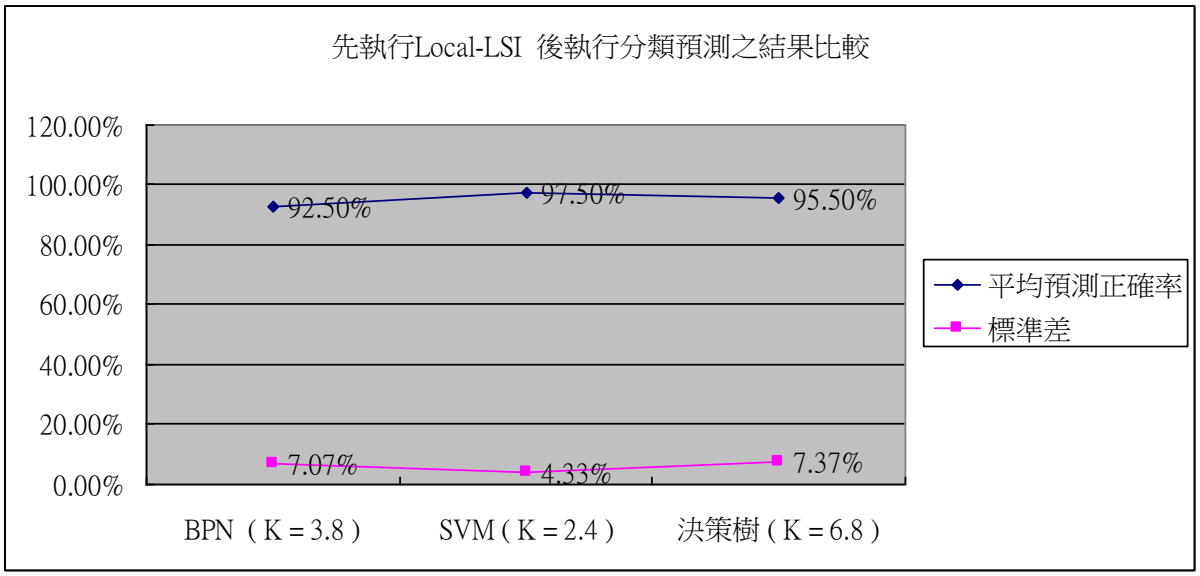
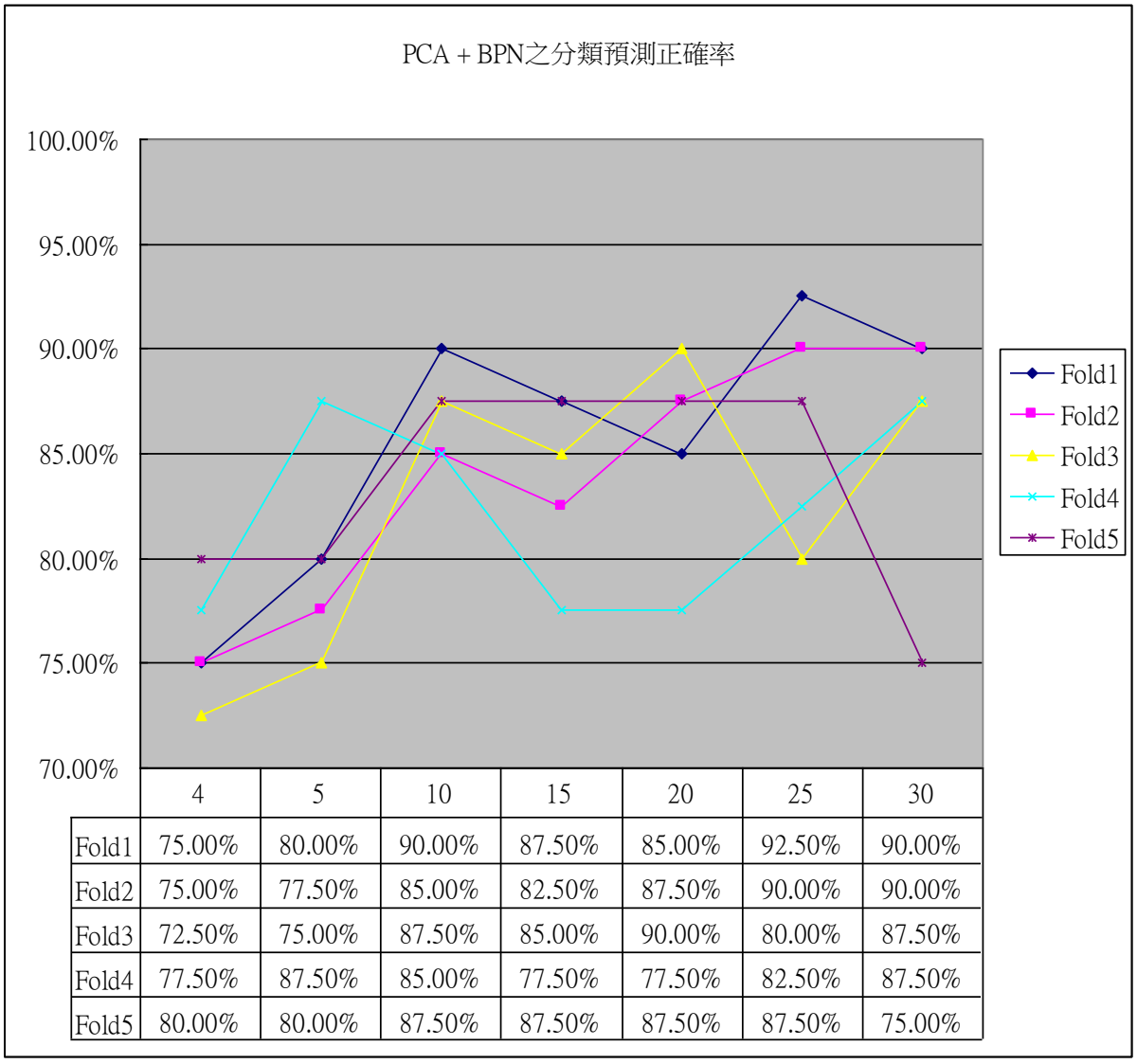


圖 4-12 先執行 Local-LSI　後執行分類預測之結果比較

表 4-23　PCA + BPN 之分類預測正確率

	PCA + BPN 分類預測正確率	k
Fold 1	92.50%	25
Fold 2	90.00%	25
Fold 3	90.00%	20
Fold 4	87.50%	5
Fold 5	87.50%	5
平均	89.50%	16
標準差	2.09%	

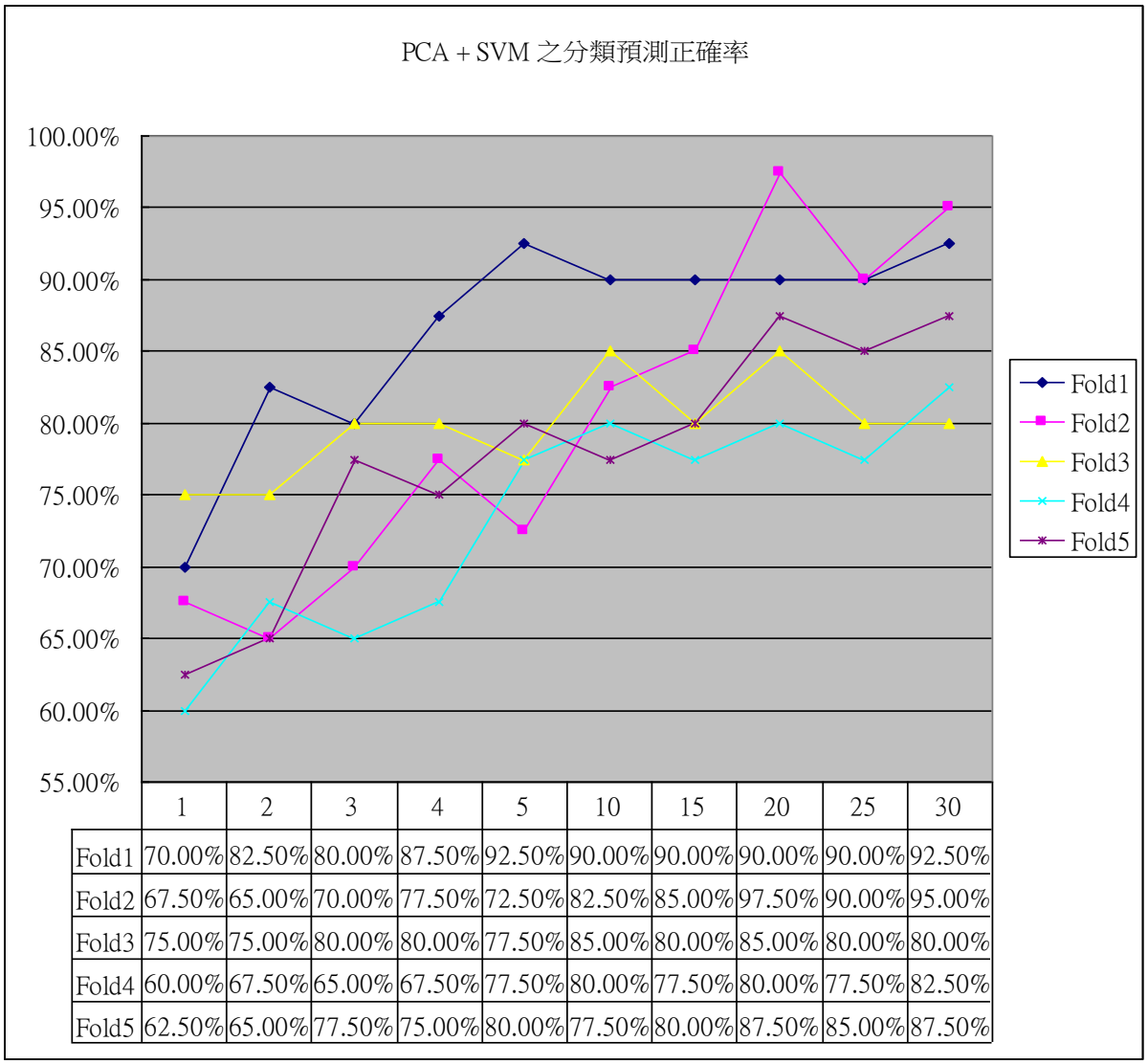


	4	5	10	15	20	25	30
Fold1	75.00%	80.00%	90.00%	87.50%	85.00%	92.50%	90.00%
Fold2	75.00%	77.50%	85.00%	82.50%	87.50%	90.00%	90.00%
Fold3	72.50%	75.00%	87.50%	85.00%	90.00%	80.00%	87.50%
Fold4	77.50%	87.50%	85.00%	77.50%	77.50%	82.50%	87.50%
Fold5	80.00%	80.00%	87.50%	87.50%	87.50%	87.50%	75.00%

圖 4-13　PCA + BPN 之分類預測正確率

表 4-24　PCA + SVM 之分類預測正確率

	PCA + SVM 分類預測正確率	k
Fold 1	92.50%	5
Fold 2	97.50%	20
Fold 3	85.00%	20
Fold 4	82.50%	30
Fold 5	87.50%	30
平均	89.00%	21
標準差	6.02%	

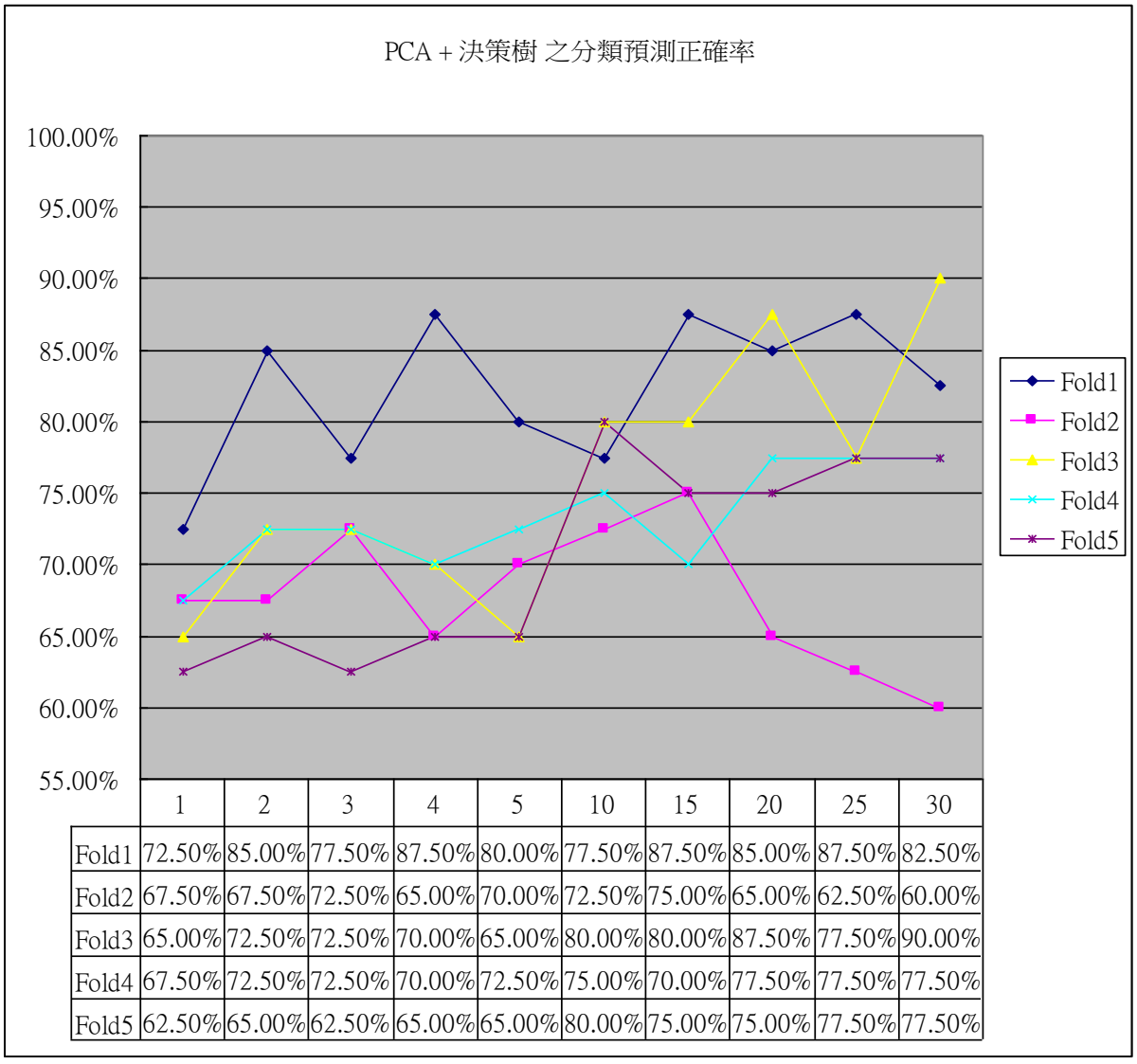


	1	2	3	4	5	10	15	20	25	30
Fold1	70.00%	82.50%	80.00%	87.50%	92.50%	90.00%	90.00%	90.00%	90.00%	92.50%
Fold2	67.50%	65.00%	70.00%	77.50%	72.50%	82.50%	85.00%	97.50%	90.00%	95.00%
Fold3	75.00%	75.00%	80.00%	80.00%	77.50%	85.00%	80.00%	85.00%	80.00%	80.00%
Fold4	60.00%	67.50%	65.00%	67.50%	77.50%	80.00%	77.50%	80.00%	77.50%	82.50%
Fold5	62.50%	65.00%	77.50%	75.00%	80.00%	77.50%	80.00%	87.50%	85.00%	87.50%

圖 4-14　PCA + SVM 之分類預測正確率

表 4-25　PCA + 決策樹 之分類預測正確率

	PCA + 決策樹分類預測正確率	*k*
Fold 1	87.50%	4
Fold 2	75.00%	15
Fold 3	90.00%	30
Fold 4	77.50%	20
Fold 5	80.00%	10
平均	82.00%	15.8
標準差	6.47%	



	1	2	3	4	5	10	15	20	25	30
Fold1	72.50%	85.00%	77.50%	87.50%	80.00%	77.50%	87.50%	85.00%	87.50%	82.50%
Fold2	67.50%	67.50%	72.50%	65.00%	70.00%	72.50%	75.00%	65.00%	62.50%	60.00%
Fold3	65.00%	72.50%	72.50%	70.00%	65.00%	80.00%	80.00%	87.50%	77.50%	90.00%
Fold4	67.50%	72.50%	72.50%	70.00%	72.50%	75.00%	70.00%	77.50%	77.50%	77.50%
Fold5	62.50%	65.00%	62.50%	65.00%	65.00%	80.00%	75.00%	75.00%	77.50%	77.50%

圖 4-15　PCA + 決策樹 之分類預測正確率

表 4-26　先執行 PCA　後執行分類預測之結果比較

	平均預測正確率	標準差	平均縮減維度 K
PCA + BPN	89.50%	2.09%	16
PCA + SVM	89.00%	6.02%	21
PCA +決策樹	82.00%	6.47%	15.8

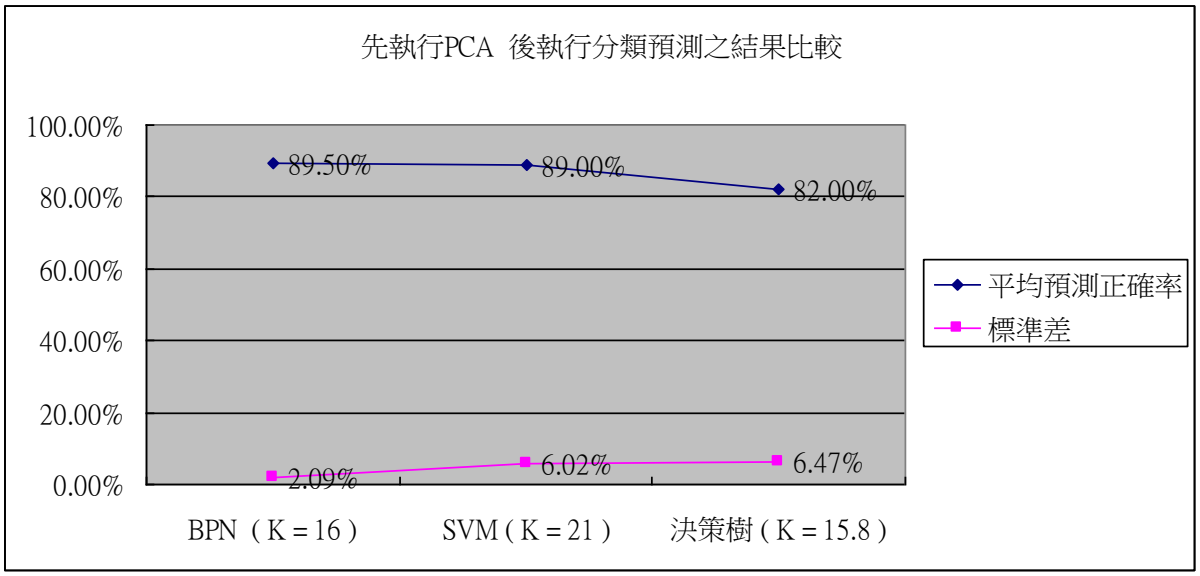
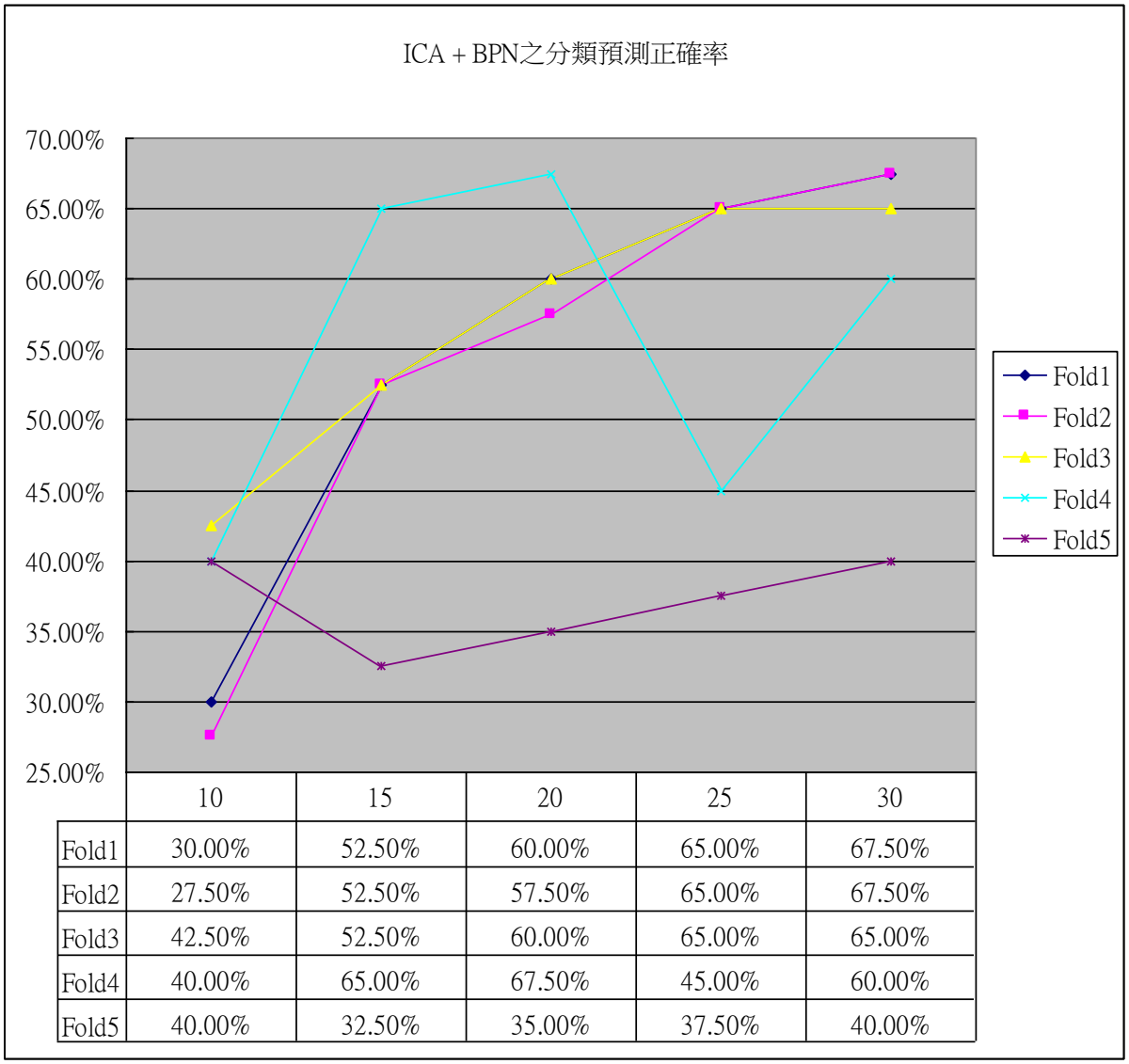


圖 4-16 先執行 PCA　後執行分類預測之結果比較

其中，圖 4-13　PCA + BPN 之分類預測正確率一圖中，缺乏 k =1、k =2 及 k =3 的資訊，是因為在 BPN 的網路學習中，k =1、2、3 時，無法收斂。

表 4-27　ICA + BPN 之分類預測正確率

	ICA + BPN 分類預測正確率	k
Fold 1	67.50%	30
Fold 2	67.50%	30
Fold 3	65.00%	25
Fold 4	67.50%	20
Fold 5	40.00%	10
平均	61.50%	23
標準差	12.07%	

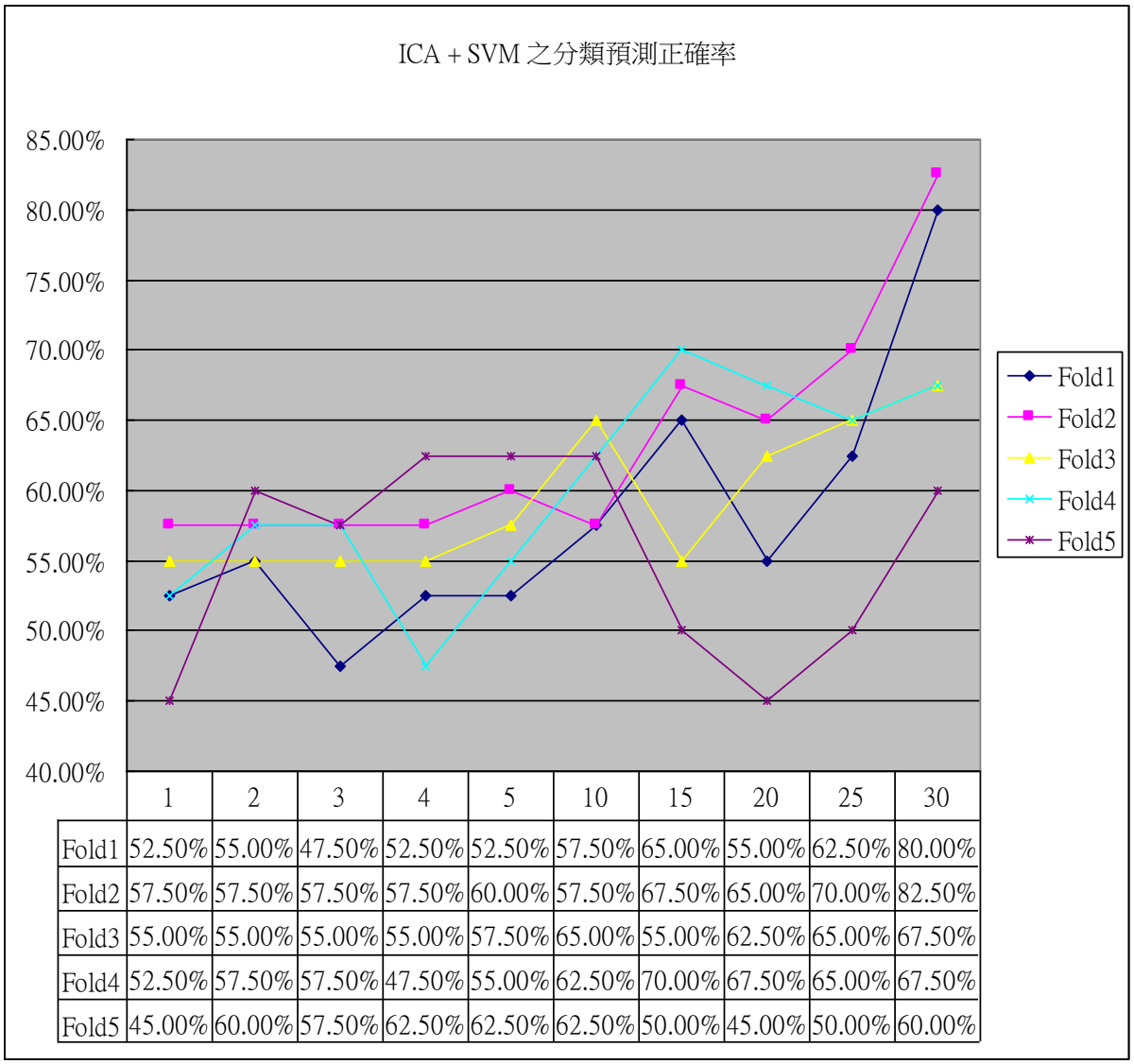


	10	15	20	25	30
Fold1	30.00%	52.50%	60.00%	65.00%	67.50%
Fold2	27.50%	52.50%	57.50%	65.00%	67.50%
Fold3	42.50%	52.50%	60.00%	65.00%	65.00%
Fold4	40.00%	65.00%	67.50%	45.00%	60.00%
Fold5	40.00%	32.50%	35.00%	37.50%	40.00%

圖 4-17　ICA + BPN 之分類預測正確率

表 4-28　ICA + SVM 之分類預測正確率

	ICA + SVM 分類預測正確率	k
Fold 1	80.00%	30
Fold 2	82.50%	30
Fold 3	67.50%	30
Fold 4	70.00%	15
Fold 5	62.50%	4
平均	72.50%	21.8
標準差	8.48%	

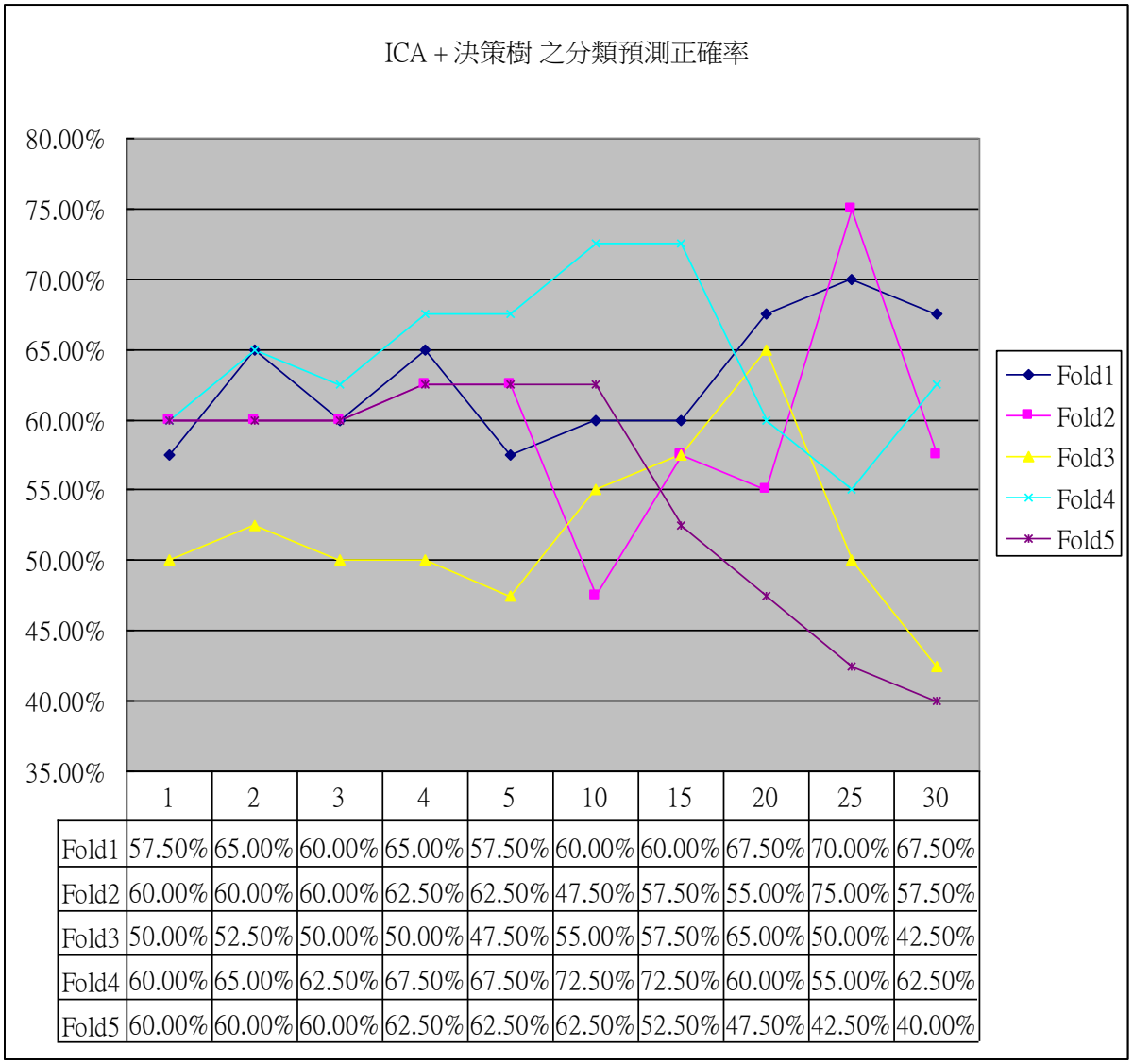


	1	2	3	4	5	10	15	20	25	30
Fold1	52.50%	55.00%	47.50%	52.50%	52.50%	57.50%	65.00%	55.00%	62.50%	80.00%
Fold2	57.50%	57.50%	57.50%	57.50%	60.00%	57.50%	67.50%	65.00%	70.00%	82.50%
Fold3	55.00%	55.00%	55.00%	55.00%	57.50%	65.00%	55.00%	62.50%	65.00%	67.50%
Fold4	52.50%	57.50%	57.50%	47.50%	55.00%	62.50%	70.00%	67.50%	65.00%	67.50%
Fold5	45.00%	60.00%	57.50%	62.50%	62.50%	62.50%	50.00%	45.00%	50.00%	60.00%

圖 4-18　ICA + SVM 之分類預測正確率

表 4-29　ICA + 決策樹 之分類預測正確率

	ICA + 決策樹分類預測正確率	*k*
Fold 1	70.00%	25
Fold 2	75.00%	25
Fold 3	65.00%	20
Fold 4	72.50%	10
Fold 5	62.50%	4
平均	69.00%	16.8
標準差	5.18%	



	1	2	3	4	5	10	15	20	25	30
Fold1	57.50%	65.00%	60.00%	65.00%	57.50%	60.00%	60.00%	67.50%	70.00%	67.50%
Fold2	60.00%	60.00%	60.00%	62.50%	62.50%	47.50%	57.50%	55.00%	75.00%	57.50%
Fold3	50.00%	52.50%	50.00%	50.00%	47.50%	55.00%	57.50%	65.00%	50.00%	42.50%
Fold4	60.00%	65.00%	62.50%	67.50%	67.50%	72.50%	72.50%	60.00%	55.00%	62.50%
Fold5	60.00%	60.00%	60.00%	62.50%	62.50%	62.50%	52.50%	47.50%	42.50%	40.00%

圖 4-19　ICA + 決策樹 之分類預測正確率

表 4-30　先執行 ICA　後執行分類預測之結果比較

	平均預測正確率	標準差	平均縮減維度 K
ICA + BPN	61.50%	12.07%	23
ICA + SVM	72.50%	8.48%	21.8
ICA +決策樹	69.00%	5.18%	16.8

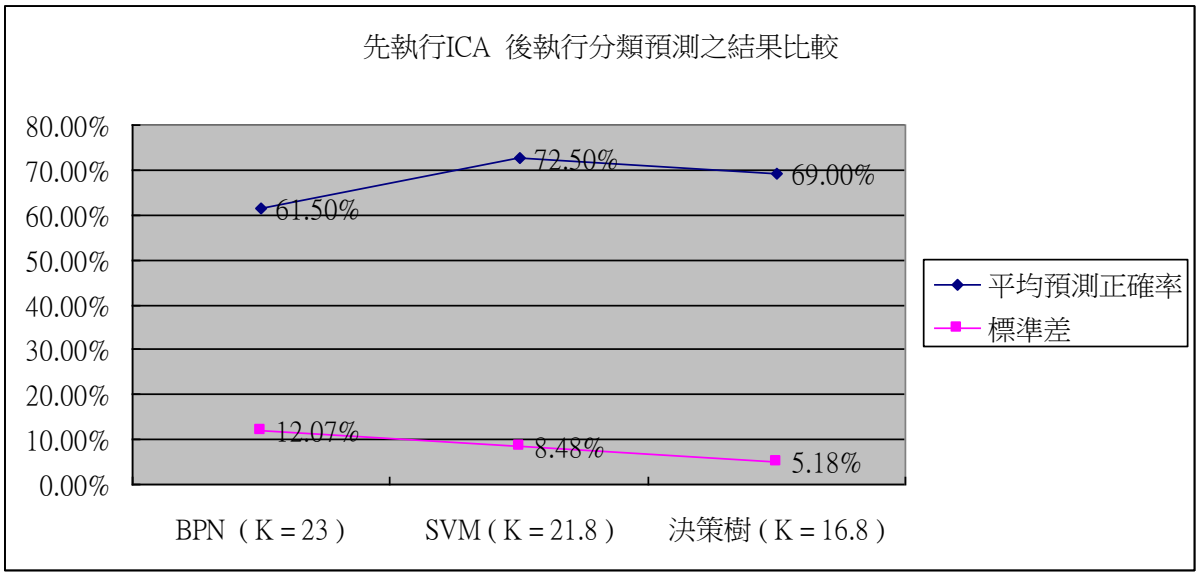


圖 4-20 先執行 ICA　後執行分類預測之結果比較

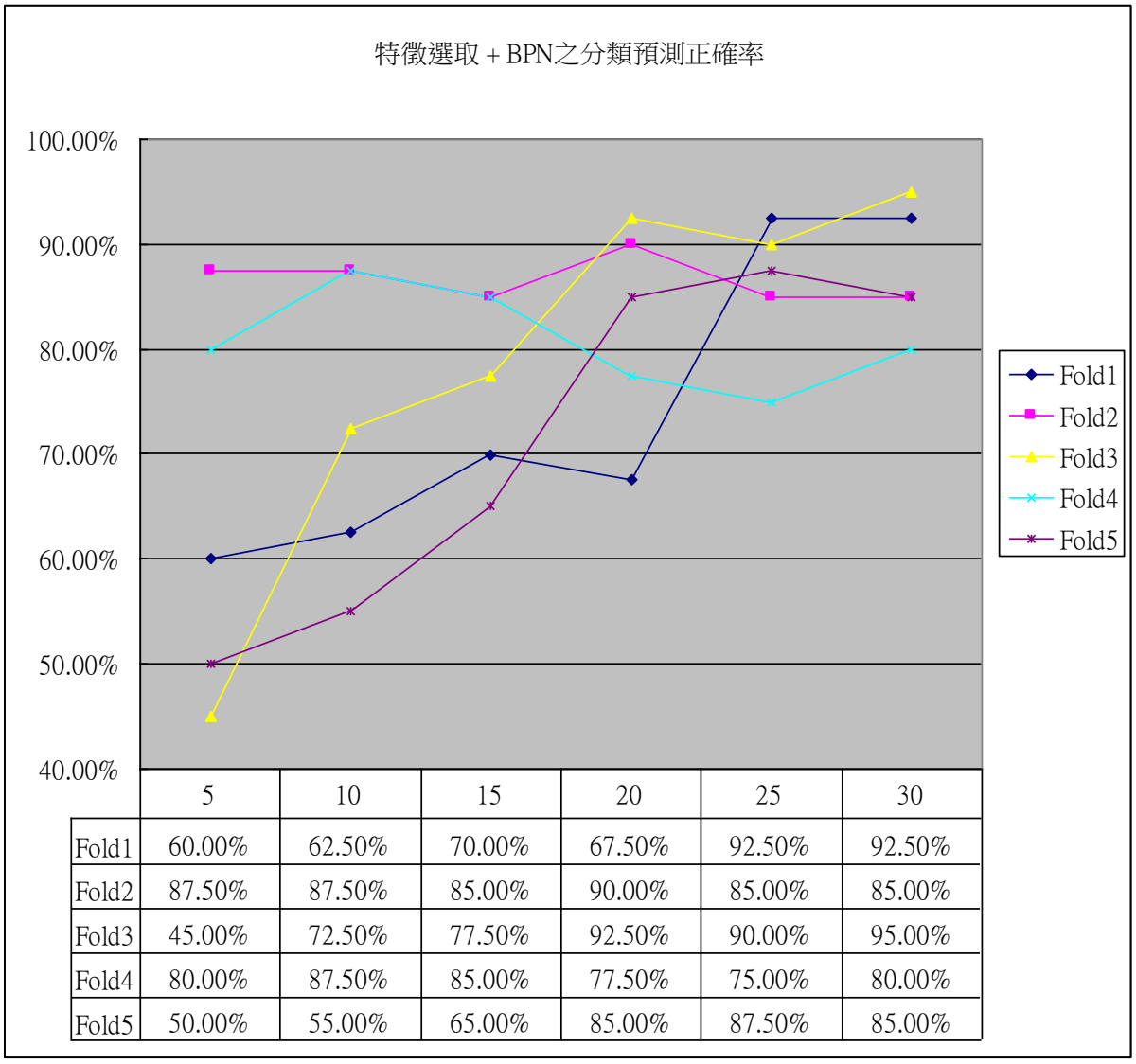
其中，圖 4-17　ICA + BPN 之分類預測正確率一圖中，缺乏 k =1、k =2、k =3、k =4 及 k =5 的資訊，是因為在 BPN 的網路學習中，k =1、2、3、4、5 時，無法收斂。

4.4.2 特徵選取

在特徵選取的實驗中，我們所選取的維度為 k =1、5、10、15、20、25 及 30。下列表 4-31~4-33 與圖 4-21~4-23，為先執行特徵選取，再執行分類預測方法，表 4-34 與圖 4-24，為先執行特徵選取，再執行分類預測之結果比較。

表 4-31　特徵選取 + BPN 之分類預測正確率

	特徵選取 + BPN 分類預測正確率	k
Fold 1	92.50%	25
Fold 2	90.00%	20
Fold 3	95.00%	30
Fold 4	87.50%	10
Fold 5	87.50%	25
平均	90.50%	22
標準差	3.26%	

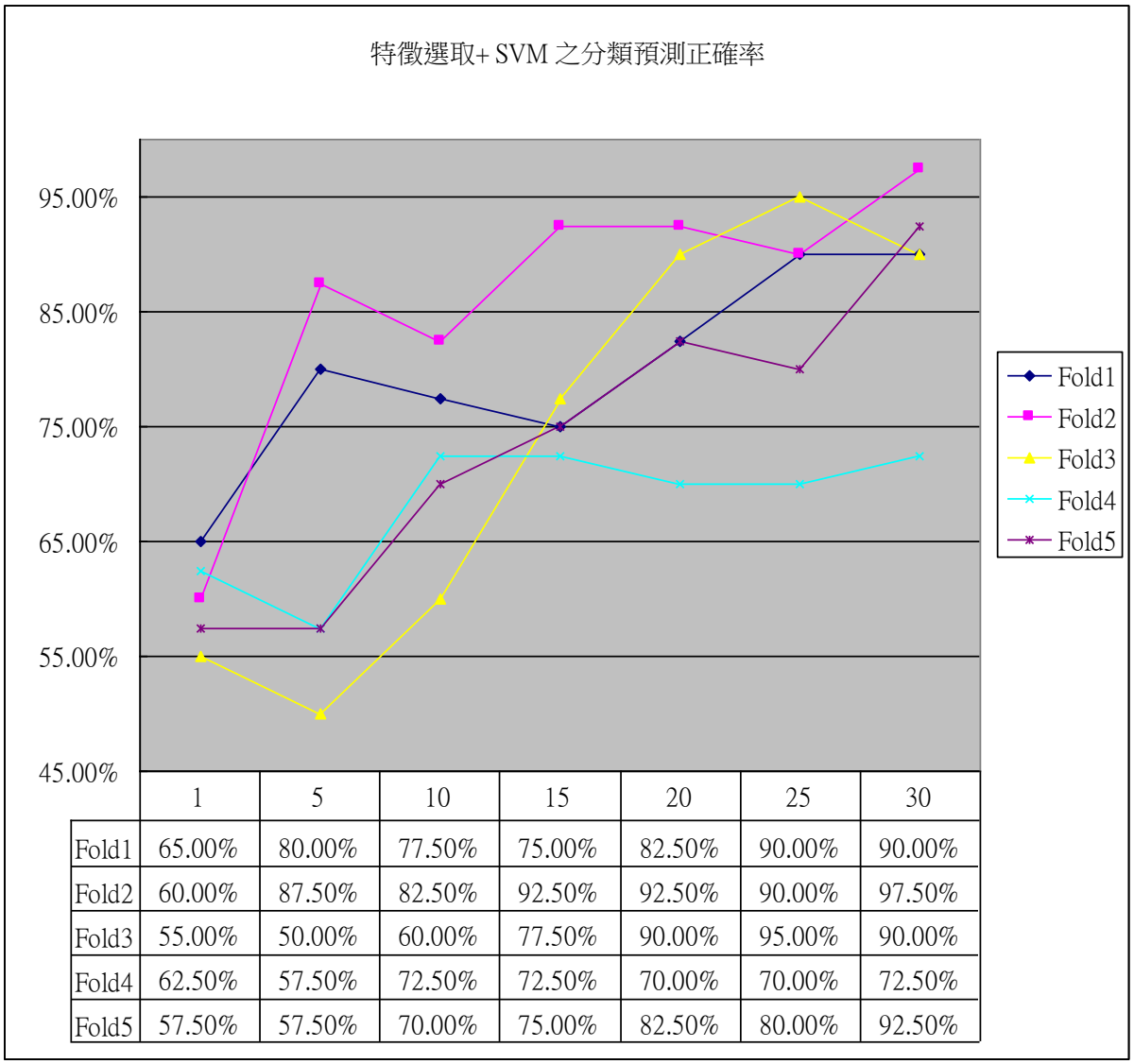


	5	10	15	20	25	30
Fold1	60.00%	62.50%	70.00%	67.50%	92.50%	92.50%
Fold2	87.50%	87.50%	85.00%	90.00%	85.00%	85.00%
Fold3	45.00%	72.50%	77.50%	92.50%	90.00%	95.00%
Fold4	80.00%	87.50%	85.00%	77.50%	75.00%	80.00%
Fold5	50.00%	55.00%	65.00%	85.00%	87.50%	85.00%

圖 4-21 特徵選取 + BPN 之分類預測正確率

表 4-32　特徵選取+ SVM 之分類預測正確率

	特徵選取+ SVM 分類預測正確率	k
Fold 1	90.00%	25
Fold 2	97.50%	30
Fold 3	95.00%	25
Fold 4	72.50%	10
Fold 5	92.50%	30
平均	89.50%	24
標準差	9.91%	

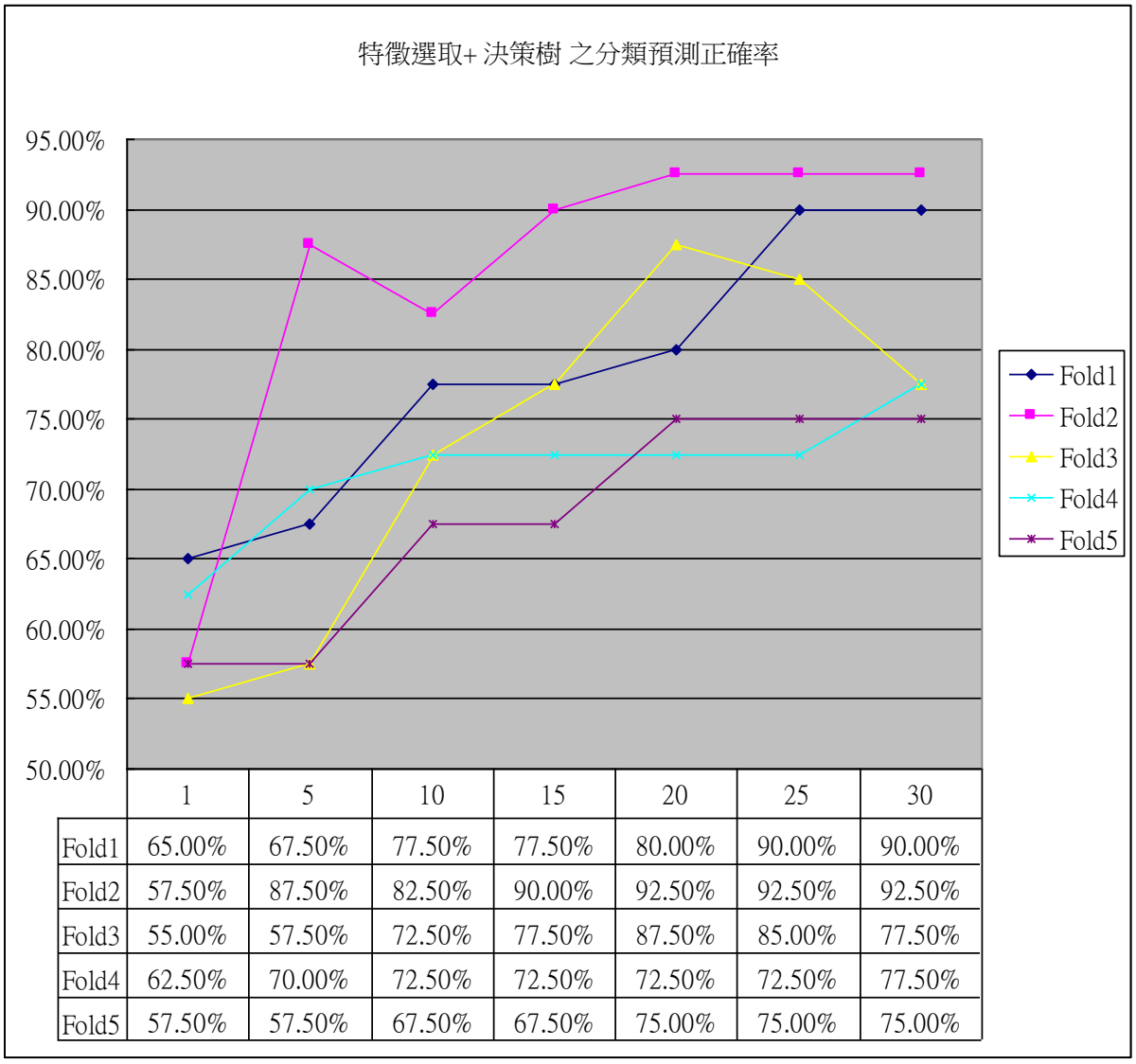


	1	5	10	15	20	25	30
Fold1	65.00%	80.00%	77.50%	75.00%	82.50%	90.00%	90.00%
Fold2	60.00%	87.50%	82.50%	92.50%	92.50%	90.00%	97.50%
Fold3	55.00%	50.00%	60.00%	77.50%	90.00%	95.00%	90.00%
Fold4	62.50%	57.50%	72.50%	72.50%	70.00%	70.00%	72.50%
Fold5	57.50%	57.50%	70.00%	75.00%	82.50%	80.00%	92.50%

圖 4-22 特徵選取+ SVM 之分類預測正確率

表 4-33　特徵選取+ 決策樹 之分類預測正確率

	特徵選取+ 決策樹分類預測正確率	*k*
Fold 1	90.00%	25
Fold 2	92.50%	20
Fold 3	87.50%	20
Fold 4	77.50%	30
Fold 5	75.00%	20
平均	84.50%	23
標準差	7.79%	



	1	5	10	15	20	25	30
Fold1	65.00%	67.50%	77.50%	77.50%	80.00%	90.00%	90.00%
Fold2	57.50%	87.50%	82.50%	90.00%	92.50%	92.50%	92.50%
Fold3	55.00%	57.50%	72.50%	77.50%	87.50%	85.00%	77.50%
Fold4	62.50%	70.00%	72.50%	72.50%	72.50%	72.50%	77.50%
Fold5	57.50%	57.50%	67.50%	67.50%	75.00%	75.00%	75.00%

圖 4-23 特徵選取+ 決策樹 之分類預測正確率

表 4-34　先執行特徵選取　後執行分類預測之結果比較

	平均預測正確率	標準差	平均縮減維度 K
特徵選取 ＋BPN	90.50%	3.26%	22
特徵選取 ＋SVM	89.50%	9.91%	24
特徵選取+決策樹	84.50%	7.79%	23

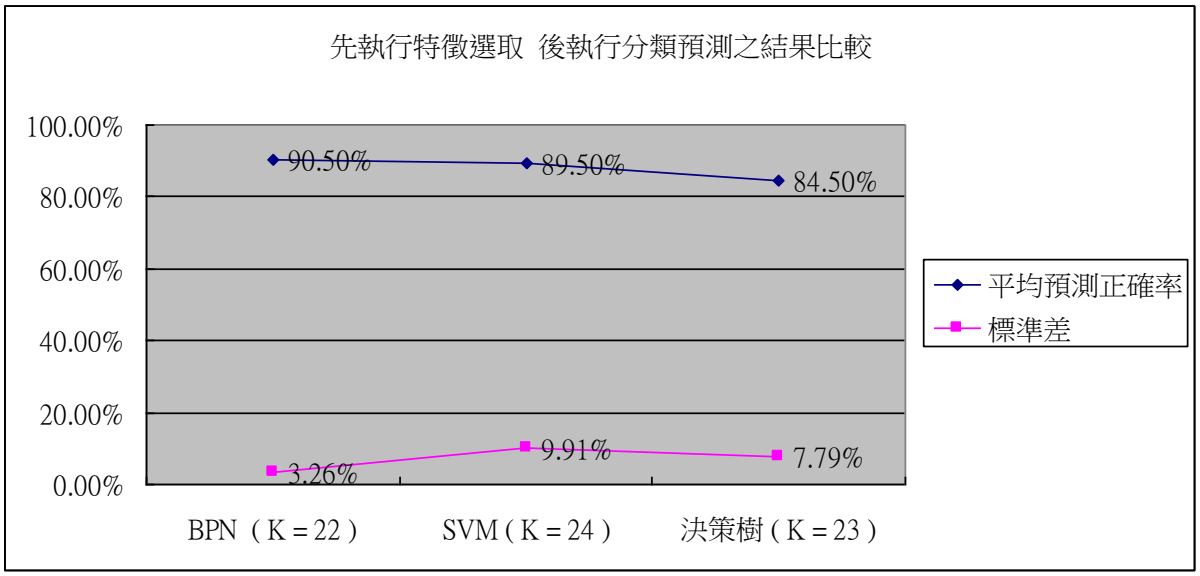


圖 4-24 先執行特徵選取　後執行分類預測之結果比較

其中，圖 4-17　ICA + BPN 之分類預測正確率一圖中，缺乏 k =1、k =2、k =3 及 k =4 的資訊，是因為在 BPN 的網路學習中，k =1、2、3、4 時，無法收斂。

經過了 4.4 節維度縮減的動作後，我們得到了上述圖、表以及數據，我們將在 4.5 節比較各方法的優劣。

4.5 方法比較

在一連串的實驗後，我們將進行分類預測方法的比較，以及進行維度縮減工作前與進行維度縮減後之方法比較。表 4-31 以及圖 4-21 為 BPN 方法比較圖表，表 4-32 與圖 4-22 為 SVM 方法比較圖表，表 4-33 與圖 4-23 為決策樹方法比較圖表。

表 4-35　BPN 方法比較表

維度縮減法	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	平均	標準差
Orginal data	82.50%	85.00%	92.50%	92.50%	85.00%	87.50%	4.68%
Global LSI	92.50%	97.50%	95.00%	95.00%	87.50%	93.50%	3.79%
	K=5	K=4	K=5	K=2	K=1	K=3.4	1.82
Local LSI	95.00%	80.00%	95.00%	95.00%	97.50%	92.50%	7.07%
	K=3	K=3	K=5	K=3	K=5	K=3.8	1.10
PCA	92.50%	90.00%	90.00%	87.50%	87.50%	89.50%	2.09%
	K=25	K=25	K=20	K=5	K=5	K=16	10.25
ICA	67.50%	67.50%	65.00%	67.50%	40.00%	61.50%	12.07%
	K=30	K=30	K=25	K=20	K=10	K=23	8.37
Feature Selection	92.50%	90.00%	95.00%	87.50%	87.50%	90.50%	3.26%
	K=25	K=20	K=30	K=10	K=25	K=22	7.58

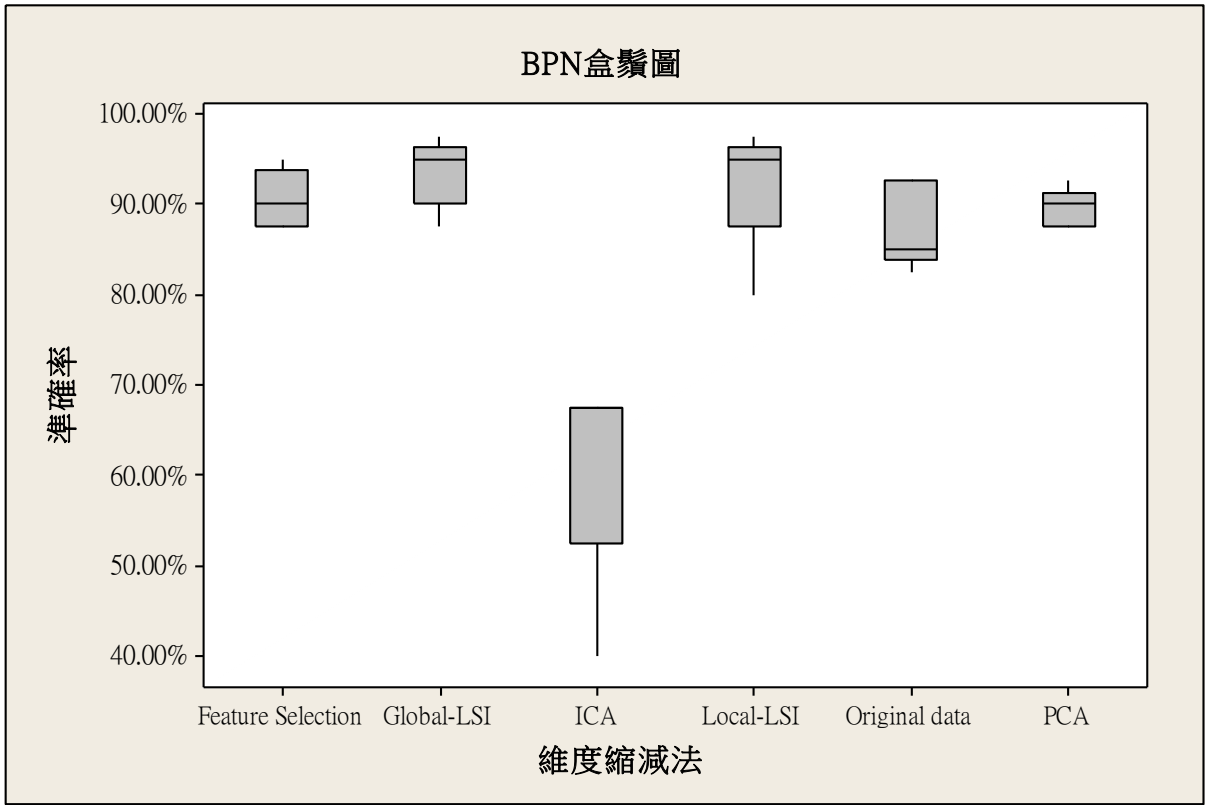


圖 4-25　BPN 方法比較盒鬚圖

從表 4-31 以及圖 4-21 來看，我們可以清楚的觀察 BPN 分類預測方法，應用於未經過維度縮減以及應用於維度縮減後的結果。其中我們發現，未經過維度縮減之平均預測正確率為 87.50%，而 Global-LSI 為 93.50%、Local-LSI 為 92.50%、PCA 為 89.50%、ICA 為 61.50%、Feature Selection 為 90.50%。

因此，就 BPN 分類預測方法而言，我們可以推論出四種有效改善稀疏性資料分類器為：

1. 先執行 Global-LSI，之後執行 BPN，能有效改善稀疏性資料分類預測正確率，平均維度縮減可以達到 k 為 3.4。
2. 先執行 Local-LSI，之後執行 BPN，能有效改善稀疏性資料分類預測正確率，平均維度縮減可以達到 k 為 3.8。
3. 先執行 PCA，之後執行 BPN，能有效改善稀疏性資料分類預測正確率，平均維度縮減可以達到 k 為 16。
4. 先執行 Feature Selection，之後執行 BPN，能有效改善稀疏性資料分類預測正確率，平均維度縮減可以達到 k 為 22。

表 4-36　SVM 方法比較表

維度縮減法	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	平均	標準差
Original data	92.50%	95.00%	87.50%	82.50%	92.50%	90.00%	5.00%
Global LSI	92.50%	97.50%	97.50%	95.00%	97.50%	96.00%	2.24%
	K=3	K=4	K=2	K=2	K=3	K=2.8	0.84
Local LSI	100.00%	90.00%	100.00%	100.00%	97.50%	97.50%	4.33%
	K=2	K=4	K=2	K=2	K=2	K=2.4	0.89
PCA	92.50%	97.50%	85.00%	82.50%	87.50%	89.00%	6.02%
	K=5	K=20	K=20	K=30	K=30	K=21	10.25
ICA	80.00%	82.50%	67.50%	70.00%	62.50%	72.50%	8.48%
	K=30	K=30	K=30	K=15	K=4	K=21.8	11.88
Feature Selection	90.00%	97.50%	95.00%	72.50%	92.50%	89.50%	9.91%
	K=25	K=30	K=25	K=10	K=30	K=24	8.22

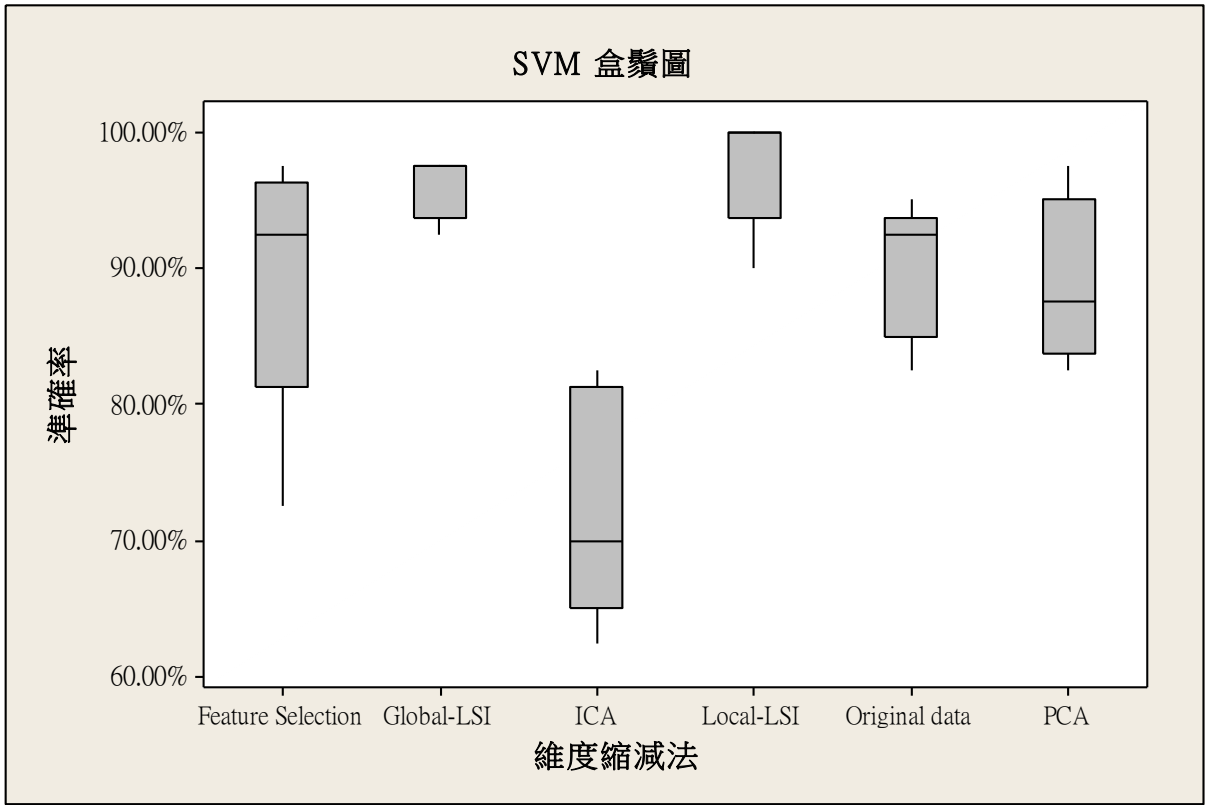


圖 4-26　SVM 方法比較盒鬚圖

從表 4-32 以及圖 4-22 來看，我們可以清楚的觀察 SVM 分類預測方法，應用於未經過維度縮減以及應用於維度縮減後的結果。其中我們發現，未經過維度縮減之平均預測正確率為 90.00%，而 Global-LSI 為 96.00%、Local-LSI 為 97.50%、PCA 為 89.00%、ICA 為 72.50%、Feature Selection 為 89.50%。

因此，就 SVM 分類預測方法而言，我們可以推論出兩種有效改善稀疏性資料分類器為：

1. 先執行 Global-LSI，之後執行 SVM，能有效改善稀疏性資料分類預測正確率，平均維度縮減可以達到 k 為 2.8。
2. 先執行 Local-LSI，之後執行 SVM，能有效改善稀疏性資料分類預測正確率，平均維度縮減可以達到 k 為 2.4。

表 4-37　決策樹方法比較表

維度縮減法	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	平均	標準差
Original data	90.00%	92.50%	77.50%	77.50%	75.00%	82.50%	8.10%
Global LSI	90.00%	97.50%	95.00%	87.50%	80.00%	90.00%	6.85%
	K=3	K=3	K=4	K=2	K=15	K=5.4	5.41
Local LSI	100.00%	82.50%	97.50%	100.00%	97.50%	95.50%	7.37%
	K=15	K=10	K=2	K=2	K=5	K=6.8	5.63
PCA	87.50%	75.00%	90.00%	77.50%	80.00%	82.00%	6.47%
	K=4	K=15	K=30	K=20	K=10	K=15.8	9.91
ICA	70.00%	75.00%	65.00%	72.50%	62.50%	69.00%	5.18%
	K=25	K=25	K=20	K=10	K=4	K=16.8	9.42
Feature Selection	90.00%	92.50%	87.50%	77.50%	75.00%	84.50%	7.79%
	K=25	K=20	K=20	K=30	K=20	K=23	4.47

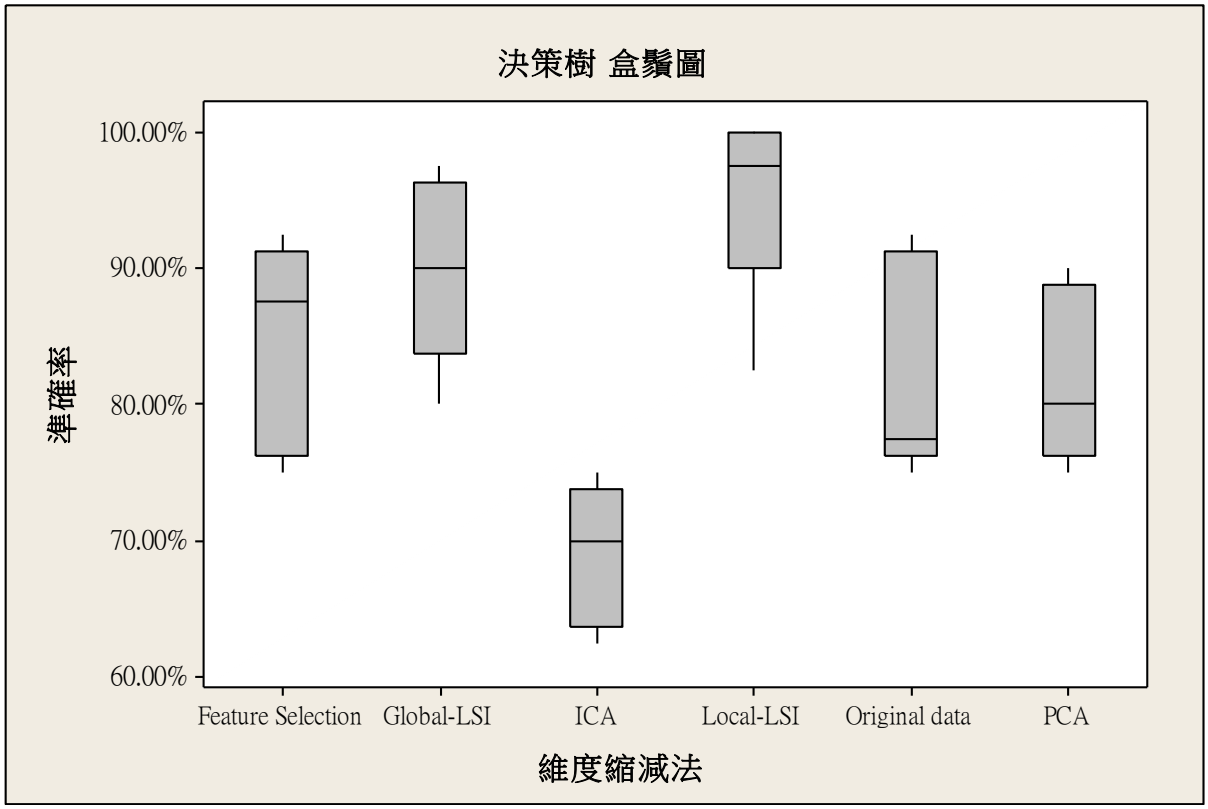


圖 4-27　決策樹方法比較盒鬚圖

從表 4-33 以及圖 4-23 來看，我們可以清楚的觀察決策樹分類預測方法，應用於未經過維度縮減以及應用於維度縮減後的結果，其中我們發現，未經過維度縮減之平均預測正確率為 82.50%，而 Global-LSI 為 90.00%、Local-LSI 為 95.50%、PCA 為 82.00%、ICA 為 69.00%、Feature Selection 為 84.50%。

因此，就決策樹分類預測方法而言，我們可以推論出三種有效改善稀疏性資料分類器為：

1. 先執行 Global-LSI，之後執行決策樹，能有效改善稀疏性資料分類預測正確率，平均維度縮減可以達到 k 為 5.4。
2. 先執行 Local-LSI，之後執行決策樹，能有效改善稀疏性資料分類預測正確率，平均維度縮減可以達到 k 為 6.8。
3. 先執行 Feature Selection，之後執行決策樹，能有效改善稀疏性資料分類預測正確率，平均維度縮減可以達到 k 為 23。

最後，我們總結三種分類預測方法 BPN、SVM 及決策樹。同時能夠在三種方法的實驗中，有效改善稀疏性分類預測正確率的為 Global-LSI 以及 Local-LSI。此兩種方法不僅能夠提升預測的準確率，並且能夠有效的達到維度縮減，略優於其他三種維度縮減的方法。其中又以 Local-LSI 預測準確率表現優於 Global-LSI，是因為 Local-LSI 有考慮分類目標群的資訊。

而在維度縮減的表現上，我們從表 4-38 可以發現 Global-LSI 與 Local-LSI 的維度縮減效果，都能將矩陣的平均特徵個數縮減至個位數，而 PCA 與 ICA 之矩陣的平均特徵個數最佳分別縮減為 15.8 與 16.8 個，至於特徵選取的矩陣平均選取個數都在 20 個以上。

總結各樣維度縮減方法之平均預測正確率，以及平均維度縮減特徵個數兩個數據結果，我們可以知道，Global-LSI 及 Local-LSI 在本研究中，是最適合拿來作為發展改善稀疏性資料分類器的維度縮減方法。

表 4-38　矩陣維度縮減比較表

維度縮減法	平均	標準差	維度縮減法	平均	標準差	維度縮減法	平均	標準差
Global-LSI with BPN	3.4	1.82	PCA with BPN	16	10.25	特徵選取 with BPN	22	7.58
Global-LSI with SVM	2.8	0.84	PCA with SVM	21	10.25	特徵選取 with SVM	24	8.22
Global-LSI with 決策樹	5.4	5.41	PCA with 決策樹	15.8	9.91	特徵選取 with 決策樹	23	4.47
Local-LSI with BPN	3.8	1.10	ICA with BPN	23	8.37			
Local-LSI with SVM	2.4	0.89	ICA with SVM	21.8	11.88			
Local-LSI with 決策樹	6.8	5.63	ICA with 決策樹	16.8	9.42			

第五章 結論與建議

部落格是一個新的媒介，特別是對於網際網路用戶，更是一個普遍且受到歡迎的資訊平台。不管是企業利用部落格來和消費者溝通，或者是用戶自行設立的個人部落格，在部落格充斥著無限的知識以及資訊。因此如何應用資料探勘的技術，將部落格裡的文章，轉為有用的知識是非常重要的。特別是對企業而言，若是能在部落格挖掘出商業資訊，更是能提供企業作為行銷決策支援。為此，本文提出一個部落格探勘模組，並以網路電話產品為對象，實際探勘部落格裡的知識及資訊。

5.1 研究結論

網際網路科技的進步日新月異，過去類似個人網頁的行銷媒體已逐漸的被部落格取代，部落格儼然成為網路行銷媒體的最佳選擇。因此，本研究提出之部落格探勘模組，利用 SOM 將討論網路電話之部落客適當的做出顧客區隔，並以 BPN、SVM、決策樹等分類方法來預測未知之新進顧客所屬顧客群，並輔以維度縮減的技術來發展改善稀疏性資料分類器，經過了實際的進行研究後，整理出下列的研究結論與貢獻：

1. 應用 SOM 將討論網路電話之部落客，適當的做出顧客區隔方面：我們透過 SOM 的群聚方法，將發表網路電話文章於部落格之部落客，區隔為三類用戶。第一類為著重網路電話多功能面之玩家級網路電話部落

客，第二類為重視通話品質之企業用戶網路電話部落客，以及第三類重視通話費用之經濟實惠網路電話部落客。因此，企業如果要針對網路電話產品進行行銷決策，可以針對每一類不同需求的用戶，推出適合的促銷決策。

事先利用 SOM 來針對發表文章之部落客進行顧客區隔，可以提供企業在行銷決策上之支援，或者直接針對特定客戶，提供客製化行銷。

2. 以 BPN、SVM 及決策樹將未知新進顧客進行分類預測方面：針對新發表有關網路電話文章之部落客，我們可以利用 BPN、SVM、及決策樹等分類預測方法進行，將未知的新進顧客劃分在適合的顧客群之中，以利企業進行行銷決策之應用。為此，我們進行分類預測實驗，得到 BPN 之平均預測正確率達到 87.50%，SVM 之平均預測正確率達到 90.00%，決策樹之平均預測正確率達到 82.50%。

因此，就預測正確率而言，BPN 和 SVM 擁有較高的正確率，但是並無法提供分類的規則。而決策樹雖然預測正確率的表現不如 BPN 及 SVM，卻擁有讓人一目了然的分類規則。

3. 輔以維度縮減方法，發展改善稀疏性資料分類器方面：我們使用了 PCA、ICA、LSI 等三種屬性擷取的工具，以及特徵選取的方法。先將欲進行分類預測的資料維度進行縮減，於資料維度縮減之後再進行 BPN、SVM 及決策樹分類預測。我們研究的結果顯示，同時能夠在三

種方法的實驗中，有效改善稀疏性分類預測正確率的為 Global-LSI 以及 Local-LSI。此兩種方法不僅能夠提升預測的準確率，並且能夠有效的達到維度縮減，略優於 PCA、ICA 及特徵選取等維度縮減方法。因此，在發展改善稀疏性資料分類器時，我們建議導入 LSI 方法，效果會較為顯著。

4. 特徵選取雖然在空間維度縮減的表現上，比屬性擷取不理想，但是特徵選取的新集合特徵變數中，仍保有原始特徵集合的意義，而屬性擷取後的新集合特徵變數卻無法保有原始特徵集合的意義。因此，在進行維度縮減時，需看研究目的來選擇維度縮減的方法。
5. ICA 在本研究的維度縮減表現並不理想，原因可歸咎於本研究之資料型態為文章-詞彙向量矩陣，文章中的關鍵字彼此具有高度關聯性，故以 ICA 來進行維度縮減時，效果不彰。

原本看似雜亂的部落格文章，只是部落客很單純的描述使用的感想，透過我們提出的部落格探勘模組，成為能提供企業作為行銷決策之有用的知識以及資訊，不僅可以達到顧客區隔，亦能有效的將新進顧客分類至適當的顧客群。我們的部落格探勘模組，對於新興的行銷媒體部落格而言，確實能夠挖掘出有用的知識與資訊。

5.2 未來建議

本研究之部落格探勘模組，是以網路電話產品為例，對發表文章之部落客做出顧客區隔及未知新進顧客分類預測，並且利用維度縮減來改善文件檢索所面臨的稀疏性資料問題。針對本研究的後續研究，有下列幾點建議：

1. 本研究是以兩百筆網路電話部落格文章為分析資料，建議後續者可以利用別的產品以及增加文章筆數。例如：五百筆討論數位相機部落格之文章，或一千筆討論汽車之部落格文章。
2. 本研究是以 SOM 來進行顧客區隔，建議後續之研究可以採用其他不同的分群方法來進行顧客區隔。
3. 本研究以 BPN、SVM 及決策樹進行未知新進顧客之分類預測，建議後續者可以採用其他分類預測方法來進行。
4. 本研究利用 Local-LSI 將 SVM 之平均正確預測率，從 90.00%提升至 97.50%，大幅的改善稀疏性資料分類預測的問題，建議其他後續者嘗試不同的維度縮減方法進行，以求能有更好的效果。
5. 由於中文斷詞頗難，建議後續研究者在定義關鍵字上，可採用不同的中文斷詞系統或方法。

參考文獻

中文部分

[1] 傅大煜 (2005)，高度酒消費行為及行銷策略之研究—以金門高粱酒為例，碩士論文，銘傳大學管理科學研究所，台北。

[2] 粘志鵬 (2006)，基於支援向量機之中文自動作文評分系統，碩士論文，國立交通大學資訊科學與工程研究所，新竹。

[3] 李銘浚 (2007)，應用獨立成分分析、對數頻譜預估、及頻率成分調整技術做語音增強之研究，碩士論文，國立清華大學電機工程學系碩士班，新竹。

[4] 李韋承 (2005)，建構階層式知識地圖及其知識搜尋法之研究，碩士論文，國立成功大學資訊管理研究所，台南。

[5] 林家民 (2005)，基於潛藏語意分析之多語言文件自動分群技術，碩士論文，國立中山大學資訊管理學系研究所，高雄。

[6] 林巧苑 (2001)，獨立成分分析法應用於磁振腦血流灌注研究之評估，碩士論文，國立陽明大學放射醫學科學研究所，台北。

[7] 林昭妘 (2006)，Blog 商業模式之研究，碩士論文，國立政治大學國際貿易研究所，台北。

[8] 林長富 (2006)，自組織映射圖網路於碎形影像壓縮之研究，碩士論文，義守大學資訊工程學系碩士班，高雄。

[9] 林士玄 (2006)，台灣網路電話消費行為研究，碩士論文，國立交通大學傳播研究所，新竹。

[10] 林盈源 (2003)，決策樹在資料庫行銷決策之應用，碩士論文，國立成功大學工業管理科學系專班，台南。

[11] 劉麗蘭 (2006)，以決策樹分析台灣上市櫃紡織業公司的財務危機，碩士論文，逢甲大學經營管理碩士在職專班，台中。

[12] 戈立秀 (2007)，部落格之資訊蒐集與分享行為之研究，碩士論文，國立台灣大學圖書資訊學研究所，台北。

[13] 郭芷婷 (2005)，Blog、BBS、個人網頁自助式成名 3 種方法，e 天下雜誌，51 期，第 130 頁。

[14] 黃敏菁 (2005)，支援向量機在財務時間序列預測之應用，碩士論文，輔仁大學金融研究所，台北。

[15] 黃鈁嫕 (2005)，運用類神經網路預測新進顧客產品喜好之個人化商品推薦技術，碩士論文，朝陽科技大學資訊管理系，台中。

[16] 黃承龍、陳穆臻、王界人 (2004)，「支援向量機於信用評等之應用」，計量管理期刊，第一卷，第二期，第 155-172 頁

[17] 黃應欽 (2006)，創新科技產品採用之研究—以網路電話為例，碩士論文，國立成功大學企業管理學系，台南。

[18] 何鴻聖 (2004)，自我組織神經網路在選股策略的應用，碩士論文，國

立東華大學國際經濟研究所，花蓮。

[19] 洪淑芬 (2006)，潛在語意索引在生醫文件分類之應用，碩士論文，樹德科技大學資訊管理研究所，高雄。

[20] 徐豐智 (2005)，Support Vector Machines 分類技術應用於文件相關性量測之探討，碩士論文，國立高雄應用科技大學電機工程系，高雄。

[21] 許邦輝 (2006)，以主成分分析法為基礎之文件自動分類模式，碩士論文，國立清華大學工業工程與工程管理系，新竹。

[22] 蕭宇翔 (2005)，應用 MTS 於非平衡資料分析之穩健性研究—以行動電話檢測流程為例，碩士論文，國立交通大學工業工程與管理學系，新竹。

[23] 張斐章、張麗秋、黃浩倫 (2003)，類神經網路：理論與實務，台灣東華書局股份有限公司。

[24] 張錦村 (2005)，台灣中小尺寸液晶顯示器產業的經營策略分析，碩士論文，國立交通大學管理學院高階主管碩士學程，新竹。

[25] 張結雄 (2002)，使用主成分分析及貝氏網路方法於離子植入製程之錯誤偵測與診斷，碩士論文，國立交通大學電機與控制工程學系，新竹。

[26] 鄭志強 (2006)，以決策樹演算法建構台灣企業財務危機預警模式，碩士論文，銘傳大學資訊管理學系碩士班，台北。

[27] 陳品均 (2006)，Web2.0 應用服務策略行動之研究—以 Yahoo！、

Google、MSN 為例，碩士論文，國立台灣大學商學研究所，台北。

[28] 陳穆臻 (2005)，Blog 商業模式，管理雜誌，第 375 期，第 112-114 頁。

[29] 陳穆臻 (2005)，Blog 行銷有多行?!，管理雜誌，第 376 期，第 114-117 頁。

[30] 陳昭穎 (2006)，資料探勘技術於超音波旋轉肌肌群影像之診斷應用，碩士論文，國立屏東商業技術學院資訊管理系，屏東。

[31] 曾韋榮 (2006)，結合潛在語意檢索及資訊粒化於資料探勘，碩士論文，國立台北科技大學商業自動化與管理研究所，台北。

[32] 蔡正修 (2007)，台灣上市電子類股價指數走勢預測之研究，碩士論文，國立成功大學統計學系，台南。

[33] 孫明源 (2003)，服務品質、服務價值、滿意度與顧客行為意向關係之研究—以固網寬頻上網服務為例，碩士論文，國立成功大學電信管理研究所，台南。

[34] 楊敦翔 (2003)，以類神經網路與特徵選取技巧處理空氣能見度預測問題之研究，碩士論文，國立中山大學機械與機電工程學系，高雄。

[35] 楊東昌 (2004)，自組織映射圖神經網路改善模式與分群應用之回顧研究，碩士論文，華梵大學工業管理學系，台北。

[36] 楊啟洲 (2005)，以倒傳遞類神經網路作為授信風險預測之研究，碩士論文，中華大學科技管理研究所，新竹。

[37] 尤春惠 (2004)，資料探勘在用藥安全上的應用：預測泛可黴素在腎衰竭病患上的用量適當性，碩士論文，國立中山大學資訊管理學系，高雄。

[38] 葉怡成 (2002)，類神經網路模式應用與實作，第七版，儒林圖書有限公司。

[39] 姚力群、陶卿 (2005)，「局部線性與 One-Class 結合的科技文本分類方法」，計算機研究與發展，42（11），第 1862-1869 頁

[40] 吳忻萍 (1997)，以隱藏語意索引為基礎之中文資訊檢索，碩士論文，國立台灣大學資訊管理學系，台北。

[41] 王彥翔 (2003)，自組特徵映射與學習向量量化神經網路於河川流量之預測，碩士論文，國立台灣大學生物環境系統工程學系，台北。

西文部分

[1] A. K. Jain, R. P. W. Duin, and J. Mao (2000), "Statistical Pattern Recognition：A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, pp. 4-37.

[2] A. Kaban,and M. A. GiroLami (2002), "Fast Extraction of Semantic Features from a Latent Semantic Indexed Text Corpus," *Neural Processing Letters*, Vol. 15, No. 1, pp. 31-43.

[3] A. Kontostathis and W. Pottenger (2006), "A Framework for Understanding Latent Semantic Indexing (LSI) Performance," *Information Processing and Management*, Vol. 42, Issue 1, pp. 56-73.

[4] A. Rosenbloom (2004), "The Blogosphere," *Communications of the ACM*, Vol. 47, Issue 12. pp.31-33.

[5] A. Selamat and S. Omatu (2004), "Web page feature selection and classification using neural networks," *Information Sciences*, Vol. 158, pp. 69-88.

[6] B. Tseng, J. Tatemura, and Y. Wu (2005), "Tomographic Clustering to Visualize Blog Communities as Mountain Views," *In WWW 2005 Workshop on the Weblogging Ecosystem*.

[7] C. Cortes and V. Vapnik (1995), "Support-vector Networks," *Machine Learning*, Vol. 20, No. 3. pp. 273-297.

[8] C. T. Su and C. H. Yang (2008), "Feature selection for the SVM:An application to hypertension diagnosis," *Expert Systems with Applications*, Vol. 34, Issue. 1, pp. 754-763.

[9] F. M. Facca and P. L. Lanzi (2005), "Mining Interesting Knowledge from Weblogs: A Survey," *Data and Knowledge Engineering*, Vol. 53, Issue 3, pp. 225-241.

[10] G. K. F. Tso and K. K. W. Yau (2007), "Predicting Electricity Energy Consumption- A Comparison of Regression Analysis, Decision Tree and Neural Networks," *Energy*, Vol. 32, Issue 9, pp. 1761-1768.

[11] H. K. Ekenel and B. Sankur (2004), "Feature selection in the independent component subspace for face recognition," *Pattern Recognition Letters*, Vol. 25, Issue. 12, pp. 1377-1388.

[12] H. L. Garcia and I. M. Gonzalez (2004), "Self-organizing Map and Clustering for Wastewater Treatment Monitoring," *Engineering Applications of Artificial Intelligence*, Vol. 17, Issue 3, pp. 215-225

[13] J. Fortuna and D. Capson (2004), "Improved support vector classification using PCA and ICA feature space modification," *Pattern Recognition*, Vol.

37, Issue. 6, pp. 1117-1129.
[14] J. Gao and J. Zhang (2005), "Clustered SVD Strategies in Latent Semantic Indexing," *Information Processing and Management*, Vol. 41, Issue 5, pp. 1051-1063.
[15] J. Han and M. Kamber (2001), Data Mining：Concepts and Techniques, *Mogan Kaufmann Publishers*.
[16] K. S. Shin, T. S. Lee, and H. J. Kim (2005), "An Application of Support Vector Machines in Bankruptcy Prediction Model," *Expert Systems with Applications*, Vol. 28, Issue 1, pp. 127-135.
[17] K. Shima, M. Todoriki, and A. Suzuki (2004), "SVM-based Feature Selection of Latent Semantic Features," *Pattern Recognition Letters*, Vol. 25, Issue 9, pp. 1051-1057.
[18] M. C. Wu, S. Y. Lin, and C. H. Lin (2006), "An Effective Application of Decision Tree to Stock Trading," *Expert Systems with Applications*, Vol. 31, Issue 2, pp. 270-274.
[19] M. Ture, I. Kurt, and Z. Akturk (2007), "Comparison of dimension reduction methods using patient satisfaction data," *Expert Systems with Applications*, Vol. 32, Issue 2, pp. 422-426.
[20] M. Y. Kiang, M. Y. Hu, and D. M. Fisher (2006), "An Extended Self-organizing Map Network for Market Segmentation — A Telecommunication Example," *Decision Support Systems*, Vol. 42, Issue 1, pp. 36-47.
[21] M. Y. Kiang, M. Y. Hu, and D. M. Fisher (2007), "The Effect of Sample Size on The Extended Self-organizing Map Network—A Market Segmentation Application," *Computation Statistics and Data Analysis*, Vol. 51, Issue 12, pp. 5940-5948.
[22] P. Husbands, H. Simon, and C. Ding (2005), "Term Norm Distribution and Its Effects on Latent Semantic Indexing," *Information Processing and Management*, Vol. 41, Issue 4, pp. 777-787.
[23] P. Y. Hao, J. H. Chiamg, and Y. K. Tu (2007), "Hierarchically SVM Classification based on Support Vector Clustering Method and Its Application to Document Categorization," *Expert Systems with Applications*, Vol. 33, Issue 3, pp. 627-635.
[24] R. A. Kumar, V. Sugumaran, B. H. L. Gowda, and C. H. Sohn (2008), "Decision Tree- A Very Useful Tool in Analysing Flow-induced Vibration Data," *Mechanical Systems and Signal Processing*, Vol. 22, Issue 1, pp. 202-216.
[25] R. Huang, L. Xi, X. Li, H. Qiu, amd J. Lee (2007), "Residual Life Predictions for Ball Bearings based on Self-organizing Map and Back Propagation Neural Network Methods," *Mechanical Systems and Signal Processing*, Vol. 21, Issue 1, pp. 193-207.

[26] R. Kumar, J. Novak, P. Raghavan and, A. Tomkins (2005), "On The Bursty Evolution of Blogspace," *World Wide Web*, Vol. 8, No. 2, pp. 159-178.
[27] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990), "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, Vol. 41, Issue 6, pp. 391-407.
[28] S. Lee, S. Lee, and Y. Park (2007), "A Prediction Model for Success of Services in E-commerce Using Decision Tree—E-customer's Attitude Towards Online Service," *Expert Systems with Applications*, Vol. 33, Issue 3, pp. 572-581.
[29] T. C. Chang and R. J. Chao (2006), "Application of Back-propagation Networks in Debris Flow Prediction," *Engineering Geology*, Vol. 85, Issue 3-4, pp. 270-280.
[30] T. Kohonen (1982), "Self-organizing Formation of Topologically Correct Feature Maps," *Biological Cybernetics*, Vol. 43, No. 1, pp. 59-69.
[31] T. Kohonen (1998), "The Self-organizing Map," *Neurocomputing*, Vol. 21, No. 1, pp. 1-6.
[32] T. L. Lee (2008), "Back-propagation Neural Network for The Prediction of the Short-term Storm Surge in Taichung Harbor, Taiwan," *Engineering Applications of Artificial Intelligence*, Vol. 21, Issue 1, pp. 63-72.
[33] V. Sugumaran, V. Muralidharan, and K. I. Ramachandran (2007), "Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing," *Mechanical Systems and Signal Processing*, Vol. 21, Issue. 2, pp. 930-942.
[34] Y. C. Lee (2007), "Application of Support Vector Machines to Corporate Credit Rating Prediction," *Expert Systems with Applications*, Vol. 33, Issue 1, pp. 67-74.
[35] Y. Liu and Y. F. Zheng (2006), "FS_SFS:A novel feature selection method for support vector machines," *Pattern Recognition*, Vol. 39, Issue 7, pp. 1333-1345.