國立臺灣體育學院體育研究所
碩士學位論文

植基於資料挖掘技術的體適能評級模式建構之研究

The Study of the Classified Model for Physical Fitness Based on Data mining Techniques

研 究 生：謝俊宏　撰
指導教授：陳定雄　教授

中 華 民 國 九十三年六月

論文名稱:植基於資料挖掘技術的體適能評級模式建構之研究
總頁數：
院校組別：國立台灣體育學院體育研究所
畢業時間及提要別：九十二學年度第二學期碩士學位論文提要
研究生：謝俊宏　　**指導教授：陳定雄教授**

中 文 摘 要

資料挖掘(Data mining) 是最近被提出來的一種資料處理的技術，它結合了統計理論及機器學習技術。本文嘗試使用資料探勘技術以建構一女性青少年的體適能評估模式，藉以科學化的區別青少年的體適能之等級。由於本模式係針對女性青少年所建構，因此樣本選用台中技術學院五專部的新生女學生。模式建構之前，首先選擇最能代表體適能的各項測驗指標，此項測驗指標的選擇是以教育部所公佈的體適能評鑑項目為依據(包含：身體組成、柔軟度、腹肌肌力或肌耐力、瞬發力、心肺耐力等各項測驗)。

模式建構的程序如下：首先，收集樣本，續而依據各學生的體適能資料，將所有學生進行群組分析(形成不同類別體適能的群組)。再將各體適能群組分別利用區別分析、神經網路、支援向量機等三項技術建構體適能評估模式，作為青少女體適能評級分類的基礎，最後再以用此體適能評估模式，以協助學生了解自己的體適能，藉以提升國民健康程度，此亦為本文最大之目的。

關鍵字：群集分析、區別分析、體適能、測驗指標、體適能等級

The Study of the Classified Model for Physical Fitness Based on Data mining Techniques

Abstract

Data Mining have been recently introduced as a new technique for data processing field. This research is to construct a fitness model of female teenagers, which is based on the data mining techniques. By using this model, we can classify the physical fitness scales of teenagers. The samples for this model are 15 years old National Taichung Institute of Technology freshmen. Before constructing this model, we have chosen the most reputable exercise index of physical fitness. The selected criteria are based on the data of The Department of Education. The assessment of physical fitness includes body composition, flexibility, abdominal strength or muscle endurance, anaerobic power and cardiorespiratory endurance.

The fitness models have been constructed as follow: First, collecting the sample and using the Clustering method to form the different cluster of fitness. Second, constructing the fitness models based on the Discriminant Analysis、Neural Netwoks、Support Vector Machine. Finally, utilizing these models to guide students not only to understand their physical fitness scales, but also to improve their health.

Key words: Cluster Analysis, Discriminant Analysis, physical fitness, exercise index, physical fitness scales

目　　錄

表　目　錄

圖　目　錄

第一章　緒　論

1-1 前言

資料探勘（Data Mining）是近年來資料處理應用領域中的一項新技術，其目的是找尋隱藏在資料中的訊息，如趨勢（Trend）、特徵（Pattern）及相關性（Relationship），亦即由資料中發掘資訊或知識。有的文獻稱之為「資料發掘」(Knowledge Discovery in Databases, KDD)，亦有文獻稱為「資料考古學」(Data Archaeology)、「資料樣型分析」（Data Pattern Analysis）或「功能相依分析」（Functional Dependency Analysis）。資料探勘目前已被許多研究人員視為結合資料庫系統與機器學習技術的重要領域。其使用的分析方法，如：預測模式（迴歸分析、時間數列）、資料分群及分類（Data Clustering and Classification）、偏差偵測（Deviation Detection）等。一般而言，Data Mining 內容可包含下列五項：

1. 分類（Classification）

   按照分析對象的屬性分門別類加以定義，建立類別（class）。例如，將受試者的體適能等級，區分為良好體適能者，普通體適能者及不良體適能者。使用的技巧有決策樹（decision tree），記憶基礎推理（memory-based reasoning）等。

2. 估計（Estimation）

   根據既有之相關屬性資料，以獲致某一屬性未知之值。例如：按照受試者的各項體適能測試成績以推估其體適能等級。使用的技巧包括統計方法上之相關分析、迴歸分析及類神經網路。

3. 預測（Prediction）

根據對象屬性之過去觀察值來推估該屬性未來之值。例如由已建構的體適能預測模式以預測認任一受試者的體適能等級。使用的技巧包括迴歸分析、時間數列分析及類神經網路。

4.關聯分群（Affinity grouping）

從所有物件中決定那些相關物件應該放在一起。例如：超市中相關之體育用品（球鞋、球襪、護膝），可放在同一貨架上。在運動行銷系統上，此項功能可用來確認交叉銷售（cross-selling）的機會以設計出吸引人的產品群組。

5.同質群集（Clustering）

將異質母體中區隔為較具同質性之群組（clusters），換言之，其目的是要將組與組之間的差異辨識出來。同質分組相當於行銷術語中的區隔化（segmentation），其強調：假定事先未對於區隔加以定義，而是由資料中自然產生區隔。使用的技巧包括 k-means 法及 agglomeration 法。

目前資料探勘已被廣泛應用於各項領域如: 心理學，疾病分類學，圖樣辨識，生物資訊、、等，尤其商業管理。由於現代的企業體經常蒐集了大量資料，包括市場、客戶、供應商、競爭對手以及未來趨勢等重要資訊，但是資訊超載與無結構化，使得企業決策單位無法有效利用現存的資訊，甚至使決策行為產生混亂與誤用。而資料探勘技術可從巨量的資料庫中，發掘出不同的資訊與知識以作為決策支援之用，使得企業產生競爭優勢。在體育科技與管理領域中，資料探勘的應用尚屬少見。

1-2 研究動機與目的

近幾年政府大力倡導民眾從事體育活動，藉以增進國民之健康，如：333 活動，強調民眾身體要健康，必須每週運動至少三天，每次時間為 30 分鐘，且心跳達到 130 以上。而教育部亦推行學生之體適能檢測，並盼經由此檢測，以評估每一位學生的體適能力，以便提出建議，進而提升學生之體能素質。要檢測學生的體適能必須建構一個體適能評估模式，藉由此評估模式，可以正確評估出學生的體適能之等級。目前教育部的體適能體育獎章之頒發，是依據受試者的身體組成(身體質量指數)、柔軟度(坐姿體前彎)、腹肌力或肌耐力(仰臥起坐)、瞬發力(立定跳遠)、心肺耐力 (800m 女)等五項的「運動體能」表現。其頒發方式是採齊頭的評級模式，亦即要求柔軟度、腹肌肌力或肌耐力、瞬發力、心肺耐力等的各項測驗分數必須同時達到最低標準才可獲獎(獎項分成金、銀、銅三等級)。此種給獎方式並非公平的方式，其不但缺乏彈性，且假設各項測驗分數與體適能等級呈線性平面的關係。然而實際的資料顯示，各項測驗分數與體適能等級之間的關係相當複雜，必須具有乏晰判斷(或稱模糊判斷)的理念，方可作出正確的體適能等級之歸屬。因此使用齊頭式評級模式進行體適能的等級分類，將可能造成誤判。

國內目前所使用的體適能評估模式；大都使用常模(Normal Distribution)的建構，並將學生的各項測驗分數與各項分數的常模進行比對，以進行齊頭式的評估學生的體適能，此項評比方式在若干情況會產生偏誤。因此引進資料發掘的群集、分類與預測分析以建構體適能評級模式，以增進體適能評估之精準度，為本文中最主要之動機。

由於群集、分類與預測分析都分別包含了傳統多變量統計、人工智慧、類神經網路等技術，而這些不同技術均有其特色。鑑于此，本文有兩個目的：

1. 嘗試分別使用多變量統計(Cluster Analysis、Discriminant Analysis)、人工智慧(Support Vector Machine)、類神經網路(Neural Network)等技術建構三個體適能評級模式，以提供不同情況之使用。
2. 比較三項體適能評級模式的分類正確率。

1-3 本文結構

本文內容共分成五章，第一章為緒論，包含研究動機與目的；第二章為文獻探討，包括資料探勘相關技術及過去有關體適能評級的回顧。第三章至第五章章分別使用多變量統計、類神經網路、支援向量機等技術建構三個體適能評級模式。第五章**結論與建議**。

第二章文獻探討

2-1.資料探勘之理論與技術

隨著資訊科技的進展，許多新的資訊分析工具問世，例如：關聯式資料庫、模糊計算理論、基因演算法則以及類神經網路等，使得從資料中發掘有用的知識成為一種系統性且可實行的程序。

一般而言，資料探勘(Data Mining)[5,11,19,30]的理論技術可分為傳統技術與改良技術兩支。傳統技術以統計分析為代表，舉凡統計學內所含之敘述統計、機率論、迴歸分析、類別資料分析等皆屬之，尤其資料探勘對象多為變數繁多且筆數龐大的資料，是以高等統計學裡所含括之多變量分析中用來精簡變數的因素分析（Factor Analysis）、用來分類的區別分析（Discriminant Analysis），以及用來區隔群體的分群分析（Cluster Analysis）等，在 Data Mining 過程中特別常用。

在改良技術方面，應用較普遍的有決策樹理論（Decision Trees)、規則歸納法(Rules Induction)、類神經網路(Neural Network)以及支援向量機(Support Vector Machine)等。決策樹是一種用樹枝狀展現資料受各變數的影響情形之預測模型，根據對目標變數產生之效應的不同而建構分類的規則，一般多運用在對顧客資料的區隔分析上，例如針對「會至現場觀賞職棒」與「未至現場觀賞職棒」的觀眾群對象找出影響其分類結果的變數組合，常用分類方法為 CART（Classification and Regression Trees）及 CHAID（Chi-Square Automatic Interaction Detector）兩種。

類神經網路[9,18]是一種模擬人腦思考結構的資料分析模式，經由輸入之變數與數值中自我學習，並根據學習經驗所得之知識不斷調整參數以期建構資料的型樣（patterns）。類神經網路為非線性

的設計，與傳統迴歸分析相比，好處是在進行分析時無須限定模式，特別當資料變數間存有交互效應時可自動偵測出；缺點則在於其分析過程為一黑盒子，故常無法以可讀之模型格式展現，每階段的加權與轉換亦不明確，是故類神經網路多利用於資料屬於高度非線性且帶有相當程度的變數交互效應時。

規則歸納法是知識探勘領域中最常用的格式，這是一種由一連串的「如果…/則…（If / Then）」之邏輯規則對資料進行細分的技術，在實際運用時如何界定規則為有效是最大的問題，通常需先將資料中發生數太少的項目先剔除，以避免產生無意義的邏輯規則。

運用以上各式理論技術，Data Mining 可以建立下列多項應用模式：

1. Classification 模式

是根據一組自變數的數值進行計算，再依照計算結果作分類。計算的結果最後會被分類為幾個少數的離散數值，例如將一組職棒觀眾分為「會至現場觀賞職棒」與「未至現場觀賞職棒」兩種觀眾群。其作法是使用已經分類的資料來研究它們的特徵，然後再根據這些特徵對其他未經分類或是新的資料做預測。這些我們用來尋找特徵的已分類資料可能是來自我們的現有的歷史性資料，或是將一個完整資料庫做部份取樣，再經由實際的運作來測試；譬如利用一個大的職棒對象資料庫的部份取樣來建立一個 Classification Model，以後再利用這個 Model 來對其他觀眾作預測。Classification 通常會牽涉到兩種統計方法：Logistic Regression 以及 Discriminant Analysis。由於 Data Mining 已漸普遍，使得 Neural Networks 以及 Decision Tree 也漸漸受到採用。

雖然這些統計方法本身都十分複雜，但使用者並不須要牽涉到這些繁雜的統計。

2. Neural Networks 模式

神經網路使用許多參數（每個參數代表網路上的一個 Node）來建立一個模式。此模式可接受一組輸入值並預測出一個連續值或分類值。每一個節點（Node）都是一個函數，這個函數是使用輸入該節點的相鄰節點值的加權總和（Weighted Sum）做運算。在建立一個模式的過程中，須使用一些資料來'餵'給這個網路，'訓練'它來找到一組能夠產生最佳輸出結果的加權值（Weights）。一般最常用的訓練演算法為 Back-Propagation，它是把輸出結果與一個已知的正確結果相比。每次相比之後就產生另一組調整過的 Weights，然後再產生一個新的輸出值再與該已知值相比。這個過程經過反覆的執行後，此 Neural Networks 就被訓練得能夠相當正確的做預測。

然而 Neural Networks 存在有兩個問題。首先，Neural Networks 最受質疑的是它的'曖昧不明'的特性，也就是它做的預測所根據的因素並不明確。第二，Neural Networks 對原始測試資料可以做相當正確的預測，但是對真實資料預測的準確性則較差。目前已有一些新的技術可以改正這個缺點。

3. Decision Tree 模式

利用一系列的規則來得到一個類別或數值。例如，想把體適能受試者歸類成'良好體適能'與'不良體適能'兩種。有了這個 Decision Tree，一體適能受試者即可容易被判定為是屬於良好體適能群或不良體適能群。

4. Regression 模式

是使用一系列的現有自變數值來預測一個連續型應變數的可能值。若將範圍擴大亦可利用 Logistic Regression 來預測類別型應變數。前面所提到的類神經網路或決策樹理論等分析工具，亦可用於建構 Regression 模式，使得在預測的功能上大大增加了選擇工具的彈性與應用範圍的廣度。

5. Time-Series Forecasting 模式

與 Regression 模式近似，時間數列模式是以現有的數值來預測未來的數值。Time-Series Forecasting 與 Regression 不同點在於它所分析的數值都與時間有關。

6. Clustering 模式

將一群資料分成若干組，其目的是要將組與組之間的差異找出來，同時也要將一個組之中的成員的相似性找出來。

7. Association 模式

找出與某一事件會同時出現的另一事件。Association 主要是要找出下面這樣的資訊：如果 Item A 是某一事件的一部份，則 Item B 也出現在該事件中的機率有 X%。（例如：如果一個顧客買了一雙球鞋，則此個顧客同時也買球襪的機率是 90%。）

8. Sequence Discovery 模式

與 Association 模式很相似，所不同的是 Sequence Discovery 中相關的 Item 是以時間區分開來（例如：如果一球迷今天在此觀賞這

場職棒，則明天在此觀賞職棒的機率是 80%。

由以上可以發現資料探勘技術的多樣化，有傳統分析工具，例如：統計迴歸預測模型、資料庫分割、連接分析、偏差偵測等；亦有新的應用技術，如：類神經網路、機器學習、專家系統等人工智慧的工具。文獻顯示一些新的演算法技術如：基因演算法（Genetic algorithms）亦逐漸被應用於資料探勘領域。

基因演算法是一種全新的最佳化空間搜尋法，其最初概念是由 John Holland 於 1975 年提出，其主要目的如下：1.以嚴密的科學方法解釋自然界中「物競天擇、適者生存」的演化過程。2.將生物界中基因演化重要機制以資訊科學軟體實作模擬。近年來，資訊科技的長足進步，在更快穩定的系統支援下，基因演算法被各領域廣泛應用。基因演算法屬於人工智慧領域中的自我學習機制，可提供各類最佳化問題的快速求解，它亦提供了一種不同以往的思考模式，運用在 Data Mining 上，可以在巨量資料中快速搜尋、比對、演化出最佳點，尤其具有學習機制，可在 Data Mining 領域綻放光芒。

基因演算法是應用演算法的適應函數來決定搜尋的方向，再運用一些擬生物化的人工運算過程，例如選擇（selection）、複製（reproduction）、交配（crossover）和突變（mutation）等進行演化，週而復始地進行一代一代的演化，以求得一個最佳的結果。它具有強固性（robustness）與求值空間的獨立性（domain independence）。強固性使問題的限制條件降到最低，並大幅提高系統的容錯能力；而求值空間的獨立性則使基因演算法的設計單一化，且適用於多種不同性質、領域的問題。因此，利用它於 Data Mining 領域中，可以挖掘出不同的資訊、別人看不出的資訊，必然帶給企業體巨大的商機。

2-2 體適能之評級

體適能(Physical Fitness)是從事各項運動的基礎，更是一個人的健康指標。一個具有良好體能的人，不僅可以增加其學習各項運動技巧的效率，同時更是健康身體的指標。一般人可以經由體適能的檢測結果，了解自己身體運動體能的趨向，並加強不足之項，以增進身體之健康。

青少年是國家的棟樑，也是體適能最易琢磨的年齡層。若能建構一簡易且精確的體適能評級系統，將有助於青少年的體適能訓練。目前國內有關體適能的相關研究均側重於全體學生的體適能調查，其作法大部份都是收集受試者的各項體適能的指標資料，並將各項指標分別建立常模，最後再以各項常模作為體適能的評價水準之基礎。另外有部份研究則偏向於不同群體的體適能比較，其作法是將不同的學校學生體適能資料收集之後，再以單因子變異數分析、積差相關及逐步迴歸分析等統計方法進行體適能各項指標的比較與分析。而有的研究則是將回收之有效樣本以敘述性統計、因素分析、獨立樣本 T 考驗等統計方法進行資料分析。部份研究亦採用歷史研究法及文件分析法，經過分類整理、考證、分析、討論並作出簡單之評論，而對於建構一可評估體適能的機制，則甚少探討。

至於國外相關研究是利用多元迴歸模式建立分類模式[31]，由於多元迴歸線性模式必須假設所有變項均呈多元常態分配，且所有自變項與應變項之間均假設存在線性關係。若違反此項假設，則可能造成預測或分類的誤差。

第三章應用多變量統計建構體適能評級模式

3-1.前言

本章嘗試使用多變量統計中的群集分析 (Cluster analysis)及區別分析(Discriminant analysis)技術以建構一女性青少年的體適能評級模式，藉以科學化的區別青少年的體適能之等級。由於本模式係針對女性青少年所建構，因此樣本選用台中技術學院五專部的新生女學生。模式建構之前，首先選擇最能代表體適能的各項測驗指標，此項測驗指標的選擇是以教育部所公佈的體適能評鑑項目為依據(包含：身體組成、柔軟度、腹肌肌力或肌耐力、瞬發力、心肺耐力等各項測驗)。模式建構之步驟：

(1) 選擇體適能的各項檢測指標

本文所選擇的各項體適能指標，是以教育部所公佈的體適能評鑑項目作為分類依據，其評鑑項目包含：身體組成(身體質量指數)、柔軟度(坐姿體前彎)、腹肌力或肌耐力(仰臥起坐或一分鐘引體向上)、瞬發力(立定跳遠)、心肺耐力 (800m女) 等五項指標。

(2) 收集青少年之各項體適能測驗成績

由於本研究屬實驗性質，樣本來源僅由國立台中技術學院九十二學年度五專部新生取得。首先從體育組取得新生體適能資料，剔除不完整及不足之資料，作為本研究之觀測樣本。

(3) 建構分類模式

本步驟是整個模式建構的重點，首先利用群集分析將樣本依據其五項體適能成績進行群集(cluster)。針對

所產生的若干個群集，分別進行區別模式的建立，如此則可以建構一項適合國內青少年的體適能分類模式。

(4) 驗證模式的準確率

將建構的體適能評估模式，加以驗證其準確度。以下為體適能模式建構之流程：

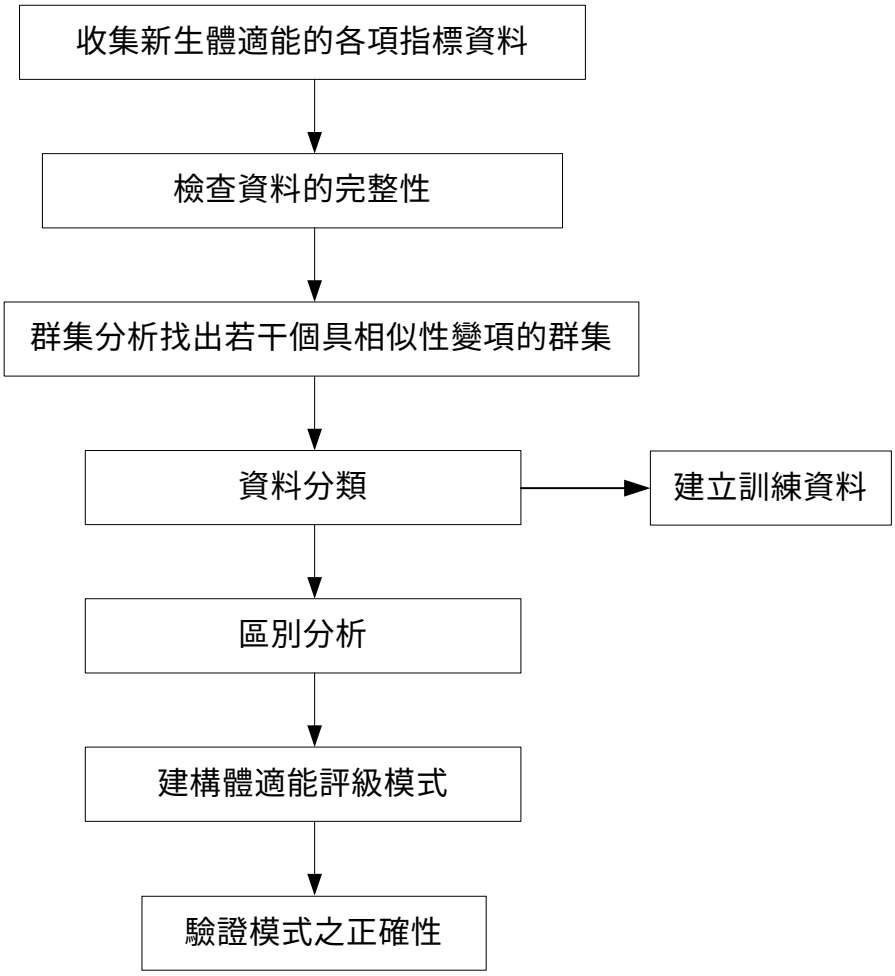

圖 3-1、模式建構流程

3-2.多變量統計的區別分析模式

區別分析模式[係依據 Fisher 程序而來。此方法強調使得組間變異與組內變異之比為最大。由 Fisher 程序所獲取之分類規則將可使得“錯誤分類”的成本最低，但使用區別分析模式，必須滿足下列三項假設前提：

1.每一組群均須滿足多元常態分配。

2.每一組群的共變數矩陣均一致。

3.平均數，變異數，事前機率，每一組群的錯誤分類的成本需已知。

對於兩組群的區別函數，可寫成如下：

$D(X) = X' \Sigma^{-1} (\mu_1 - \mu_2) - {}^1/_2 (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 + \mu_2)$

其中：μ_1, μ_2 為兩群體之平均數矩陣。

Σ^{-1} 為共變數矩陣之反矩陣。

X 為變數矩陣。

其分類規則如下：

1.若滿足 $D(X) >= Ln(c_{21}P_1 / c_{12}P_2)$，則 X 歸屬於組群 1。

2.若滿足 $D(X) < Ln(c_{21}P_1 / c_{12}P_2)$，則 X 歸屬於組群 2。

其中：

(1).P_1, P_2 為每一組群的事前機率。

(2).c_{21} 表示應屬於組群 1，卻被誤分類為群 2 的成本。

(3).c_{12} 表示應屬於組群 2，卻被誤分類為群 1 的成本。

3-3.多變量統計的群集分析模式

當隨機變數的分配未知，則此時必須使用無母參數模式來分類樣本。較知名的無母數方法包含：K最鄰近模式（K Nearest Neighbor）及數學規劃模式（Mathematic Programming Models)。K最鄰近模式可用於當群體不屬於常態分配的情況，尤其同一群體內的觀測樣本形成群聚（cluster）的情況更適用於此方法。與母參數方法比較，K最鄰近模式更適用於“對母體群有較少限制假設”的情況。　　K最鄰近模式不但不需滿足常態分配的假設，且其分類時並不需要一區別函數格式，其分類一樣本的歸屬是取決於其鄰近的K個觀測樣本是否屬於該群體。當K最鄰近模式將一樣本歸屬於一群聚時（該群聚是由K個距離最近的相鄰樣所形成)，係依據其距離。該距離的量測是採用歐幾里得距離測度（Euclidean distance)。

3-4.實驗結果與分析

本研究的樣本是由台中技術學院體育室所取得的200筆五專一年級女生新生的體適能資料，將此樣本經群集分析後，可得到三個組別，分別定義為良好、中等、不良(如表3-1)。再經由區別分析，為此三個組別，分別建構區別函數：

表3-1 組別統計量

組別		平均數	標準差
良好(1)	身體組成(身體質量指數)	22.92	1.45
	柔軟度(坐姿體前彎)	44.33	5.42
	腹肌力或肌耐力(仰臥起坐)	39.91	4.37
	瞬發力(立定跳遠)	172.46	5.6
	心肺耐力 (800m 女)	215.83	26.89
中等(2)	身體組成(身體質量指數)	19.14	1.52
	柔軟度(坐姿體前彎)	31.61	5.16
	腹肌力或肌耐力(仰臥起坐)	29.43	4.18
	瞬發力(立定跳遠)	156.81	6.16
	心肺耐力 (800m 女)	312.94	22.88
差(3)	身體組成(身體質量指數)	14.60	1.02
	柔軟度(坐姿體前彎)	17.07	3.27
	腹肌力或肌耐力(仰臥起坐)	15.57	3.08
	瞬發力(立定跳遠)	139.86	3.80
	心肺耐力 (800m 女)	375.29	20.90
總和	身體組成(身體質量指數)	19.69	2.57
	柔軟度(坐姿體前彎)	33.52	8.63
	腹肌力或肌耐力 (仰臥起坐或一分鐘引體向上)	30.87	7.35
	瞬發力(立定跳遠)	159.22	10.28
	心肺耐力 (800m 女)	303.25	40.00

表 3-2、典型區別函數之特徵值

函數	特徵值	解釋變異量%	累積%	典型相關
1	2047.564*	99.9	99.9	1.000
2	1.873*	0.1	100.0	0.807

*為分析時會使用的前二個典型區別函數

在表 3-2 中有二項典型區別函數，其特徵值愈大，代表區別函數的辨別能力愈好。由上表得知，第 1 個區別函數可解釋 99.9%的變異量。而第 2 個區別函數則具有 0.1%的變異量之解釋能力。

表 3-3 標準化的典型區別函數係數

	函數	
	1	2
身體組成(身體質量指數)	0.990	0.354
柔軟度(坐姿體前彎)	1.439	0.462
腹肌力或肌耐力 (仰臥起坐或一分鐘引體向上)	1.373	4.449
瞬發力(立定跳遠)	1.372	-5.161
心肺耐力 (800m 女)	-5.084	0.164

表 3-3 為二個典型區別函數的標準化係數，分別為：

區別函數 1：

F_1=0.990*身體組成+1.439*柔軟度+1.373*腹肌力或肌耐力+1.372*瞬發力-5.084*心肺耐力

區別函數 2：

F_2=0.354*身體組成+0.462*柔軟度+4.449*腹肌力或肌耐力-5.161*瞬發力+0.164*心肺耐力

從標準化典型區別函數係數值的大小，可看出“柔軟度(坐姿體前彎)”對區別函數 1 的影響最大，而“腹肌力或肌耐力”對區別函數 2 的影響亦最大。

表 3-4　結構矩陣

	函數	
	1	2
腹肌力或肌耐力(仰臥起坐)	0.094	0.331*
身體組成(身體質量指數)	0.093	0.242*
瞬發力(立定跳遠)	-0.102	0.231*
柔軟度(坐姿體前彎)	0.093	0.207*
心肺耐力 (800m 女)	0.093	0.134*

*每個變數和任一區別函數之間的最大絕對關係

區別變數和標準化典型區別函數之合併組內矩陣稱為「結構矩陣」，若相關係數之絕對值愈大，表示變數對區別函數具有較大的影響力。從表 3-4 中可看出腹肌力或肌耐力(仰臥起坐)、身體組成(身體質量指數)、瞬發力(立定跳遠)、柔軟度(坐姿體前彎)、心肺耐力 (800m 女)對區別函數 2 的影

響較大。

表3-5 Fisher's 線性區別函數之分類函數係數

	組別		
	1(良好)	2(中等)	3(差)
身體組成(身體質量指數)	-213.853	-150.119	-81.515
柔軟度(坐姿體前彎)	-448.033	-420.773	-391.317
腹肌力與肌耐力 (仰臥起坐或一分鐘引體向上)	-188.288	-153.818	-126.319
瞬發力(立定跳遠)	506.029	523.106	551.616
心肺耐力 (800m 女)	15.878	-2.993	-24.147
(常數)	-31641.493	-30214.258	-32860.378

表3-5中，每一個運動項目都有一組函數的係數，本研究共有良好、中等、差等三項體適能等級，因此產生三項分類函數，分別為：

群組分類1（良好）：

D_1=-31641.493-213.853*身體組成-448.033*柔軟度-188.288*腹肌力或肌耐力+506.029*瞬發力+15.878*心肺耐力

群組分類2（中等）：

D_2=-30214.258-150.119*身體組成-420.773*柔軟度-153.818*腹肌力或肌耐力+523.106*瞬發力-2.993*心肺耐力

群組分類3（不良）：

D_3=-32860.378-81.515*身體組成-391.317*柔軟度-126.319*腹肌力或肌耐力+551.616*瞬發力-24.147*心肺耐力

此三個分類函數模式可作為青少年體適能評級分類模式。若有新的學生體適能資料時，只須將青少年的各項體適能項目分別輸入這三個群組分類函數式子，再依照函數所產生的數值大小進行比較，若群組函數值最大者，則表示該青少年屬於該體適能等級。

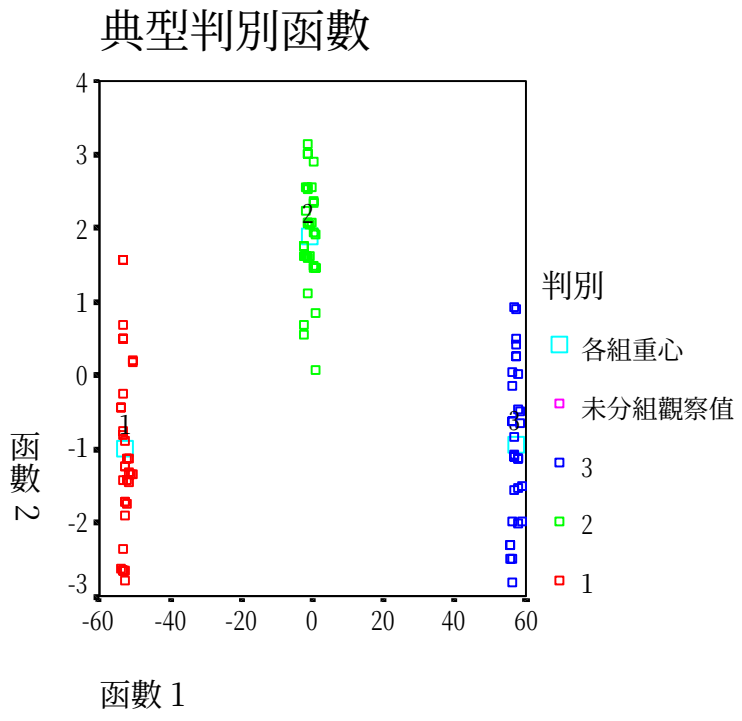

圖 3-2、典型判別函數之合併組散佈圖

圖 3-2 為樣本代入兩個典型判別函數中，所得到函數值的散佈圖。由圖看出樣本形成三個不同組群區域。

表 3-6 組群分類的正確性

組別			預測的各組成員			總和
			1	2	3	
原始的	個數	1	45	0	1	46
		2	15	118	7	140
		3	0	0	14	14
	%	1	97.8	0	2.2	100.0
		2	10.7	84.3	5.0	100.0
		3	0.0	0.0	100.0	100.0
88.5%個原始組別觀察值已正確分類。						

在表 3-6 中，第 1 組別(良好)中的 46 位青少年，有 46 位被正確分類，1 位被分類到第 3 組別(差)。而第 2 組別(中等)中的 140 位青少年，有 118 位被分類正確，有 15 位被分類到第 1 組別(良好)，而有 7 位被分類到第 3 組別(差)。第 3 組別(差)中的 14 位青少年，有 14 位被分類正確。

而在表 3-6 中的下半部百分比部分，此對角線呈現的數值是為正確率，例如：第 1 組別(良好)的正確率是 45/46=97.8%，第 2 組別(中等)的正確率是 118/140=84.3%，而第 3 組別(差)的正確率是 14/14=100%。所以整體的準確率是：177/200=88.5%。也就是說，有 200 位青少年，其中 177 位被正確分類，所以有精確的準確率。

第四章應用類神經網路建構體適能評級模式

4-1 前言

過去十年許多文獻探討類神經網路在預測分類領域的應用。類神經網路(Neural Network)係以調整網路上的連結權值，以進行學習，兼具有歸納學習及統計方法的潛力，其特點是將學習的知識分散儲存於網路中，因此可容忍雜訊。此外神經網路異具有知識增加性(Capable of learning in crementlly)的能力，當一新的範例加入網路時，可以很容易的經由學習而更新知識。在類神經網路拓撲中，自組織特徵映射網路(Self-Organizing Feature Map)適合用於分群(Clustering)作業，而倒傳遞網路(Bropagation)則適用於分類(Classification)及預測(Forcasting)作業。

自組織特徵映射網路簡稱為 SOFM，或稱為 Kohonen Map 係屬於一非監督式學習網路，可以將多維度空間的特徵投射於較少維度的輸出空間(通常唯一維或二維)亦即將圖樣(pattern)的內在關係展限於平面圖形上，可達成簡化資料分群的目的。倒傳遞神經網路則屬於一監督式網路，以網路層級(Layer)間節點的連結形式，形成複雜的決策區間(Region)，足以解決各種非結構性的分類及預測問題。

本章主要目的使用 SOFM 網路及倒傳遞網路建構一體適能分類模式。首先利用 SOFM 網路將受試者分群，再利用倒傳遞網路將分群後的樣本建構一體適能評級模式。模式建構流程與第三章相似，步驟如下：

(1)選擇柔軟度(坐姿體前彎)、腹肌力或肌耐力(仰臥起坐或一分鐘引體向上)、瞬發力(立定跳遠)、心肺耐力

(800m 女) 等四項指標做為體適能的各項檢測指標。

(2)收集受試者之各項體適能測驗成績。

(3)建構評級模式

首先利用 SOFM 網路將受試者依據其四項體適能成績進行群集(cluster)。群集後每一受試者將會隸屬於一群集，再將此成對的資料輸入倒傳遞網路則可以建構一值基於類神經網路的體適能評級模式。

(4) 驗証模式的準確率

將建構的體適能評估模式，加以驗證其準確度。

4-2.倒傳遞網路模式

類神經網路是一個類似大腦資料處理的計算機結構，它有許多名稱如：連結模式（Connectionist Model），平行分散處理模式（Parallel Distributed Processing Model）。不若傳統的計算機必項依序的執行程式指令，類神經網路係基於同步處理之觀念，使用大量平行網路，而網路上佈滿許多非線性的處理單元（節點）。網路上的節點以連結相接，而連結上賦有權值。類神經網路的知識即以權值方式儲存於多層次（Multi-layer）的網路上。

類神經網路由訊號運算特性，網路拓撲以及學習演算法組成，目前一般常見的網路有：ADALINE、ART、Hamming Net、Hopfield Net、Kohonen Feature Map 及 Multi-layer Perceptron 等。通常類神經網路的第 i 個計算節點接收了一組輸入值（X1,X2,…,Xn），並伴隨著一組權值（Wj1,Wj2,…,Wjn）其 Wji 表示連接節點 i 與節點 j 的權值。節點上的啟動機制（Activation machanism）會計算輸入值與對應權值的乘積總和並將之與該節點的閥值（Threshold value）作比較，得一比較值，以決定激發程度。此比較值再經轉換機制（Transfer machanism）將其轉換成輸出值，此輸出值又成為其他節點的輸入值，如圖 4-1。一般常見的轉換機制有：Sigmod、Hard limited 及 Threshold logic function 。

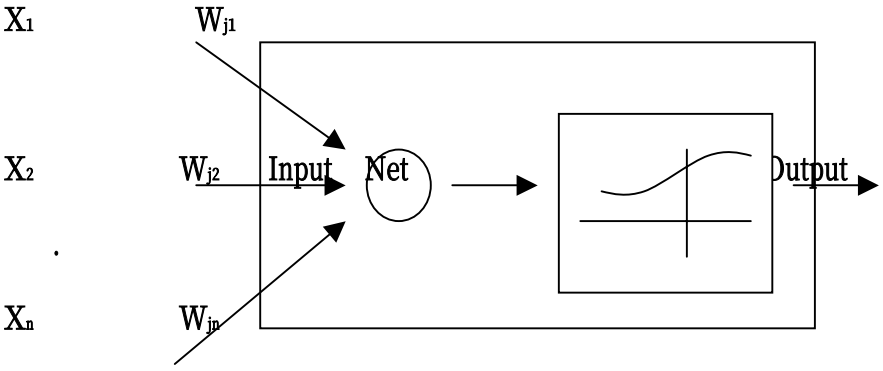

圖 4-1　非線性的處理單元

拓撲（Topology）係指網路層級（Layer）間節點的連結型式，其中以倒傳遞網路最適合用於分類問題。一倒傳遞網路係由輸入層、隱藏層及輸出層所構成。在網路上，同一層內的節點不連結，而相鄰層的節點則完全連結。三層倒傳遞網路是一最常使用的倒傳遞網路，如圖 4-2。類神經網路最重要的部份是其學習演算法，學習演算法可以依據目前的結果，經由多次的權值調整，以改善其效益，亦即學習演算法可賦與類神經網路調適能力（Adaptability）經由權值調整，可使類神經網路趨近穩健（Robustness）程度。

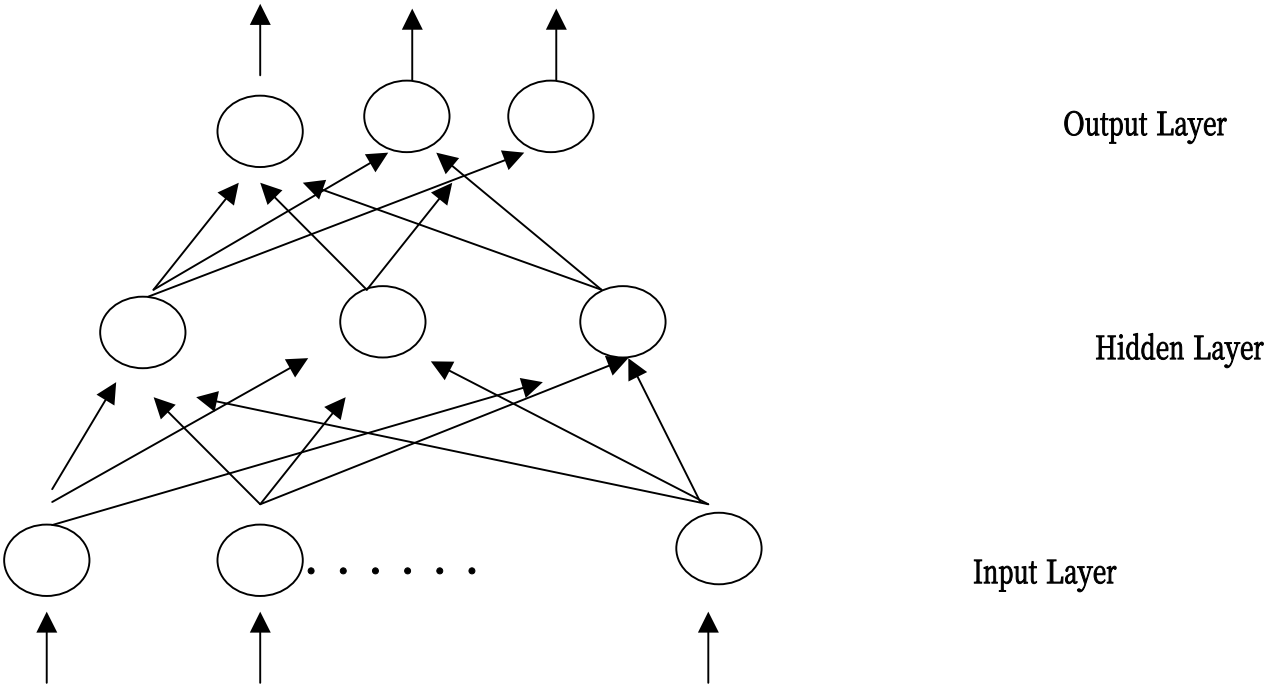

圖 4-2 三層網路架構

4-3.自組織特徵映射網路模式

生物資訊系統的研究者曾經推測，為了在大腦中有效的表達資料，人類大腦的資訊處理方法是採取縮小範圍只表達重要事實，以達成資訊壓縮，但不遺漏原始資料的拓撲關係，此為之拓撲保留(Topology preservation)。自組織特徵映射網路即是模擬此種資訊處理的過程。

Kohonen根據早期Willshaw即Vonder Malsburg的研究工作，在1979-1982之間發展了自組織特徵映射網路，此網路模式屬於非監督式競爭學習網路，可用與萃取高維度輸入空間的拓撲及程結構，而以二維圖示呈現。其主要用途為群聚(Clustering)及圖樣辨識(Pattern Recognition)。競爭式神經網路屬於非監督式學習神經網路，可視為群聚演算法在平行分散式架構上的實現。其基本理念是試圖在未經標示的樣本中，尋找相似的特徵，規則或是關係，然後將具共同特徵的樣本聚集成一群聚。然而依據生理學所得到之證據，神經元之間具有側向交互作用（Lateral Interaction)。此項側向交互作用，不只是一單純競爭關係，而是一項競爭（抑制作用）加上合作（激勵作用）之關係。而自組織特徵映射網路則可用來模擬此側向交互作用。自組織特徵映射是一種兩層順向連結的神經網路架構，其和一般神經網路架構最大不同，就是將輸出神經元依序安排在有前後關係的直線或平面上。

此種特徵映射主要目的是將高維度的圖樣特徵映至一維或二維的輸出神經元陣列。換言之，當圖樣之間存在著某一種量測或拓撲的順序關係時，吾人希望經由連結權值之學習，使得輸出單元之間也可保持此種拓撲關係，此項拓撲關係簡單而言，即是神經元之間的鄰居關係。

自組織特徵映射網路之另一項特色，就是它的競爭方式。一般競爭式學習神經網路所採用的是“贏者通吃”的方式，而自組織特徵映射網路所採用的則是“有福同享”的方式，也就是說，競爭之後，不止獲勝的神經元有資格學習，其周圍的神經元也可以學習，此種學習方式，基本上近似墨西哥帽式的側向作用函數。如圖 4－3：

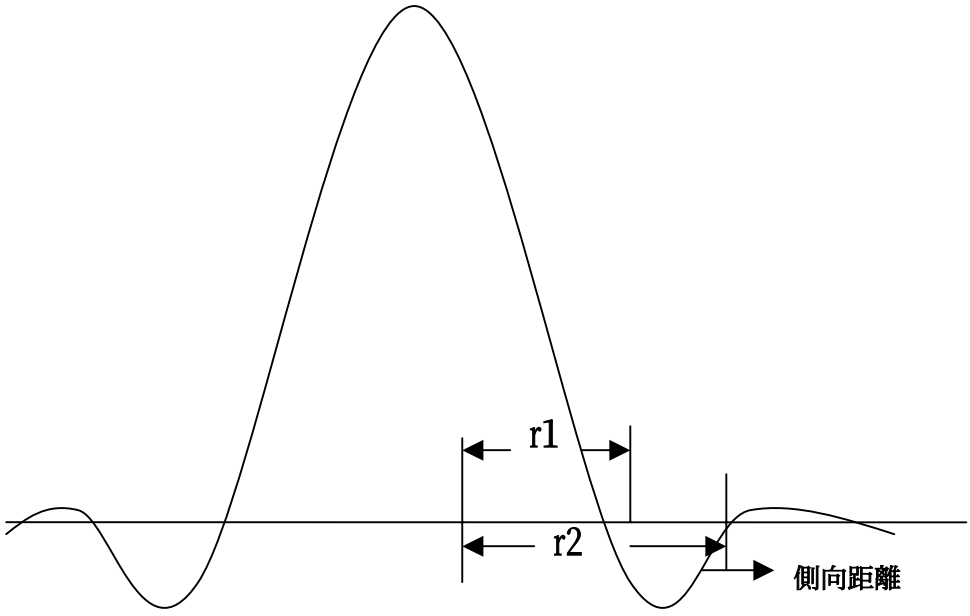

圖 4－1 墨西哥帽式的側向交互作用

（資料來源：Kohonen,1988,P11）

自組織特徵映射神經網路擁有兩層節點：輸入層及Kohonen層，兩層之間採完全連結的方式（Full Connected)。其中Kohonen層是自組織特徵映射網路的核心，其主要功能是將稀疏的高維度輸入資料加以壓縮，並以密集方式呈現在二維圖上。自組織特徵映射會自動將Kohonen層的各子區域形成代表不同的群聚，因此，Kohonen層上被啟動的處理單元係代表了具一特定性質的一組輸入圖樣。

在訓練過程中，圖樣被經由輸入層之節點輸入於網路。輸入的圖樣包含一個或多個值，每一個值餵入輸入層的各輸入節點。此具m值的輸入圖樣，可以定義成一向量，如下：$X=(X_1,X_2,.....,X_m)$，Kohonen層的每一節點均有一權值向量與它相連。假定Kohonen層有N個節點，

則與第j個節點相連的權值向量定義為：

$$W_j=(W_{j1}, W_{j2},.....,W_{jm})$$

其中：W_{j1}表示連接節點j與輸入層節點i的權值。

在訓練的初始過程中，權值的設定是採隨機賦予。當訓練過程持續，自組織特徵映射會依據輸入資料的拓撲關係調整其權值向量。對任一輸入圖樣，此Kohonen層會依據各節點的權值向量與輸入向量的距離，來決定獲勝節點。其中具有最短距離的節點為獲勝者，並且調整其權值以便接近輸入圖樣。用以測量Kohonen層的節點與輸入圖樣之距離的最常見方法是歐幾里得距離（Euclidean）測度。自組織特徵映射的學習演算法如下：

步驟 1） 初始權值設定

首先設定鄰近半徑為 $N_c(0)$，再隨機設定 Kohonen 層的各節點之初始權值 $W_j(0)$。

步驟 2） 依序輸入圖樣向量 $x(t)=[x_1,x_2,.....,x_m]^T$

其中：x（t）表第 t 次迭代輸入的圖樣向量。

步驟 3） 尋找獲勝神經元，並計算其輸出

$y_c=1$ 若 $\|x(t)-W_c(t)\|=\text{Min}\|x(t)-W_j(t)\|$ $j=1,2,\cdots,N$

$y_c=0$ 其它

其中：c 表獲勝單元

$W_j(t)$ 表經 t 次迭代後，Kohonen 層的第 j 節點之權值向量

步驟 4） 調整神經元 c 及其鄰域之連結權值

$W_j(t+1)=W_j(t)+\alpha(t)[X(t)-W_j(t)]$ 若 $j\leqq N_c(t)$

$W_j(t+1)=W_j(t)$ 若 $j=N_c(t)$

其中：學習係數 $\alpha(t)=\alpha(1-t/K)$ $0<\alpha<1$

K 為輸入圖樣的總數目

$N_c(t)$ 表經 t 次迭代後的鄰域半徑

步驟 5） 回到步驟 2），直到滿足一定的迭代數。

初始鄰域集合 $N_c(0)$ 是以神經元 c 為中心，上下左右各取 R_0 個神經元，其所構成的 $(2R_0+1)\times(2R_0+1)$ 的方形區域，通常要求必須涵蓋半數以上之神經元。為了避免某一節點獲勝次數過多，吾人加進了良心機制（Conscience）以便使得距離測量產生偏移。

為了使權值向量能夠學習到圖樣分佈之情形，吾人將圖

樣空間分成 L 個區域。希望每個區域所涵蓋之圖樣數目大致相同。良心機制的設計目的是使得 Kohonen 層的某些節點獲勝次數過多時，應拿出良心來，把機會讓給其它的節點，其最終目的是能學習到均勻的圖樣分佈，

亦即若圖樣空間可形成 L 個群聚，則每個區域所涵蓋之圖樣數目大致相同為 K／L。

4-4 實驗結果與分析

首先取得台中技術學院女生新生的體適能資料，將此樣本經自組織特徵映射神經網路的學習後，可得到三個群集(如圖4-1)，分別定義為良好、中等、不良。此三個群集的樣本敘述統計量如表 4-1。

表 4-1、群集統計量

組別		平均數	標準差
良好(1)	身體組成(身體質量指數)	23.16	0.95
	柔軟度(坐姿體前彎)	45.00	3.19
	腹肌力或肌耐力(仰臥起坐)	40.22	2.65
	瞬發力(立定跳遠)	173.28	3.70
	心肺耐力 (800m 女)	219.03	17.09
中等(2)	身體組成(身體質量指數)	19.29	0.80
	柔軟度(坐姿體前彎)	31.59	3.09
	腹肌力或肌耐力(仰臥起坐)	29.21	2.65
	瞬發力(立定跳遠)	156.82	3.72
	心肺耐力 (800m 女)	310.26	12.24
差(3)	身體組成(身體質量指數)	14.76	0.70
	柔軟度(坐姿體前彎)	16.47	1.93
	腹肌力或肌耐力(仰臥起坐)	15.00	1.95
	瞬發力(立定跳遠)	139.71	2.29
	心肺耐力 (800m 女)	374.53	12.13
總和	身體組成(身體質量指數)	18.99	3.53
	柔軟度(坐姿體前彎)	30.74	11.98

	腹肌力或肌耐力 (仰臥起坐或一分鐘引體向上)	27.90	10.61
	瞬發力(立定跳遠)	156.27	14.09
	心肺耐力 (800m 女)	302.92	65.15

再將每一受試者的體適能的各項檢測指標與對應的群集別輸入於倒傳遞網路，經迭代學習後可建構一體適能評級模式。如下圖 4-2：

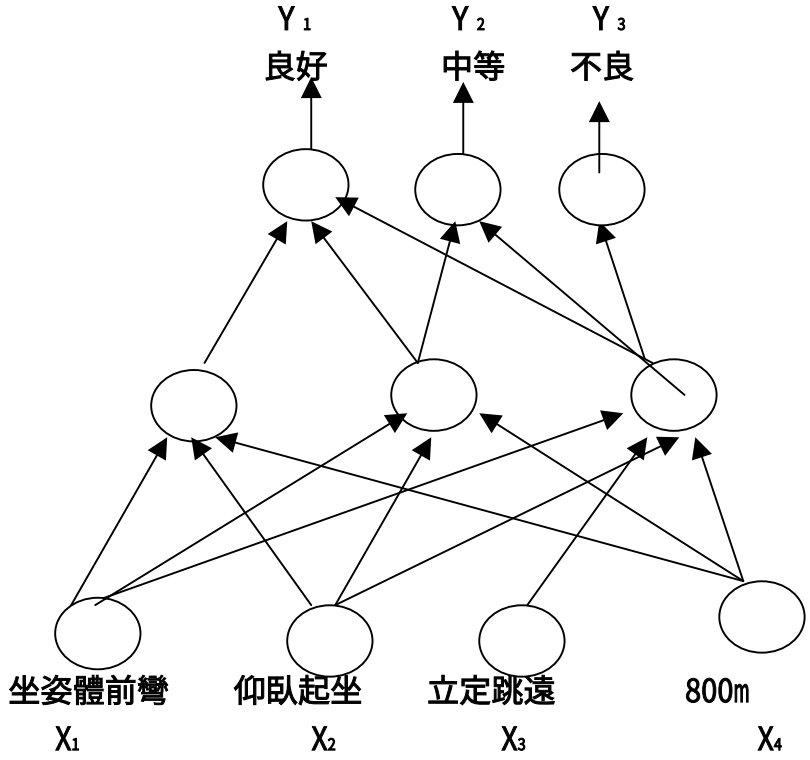

圖 4-2 倒傳遞網路的體適能評級模式

表 4-2 組群分類的正確性

組別			預測的各組成員			總和
			1	2	3	
原始的	個數	1	48	0	3	51
		2	11	117	6	134
		3	0	1	14	15
	%	1	94%	0%	6%	100%
		2	8%	87%	5%	100%
		3	0%	7%	93%	100%
a. 90%個原始組別觀察值已正確分類。						

在表 4-2 中，第 1 組別(良好)中的 51 位青少年，有 46 位被正確分類，3 位被分類到第 3 組別(差)。而第 2 組別(中等)中的 136 位青少年，有 117 位被分類正確，有 11 位被分類到第 1 組別(良好)，而有 8 位被分類到第 3 組別(差)。第 3 組別(差)中的 14 位青少年，有 14 位被分類正確，而有 1 位被分類到第 2 組別(中等)。

而在表 4-2 中的下半部百分比部分，此對角線呈現的數值是為正確率，例如：第 1 組別(良好)的正確率是 48/51=94%，第 2 組別(中等)的正確率是 117/134=87%，而第 3 組別(差)的正確率是 14/15=93%。所以整體的準確率是：179/200=90%。也就是說，有 200 位青少年，其中 179 位被正確分類，本模式具有高準確分類率。

第五章應用支援向量機建構體適能評級模式

5-1 前言

支援向量機(Support Vector Machines，簡稱 SVM)是一種以統計學習理論(statistical Learning Theory)為基礎，而發展出來的機器學習系統。支援向量機的應用領域相當多，例如：文字分類(test categorization)、影像識別(image recognition)、手寫數字辨識(Hand-written Digit Recognition)、資料探勘(Data Mining)、生物資訊(Bioinformatics)….等等。目前線性支援向量機(Linear Support vector Machines)已被証實具有高正確的分類能力。

本章主要目的使用群集分析及支援向量機建構一體適能評級模式。首先利用群集分析將受試者分群，再利用支援向量機對分群後的樣本建構一體適能評級模式。

5.2 線性支援向量機

1. 處理二類別的分類問題

線性支援向量機(Linear Support Vector Machines)如何處理可區分為二類的資料(Separable Data)。首先對每筆不同類的訓練資料(Training Data)加上標註：”+1”或是”-1”，以數學表示為$\{x_i, y_i\}$, $i=1,\cdots,l$, $y_i \in \{-1,1\}$, $x_i \in R^d$。假設有一個超平面可以將標註為”+1”和標註為”-1”之二類資料區分，則此超平面稱為區分平面(Separating Hyperplane)；落在此平面上的所有 x 必須滿足 $w.x+b=0$，w 為超平面之法向量(Normal Vector)。

定義區分平面之邊界(Margin)為 $d_+ + d_-$，$d_+(d_-)$為所有標註為”+1(-1)”的訓練資料和區分平面之最短距離。處理可區分為二類的資料時，支援向量機會找尋一個具有最大邊界的區分平面。此類型資料必須符合以下二個限制式：

$$x_i . w+b \geqq +1 \qquad \text{for } y_i=+1 \qquad (1)$$

$$x_i . w+b \leqq -1 \qquad \text{for } y_i=-1 \qquad (2)$$

可將(1)(2)二式結合為以下不等式：

$$y_i (x_i . w+b)-1 \geqq 0 \quad \forall i \qquad (3)$$

由(1)(2)可得知 $d_+=d_-=1/\|w\|$，所以邊界為 $2/\|w\|$。因此欲求得具有最大邊界的區分平面；可在符合限制式(3)的條件下，求 $\|w\|^2$ 的最小值；當符號成立時，x_i 稱為支持向量(Support Vector)。以維度為 2 的訓練資料為例，圖 5-1 表示所有的資料符合限制式(3)，⊙和◎為支持向量。

圖 5-1 支援向量機：處理可分為二類別的資料

在限制式為(3)的情況下，求$\|w\|^2$的最小值；將這個問題轉換為朗格朗吉(Lagrange)問題：

$$L_P \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^{l} \alpha_i \qquad (4)$$

其中：朗格朗吉係數(Lagrange multipliers) α_i , $i=1,\ldots,l$，對應到(3)式中的每一個不等式，且 $\alpha_i \geqq 0$。原本面對的問題變成求 L_P 的最小值，限制式為 $\alpha_i \geqq 0$。原本面對的問題變成求 L_P 的最小值，限制式仍為 $\alpha_i \geqq 0$ 的最佳化問題。但是此模式在使用非線性支援向量機時，仍不易求出最佳解；解決之道是找出其對偶問題(dual problem)。

由 $\frac{\partial}{\partial \omega} L_P = 0$，所以 $w = \sum_i \alpha_i y_i x_i$ (5)

由 $\frac{\partial}{\partial \omega} L_P=0$，所以 $w=\sum_i \alpha_i y_i=0$ (6)

將(5)(6)代入(4)式後，得到新的 L_P 給予其另一個符號 L_D 以避免混淆：

$$L_D=\sum_i \alpha_i-\frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j x_i \cdot x_j \quad (7)$$

原本求 L_P 的最小值問題，其對偶問題變成求 L_D 的最大值，限制式為(6)和 $\alpha_i \geqq 0$。

在求對偶問題最佳解時，每一個朗格朗吉係數 α_i 都對應到每一筆訓練資料；如果 $\alpha_i>0$ 表示該資料是此問題的支持向量，會落在區分平面的邊界上(如圖 5-1 所示)，將 α_i 代入(5)式可求得 w。為了求得 b，可以利用 Karush-kuhn-Tucker complementality conditions：

$$\alpha_i(y_i(w \cdot x_i+b)-1)=0 \quad \forall i \quad (8)$$

最後得到一個可以處理分類的問題的函數：

$$f(x)=sgn(\sum_{i=1}^{l} y_i \alpha_i \cdot (x \cdot x_i)+b) \quad (9)$$

當 $f(x)>0$ 時，表示該資料與標註為"+1"的資料屬於同一類；反之則屬於另一類。

2.處理不可區分為二類別的問題

線性支援向量機在處理不可區分(non-separable)為二類的資料時(如圖 5-2)，Cortes 等人提出在限制式中加入 Slack 變數 ξ_i，$i=1,\cdots,l$，原本的限制式變成：

$$x_i \cdot w+b \geqq +1-\xi_i \quad \text{for } y_i=+1 \quad (10)$$

$$x_i \cdot w+b \leqq -1+\xi_i \quad \text{for } y_i=+1 \quad (11)$$

$$\xi_i \geqq 0 \quad \forall i \quad (12)$$

由上式得知，當訓練資料在分類發生錯誤時，ξ_i 就會大於零。

因此在求區分平面的同時，$\sum i\ \xi_i$的值也希望愈小愈好。所以原本目標函數是求$\|w\|^2/2$的最小值，會變成求目標函數$\|w\|^2/2+C(\sum i\ \xi_i)$的最小值；成本(Cost)參數 C 由使用者自行決定，當 C 愈大時表示錯誤發生時對目標函數影響愈大，反之影響愈小。再利用前節所介紹的觀念，最佳化的問題可轉換成：

Maximize：

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \qquad (13)$$

Sub to ： $0 \leqq \alpha \leqq C,$ (13a)

$$\sum_i \alpha_i y_i = 0 \qquad (13b)$$

在處理不可區分的資料時，和上節不同之處在於朗格朗吉係數多了上限值 C。最後 α_i代入(5)式可求得 w，再利用 Karush-kuhn-Tucker complementarity conditions 求得 b。

圖 5-2 支援向量機：處理不可分為二類別的資料

5.3 非線性線性支援向量機

1.特徵空間

線性支援向量機用來處理區分為二類的資料時，是以一個線性函數來區分這二類不同的資料。但是資料可能無法用線性的函數完全區分開來，所以在 5.2 節加入成本參數 C 來控制錯誤。除了控制錯誤以外，使用非線性的函數來區分資料可以大幅減少錯誤的出現。以圖 5-2 為例，原本資料無法用線性的函數來區分；但是如果使用非線性的函數就可以將資料區分開來，如圖 5-3 所示，$g(x)=0$ 是非線性函數。

圖 5-3 非線性函數處理不可分為二類別的資料

根據 Boser 等人針對以非線性函數區分資料之研究，如果將原始資料透過一函數轉換到另一個較高維度的特徵空間(Feature Space)：

$$\Phi : R^d \rightarrow F \qquad (14)$$

原本不能以線性函數區分的資料，在高維度的特徵空間中較可能用

線性函數來區分不同類別的資料。在接下來的部分將用一個實例來說明特性。

實例：假設訓練資料有 6 筆，如圖 5-4 所示(白點表示標註+1，黑點表示標註-1)：

圖 5-4 空間維度為 2 的訓練資料

很明顯的，此 6 筆資料在維度為 2 的空間中，不可能用線性函數將二類資料區分；因此我們選用 $\Phi : R^2 \rightarrow R^3$ 轉換此 6 筆資料到維度為 3 的空間：

$$\Phi(x) = \begin{pmatrix} X^2_1 \\ \sqrt{2}X_1X_2 \\ X^2_2 \end{pmatrix} \qquad (15)$$

資料轉換到維度為 3 的特徵空間後(如表 5-1)，將可以用線性函數將

二類資料區分(如圖 5-5)。

表 5-1 訓練資料整理

資料維度為 2	資料維度為 3
$x_1=(1,1)\rightarrow y_1=+1$	$t_1=\Phi(x_1)=(1,\sqrt{2},1)\rightarrow y_1=+1$
$x_2=(-1,-1)\rightarrow y_2=+1$	$t_2=\Phi(x_2)=(1,\sqrt{2},1)\rightarrow y_2=+1$
$x_3=(1,-1)\rightarrow y_3=+1$	$t_3=\Phi(x_3)=(1,-\sqrt{2},1)\rightarrow y_3=+1$
$x_4=(-1,1)\rightarrow y_4=+1$	$t_4=\Phi(x_4)=(1,-\sqrt{2},1)\rightarrow y_4=+1$
$x_5=(0,0)\rightarrow y_5=-1$	$t_5=\Phi(x_5)=(0,0,0)\rightarrow y_5=-1$
$x_6=(1,0)\rightarrow y_6=-1$	$t_6=\Phi(x_6)=(1,0,0)\rightarrow y_6=-1$

圖 5-5 空間維度為 3 的訓練資料

2. Kernel 函數

Kernel 函數定義為：

$$k(x_i,x_j)=(\Phi(x_i)\ .\ \Phi(x_j)) \qquad (16)$$

由(13)式可知 x_i. x_j 會影響最後結果，如果將資料轉換到特徵空間中最後會影響結果則是 $\Phi(x_i)\cdot\Phi(x_j)$，並不需要知道 $\Phi(x_i)$或 $\Phi(x_j)$個別的值是多少。所以非線性支援向量機所處理的最佳化問題是：

Maximize：

$$L_D=\sum_i \alpha_i-\frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j k(x_i,x_j) \qquad (17)$$

Subject to ：

$$0\leqq\alpha\leqq C \qquad (17a)$$

$$\sum_i \alpha_i y_i=0 \qquad 17b)$$

如果 $k(x_i,x_j)$是半正定函(Symmetric Positive Definite Function)，則此 Kernel 函數會滿足 Mercer's Condition：

$$\iint k(x_i,x_j)g(x_i)g(x_j)dx_i dx_j>0, g\in L_2 \qquad (18)$$

若滿足 Mercer's Condition，(17)式則可以保證存在最佳解。

根據 1998 年 Gunn 整理許多其它常者所提出的 Kernel 函數，在

本節僅舉出其中最常見的三種：

I. Polynomial Kernel

$$k(x_i,x_j)=(x_i \cdot x_j+1)^d \qquad (19)$$

II. Multi-Layer Perceptron (MLP Kernel)

$$k(x_i,x_j)=\tanh(kx_i \cdot x_j-\delta) \qquad (20)$$

III. Radial Basis Function (RBF Kernel)

$$k(x_i,x_j)=\exp(-r\|x_i-x_j\|^2) \qquad (21)$$

在處理不同的問題的時候使用不同的 Kernel 函數，再配合不同的參數，會造成不同結果。

5-4、模式之建構

本章主要目的即在利用 SVM 建構一個適用於青少年的體適

能評級模式，以下為模式建構之步驟：

(1)選擇體適能的各項檢測指標

此項體適能指標的選擇是以教育部所公佈的體適能評鑑項目為依據，包含：身體組成(身體質量指數)、柔軟度(坐姿體前彎)、腹肌力與肌耐力(一分鐘引體向上)、瞬發力(立定跳遠)、心肺耐力 (800m 女) 等各項指標。

(2)各項體適能測驗成績與其體適能等級資料的收集

此項作業是收集參與測驗學生的各項檢測指標及體適能等級之資料。由於本研究屬實驗性質，樣本來源僅由國立台中技術學院九十一學年度五專部新生取得，吾人希望能再有更多的資料，如此才能達到更準確的分類。

首先從體育組取得新生體適能資料，剔除不完整及不足之資料。其次，將新生資料整理後，經群集分析形成三個等級(良好、中等、不良)，並分別將其量化成(1、2、3)三種等級。最後，再把所有量化資料整理後，分成模式建構的訓練資料和模式建構的測試資料。

(3)選擇合適的 SVM 軟體

本步驟是整個模式建構的重點，在此吾人選用 MYSVM 軟體對訓練樣本進行學習，以建構一項適合國內青少年的體適能評級模式。

(4) 評估模式的準確率

依據吾人所建構的體適能評估模式，將測試樣本資料輸入，以驗證此項評級模式的準確度。以下為體適能模式建構之流程：

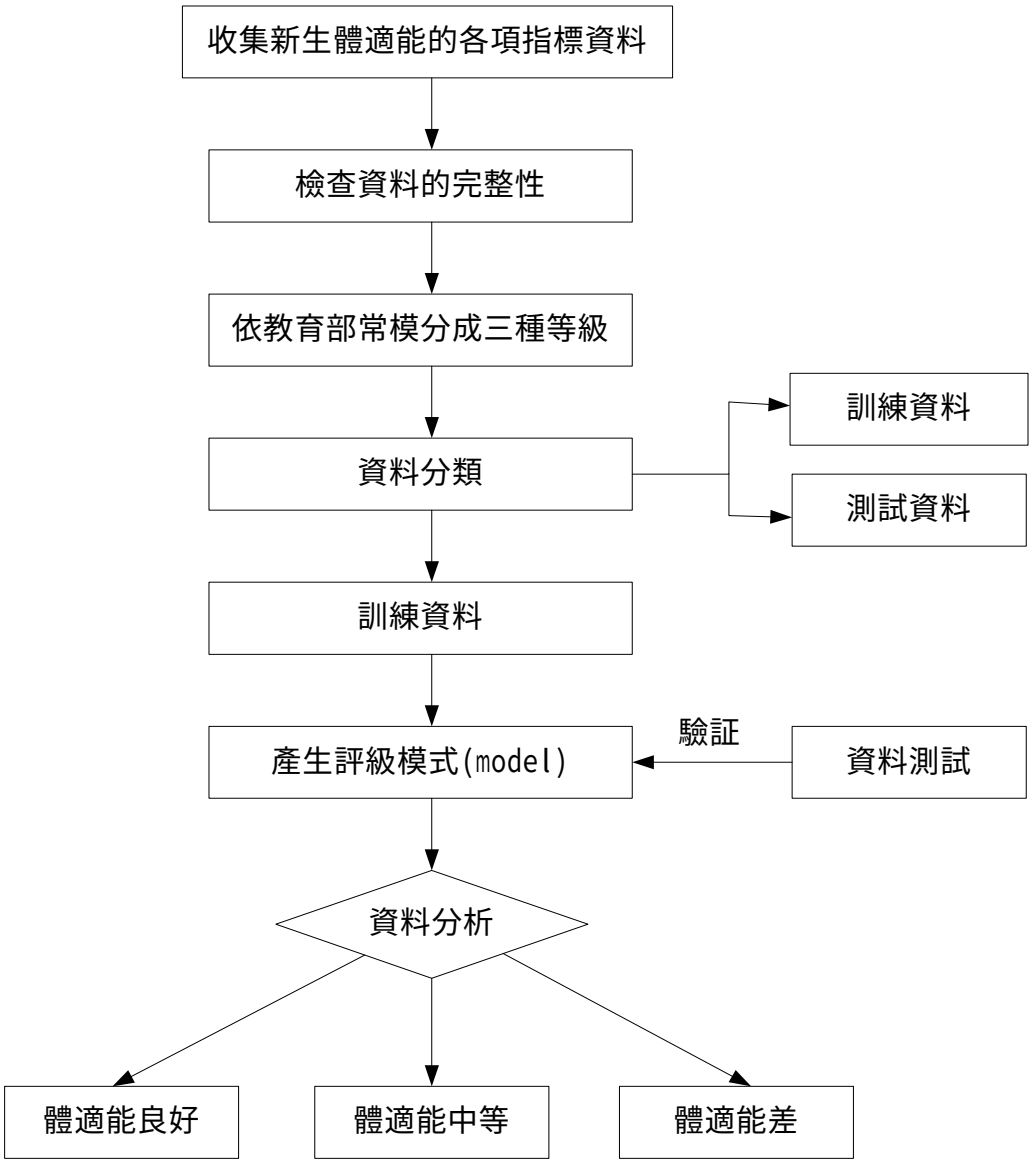

圖 5-6 模式建構之流程圖

5-5、實驗結果與分析

首先取得台中技術學院女生新生的體適能資料，將此樣本經群集分析的學習後，可得到三個群集，分別定義為良好、中等、不良。此三個群集的樣本敘述統計量如表 5-2。

表 5-2、群集統計量

組別		平均數	標準差
良好(1)	身體組成(身體質量指數)	24.14	0.92
	柔軟度(坐姿體前彎)	45.88	3.06
	腹肌力或肌耐力(仰臥起坐)	41.43	2.59
	瞬發力(立定跳遠)	174.17	3.59
	心肺耐力 (800m 女)	223.45	18.19
中等(2)	身體組成(身體質量指數)	19.35	0.92
	柔軟度(坐姿體前彎)	30.99	3.18
	腹肌力或肌耐力(仰臥起坐)	28.78	2.78
	瞬發力(立定跳遠)	157.26	3.68
	心肺耐力 (800m 女)	308.26	14.34
差(3)	身體組成(身體質量指數)	14.86	0.73
	柔軟度(坐姿體前彎)	17.37	1.85
	腹肌力或肌耐力(仰臥起坐)	15.65	1.84
	瞬發力(立定跳遠)	138.71	2.28
	心肺耐力 (800m 女)	373.65	12.04
總和	身體組成(身體質量指數)	18.89	3.39
	柔軟度(坐姿體前彎)	30.75	11.88
	腹肌力或肌耐力(仰臥起坐)	27.76	10.56
	瞬發力(立定跳遠)	156.27	14.09

	心肺耐力 (800m 女)	302.88	65.35

繼而利用 SVM 的套裝軟體 MYSVM 進行體適能評級模式的建構，其建構狀況如圖 5-7。圖 5-7 中的 Mysvm ended successfully 表示已建構完成體適能模式。

```
.....................................................................................
.....................................
***** Checking convergence for all variables
*** Convergence
Done training: 47719 iterations.
Target function: -51980.97
----------------------------------------
The results are valid with an epsilon of 0.00099247442 on the KKT conditions.
Average loss  : 0.1966315 (loo-estim: 69327.587)
Avg. loss pos : 0.15881811       (73 occurences)
Avg. loss neg : 0.23248056       (77 occurences)
Mean absolute error : 0.1966315
Mean squared error  : 0.068881853
Support Vectors : 150
Bounded SVs     : 142
min SV: -1000
max SV: 1000
|w| = 3.3962787
max |x| = 5.7279578
VCdim <= 379.44805
w[0] = -0.1238402
w[1] = 0.071032049
w[2] = -0.097197105
w[3] = -0.02676866
b = 27.131422
Time for learning:
init        : 0s
optimizer   : 0s
convergence : 0s
update ws   : 0s
calc ws     : 0s
=============
all         : 0s
Saving trained SVM to train.dat.svm
mysvm ended successfully.
```

圖 5-7 MSVM 執行畫面

(2)模式正確性的驗証分析

為驗証本研究所產生的體適能評級模式的正確性，實驗中所使用樣本資料共有 200 筆，其中 20 筆為體適能「良好」的新生，26 筆為體適能「中等」的新生，4 筆為體適能「差」的新生。將此測試資料輸入評級模式中進行正確性的測試。其正確性的分析如表 5-3：

表 5-3 組群分類的正確性

組別			預測的各組成員			總和
			1	2	3	
原始的	個數	1	52	2	0	54
		2	7	120	4	131
		3	0	0	15	15
	%	1	96%	4%	0%	100%
		2	5%	92%	3%	100%
		3	0%	0%	100%	100%
a. 93%個原始組別觀察值已正確分類。						

在表 5-3 中，第 1 組別(良好)中的 54 位青少年，有 52 位被正確分類，2 位被分類到第 2 組別(中等)。而第 2 組別(中等)中的 131 位青少年，有 120 位被分類正確，有 7 位被分類到第 1 組別(良好)，而有 4 位被分類到第 3 組別(差)。第 3 組別(差)中的 15 位青少年，有 15 位被分類正確。

而在表 5-3 中的下半部百分比部分，此對角線呈現的數值是為

正確率，例如：第 1 組別(良好)的正確率是 52/54=96 %，第 2 組別(中等)的正確率是 120/131=92%，而第 3 組別(差)的正確率是 14/15=93%。所以整體的準確率是：187/200=94%。也就是說，有 200 位青少年，其中 187 位被正確分類，本模式具有高準確分類率。

第六章結論

6-1 研究成果及限制

資料挖掘技術中的多變量統計方法，如：區別分析及多元迴歸模式為一般廣泛使用。但其使用上的缺點是：必須受制於母體為多元常態分配的假設前提。

類神經網路是一平行資料處理的技術。使用毋需任何統計分配的假定，且具備處理資料遺失或錯誤狀況之能力，同時可以認知變數間的關係及辨識內在規則，最重要一點是類神經網路具有學習的能力，可以隨時依據新的數據自我學習調適其內部儲存知識。缺點則在於其分析過程為一黑箱，故常無法以可讀之模型格式展現，每階段的加權與轉換亦不明確，是故類神經網路適用於資料屬於高度非線性且帶有相當程度的變數交互效應的情況。

支援向量機(Support Vector Machines，簡稱 SVM)則是一種以統計學習理論(statistical Learning Theory)為基礎，而發展出來的機器學習系統。目前正被廣泛討論，其優點是具高分類率，而其缺點是其數學演算過程非常複雜。

由於以上三種資料挖掘技術各有其優缺點，因此本文嘗試使用此三種技術建構體適能評級模式，並比較其分類正確率，以提供不同情況的使用。

綜合前三章的探討，可獲得以下幾項結論：

1. 在分類正確率方面，支援向量機所建構模式的分類正確率最高，依次為類神經網路，其次是多變量統計技術。
2. 多變量統計技術純屬統計理論，而類神經網路為機器學習的一種，而 SVM 的評級模式具有最高的分類正確率，究其原

因，可歸因於 SVM 結合了統計及機器學習理論。

3. 類神經網路模式的效益高於多變量技術模式，可歸因於類神經路模式較適於多變數，且具複雜的交互關係，而事實上體適能的各項資料未必滿足常態分配。
4. 神經網路模式的學習演算法計算容易，因此學習速度快(約節省 50%的時間)，且其可以經由學習累積知識資訊，實為一良好的分類工具。此外，Kohonen 網路可保留資料拓撲次序的能力，其被選作分類工具，可以將資料之間的關係以視覺方式展現於平面圖上，藉由平面圖的審視，可使得決策者獲取更多資訊，實為一良好的群集工具。

綜合文獻探討及以上的討論，本文主要貢獻如下：

1. 首度引入資料挖掘技術於體育領域，並將其應用於體適能模式之建構。
2. 兼用非監督式神經網路架構及監督式神經網路架構作為分群工具，使得在群聚過程中，可獲取更多資訊，此結果可提供體育領域其他主題的應用，例如：職棒觀眾的分群及發掘。
3. SVM 是一項非常新的分類技術，目前國內使用並不多，本文首度將其導入運動領域的分類評級。

綜合各章的實驗探究，吾人亦發現，在建構體適能評級模式中，亦有如下限制：

1. 本文中的體適能評級模式係使用實際的體適能資料，並未證實其為多元常態分配，因此使用多變量技術將可能造成模式的偏

誤。

2. 到目前為止，並無一正式理論可用以決定最適的神經網路隱藏層層數與節點數，每次學習均須經由實驗獲得，且類神經網路並無提供 How 及 Why 的能力。
3. 類神經網路所學習知識，係以編碼方式存於多層網路中的權值，屬於內在型式的決策規則，在解釋方面不若統計模式，需要更多的專家知識。
4. SVM 的研究屬於新的資訊技術，目前屬於研究階段，其效益是否確實，有待更多的實驗證實。

6.2 未來研究方向

本文主要探討以資料挖掘為基礎的體適能評級模式之建構，在探討的過程中，吾人亦發現以下幾項問題值得繼續深入探討，此亦為未來努力的方向。

1. 由於自組織特徵映射具有將高維度的輸入向量呈現於低維度空間的能力，可以將圖樣之間的關係以視覺方式展現，因此可用於運動管理上的領域，例如：職棒球迷的群集。
2. 基因演算法亦為一非常好的演算理論，若能導入體育領域，為一有趣的研究主題。
3. 實驗結果顯示以 SVM 為基礎的體適能評級模式的效益最佳，其意味著整合機器學習理論及統計技術是一項極具潛力的方向，此亦為未來的一研究課題。

參考文獻

[1] 邱義堂，2000，”通信資料庫之資料探勘：客戶流失預測之研究”，中山大學資訊管理系碩士論文。

[2] 張瑋倫，2000，”應用資料探勘學習方法探討顧客關係管理問題”，台北輔仁大學資訊管理學系。

[3] 張勳騰，1999，”資料探採在通信資料庫上目標行銷的應用”，國立中山大學。

[4] 黃彥文，1999，”資料探勘之應用-會員消費特徵之發掘”，屏東科技大學資訊管理系碩士論文。

[5] 廖雅郁，2001，”應用資料探採於我國西藥行銷之研究”，交通大學經營管理學系碩士論文。

[6] A.K. Jain, “ Data Clustering: a review,” ACM Computing Surveys, Vol.31, Issue 3, 1999, PP. 264-323.

[7] A.-O. Boudraa, “ Dynamic estimation of number of clusters in data sets,” Electronics Letters, Vol. 35, No. 19, 1999, pp. 1606-1608.

[8] B. Scholkopf, C.J.C. Burges, A.J. Smola, Introduction to Support Vector Learning, Advances in Kernel Methods-Support Vector Learning, pp. 1-15, Cambridge, MA, 1999, MIT Press.

[9] B.K. Wong T.A. Bodnovich, and Seliv, Y., “ Neural Network Applications in Business: A Review and Analysis of the Literature, “ Decision Support Systems, Vol. 19, 1997, pp. 301-320.

[10] C. Zopounidis, C., M. Doumpos, and N.F. Matsatsinis, “ On the use of Knowledge-based Decision Support Systems in Financial Management: A Survey,” Decision Support Systems, Vol. 20, 1997, pp. 259-277.

[11] C.B. Apte, B. Liu, E. Pednault and P. Smyth, “ Business applications of data mining,” Communications of the ACM, Vol. 45, No. 8, 2002, pp. 49-53.

[12] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001.

[13] C.-C. Chang, C.-W. Hsu. and C.-J. Lin. The analysis of decomposition methods for support vector machines. IEEE Transactions on Neural Network 11 (4), 1003-1008. 2000.

[14] C.J.C. Burges. A tutorial on support vector machines for pattern

recogition. Data Mining and Knowledge Discovery, 2(2):955-974, 1998.
[15] C.P. Rainsford and J.F. Roddick, " Database Issues in Knowledge Discovery and Data Mining, " Aust. J. Inf. Syst., Vol. 6, No. 2, In Press, 1999.
[16] C.W. Hsu and C.J. Lin. A Comparison of Methods for Multi-class Support Vector Machines. IEEE Transactions on Neural Networks, 13, 415-425, 2002.
[17] C.W. Hsu and C.J. Lin. A simple decomposition method for support vector machines. Machines Learning 46, 291-314, 2002b.
[18] Carven, M.W. and Shavlik, J.W., 1997, " Using neural network for data mining " , Future Generation Computer System, 13, 221-229.
[19] D. Pyle, Data preparation for data mining, Morgan Kaufmann publishers, Inc. San Francisco, California, 1999.
[20] F. Murtagh and A. Aussem, " Using the Wavelet Transform for Multivariate Data Analysis and Time Series Forecasting " , Proc. IFCS ' 96, Kobe, Springer-Verlag.
[21] Fu Y., 1997, " Data mining task, technique and applications " , IEEEE POTENTIALS.
[22] H. He and S. H. Chen, " Searching Financial Patterns with Self-Organizing Maps, " in processing of 3rd International Conference on Computer Vision, Pattern Recognition, and Image Processing, volume II, 2002, pp. 968-976.
[23] J.P. Bigus, " Data Mining with Neu ral Networks " , McGraw-Hill, 1996.
[24] Keerthi, S.S., E.G. Gilbert. Convergence of a generalized SMO algorithm for SVM classifier design. Machine Learning 46, 351-360, 2002.
[25] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data-An Introduction to Cluster Analysis, John Wiley &Sons, New York. 1990.
[26] N. Cristianini, J. Shawf-Taylor. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.
[27] N.R. Pal and J.C Bezdek, " On Cluster Validity for the Fuzzy c-Means Model, " IEEE Transactions on Fuzzy Systems, Vol. 3, No. 3, August 1995, pp. 370-379.

[28] S.-T. Li, “ A Web-aware Interoperable Data Mining System, ” Expert Systems with Applications, Vol. 22, pp. 135-146, 2002. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[29] T. Kohonen, 1997, Self-Organizing Maps, Springer, Germany.
[30] U.M. Fayyad, G.P. Shapiro, and P. Smyth, “ From data Mining to Knowledge Discovery: An Overview ” , 1996, pp. 1-36.
[31] Gary E. Larsen, James D. George, Jeffrey L. Alexander, Gilbert W. Fellingham, Steve G. Aldana, and Allen C. Parcell, “ Prediction of Maximum Oxygen Consumption From Walking, Jogging, or Running ” , Research Quarterly for Exercise and Sport c2002 by the Alliance for Health, Physical Education, Recreation and Dance Vol. 73,No. 1, pp. 66-72.