第三章　研究方法與步驟

本章主要目的在於敘述研究的設計與實施的方式，全章共分為六節：第一節、研究方法；第二節、研究架構與流程；第三節、研究對象；第四節、資料分析方法；第五節、實施程序；第六節、資料處理工具，分別敘述於後。

第一節 研究方法

本研究旨在運用資料探勘分類技術，建構台灣師大游泳會員流失區別模式，並從中了解流失顧客的重要特徵。經由文獻分析顯示，傳統技術之鑑別分析與改良技術中之 MARS、類神經網路中的 BMP 是較常被討論、應用的資料探勘分類方法。又因 BPN 有收斂緩慢、易落入局部最小值的缺失，研究中另提出整合 MARS 與 BPN 的分析模式，期能發展出一個更為精確、快速的分析模式。

因此，本研究擬以資料探勘中的鑑別分析、MARS、BPN、整合 MARS 與 BPN 等分類技術對台灣師大游泳會員資料庫進行分析探討，期能建構一精確有效之游泳會員流失分析模式，達到減少游泳會員流失之目的。

第二節 研究架構與流程

一、研究架構

依據研究目的、研究問題，並參酌相關之文獻內容，設計本研究之架構，共分為三個階段，說明於下：

(一)資料前置處理階段

前置處理作業包含對原始資料庫之重建，並做選取、轉換、淨化、合併、衍生等工作，並將其結果儲存於新的資訊貯藏空間，此貯藏空間即為資料倉儲。

(二)模式訓練階段

訓練與測試的樣本比例方面，為不使訓練資料筆數與測試資料筆數太過接近或太過極端而使最後結果產生偏誤，本研究採用陳麒文(民 91)之建議，以 80：20 的比例隨機抽出訓練模式的樣本及測試模式的樣本個數。因此，本研究將資料倉儲中之樣本隨機抽取 80%做為訓練樣本，並經由鑑別分析、MARS、BPN、整合 MARS 與 BPN 等方法，建立四種游泳會員流失區別模型。

(三)模式測試階段

將倉儲中剩餘 20%的樣本，輸入上階段所建立的模型中，以考驗各模型的整體判別率，並比較其精確度，以達研究之目的。

為讓架構全貌更為清晰，茲以圖 3-1 呈現本研究之架構。

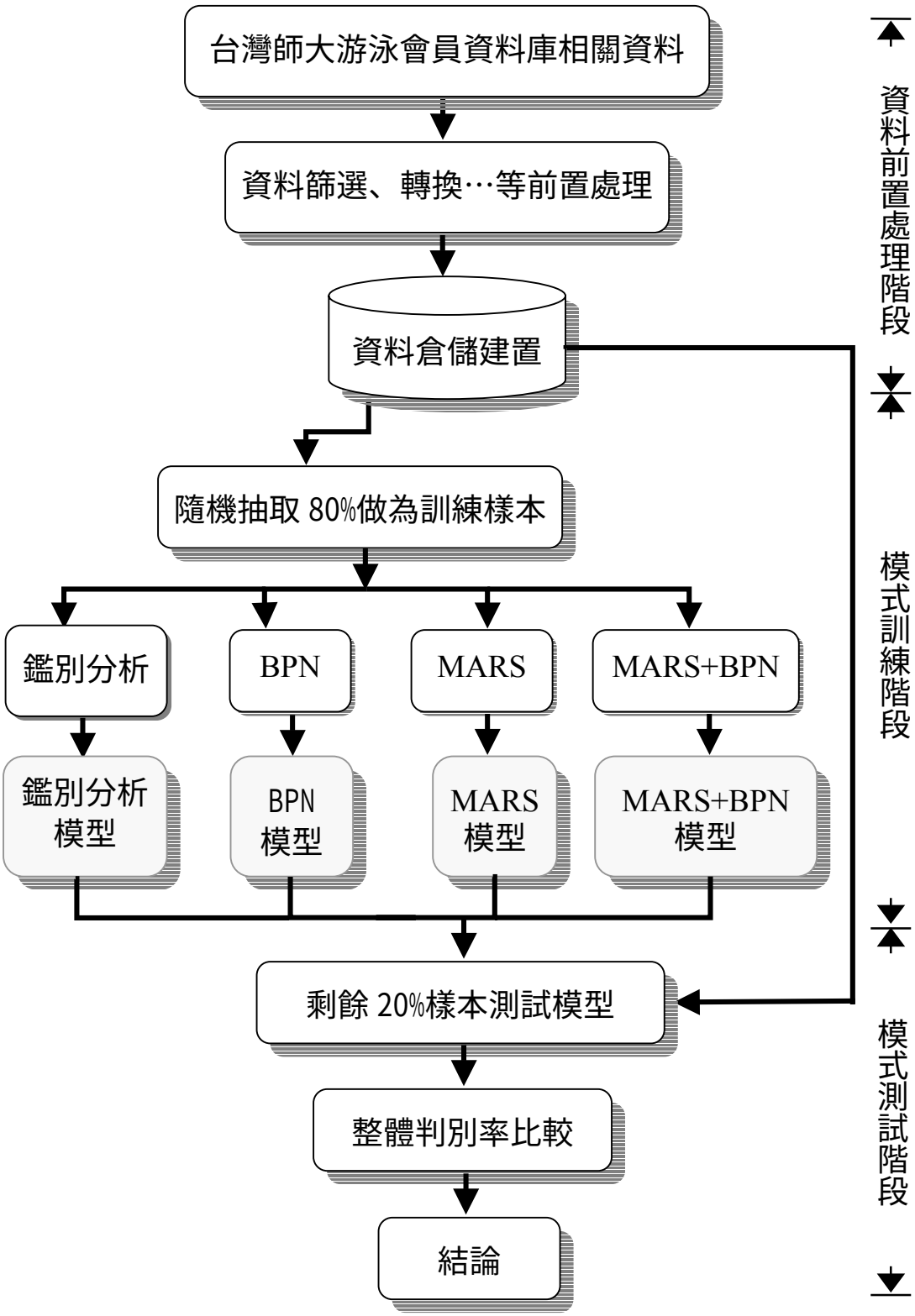

圖 3-1　研究架構圖

二、研究流程

本研究於研究主題確定後，即著手蒐集相關文獻，經過整理與探討後，擬定研究目的、研究問題等內容，並建立本研究之架構。在取得研究對象之資料後，加以整理彙總，並進行實證分析。最後，比較討論研究結果並提出具體建議。本研究流程如圖 3-2 所示。

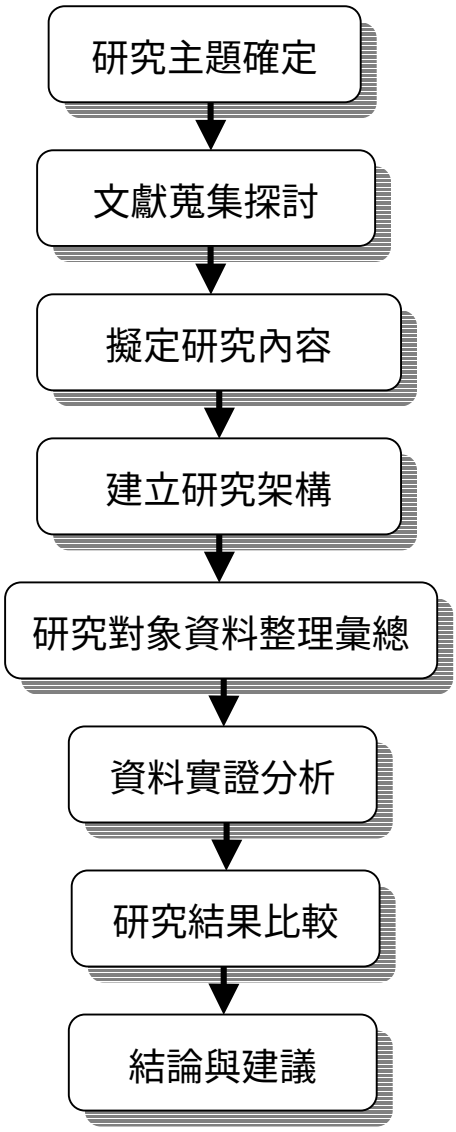

圖 3-2　研究流程圖

第三節 研究對象

本研究係以民國 90 年 5 月 1 日至民國 93 年 4 月 1 前近兩年間，師大本部游泳池所記錄之校外人士會員相關資料所建置之資料庫做為研究對象。

第四節 資料分析方法

本研究以鑑別分析、多元適應性雲形迴歸(MARS)、倒傳遞類神經網路(BPN)及整合 MARS 與 BPN 做為分析方法，茲分述於下：

一、鑑別分析

本研究實施鑑別分析的程序如下：

(一)名目變數的轉換

Jobson(1992)指出，從衡量的尺度差異，可以將變數分為兩種類型：1.量的變數(Quantitative Scale)；2.質的變數(Qualitative Scale)，也稱為類別變數或名目變數。由於質的變數有時用數字代表類別，其值表達的是種類的不同，並不表示彼此間有量的關係。因此，使用質的變數時，必須用虛擬(Dummy)變數表示。若變數值的類別多於二個以上時，可以用 1-of-(c-1)編碼或效果編碼(Effect Coding)(丁玉成，民 89)。所謂 1-of-(c-1)編碼，即是以 c-1 個虛擬變數來代表 c 個類別，以避免線性相依(Linear Dependency)的問題。

而效果編碼類似 1-of -(c-1)編碼，但是省略的值之虛擬變數全設為 -1。例如當表達色彩的預測變數值有三個顏色時，可以用二個虛擬變數表示，而其 1-of -(c-1)編碼與效果編碼如表 3-1。

表 3-1 質的變數編碼原則表

類 別	1-of-(c-1)編碼		效果編碼	
紅	1	0	1	1
藍	0	1	0	1
綠	0	0	-1	-1

資料來源：本研究整理

本研究採取 1-of-(c-1)編碼原則，將資料中的名目變數，重新編碼成為於鑑別分析可採用的虛擬變數。

(二)預測變數的篩選

進行鑑別分析線性函數時，預測變數的重要性是以標準化係數(Standardized Coefficients)絕對值大小來衡量。標準化係數即為函數的係數估計值，標準化係數的絕對值愈大表示預測變數的影響力愈大，該預測變數也愈重要。

SPSS 統計軟體提供了二種選擇預測變數的方法，這二種方法的選擇標準容忍度均為 0.001，分述如下：

1.強迫進入法(Enter independent together)：此方法是將所有符

合容忍度標準的自變數同時選進鑑別函數中。

2.逐步選取變數法(Stepwise)：依五種不同的變數選取原則將重要的鑑別因子保留在鑑別函數中。

本研究採取逐步選取變數法建立鑑別函數方程式，並以最小的 Wilks’ Lambda 值作為變數選擇的原則。

(三)鑑別函數適合度檢定

鑑別分析的適合度檢定，是以檢定鑑別函數的預測變數係數均為零當作虛無假說，若拒絕虛無假說則模式具有顯著性，鑑別函數的係數估計值是有意義的。檢定時以 Wilks’ Lambda 統計量與 F 分配做為檢定的依據。

(四)建構典型鑑別模式

若鑑別函數適合度檢定為顯著，即可建立典型鑑別模式，其形態如式 3-1 所示。

$$F = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (3\text{-}1)$$

其中 F 為判別分數；β_0 為函數常數項；β_i 為標準化係數；X_i 為預測變數。

(五)函數的鑑別力檢定

由典型鑑別函數的標準化係數可判斷各預測變數的影響程度大小，標準化係數的絕對值愈大，則影響力愈大。但典型鑑別函數僅可

觀察選入鑑別函數的預測變數觀察，並無法顯示模式的整體鑑別能力，透過典型相關係數的判讀可達到上述目的。

所謂典型相關是判別分數與組別間關聯程度的量數，即總變異中由組間變異可解釋的比例之開方根。本研究之依變數有二種類型，故典型相關函數只有一個。在一般論文中，典型相關係數的絕對值>0.3，則表示模式有別能力；若係數的絕對值>0.45 則表示該模式的鑑別能力強(張紹勳、張紹評、林秀娟，民 89)。

二、倒傳遞類神經網路(BPN)

由文獻討論中可得知，倒傳遞類神經網路為目前應用最普遍的類神經網路。又根據陳瑞龍(民 90)整理葉怡成等學者的建議指出，最適合「分類型」問題的網路模式的優先順序為：(一)倒傳遞類神經網路；(二)多層函數連結網路；(三)通用迴歸網路；(四)學習向量量化網路；(五)半徑式函數網路。因此，本研究選取 BPN 做為網路建構的模式。以下針對 BPN 的架構、演算過程及參數設定做說明。

(一)倒傳遞類神經網路架構

倒傳遞類神經網路架構，包括輸入層、隱藏層及輸出層(葉怡成，民 91)，如圖 3-3 所示，簡述於下。

1.輸入層：

是用來表現網路的輸入變數，而其處理單元數目依其所遭遇的問

題而定，一般而言，是欲研究問題的輸入變數。輸入層所使用的函數是線性轉換函數。

2.隱藏層：

是用來表現處理單元間的交互影響，其處理單元的數目沒有標準的方法可以決定。因此，須以試驗的方式才能決定出最佳的數目。而在層數方面，隱藏層可以是一層以上，也可以沒有隱藏層。隱藏層所使用的函數是非線性轉換函數。

3.輸出層：

用來代表網路的輸出變數，其處理單元的數目則必須要依照問題的特性來決定。一般而言，是欲研究問題的輸出變數，輸出層也是使用非線性的轉換函數。

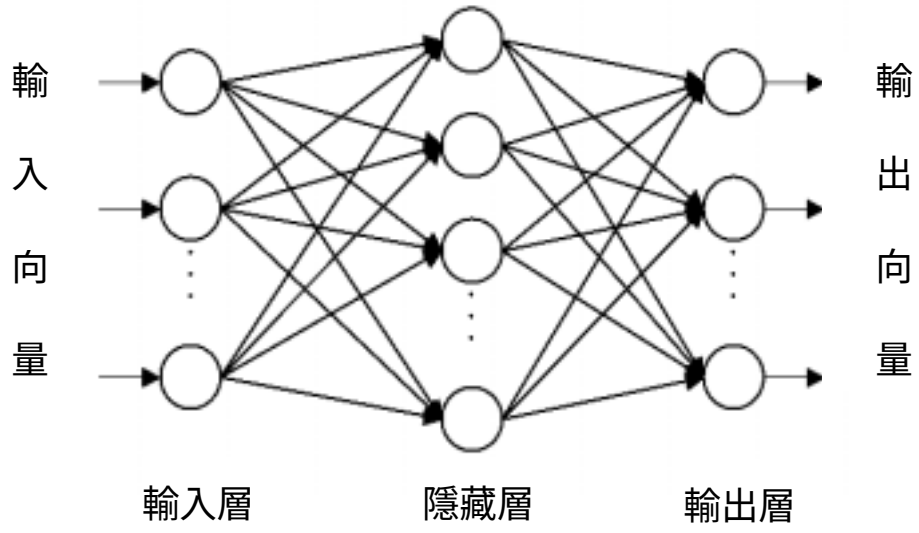

圖 3-3　倒傳遞類神經網路模式架構

資料來源：葉怡成(民 91)：*類神經網路模式應用與實作*。台北：儒林圖書有限公司。

(二)網路演算過程

葉怡成(民 91)指出，倒傳遞網路的演算過程主要可以分為兩個階段，分別是學習(Training)過程與回想(Recalling)過程。簡述於下。

1.學習過程

即網路的訓練過程，主要是針對權重值的更新與調整。當網路的學習誤差低於容許誤差時，網路即完成學習過程，也就是網路模式訓練完畢。於此過程中，網路會反覆執行順向傳遞(Forward Pass)與逆向傳遞(Reverse Pass)兩個步驟。網路一開始先給一些隨機數值當作啟始權值及閥值；在順向傳遞時，先一層一層向前傳遞並計算各處理單元的輸出值，直到輸出層；在逆向傳遞時，計算輸出層處理單元的結果與目標的差異，依據誤差法則倒傳遞回整個網路，以更新每個權值及閥值。上述步驟會反覆執行，直至誤差收斂至預設的條件為止，最後得到一個最終的預測模式。

2.回想過程

即網路的測試過程，網路訓練完畢之後，緊接著進行網路測試。此階段是以不改變網路中已經訓練完成的各個權重值，輸入測試樣本，再經由層層的權值加權、總和計算、函數轉換等的處理，最後可以由輸出層處得到輸出值。網路的表現成效則可藉由輸出值與期望輸

出值的比較來得知。

(三)網路參數設定：

Davies (1994)認為沒有絕對的法則可以決定網路中應包含的參數，唯有透過試誤法(Trial and Error)才能找到相對較佳的結果。雖然如此，相關參數的選擇仍有許多基本原則可以在建立網路模式時運用，敘述於下。

1.輸入層與輸出層處理單元數

在輸入層處理單元數的決定上是以輸入變數為根據，其數目就是輸入變數的數目。至於在輸出層處理單元數方面，由於本研究是以區分游泳會員流失與否為主要目標，因此輸出層只有一個處理單元。

2.隱藏層層數及處理單元數目

在隱藏層層數方面，根據葉怡成(民 91)的看法，一般問題使用隱藏層數一或二時可得到最佳結果。而 Zhang 等人(1998)則提出具有單一層隱藏層的神經網路就能達到所需要的精確度。故本研究將隱藏層層數設定為一層。

在隱藏層處理單元數目方面，並沒有明確的理論說明如何決定，必須經過一次又一次的測驗，以決定最佳的處理單元個數，來達到最佳的收斂效果。以下彙整周慶華(民 90)、葉怡成(民 91)等學者，於文獻中計算隱藏層單元數之公式。

(1) (輸入單元數 ＋ 輸出單元數)/ 2

(2) (輸入單元數 ｘ 輸出單元數)/ 2

(3) (輸入單元數 ＋ 1)ｘ 2

(4) (輸入單元數 ＋ 輸出單元數)ｘ 2

(5) (輸入單元數 ＋ 輸出單元數)/ 2 ＋(訓練範例)/ 2

(6) (輸入單元數 ＋ 輸出單元數)

(7) (2 ｘ 輸入單元數)- 2 至(2 ｘ 輸入單元數)＋ 2

因第七種方法，測試次數較多，故本研究將隱藏層處理單元數目設定為(2 ｘ 輸入單元數)- 2 至(2 ｘ輸入單元數)＋2。

3.起始權數

依據一般研究中的處理方式，網路起始權數是以均勻分佈的隨機亂數值來加以設定，或是以 0 來加以設定。這些權數可使用不同的機率分配形式(最常用的是均勻分配及常態分配)，而數值的範圍不宜過大，Hush 等人(1992)的研究證實，較小的初始權數可獲得較佳之學習效果。施柏屹(民 89)指出，為獲得較佳之收斂結果，一般可嘗試多組起始權數值。若收斂結果相差不多，則取收斂結果最佳的那組權數，若收斂結果相差很大，表示尚未找到最佳的起始權數，此時應再嘗試其它不同之起始權數。本研究設定網路的初始值是由軟體以均勻分佈的隨機亂數值來加以設定。

4.學習速率

學習速率又稱為步距(Step size)，通常學習速率太大或太小對網路的收斂性質均不利。葉怡成(民 91)指出，依據經驗及軟體特性取 0 到 1.0 間的值作為學習速率的值，大都可得到良好的收斂性。

在本研究中，參照使用相同分析軟體之文獻(許俊源，民 90；陳麒文，民 91；黃正鳳，民 91；陳靜怡)，將學習速率以 0.002、0.004、0.006、0.008、0.010 等多組數值設定之。

5.網路學習終止條件

倒傳遞網路之學習是依據訓練樣本來調整網路之連結權數，而網路學習的停止條件，可以梯度法(Gradient)、誤差均方根(RMS)、學習次數、交互驗證法(Cross-validation)等四種方式設定中止條件(Hush et al., 1992)。分述於下：

(1)梯度法：倒傳遞網路的學習是向最大斜率方向改變，當梯度不變時，表示斜率亦不改變，此時權數不會再改變，網路便可以終止學習。

(2)誤差均方根:若網路之 RMS 誤差值小於某特定收斂值時，表示網路已達某一特定收斂程度，則可停止學習。

(3)學習次數：當網路完成所設定之學習次數後，即自行停止網路之學習過程。

(4)交互驗證法:將數據分為訓練及測試兩組樣本，一組訓練一組測試，當訓練與測試誤差水準同時小於設定值時就可以停止學習。若網路訓練好而測試不好，表示過度學習，若訓練不好而測試好，表示學習不足。

為避免網路過度學習或學習不足，本研究參照使用相同分析軟體之文獻(陳麒文，民 91)，以誤差均方根小於 0.0001 與學習次數小於 10,000 次兩種終止網路學習之設定方式。

6.轉換公式

葉怡成(民 91)指出，在轉換的過程中，倒傳遞類神經網路模式最常用的轉換函數是雙彎曲函數(Sigmoid Function)，其型式如式 3-2 所示。

$$f(x)=1/(1+e^{-X}) \qquad (3\text{-}2)$$

而最常用的訓練過程則是最陡坡降法(Gradient Steepest Descent Method)，它是用來調整權數變動的幅度。在本研究所使用的軟體中，對於相關參數之設定提供了簡便的設定功能，故由軟體中所提供的功能來加以設定之。

三、多元適應性雲形迴歸(MARS)

MARS 是一個新興的多變量無母數迴歸程序技術，藉由數個線段基本函數(Basic Funtion, BF)的累加模型組合出一個較具彈性的模

式，用以解釋非線性狀態的工具(Friedman, 1991)。其通用模型(General Model)如式 3-3 所示。而後段累乘的部分即是 BF, 如式 3-4 所示。

$$\hat{f}(x) = a_0 + \sum^{M} a_m \prod^{k_m} \left[s_{km} \cdot \left(x_{v(k,m)} - t_m \right) \right] \qquad (3\text{-}3)$$

$$\sum^{M} a_m \prod^{k_m} \left[s_{km} \cdot \left(x_{v(k,m)} - t_m \right) \right] \qquad (3\text{-}4)$$

其中

a_0 與 a_m 皆為參數值，功能類似迴歸係數。

M 為 BF 的個數，經由評估準則決定。

k_m 切割的折點數，即該 BF 中預測變數的個數。

$v(k,m)$ 是對預測變數的標示。

t_m 為柵欄值(Threshold Value)。

s_{km} 之值限定為＋1 或-1，其作用為顯示方向。

在 MARS 2.0 應用軟體中的 Basis Function 是以 $Max(0, X-C)$ 或 $Max(0, C-X)$ 的形式表現，因此，當 $s_{km}=1$ 時，模型內之中括弧即為 $Max(0, x_{v(k,m)} - t_m)$；當 $s_{km}= -1$ 時，模型內之中括弧即為 $Max(0, t_m - x_{v(k,m)})$，由此特性，兩個互相對應的 BF 在變數折點的上、下方產生不同的斜率，使模型結構貼近非線性資料型態。

建立 MARS 的最佳模型須執行以下二步驟：第一個步驟使用前推

式(Forward)演算法將所有的 BF 放入模型中，讓所有的 BF 做加總，產生過度配適模型(Overfit Model)；第二步驟則是使用後推式(Backward)演算法修剪不適合的 BF，BF 的修剪主要是評估各個 BF 在過度配適模型中的損適性(Loss Of Fit, LOF)後，再一一剔除掉對模型貢獻度最少的 BF，直到找到一個偏誤與變化達到最適當(Optimal)的模型為止。而 BF 之貢獻度則是依據 Spline 的研究先趨 Craven 與 Wahba(1979)所提出的 GCV 值(Generalized Cross Validation)作為判斷準則，其原則是判斷當一變數由模型中去除時，若 GCV 值顯著降低時，則可知此變數為重要的關鍵要素。GCV 的計算公式如式 3-5 所示。

$$LOF\left(\hat{f}_M\right) = GCV(M) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}_M(x_i)\right]^2 \Big/ \left[1-\frac{C(M)}{N}\right]^2 \qquad (3\text{-}5)$$

綜合上述，本研究依據 LOF-GCV 值的評估準則，篩選出顯著的 BF，並根據包含在其中的變數重要性，來決定所需的變數個數，以得到較具貢獻的輸入變數，並依此建立 MARS 區別模型。

四、整合 MARS 與 BPN 之分析技術

由於 BPN 無法由眾多的變數中挑選出重要影響的變數，且有學習時間冗長、收斂緩慢、易落入局部最小值的缺失。有鑑於此，本研究將 MARS 與 BPN 兩階段模式之建構程序做一整合，期能夠發展出一個

更快速、精確度更高的區別模型。整合模型的主要流程為：一、先經由 MARS 進行變數的篩選，將其篩選出來的重要影響變數做為 BPN 的輸入資訊；二、再透過 BPN 的學習、辨識能力，發展出一個整合的區別模式。

第五節 實施程序

本研究實施程序包括資料取得、前置處理與探勘等三個步驟，分述如下：

一、資料取得

經台灣師大游泳池管理單位同意後，於民國 93 年 3 月取得該校游泳會員之相關資料。

二、資料庫建置與資料前置處理

於資料取得後，進行資料庫建置與資料選取、轉換、合併、衍生...等前置處理，並將處理後的資料建製成資料倉儲，以供探勘之用。

三、資料探勘分析

於資料倉儲建置完成後，進行資料探勘分析作業，最後整理分析結果並做成結論與建議。

第六節 資料處理工具

一、分析軟體方面

本研究所使用之資料分析軟體，係與輔仁大學管理學研究所李天行教授商請借用，經同意授權後，使進行資料分析，茲將各分析軟體羅列於後。

(一)鑑別分析

使用 SPSS 公司所出版的 SPSS for Windows 10.0 版統計套裝軟體(SPSS Inc.,2001)。

(二)類神經網路

採用 Vesta 公司出版的 Qnet 97 軟體(Vesta Service,1998)。

(三)多元適應性雲形迴歸

採用 Salford Systems 公司所出版的 MARS 2.0 版分析軟體(Salford Systems, 2000)。

二、作業環境方面

CPU 使用 P-IV 1600MHz 處理器，搭配 512MB 記憶體、80GB 硬碟、Windows XP Professional 作業系統。