管 理 與 系 統
第十三卷 第二期 民國九十五年四月
221-240頁

Journal of Management & Systems
Vol. 13, No. 2, April 2006
pp. 221-240

利用混合整數規劃處理多類別分類

A Multiple Group Classification Method by Mixed Integer Programming

余菁蓉 Jing-Rung Yu 陳家豪 Chia-Hao Chen

國立暨南國際大學資訊管理學系

Depart of Information Managememt, National Chi Nan University

(Received July 26, 2005; Final Version December 16, 2005)

摘要：本研究提出延伸 Sueyoshi 兩階段混合整數規劃，利用樹狀圖逐步分割概念提出可分多類別的方法；並加入多變量主成份分析做資料的前處理，提升分類能力。兩階段混合整數規劃的分類方法，主要用於兩類別資料判別，其優點有二：一是利用兩階段的概念對重疊區域作分類，可有效降低誤判率；另一則是用較少的二元變數 (binary variables) 處理分類問題，降低運算時的複雜度，也因此會比一般的分類方法來得有效能，但此法不能用於多組資料分類。故本研究以兩階段混合整數規劃爲基礎，利用樹狀圖來做逐步分割，藉由兩兩類別間的中心點距離求出最佳分割順序樹狀圖來分多類別資料，同時減少誤判的發生；並藉由主成份分析對原始變數作前處理，使其轉換後的主成份變數間具有相互獨立的特性，進而提升分類時的正確率。另外，由於支持向量機是當前相當受到歡迎的分類方法，不僅可以分類多類別的資料，且在大量樣本上有良好的判別能力，因此，最後以兩個範例來做比較，結果顯示本研究所提出的多類別分類方法比支持向量機及統計上的判別分析法，更適用在小樣本上，驗證本研究的方法在小樣本上確實比支持向量機有較高的效能及可用性，且與支持向量機有相輔的特性。

關鍵詞：混合整數規劃、樹、判別分析、主成份分析、支持向量機

Abstract：This paper proposes a multiple group classification method which adopts principal components analysis as the data preprocessing and then extends Sueyoshi's two-stage mixed integer

programming by using the tree concept to enhance the discriminating capability. The two-stage mixed integer programming which is usually applied to two-group classification has two main advantages: (i) It deals with overlap area by using the two-stage approach, thus it is a more effective method for reducing the misjudgments; (ii) To reduce the complexity, it uses less binary variables than other mixed integer programming methods. However, the two-stage mixed integer programming cannot deal with multiple group discrimination. In order to overcome this problem, a mixed integer programming with a tree concept and principal components analysis is proposed. A tree is generated according to the center distance of each pair groups. Then the original variables are transformed into new ones by principal components analysis, which makes the new variables independent, classifies easily and enhances the hit rate. The proposed method is compared with support vector machine (SVM), a popular classification method in large sample size, and statistical discriminant analysis by using two examples. The proposed method outperforms SVM and statistical discriminant analysis. Our approach can be a good alternative method of SVM especially in handling small sample size.

Keywords **:** Mixed integer programming, Tree, Discriminant analysis, Principal components analysis, Support vector machine

1. 緒論

分類問題一直是學者們所致力研究的一門領域，尤其在資料採礦的領域不斷地有新的分類方法被提出，分類方法最早的文獻可追朔至 Fisher (1936) 提出的用線性規劃的方法判別鳶尾花的種類，此外，Kendal *et al*. (1983) 及 Mclachlan (1992) 也將相關的分類方法作彙整。然而在傳統上，此類的方法通常得先假設分類的資料須符合常態分配，但在現實的情況下能滿足此一條件的資料卻不多，也因此造成此類方法的限制。

爲了解決前述方法中的限制條件，Charnes *et al*. (1955) 提出用目標規劃的方法建立一分類模型，並利用線性演算法的方式處理分類問題，其優點有二：一是不需假設分類的資料須符合常態分配，因此可以應用在更多的分類問題上；二是利用最佳化的方式求解，可以得到較佳的分類正確率。但直到 Freed and Glover (1981) 提出了建立在數學規劃上的分類方法後，才開始有許多學者致力於此方面的研究，並提出眾多利用目標規劃來分類的方法。目前多數的分類方法皆以分兩類別資料爲主，三類別以上的分類方法大多是以分兩類別爲基礎再加以延伸，當然亦有直接分多類別的方法，不過在眾多的分類方法中，雖然每個方法都仍有其獨特性，也都有相當不錯的分類能力，但卻也有相對的弱點，如 Glen (2003) 提出的疊代混合整數規劃雖然有不錯的分類能力，但卻無法處理多類別的分類；另外，Loucopoulos (2001) 提出利用不同誤判成本的數學規劃方法

分三類別資料，方法雖有獨特之處，卻因限制式中太多的二元變數造成運算上的複雜，導致耗費了過多的時間及成本；而 Gochet (1997) 所提出的判別方法，雖然藉由向量的方式區分多類別，但卻因限制式過於複雜，也導致計算時需耗費大量的時間成本；而 Sueyoshi (1999; 2001; 2004) 的兩階段混合整數規劃與其他方法比較起來，不僅在處理重疊區域內對難分的觀察樣本有良好的分類能力，且減少了二元變數過多的缺點，唯獨目前該方法尚無法處理多類別的分類問題。爲了解決 Sueyoshi (2004) 的兩階段混合整數規劃無法分多類別的問題，本研究提出利用樹狀圖做逐步分割的基礎來處理多類別分類，並加入主成份分析作資料前處理，轉換原使始變數成獨立的變數，來提升分類時的正確率，使本研究的方法有更高的可用性。

本研究方法以 Sueyoshi 的兩階段混合整數規劃爲基礎，藉由樹狀圖做逐步分割的概念對多類別資料建立一分割順序樹狀圖，延伸兩階段法爲可分多類別的方法。此外，本研究在提升分類效能的同時，發現分類資料的變數間的相關性會對分類結果產生相當程度的影響，變數間的相關程度越低，分類的正確率會越高；反之，相關程度越高，分類的正確率則越低。所以本研究又加入了主成份分析轉換原始變數，使得轉換後的主成份變數之間彼此具有獨立的特性，並提高分類時的判別率。故本研究所提出的多類別分類方法特色有二：一是利用樹狀圖做逐步分割的概念，逐步分割出距離相對較遠的類別，並依分割順序樹狀圖分類，減少誤判發生的機率，進而提升分類能力；二是對原始變數做資料的前處理，經由主成份分析轉換原始變數成彼此獨立的主成份變數，避免因變數間的相關性影響分類結果。本研究將兩階段混合整數規劃結合了樹狀圖做逐步分割的概念及主成份分析，不僅可分多類別的資料，並可減少誤判率，故在分類正確率上更能得到令人滿意的結果。此外，由於支持向量機是近幾年相當受到歡迎的分類方法，同樣也是以目標規劃爲基礎而衍生出來的分類方法，加上其分類正確率是相當不錯的，故本研究將以支持向量機爲比較的方法，並驗證本研究所提出的多類別分類方法是非常有效能及可用性。

1.1 支持向量機

支持向量機是近幾年相當受到歡迎的分類方法 (Cortez, 1995; Boser *et al*., 1992; Kim *et al*., 2003)，主要概念是利用在特徵空間中建立一分割超平面 (Separable hyperplane)，分類低維度空間中無法用線性去區隔的類別，並經由二次規劃 (Quadratic programming) 的方式，最佳化相對觀察值間的邊際距離，所以分類能力相對地較爲顯著。支持向量機是以學習演算法先對資料作訓練，在分類之前，必須先定義訓練的資料，x_i 是變數值，z_i 是相對應的類別，通常以 +1、−1 分別表示：

$$(x_1, z_1), (x_2, z_2), (x_3, z_3), ..., (x_k, z_k), x_i, i = 1, ..., k,\ x_i \in R^n, z \in \{+1, -1\}$$

目標式是要找出一最大邊際的分割超平面 $w \cdot x + b = 0$，w 爲超平面之法向量 (normal

vector) ，定義區分平面之邊界爲 (*d*+)、(*d*-) 爲訓練資料與區分超平面的最短距離，在處理資料時支持向量機會尋找一具有最大邊際的分割超平面，而資料需符合以下限制式：

$$x_i \cdot w + b \geq +1 \quad \text{for} \quad z_i = +1 \tag{1}$$

$$x_i \cdot w + b \leq -1 \quad \text{for} \quad z_i = -1 \tag{2}$$

若最大邊際爲 (*d*+)、(*d*-)，則由限制式 (1)、(2) 可得 (*d*+)、(*d*-)及$1 / \|w\|$ 相等，所以最大邊界爲 $2 / \|w\|$。故欲求最大邊際的分割超平面，可在符合限制式的條件下，求 $\|w\|^2$ 的最小值，而當等號成立時，x_i 則稱爲支持向量，以下是支持向量機的公式：

$$\begin{aligned} &\min_{w,b,\zeta} \text{imize} \quad \frac{1}{2}\|w\|^2 + C(\sum_{i=1}^{l} \zeta_i) \\ &\text{subject to} \quad z_i(x_i \cdot w + b) - (1 - \zeta_i) \geq 0,\ \zeta_i \geq 0,\ \forall i \end{aligned} \tag{3}$$

公式 (3) 中，$C\sum_{i=1}^{l}\zeta_i$ 是訓練資料時所允許的誤差，若 C 的值夠大，且資料可以被線性分割，則公式 (3) 中 ζ_i 的值會趨近於零。

由於支持向量機有相當不錯的分類能力，故已漸漸地被廣泛應用，如破產預測 (Shin *et al*., 2005)、財務時間序列預測 (Tay and Cao, 2001)、紋理分類(Kim *et al*., 2003)、文字識別 (Li *et al*., 2003)等等，所以支持向量機在近幾年越來越受到矚目，因此本研究才會以支持向量機做爲比較的分類方法。

2. 兩階段混合整數規劃及主成份分析

2.1 兩階段混合整數規劃

Sueyoshi (1999) 先提出以目標規劃的觀點所建立的分類方法，後來Sueyoshi (2001; 2004) 將舊方法改良爲可處理混合整數的分類方法。此方法最大的優點在於利用兩階段的概念來分類，在階段一先整體分類並界定重疊區域，並於階段二針對重疊區域內的觀察樣本作更詳細的分類以減少誤判，有效地提升其分類能力。另外，若在階段一中已能將所有的觀察樣本正確地分類出來，則就不需要再藉由階段二去處理重疊區域內的觀察樣本，提高分類時的效率。此外，兩階段混合整數規劃同時減少了限制式中二元變數的數量以降低運算上的複雜，更可提升分類時的效率，其方法的分類步驟如下：

階段一　分類及界定重疊區域 (Classification and overlap identification, COI)

$$
\begin{aligned}
&\text{minimize} \quad s \\
&\text{subject to} \quad \sum_{i=1}^{k}(\lambda_i^+ - \lambda_i^-)x_{ij} - d + s \geq 0, \quad j \in G_1 \\
&\qquad \sum_{i=1}^{k}(\lambda_i^+ - \lambda_i^-)x_{ij} - d - s \leq 0, \quad j \in G_2 \\
&\qquad \sum_{i=1}^{k}(\lambda_i^+ + \lambda_i^-) = 1 \\
&\qquad \sum_{i=1}^{k}(\zeta_i^+ + \zeta_i^-) = k \quad \textit{i.e.}\ \zeta_i^+ + \zeta_i^- = 1 \ \forall\, i \\
&\qquad \zeta_i^+ \geq \lambda_i^+ \geq \varepsilon\zeta_i^+ \\
&\qquad \zeta_i^- \geq \lambda_i^- \geq \varepsilon\zeta_i^- \\
&\qquad \zeta_i^+ + \zeta_i^- \leq 1 \ (i = 1, ..., k) \\
&\qquad \zeta_i^+, \zeta_i^- \in \{0, 1\}
\end{aligned}
\tag{4}
$$

上式中，s、d 的值並無限制，目的是要界定出重疊區域的範圍，G_1、G_2 分別是觀察樣本 j 所在的類別，i 是觀察樣本的變數項指標共 k 個，x_{ij} 則是觀察樣本的值，ζ_i^+、ζ_i^- 都是二元變數 (binary variables)，是 λ_i^+、λ_i^- 的上下界，而其他的變數值都大於零。

若結果中 s 的值小於零，則無重疊區域的存在；若 s 的值大於零，則表示存在一重疊區域。接著將兩類別內觀察樣本利用判別式 (5) 分類：

$$
\begin{aligned}
C_1 &= \left\{ j \in G_1 \mid \sum_{i=1}^{k} \lambda_i^* x_{ij} > d^* + s^* \right\} \\
C_2 &= \left\{ j \in G_2 \mid \sum_{i=1}^{k} \lambda_i^* x_{ij} < d^* - s^* \right\} \\
D_1 &= G_1 - C_1 \ \text{ and } \ D_2 = G_2 - C_2
\end{aligned}
\tag{5}
$$

D_1、D_2 即爲界定的重疊區域範圍，並將範圍內的觀察樣本留至階段二再分類。在階段一的分類若以圖形表示則如圖 1，利用線段一及線段二界定出的三個區域。線段一的上方的觀察樣本屬於判別式 (5) 中的 C_1；線段二的下方的觀察樣本則屬於 C_2；而兩線段包圍的區域便是重疊區域，並將其區域內的觀察樣本於階段二再次分類。

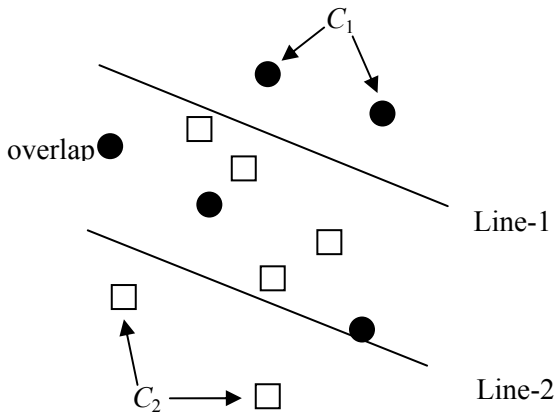

圖 1 階段一分類

階段二 處理重疊區域 (Handling overlap, HO)

$$
\begin{aligned}
\text{minimize} \quad & \sum_{j\in D_1} y_j + \sum_{j\in D_2} y_j \\
\text{subject to} \quad & \sum_{i=1}^{k}(\lambda_i^+ + \lambda_i^-)\, x_{ij} - c + \mathrm{M}y_j \geq 0, \quad j \in D_1 \\
& \sum_{i=1}^{k}(\lambda_i^+ - \lambda_i^-)\, x_{ij} - c - \mathrm{M}y_j \leq 0, \quad j \in D_2 \\
& \sum_{i=1}^{k}(\lambda_i^+ + \lambda_i^-) = 1 \\
& \sum_{i=1}^{k}(\zeta_i^+ + \zeta_i^-) = k \quad \textit{i.e.}\ \zeta_i^+ + \zeta_i^- = 1 \ \forall\, i \\
& \zeta_i^+ \geq \lambda_i^+ \geq \varepsilon\zeta_i^+ \\
& \zeta_i^- \geq \lambda_i^- \geq \varepsilon\zeta_i^- \\
& \zeta_i^+ + \zeta_i^- \leq 1 \ \ (\, i = 1, ..., k\,) \\
& \zeta_i^+, \zeta_i^- \in \{\,0\,,1\,\}
\end{aligned}
\tag{6}
$$

上式中，y_j 是二元變數，若其值爲 1，則表示有誤判；若其值爲 0，則表示無誤判。另外，觀察樣本 j 只以重疊區域內 (D_1、D_2) 爲範圍，ζ_i^+ 、ζ_i^- 同樣是二元變數，也是 λ_i^+ 、λ_i^- 的上下界，c 的值並無限制，而 M 的值是一大的常數，其他的變數值都大於零。將重疊區域內的所有觀察樣本 (R_O) 經運算後的結果同階段一的步驟，再經由判別式 (7) 作分類：

$$R_{O1} = \left\{ j \in D_1 \left| \sum_{i=1}^{k} \lambda_i^* x_{ij} \geq c^* \right. \right\} \cap R_O$$
$$R_{O2} = \left\{ j \in D_2 \left| \sum_{i=1}^{k} \lambda_i^* x_{ij} \leq c^* - \varepsilon \right. \right\} \cap R_O \quad (7)$$

階段二分類若以圖形表示如圖 2，利用公式(7)所建立的兩線段將重疊區域內的觀察樣本再次分類，減少階段一中的誤判情形。

由於 Sueyoshi 的方法利用兩階段的概念處理重疊區域，並減少二元變數過多的缺點，也因此會得到令人滿意的分類結果，但遺憾的是此一方法目前尚無法處理多類別的分類，故本研究提出以兩階段混合整數規劃結合樹狀圖逐步分割的概念延伸爲可處理多類別的分類。

2.2 主成份分析

主成份分析在 1901 年被 Pearson 最先提出，再由 Hotelling 加以延伸發展成的一種統計方法 (Gochet *et al*., 1997; Johnson and Wichern, 2002; 陳順宇，民 87)。主要的目的是要將原始資料中的變數，經由主成份分析找出比原始變數數目少的基礎變數 (fundamental variables)，進而利用這些主成份變數去解釋原始變數間的變異，也就是希望用較少的主成份變數去解釋大量原始變數的變異，如此等於是降低了原始變數的維度，同時避免運算上的複雜度。主成份分析的另一個優點就是可將原始變數轉換成彼此獨立的主成份變數，避免變數對分類結果的影響，而本研究也就是利用主成份分析對原始變數作前處理，轉換原始變數爲彼此獨立的主成份變數，進而提升分類時的正確率。此外，分析結果確實會受到單位的影響，故本研究建議將資料標準化後再做主成份

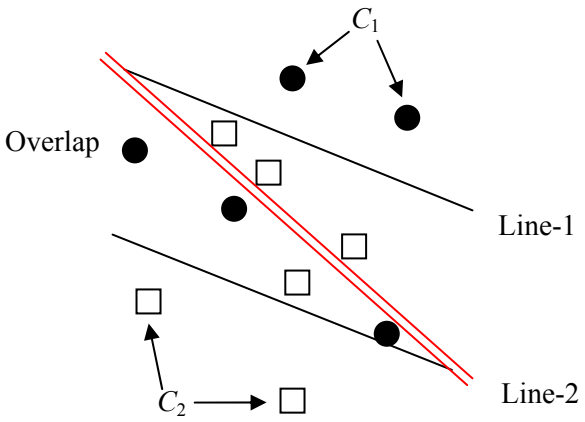

圖 2　階段二分類

分析。公式(8)是標準化後的原始變數與主成份變數間的關係式，其中 $x_1, ..., x_k$ 是原始變數，$x_1^*, ..., x_k^*$ 是標準化後的原始變數，而 $y_1, ..., y_k$ 是其主成份變數，且彼此相互獨立。

$$\begin{aligned} y_1 &= a_{11}x_1^* + a_{12}x_2^* + a_{13}x_3^* + \cdots + a_{1k}x_k^* \\ y_2 &= a_{21}x_1^* + a_{22}x_2^* + a_{23}x_3^* + \cdots + a_{2k}x_k^* \\ &\vdots \\ y_k &= a_{k1}x_1^* + a_{k2}x_2^* + a_{k3}x_3^* + \cdots + a_{kk}x_k^* \end{aligned} \tag{8}$$

主成份分析具有以下特質 (Sharma, 1996)：

(1) 所有主成份變數皆爲原始變數的線性組合；

(2) 第一主成份可解釋最大的資料變異量；

(3) 第二主成份可解釋扣除第一主成份外最大的資料變異量；

(4) 第 k 主成份可解釋扣除前 k-1 個主成份外最大的資料變異量；

(5) k 個主成份變數間是彼此獨立不相關的。

藉由主成份分析能將原始變數轉爲具相互獨立的主成份變數，因爲在研究的過程中發現，原始變數間的相關性，會影響到分類的正確率，相關性越高，分類正確率越低；相關性越低，則分類正確率越高，故本研究才會藉由主成份分析轉換原始變數爲主成份變數來分類。

3. 多類別分類方法

由於 Sueyoshi (2004) 的兩階段混合整數規劃目前尚無法處理多類別的分類，所以本研究爲了解決此一問題，故提出了利用分割順序樹的概念，使兩階段混合整數規劃能處理多類別的分類。圖 3 是本研究所提出的多類別分類方法流程圖，共分三步驟：

步驟一　首先利用兩點歐幾里得距離公式一一求出兩兩類別中心點的距離，並建立距離矩陣，依距離遠近將最近的兩類別當作一大類別並與其他類別作分割，依序將所有類別分割出來，最後建立一分割順序樹狀圖。因爲歐幾里得距離在變數間獨立的情況下，是馬氏距離的特例，爲簡化處理程序，只使用歐幾里得距離，但若處理相關的變數，則交由步驟二主成份分析來處理 (Sharma, 1996)；

步驟二　接下來觀察原始變數間是否相互獨立，相關係數 r 值界於-1 到+1 之間，r 爲正值時代表兩變數之間是正相關；r 爲負值時則是負相關。$|r|$ 越接近 1 時，表示兩變數線性相關越強；越接近於 0 表示兩變數線性相關越弱。一般認爲 $|r|$ 在 0.2 以下爲不相關；$0.2 < |r| \le 0.5$ 是低度相關；$0.5 < |r| \le 0.8$ 爲相關；$|r| > 0.8$ 以上爲高度相關。可依變數間的相關係數判別變數間是否相互獨

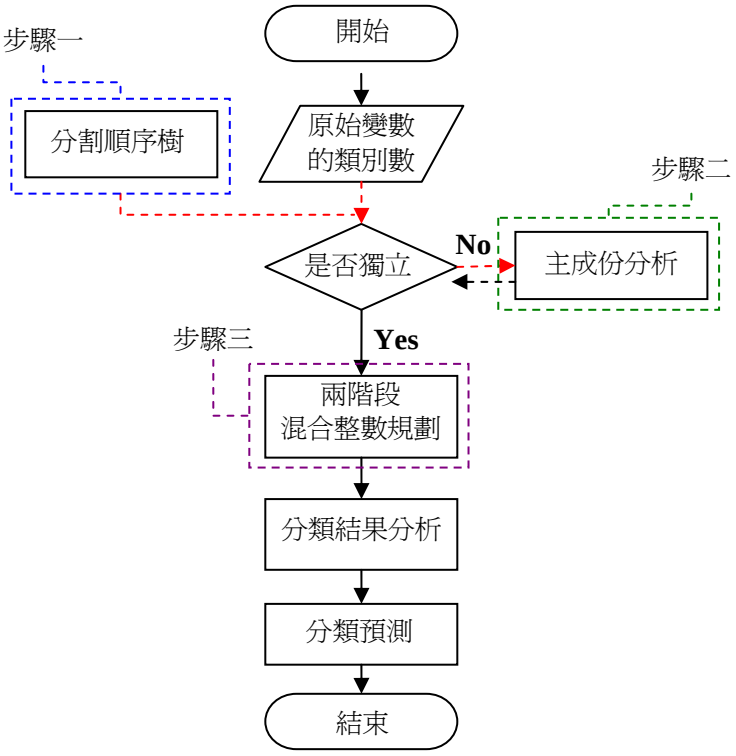

圖3　多類別分類方法流程圖

立，相關係數越低 (≤ 0.2) 表示變數間越獨立 (葉能哲，民 92)；若相關係數高(>0.2) 則需則藉由主成份分析來轉換原始變數爲主成份變數，使其變數間具有相互獨立的特性，以利後續分類。

步驟三　最後用兩階段混合整數規劃依分割順序樹狀圖作分類。

詳細說明如下：

步驟一　多類別樹狀展開圖

本研究提出的多類別分類方法是利用樹狀圖逐步分割的概念，一次分割出兩個相近的類別。因爲在研究過程的經驗中發現，距離越相近的兩類別越不易分類，故一次取最相近的兩類別並與其他類別作分割，直至將所有的類別分割出來，最後再以兩階段混合整數規劃對每一個分割處的兩兩類別做分類。

建立分割順序樹方法的四個步驟：

步驟 1.1. 先求出兩兩類別間的中心點距離，並建立一類別距離矩陣；

步驟 1.2. 依距離遠近，最相近的兩類別先放一起，並與其他類別作分割；

步驟 1.3. 對其他的類別重複步驟 1.2，直到將所有類別分割出來；

步驟 1.4. 依分割的先後順序建立一分割順序樹狀圖。

以下以五類別的範例說明，五個類別分別表示爲 G_A、G_B、G_C、G_D、G_E。

步驟 1.1. 先求出五類別的距離矩陣如表1；

步驟 1.2. 首先，在距離矩陣中找到最短距離是60.25，故 G_A、G_B 是最相近的兩類別，則將 G_A、G_B 當作一大類別組並與 G_C、G_D、G_E 做第一次分割；

步驟 1.3. 接著對 G_C、G_D、G_E 做第二次分割，其中最短距離是130，故 G_C、G_D 是最相近的兩類別，則將 G_C、G_D 當作一大類別組並與 G_E 做第二次分割；

步驟 1.4. 圖4是五類別的分割順序樹，最後，則是在每個分割處做一次兩階段混合整數規劃，就完成五類別的分類樹狀展開圖。

歸納以上範例，本研究是將最不易分類的兩類別先選取出來，並與其他的類別做分割，接著又從其他的類別中選取最不易分類的兩類別再與其他的類別做分割，如此便能將所有的類別分割出。故建立一 N 類別的分割順序樹需做 N-1 次的分割，亦即也需做 N-1 次的兩階段混合整數規劃，這就是本研究的基本概念。接著便是要依距離的遠近建立一距離矩陣求出類別間的相對距離，但在求類別變數平均值時，需注意到到離群值 (outlier) 對變數平均值的影響，因爲平均值對離群值很敏感，所以建議要再確認離群值發生的原因爲何。本研究先利用分割順序樹狀圖可延伸兩階段混合整數規劃處理多類別的分類，經由樹狀圖逐步分割的概念依序對距離較近的先分在一起，並與其他的類別作分割，這樣的方法比藉由逐一分割的概念，逐一地分割出單一類別來得要好，因爲在研究的過程中發現，分類的規則是以相對應的兩類別資料運算而得出的，故若只是隨機地逐一分割出任一類別，很容易因分割的位置不佳而導致大量的誤判發生。因此，本研究利用所有類別間兩兩類別的中心點距離建立分割順序樹，並依樹狀圖逐步地分割出兩類別直到將所有的類別分割出來，這樣的方法不僅是已做了一次簡單的分類整理，在接下來分類時更能減少誤判的機率。

步驟二　主成份分析

在研究的過程中發現，低度相關的變數其判別正確率會比高度相關的變數來得要高，所以資料的前處理有其必要性。本研究利用多變量主成份分析的特性，將原始具相關可能性的變數轉換

表 1　五類別的距離矩陣

	G_A	G_B	G_C	G_D	G_E
G_A	0	60.25	352.25	673.25	498.25
G_B		0	281	425	212
G_C			0	130	481
G_D				0	293
G_E					0

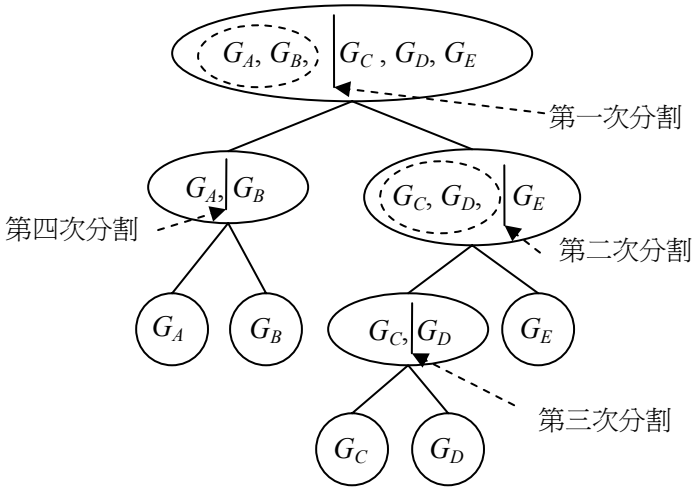

圖 4　五類別分割順序樹

成主成份變數，因爲在轉換的過程中，已將原始變數間的相關性經由主成份分析轉變成具彼此相互獨立的主成份變數，由於主成份變數間是彼此獨立的，故能提升其分類能力。另外，若原始變數的相關性已經很低或趨近於相互獨立，則可省略掉此步驟以節省處理成本，提高分類時的效率。

步驟三　兩階段混合整數規劃分類

由於 Sueyoshi 的兩階段混合整數規劃的優點是利用兩階段的概念對重疊區域作二次分類，降低誤判的可能性，並提高分類時的正確率；同時，又減少了限制式中二元變數的數量，以降低運算時的負擔，提高分類時的效率。故最後依分割順序樹在每個分割處作一次分類，以 N 類別而言，需做 N-1 次的兩階段混合整數規劃分類。在分類時，若階段一中 s^* 的值是非正的話，則表示無重疊區域的存在，便不需再藉由階段二作更詳細的分類，以提升分類的效率。另外，本研究可以將階段一、二中限制式的範圍 0 改設爲一大於零的極小值 ε，便可減少觀察樣本落在邊界上的機率，間接地降低誤判的可能性。

本研究利用樹狀圖逐步分割的概念可延伸爲多類別分類方法，並減少誤判的發生、且利用主成份分析轉換原始變數爲具有相互獨立的特性，及混合整數規劃兩階段分類的概念處理重疊區域，提升其分類時的效能。因此本研究結合了分割順序樹、主成份分析及兩階段混合整數的各個優點，使得所提出的多類別分類方法有很高的效能及可用性。在下一節中，本研究將以二個範例詳細說明方法的步驟並與支持向量機及統計判別分析作比較。

4. 範例說明

在這一節中，本研究以二個範例作說明，範例一中以美國某大學企業管理碩士的入學核准的八十五筆資料爲例 (Johnson and Wichern, pp. 620-626, 2002; 陳順宇，4-32~4-34頁，民87)，詳細說明本研究所提出的多類別分類方法的分類步驟；範例二是 Fisher 分類鳶尾花的一百五十筆資料 (Fisher, pp. 179-188, 1936; 陳順宇，4-45~4-59頁，民87)。從範例一的結果驗證本研究所提出的多類別分類方法確實比用原始變數分類的結果來得佳，而從兩個範例的結果中也顯示本研究的多類別分類方法有相當不錯的分類正確率。

(1) 範例一：企業管理碩士的入學錄取核准 (Johnson and Wichern, pp. 620-626, 2002; 陳順宇，4-32~4-34頁，民87)

資料的來源是美國某大學企業管理碩士的入學錄取核准資格，共有八十五筆資料，分成三類別(G_A：允許、G_B：不允許及 G_C：再評估)，每一類別各有三十一、二十八及二十六筆資料，包含兩項衡量變數 x_1：學業成績平均點數 (GPA)；x_2：研究所入學考試成績 (GMAT)。在此範例中，將資料分爲訓練樣本與測試樣本，每類別的前 2 / 3 作爲訓練的樣本，後 1 / 3 作爲測試的樣本，故 G_A 取前二十筆當訓練樣本，後十一筆當測試樣本；G_B 取前十九筆當訓練樣本，後九筆當測試樣本；G_C 取前十八筆當訓練樣本，後八筆當測試樣本，故訓練量本有 57 筆，測試樣本有 28 筆。

步驟一　多類別分類樹狀展開圖

首先，利用三類別資料中的訓練樣本的原始變數平均值算出兩兩類別的中心點距離 (表 2)。這樣的步驟是爲了要求出每一類別的相對距離，以建立分割樹狀圖。另外要注意的是變數平均值是以原始變數運算後得出的，因爲本研究的目的只是要比較出類別組間距離的遠近，所以用原始變數運算的結果會比用主成份變數的值計算來得明顯。從表 2 顯示出類別 G_B、G_C 的距離是 132.2651，比類別 G_A、G_B 的距離 10742.0021 與類別 G_A、G_C 的距離 13254.5642 還相近，所以將 G_B、G_C 這兩類別暫時當成新的一組類別並與類別 G_A 作分割，故共需作兩次分割，即第一次對類別 G_B、G_C 與類別 G_A 作分割；第二次對類別 G_B 與類別 G_C 作分割，分割順序樹狀圖如圖 5。

表 2　兩兩類別間距離

類別	兩類別距離
G_A, G_B	10742.0021
G_A, G_C	13254.5642
G_B, G_C	132.2651

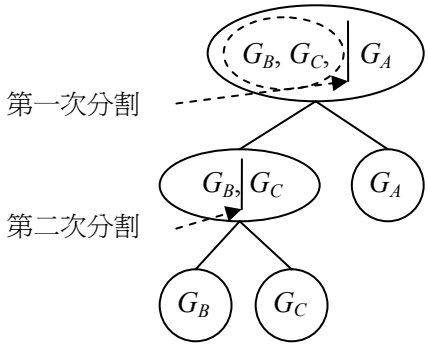

圖 5　三類別分割樹狀圖

步驟二　主成份分析

因爲研究過程中發現變數間的相關性會影響到分類時的正確率，故本研究藉由主成分分析的特性來轉換原始變數爲具有相互獨立特性的主成份變數，以提高分類時的正確率。另外，在轉換原始變數前，可以先觀察是否有轉換原始變數的必要性，若原始變數間的相關性已驅近於獨立，便可不必將原始變數轉換成主成份變數以節省時間及成本。由於變數 x_1 與 x_2 相關係數爲 0.3976，屬低度相關，所以需藉由主成份分析轉換原始變數，其轉換的公式如下，其中 x_1 與 x_2 是原始變數，y_1 與 y_2 是主成份變數。表 3 是原始變數轉換成主成份變數的比較，只取 5 個樣本比較。

步驟三　兩階段混合整數規劃分類

依步驟一的分割順序樹 (圖5)，在每個分割處做一次兩階段混合整數規劃，則共需作兩次分類，第一次對類別 G_B、G_C 與類別 G_A 作分類；第二次對類別 G_B 與類別G_C 作分類。

第一次分割：階段一分類

階段一先將類別 G_B、G_C 中的觀察樣本暫時放在同一類別並與類別 G_A 作分類，表4是從兩相對類別中各列舉一個說明，其中 λ_1^* 的值是0.9806，λ_2^* 的值是0.0194，d^* 的值是 0.1555，

表 3　原始變數轉換成主成份變數

原始變數 觀察樣本	x_1 (GPA)	x_2 (GMAT)	主成份變數 觀察樣本	y_1 (Prin$_1$)	y_2 (Prin$_2$)
1	2.96	596	1	0.9589	-0.9168
2	3.14	473	2	0.1574	0.4779
3	3.22	482	3	0.3696	0.5293
4	3.29	527	4	0.8867	0.2429
5	3.69	505	5	1.3492	1.0983

表 4 第一次分類階段一結果舉例說明

觀察樣本	y_1	y_2	$\lambda_1^* = 0.9806$ $\lambda_2^* = 0.0194$	$\sum_{i=1}^{k} \lambda_i^* y_{ij}$	$> d^* + s^*$	$< d^* - s^*$	原類別	判別結果
1	0.9589	-0.9168		0.9224	0.1555		G_A	C_A
32	-0.1164	-0.1709		-0.1174		0.1555	G_B	C_{BC}

s^*的值是 -0.0081，由於 s^* 的值是負的，表示沒有重疊區域的存在，故在判別式中將 s^* 的值忽略不計。由表得知觀察樣本1代入判別式(9) 運算後的值是0.9224，大於 $d^* + s^*$ 的值0.1555，故分類到 C_A；觀察樣本32代入判別式(10) 運算後的值是 -0.1174，小於 $d^* - s^*$ 的值0.1555，故分類到 C_{BC}。由於在階段一的分類中，並無觀察樣本屬於重疊區域，如圖1 落在不重疊區域C_1及C_2，故不需再利用階段二分類。

$$C_A = \left\{ j = G_A \mid \sum_{i=1}^{k} \lambda_i^* y_{ij} > 0.1555 \right\}, \tag{9}$$

$$C_{BC} = \left\{ j = G_{BC} \mid \sum_{i=1}^{k} \lambda_i^* y_{ij} < 0.1555 \right\}, \tag{10}$$

$$D_A = G_A - C_A \text{ and } D_{BC} = G_{BC} - C_{BC}$$

第二次分割：階段一分類

階段一將類別 G_B 與類別 G_C 作分類，表 5 是將分類的 B、C 兩類別中各列舉一個說明，其中 λ_1^* 的值是 -0.3062，λ_2^* 的值是 -0.6937，d^* 的值是 0.3347，s^* 的值是 0.1805，由於 s^* 的值是正的，表示有重疊區域的存在，在此階段共有 8 個觀察樣本落在重疊區域，故需將重疊區域內的觀察樣本留至階段二再次分類。由表得知觀察樣本 32 代入判別式(11) 運算後的值是 0.1541，不大於 $d^* + s^*$ 的值 0.5152，故分類到 D_B，並留至階段二再分類；觀察樣本 60 代入判別式(12) 運算後的值是 0.5152，不小於 $d^* - s^*$ 的值 0.1542，故分類到 D_C，同樣也是留至階段二再分類；在階段一中，共有八個觀察樣本屬於重疊區域，故需再利用階段二做更詳細的分類。

$$C_B = \left\{ j = G_B \mid \sum_{i=1}^{k} \lambda_i^* y_{ij} > 0.3347 + 0.1805 \right\} \tag{11}$$

$$C_C = \left\{ j = G_C \mid \sum_{i=1}^{k} \lambda_i^* y_{ij} < 03347 - 0.1805 \right\} \tag{12}$$

$$D_B = G_B - C_B \text{ and } D_C = G_C - C_C$$

表 5 第二次分割階段一分類結果舉例說明

觀察樣本	y_1	y_2	$\lambda_1^* = -0.3062$ $\lambda_2^* = -0.6937$	$\sum_{i=1}^{k} \lambda_i^* y_{ij}$	$> d^* + s^*$	$< d^* - s^*$	原類別	判別結果
32	-0.1164	-0.1709		0.1541	0.5152		G_B	D_B
60	-1.0720	-0.2696		0.5152		0.1542	G_C	D_C

第二次分割：階段二分類

在階段一中的分類時有八個觀察樣本是屬於重疊區域，故需再利用階段二做更詳細的分類。表6是由重疊區域中取其中兩個觀察樣本來說明，其中 λ_1^* 的值是0.0588，λ_2^* 的值是0.9440，c^* 的值是 -0.1678。由表得知觀察樣本32原屬於 G_B ，在代入判別式(13) 運算後的值是0.1164，大於 c^* 的值 -0.1678，故分類到 R_{OB}；觀察樣本60代入判別式(14) 運算後的值是 1.0715，不小於 c^*-ε 的值 -0.1678，所以不分類到 R_{OC}，而是誤判。

$$R_{\mathrm{OB}} = \left\{ j = D_B \mid \sum_{i=1}^{k} \lambda_i^* y_{ij} > -0.1678 \right\} \cap R_{\mathrm{O}} \tag{13}$$

$$R_{\mathrm{OC}} = \left\{ j = D_C \mid \sum_{i=1}^{k} \lambda_i^* y_{ij} < -0.1678 - \varepsilon \right\} \cap R_{\mathrm{O}} \tag{14}$$

將判別式 (9)、(10)、(11)、(12)、(13)、(14) 整理後，彙整如圖 6 表示。

接著並將 28 筆測試樣本先代入 G_A 與 G_{BC} 的判別式 (9)、(10) 做第一次的分類，若分類到 G_{BC}，則再利用 G_B 與 G_C 的判別式(11)、(12) 做第二次的分類。在此例中，因爲本研究所提出的多類別分類方法在分 G_A 與 G_{BC} 分類時，由於切割的位置較佳，所以在第一次分類時測試樣本並無誤判發生；而第二次對 G_B 與 G_C 作分類時，有四個測試的觀察樣本屬於重疊區域，故再利用階段二分類，在階段二的分類中，則有一個誤判發生。另外，從結果比較如表 10 顯示出經主成份分析轉換後的主成份變數，其判別率達 96.43 %，優於利用原始變數分類正確率的 85.71 %

表 6 第二次分割階段二分類結果

觀察樣本	y_1	y_2	$\lambda_1^* = 0.0588$ $\lambda_2^* = 0.9440$	$\sum_{i=1}^{k} \lambda_i^* y_{ij}$	$> c^*$	$< c^* - \varepsilon$	原類別	判別結果
32	-0.1164	-0.1709		0.1164	-0.1678		G_B	R_{OB}
60	-1.0720	-0.2696		1.0715		<-0.1678	G_C	Not R_{OC}

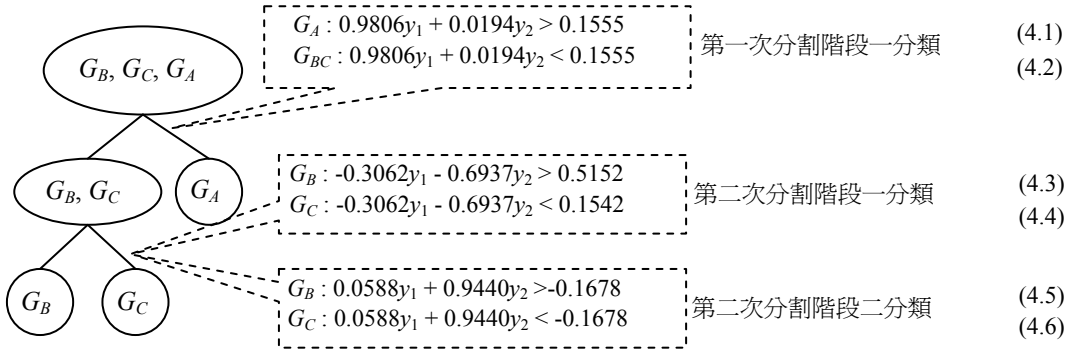

圖 6　三類別分類判別圖

及支持向量機的 50.00 %；也高於統計判別分析法所做的結果，其判別正確率爲 91.7% (陳順宇，4-34 頁，民 86) ，此結果驗證了利用主成份分析做資料前處理確實可以提升分類正確率，且分類能力比支持向量機更適用在小樣本上。

以範例一的八十五筆企管碩士入學核准爲例，若有一新的觀察樣本需分類，則將新觀察樣本從分割順序樹狀圖的最上層開始，利用分類規則判別新觀察樣本是該分類到類別 G_A 亦或G_{BC} 。若是分類到G_{BC} ，則需再作一次的判別新觀察樣本所在的類別是 G_B 亦或G_C，所以共需做兩次的判別。若分割順序樹狀圖的**類別**爲N，則最多需判別 N-1次便可分類出新觀察樣本所在的類別。但有一點需注意的是，若資料的前處理利用到主成份分析，因爲判別式的依據規則是以主成份變數運算後而建立的，故在判別新觀察樣本之前，需先將新觀察樣本的變數轉換成主成份變數才能判別；若無資料前處理，則不需做變數轉換。

表 7 是入學核准資料的主成份變數與原始變數間的特徵向量，若有一新的觀察樣本(x_1, x_2)是(3.01, 453)，在標準化後，轉變爲 (x_1^*, x_2^*)，與主成份變數(y_1, y_2) 的轉換關係如 (15)。

$$\begin{cases} y_1 = 0.7071x_1^* + 0.7071x_2^* \\ y_2 = 0.7071x_1^* + (-0.7071)x_2^* \end{cases} \tag{15}$$

所以觀察樣本的原始變數值是(3.01,453)，經由主成份分析轉換之後的值爲 (-0.2792, 0.2124)。

第一次分割：階段一分類

如表8所示，λ_1^* 的值是0.9806，λ_2^* 的值是0.0194，d^* 的值是0.1555，s^* 的值是 -0.0081，將主成份變數代入後的值是-0.2697，由於s^* 的值爲負，表示無重疊區域的存在，故判別式只有 d^* 的值0.1555，而由於其值小於0.1555，故應分類到 C_{BC}，也就是G_{BC}，如圖1 落在不重疊區域C_2，故不需再利用階段二分類。

表 7　原始變數與主成份變數的特徵向量

原始變數	特徵向量	
	$y_1(\text{Prin}_1)$	$y_2(\text{Prin}_2)$
x_1^*	0.7071	0.7071
x_2^*	0.7071	-0.7071

表8　第一次分割階段一分類結果

觀察樣本	y_1	y_2	$\lambda_1^* = 0.9806$ $\lambda_2^* = 0.0194$	$\sum_{i=1}^{k} \lambda_i^* y_{ij}$	$> d^*$	$< d^*$	判別結果
new	-0.2792	0.2124		-0.2697	0.1555		
new	-0.2792	0.2124		-0.2697		0.1555	C_{BC}

第二次分割：階段一分類

階段一的判別表9中顯示，λ_1^* 的值是 -0.3062，λ_2^* 的值是 -0.6937，d^* 的值是0.3347，s^* 的值是0.1805，由於運算後的值爲-0.0619小於0.1542，故應分類到 C_C，也就是G_C。所以新觀察樣本 (3.01 , 453) 所在的類別應是 G_C，亦即該生還要再評估。因此本研究所提出的多類別分類方法，不僅在資料分類上有高度的效能，另外在新觀察樣本的預測上也有顯著的表現，也再次驗證了此多類別分類方法的可用性。

(2) 範例二：鳶尾花種類分類 (Fisher, pp. 179-188, 1936; 陳順宇，4-45~4-59頁，民87)

資料的來源是 Fisher 在分類鳶尾花的資料，樣本共有一百五十朵鳶尾花分成三類別(setosa、versicolor及virginica)，每類別有五十朵花，有四項特徵變數： x_1：萼片長度 (speal length)；x_2：萼片寬度 (speal width)； x_3：花瓣長度(petal length)；x_4：花瓣寬度(petal width)。在此範例中，將資料分爲訓練樣本與測試樣本，在每類別中各取前三十筆爲訓練的樣本，後二十筆爲測試的樣本，故訓練樣本有90筆，測試樣本有60筆，其分類的步驟同範例一。因爲相關係數偏高，如變數x_1 與 x_3相關係數爲0.8706；變數x_1 與 x_4相關係數爲0.7986；變數x_3 與 x_4相關係數爲0.8706，所以仍需藉由主成份分析轉換原始變數。

表10爲兩範例之綜合比較，在第一個例子：八十五筆企管碩士入學評估分類問題上，本研究所提出的多類別分類方法判別率是96.43 %，高於用原始變數分類的85.71 %，亦高於支持向量機的50.00 %；第二個例子：一百五十朵鳶尾花種類分類問題上，本研究所提出的多類別分類方法判別正確率是100 %，等於用原始變數分類的100 %，高於支持向量機的78.33 %。即使用統計判別分析法所做的結果 (陳順宇，民87) ，其判別正確率在第一個例子爲91.7%；第二個例子爲

表 9 第二次分割階段—分類結果

觀察樣本	y_1	y_2	λ_1^*=-0.3062 λ_2^*=-0.6937	$\sum_{i=1}^{k}\lambda_i^* y_{ij}$	$> d^* + s^*$	$< d^* - s^*$	判別結果
new	-0.2792	0.2124		-0.0619	0.5152		
new	-0.2792	0.2124		-0.0619		0.1542	C_C

表 10 二實例的各方法判別正確率綜合比較

實例類型	方法比較		
	支持向量機	原始變數 + 逐步分割樹 + 混合整數規劃	主成份變數 + 逐步分割樹 + 混合整數規劃
85筆企管碩士入學核准	50.00 %	85.71 %	96.43 %
150朵鳶尾花種類	78.33 %	100.00 %	100.00 %

98%，均比本研究所提的方法來得低。從上述的結果，驗證了本研究所提出的多類別分類方法確實有相當高的分類效能及可用性，不僅能分多類別的資料，更比支持向量機適用於小樣本上。

5. 結論

本研究所提出的多類別分類方法有以下二點貢獻:一是利用樹狀圖逐步分割概念延伸兩階段混合整數規劃法爲可處理多類別分類的新方法，並減少分類時的誤判；二是藉由主成份分析轉換變數資料爲具有彼此獨立特性的主成份變數，提高分類正確率。

在第四節比較本研究所提出的分類方法與支持向量機後，得出要提升分類效能力可朝類別最佳分割順序及原始變數的轉換這兩處著手，尤其是類別間的分割順序影響更大，如果分割的位置不佳，很容易造成大量誤判的情形發生。但在利用兩兩類別中心點距離求分割順序樹時，類別變數的平均值極易因離群值造成偏差，導致分割位置不佳而在分類時產生大量的誤判，故若能在建立分割順序樹時考量類別間與類別內的變異，則更能避免誤判的產生及提高其分類效能。故在建構分割順序圖時，可再將類別間與類別內的變異納入考量，以增加分類的效能，以取得較佳的正確判別率。

雖然支持向機量是目前相當受歡迎的分類方法，尤其是在大量的樣本上，但是在小樣本的分類問題上，本研究所提出的多類別分類方法比支持向量機有著更高的分類效能，所以在分小樣本時，可利用本研究所提出的多類別分類方法，以獲得較高的分類正確率，且本研究的方法與支持

向量機也有互補的特性，可將本研究中的主成份分析套用在支持向量機上，也會得到較高的分類正確率。

未來將尋求解決兩階段混合整數規劃的不足處的方法，如非線性資料的分類；另外，其他諸如分類的時間、成本、方便性及分類結果的可接受度都是未來在分類時會考量到的因素，唯有將這些因素作整體性的考量後，才能發揮出分類方法最大的效能。

參考文獻

葉能哲，統計學，第二版，台北：華泰書局，民國 92 年。

陳順宇，多變量分析，台北：華泰書局，民國 87 年。

Boser, B. E., Guyon, I. M., and Vapnik, V. N., "A Training Algorithm for Optimal Margin Classifiers," *Computational Learning Theory*, 5th ed., Pittsburgh: ACM 1992, pp.114 - 152.

Charnes, A., Cooper, W. W., and Ferguson, R. O., "Optimal Estimation of Executive Compensation by Linear Programming," *Management Science*, Vol. 1, 1955, pp. 138 - 151.

Cortez, C., and Vapnik, V., "Support Vector Networks," *Machine Learning*, Vol. 20, 1995, pp. 273 - 279.

Fisher, R. A., "The Use of Multiple Measurement in Taxonomy Problems," *Annals of Eugenics*, Vol. 7, 1936, pp. 179 - 188.

Freed, N., and Glover, F., "A Linear Programming Approach to the Discriminant Problem," *Decision Sciences*, Vol. 12, 1981, pp. 68 - 74.

Glen, J. J., "An Iiterative Mixed Integer Programming Method for Classification Accuracy Maximizing Discriminant Analysis," *Computers & Operations Research*, Vol. 30, 2003, pp. 181 - 198.

Gochet, W., Stam, A., Srinvasan, V., and Chen, S., "Multigroup Discriminant Analysis Using Linear Programming," *Operations Research*, Vol. 45, 1997, pp. 213 - 225.

Johnson, R. A., and Wichern, D. W., *Applied Multivariate Statistical Analysis*, 5th Edition, Upper Saddle River, NJ: Prentice-Hall, 2002.

Kendal, S. M., Stuart, A., and Ord, J. K., *The Advanced Theory of Statistics*, London: Charles Grifin, 1983.

Kim, H. -C., Pang, S., Je, H.-M., Kim, D., and Bang, S. Y., "Constructing Support Vector Machine Ensemble," *Pattern Recognition*, Vol. 36, 2003, pp. 2757 – 2767.

Li, S., Kwok, J.T., Zhua, H., and Wang, Y., "Texture Classification Using the Support Vector Machines," *Pattern Recognition*, Vol. 36, 2003, pp. 2883 – 2893.

Loucopoulos, C., “Three-Group Classification with Unequal Misclassification Costs：a Mathematical Programming Approach,” *Omega*, Vol. 29, 2001, pp. 291 - 297.

McLachlan, G. L., *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley, 1992.

Sharma, S., *Applied Multivariate Techniques*, New York: Wiley, 1996.

Shin, K. -S., Lee, T. S., and Kim, H. -J., “An Application of Support Vector Machines in Bankruptcy Prediction Model,” *Expert Systems with Applications*, Vol. 28, 2005, pp. 127-135.

Sueyoshi, T., “DEA-Discriminant Analysis in the View of Goal Programming,” *European Journal of Operational Research*, Vol. 115, 1999, pp. 564 - 582.

Sueyoshi, T., “Extended DEA-Discriminant Analysis,” *European Journal of Operational Research*, Vol. 131, 2001, pp. 324 - 351.

Sueyoshi, T., “Mixed Integer Programming Approach of Extended DEA-Discriminant Analysis,” *European Journal of Operational Research*, Vol. 152, 2004, pp. 45 - 55.

Tay, F. E. H., and Cao, L., “Application of Support Vector Machines in Financial Time Series Forecasting,” *Omega*, Vol. 29, 2001, pp. 309–317.