國立臺灣大學工業工程學研究所

碩士論文

Graduate Institute of Industrial Engineering

College of Engineering

National Taiwan University

Master Thesis

多層混合分類樹研究及其腫瘤診斷之應用

Study of Multi-layer Hybrid Classification Tree with Applications to Cancer Diagnosis

曾煥澤

Huanze Zeng

指導教授：陳正剛 博士

Advisor: Argon Chen, Ph.D.

中華民國 104 年 7 月

July 2015

誌謝

陳正剛老師的實驗室「操」學生出名，其實我早有耳聞，但是那時憑著一份執著，僥倖加入了實驗室。到目前為止，我還是很慶幸當初的選擇，在台大工工所遇見了老師，還有我珍惜的同學們。

感謝陳正剛老師在這兩年來的教導。一次次的 meeting，讓我對於嚴謹和邏輯有了更深刻的體會，直到現在依舊在不斷學習和提高簡報的能力。謝謝老師這兩年耐心的指導，感激老師在百忙之中抽空指導我的論文。同時也感謝郭文宏和陳炯年醫師參與學生的碩士學位考試，並給予學生建議與指教，使論文得以更為完善。

感謝跟我一起度過這兩年辛苦 meeting 時光的柏頤和上峰。柏頤，你一直是我最好的夥伴之一，總在關鍵的時候提供最無私的幫助，做事非常可靠且極富責任心。上峰，我們以後怕是沒有辦法一邊熬夜一邊聊天了，有點可惜。另外感謝秉逸在口試當天和畢業流程中提供的無私幫助。

在工工所的兩年充滿著溫馨和感動。承翰，我們一起度過柔性和物件導向的艱難時光，你不僅聰明、善解人意且樂於助人。捷哥，你是我永遠的好夥伴，謝謝你犧牲了大量自己的時間維持大家的凝聚力。懋神總是充滿著正能量，不但對自己要求非常高，而且也樂於幫助身邊的朋友。剛毅，我們都是一起熬過柔性的好戰友，謝謝你很用心地組織我們的活動，期待不屈不撓的你早日從峰居畢業。其他還有小旻、思云、小歐、惠倫...我真的很高興能跟你們當同學，一起出遊、吃大餐。

感謝上蒼不只給我豐收的研究所生活，也給了一直支持我的家人，還有一群志同道合的所內外朋友。感謝軼韻、承翰、小倩倩、師傅和一芳在口試預報、論文摘要翻譯和內容修改及校對中提供的無私幫助。謝謝你們在關鍵時刻伸出援手，助力我完成畢業最後一里路。也非常感謝海虹，在忙碌的五月和七月給了我很多鼓勵和支持，讓我能夠有更多的時間專心做研究。

中文摘要

分類樹(Classification Tree)在資料探勘領域上被廣泛使用來探討感興趣資料的分類，並應用於醫學、工程等領域的機器學習。分類樹主要分為兩個主要的類別，即分類與迴歸樹(Classification and regression trees, C&ART) 和多變量分類樹。C&ART 常用於建構二元分類樹，一般利用 Gini index 做為分割的準則。多層判別分析有別於 C&ART，其每一層的待分割節點皆會分割成兩個或三個節點，允許其中一節點為未分類資料，未分類節點資料可繼續透過使用其他屬性進行分割展開新的一層，而已確定類別的節點，則不再分割。由於在醫學探勘（如腫瘤診斷）中，結合費雪線性判別分析(FLD)的分類樹模型不一定能夠有效提升分類樹的分類效能，本文嘗試構造更有效的演算法並加以實例驗證。

在模型構造中，本研究先通過引入參數 P_{FLD} 來調節費雪線性組合屬性方案的比例。同時，根據賴淑俐學者（2010）所進行的理論探討發現，多層判別分析與 C&ART 分類樹可以互補不足之處，本研究進而通過引入參數 $P_{critical}$ 調整多層判別分析和 C&ART 分類樹的相對比重。當每一個節點進入演算法中時，先通過 P_{FLD} 和多層組合屬性方案決定是否需要採用費雪線性組合屬性方案及相應的特徵數，再通過 $P_{critical}$ 和非參數型接受者操作特徵（NP-ROC）來決定節點和切割方案，即決定是否分割成 C&ART 的兩個節點或多層判別分析的兩個節點或三個節點。

為了驗證此模型，本研究利用臺大醫院所提供的 366 筆乳房腫瘤案例來測試，其中 266 筆做為訓練樣本用於選擇和訓練參數，而 100 筆則固定作為獨立測試樣本，從而比較多層混合分類樹與 C&ART、多層判別分析和強化多層判別分析的單一分類樹的判別結果和多階段調適樹群（莊曙詮，2012）的 BI-RADS 分級結果，驗證判別模型效能。

從案例驗證的結果中，可以看出新演算法的分類效能確實優於其他方法，且能在顯著增加多階段調適樹群 BIRADS 3 的良性個數同時，將惡性比例維持在可接受的範圍內。

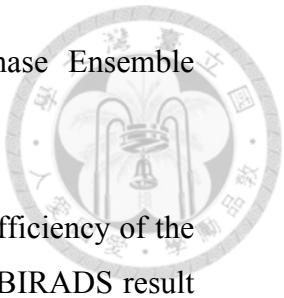
關鍵字: C&ART 分類樹；多層判別分析；多層混合分類樹;接收者操作特徵曲線；費雪線性分析

ABSTRACT

The classification decision tree is the most commonly used classification tool in data mining and machine learning in medical and engineering applications. There are mainly two types of classification trees: C&ART and multivariate classification tree. The C&ART is usually used and constructed by a hierarchical tree of decision nodes. The structure of the Multi-layer Classifier, proposed by Wu (2009), is differs from the C&ART by constructing each layer consisting of two or three nodes, of which only the node with unclassified data will be classified further into the next layer and the rest nodes contain data completely classified. The tree construction continues until a stop criterion is reached. However, the structure of the Multi-layer Classifier or C&ART combined with Fisher Linear Discriminant analysis (FLD) may not improve classification tree efficiency when it is applied to medical exploration (such as diagnosis of tumor). Hence, this thesis aims at constructing a more effective Multi-layer Hybrid Classification Tree and utilizes empirical data to validate its performance.

In the modeling of tree structure, this study first introduces a parameter, P_{FLD}, to be used to adjust the proportion of nodes constructed by FLD. At the same time, according to the theoretical discussion by Lai (2010), the multi-layer classifier and the C&ART can complement each other's insufficiency. Therefore, this study introduces a second parameter, $P_{critical}$, to be used to adjust likelihood for each tree layer of data to be classified according to the Multi-layer or C&ART decision. When a node is to be split, it needs to decide first whether to apply FLD based on the value of P_{FLD}. Then it needs to decide whether to split into two nodes with C&ART decision or three (or two) nodes with Multi-layer decision based on the value of $P_{critical}$.

In order to verify the performance of the proposed model, this study uses 366 breast cancer cases provided by National Taiwan University Hospital (NTUH) to test the proposed tree, 266 of which are taken as training samples for selection and training parameters, and the other 100 is isolated as the independent test sample. We compare this proposed Multi-layer Hybrid Classifier with C&ART, Multi-Layer Classification Tree (ML-ROC), as well as Enhanced Multi-layer Classification Tree(Enhanced-ML-ROC) proposed by Lai (2010) based on results of single tree



performance and BI-RADS results generated by Adaptive Multi-phase Ensemble (Chuang, 2012).

Based on the verification results, it is found that the classification efficiency of the newly proposed algorithm is indeed superior to other methods, and the BIRADS result shows that it not only increases the benign case number of BIRADS 3 by an observable size, but also maintains the number malignant cases of BIRADS 3 in an acceptable range.

Key words: Classification and regression trees; Multi-layer Classifier; Nonparametric AUC, Multi-layer Hybrid Classification Tree; Fisher discriminant analysis

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 研究背景

分類樹(Classification Tree)在資料探勘領域上被廣泛使用來探討感興趣資料的分類，主要藉由分類已知的訓練資料來建構一個樹狀結構，然後對測試資料進行判別。分類樹主要分為兩個主要的類別，即分類與迴歸樹(Classification and regression trees, C&ART) 和多變量分類樹。

C&ART 常用於建構二元分類樹，一般利用 Gini index 做為選擇屬性與每個節點分割的準則。在建立模型的過程中，C&ART 將每一層的節點分割成兩個子節點，展開新的一層，直到達成終止條件。在某些特定資料類型，C&ART 無法有效率地做分類，甚至有過度配適(over-fitting)的問題。

與 C&ART 不同，多變量分類樹在建立模型中，一般利用費雪線性判別分析(Fisher linear discriminant Analysis, FLD)以線性組合的方式結合多個屬性，計算得到區別分數(score)，再利用區別分數對結點做分類，而此分類同時考慮了多個屬性且有較好的正確度(Yildiz 與 Alpaydin, 2001)。FLD 是由 R.A. Fisher 所發展的方法，目標是尋找出一區別函數 (discriminant function) ，同時最小化組內變異和最大化組間變異。

在分類樹的判別過程中，接收者操作特徵曲線(receiver operating characteristic curve，簡稱 ROC curve)起到了重要的作用。ROC curve 是衡量分類和篩選模型的常用工具，近年來被廣泛應用於醫學、工程等領域的機器學習。藉由其對真陽性(true positive rate)與假陽性(false positive rate)的視覺化的描述，我們可以通過其線下面積來比較資料中各屬性的分類能力。一般假設各類別資料服從常態分配，進而計算擬合 ROC curve 再對其線下面積進行統計檢定，進而比較各屬性的分類能力。

1.2 研究動機與研究目標

在實際案例研究中，一些資料的屬性只能判別某個類別或者同時判別兩個類別，例如在判別甲狀腺癌腫瘤時，腫瘤內的囊腫(cyst)所占的面積只能用來判別良性腫瘤。若腫瘤內大部分是囊腫，則可能為良性腫瘤，但若腫瘤內無囊腫並無法代表它為惡性腫瘤。而腫瘤周圍環狀區域的血管多寡(ringPDVImax)是能拿來判別良性

腫瘤也能判別惡性腫瘤。若腫瘤周圍環狀區域的血管很多，則很有可能為良性腫瘤，若血管很少，則很有可能為惡性腫瘤，從而造成分類上的偏差。

根據賴淑俐學者（2010）所進行的理論探討發現，多層判別分析與 C&ART 分類樹表現可互補不足之處。並且在實際腫瘤資料研究中，結合費雪線性判別分析(FLD)也不一定能夠有效提升分類樹的分類效能。

基於以上原因，學者們基於 C&ART 和多變量分類樹模型，嘗試著提出了新的分類樹模型，如巫信融學者提出了多層判別分析模型。多層判別分析在每一層均區別一個類別或是兩個類別，並將尚未區別之樣本留置下一層做判別。雖然多層判別分析可以較清楚分類混雜不清的節點，但缺點是每一層已做判別的節點不能再繼續分類。據此，賴淑俐學者也曾提出強化多層判別分析模型。強化多層判別分析揉合 C&ART 和多層判別分析的優點，當每一個節點進入演算法時，皆有可能分割成 C&ART 的兩節點或多層判別分析的兩節點和三節點，且同一層中的所有節點皆有可能繼續分割，然而強化多層判別分析不僅運算量較大，而且依舊存在過度配適(over-fitting)的問題。

為彌補已有的分類樹方法在實際運用中的不足，本研究嘗試創建能綜合各分類樹方法優點的多層混合分類樹（Multi-layer Hybrid Classification Tree, Hybrid）模型，並通過腫瘤良惡性判別之實例驗證來驗證多層混合分類樹的效能。

1.3 論文架構

本論文由五個章節組成。在這一章，我們介紹研究背景、解釋研究動機，且說明研究目標。在第 2 章，介紹 ROC curve 與其線下面積統計、常見的分類樹、費雪線性判別分析和 BI-RADS 腫瘤分級系統。在第 3 章，介紹多層混合分類樹的模型形態、建構流程，然後詳細闡釋屬性評估方案和參數之功能與影響，最後介紹參數的設定方法。第 4 章為腫瘤的實例試驗。第 5 章為結論和未來研究建議。

Chapter 2　文獻探討

本章2.1節先介紹ROC curve與其線下面積;2.2節介紹線下面積的意義(Hanley et al, 1982)和統計檢定(DeLong et al, 1988)；2.3節將介紹C&ART(Breiman et al., 1984)，多層判別分析樹(巫信融, 2009)並比較兩者的分類能力(賴淑俐, 2010)；2.4節介紹費雪線性判別分析(Fisher, 1936)，與相對重要性指標(王彥龍, 2013)；2.5節介紹BI-RADS腫瘤分級系統。

2.1　接收者操作特徵曲線

2.1.1　ROC curve之建立

ROC curve是分析醫學影像診斷過程最常用的工具之一，近年來也被廣泛應用在機器學習等領域。ROC curve藉由敏感性(sensitivity)與特異性(specificity) (Altman et al., 1994)的變動，以視覺化的方式來呈現特定屬性的分類結果。

以二元分類問題為例：假設現有A種屬性、N筆資料。每一筆資料都包含A種屬性值與實際類別集合$\{0,1\}$中的一個元素，其中0代表實際類別為陰性；1代表實際類別為陽性。二元分類器在對每一筆資料進行分類之後，都會將其對應到預測類別集合$\{n, p\}$中的一個元素，其中n表示預測類別為陰性；p表示預測類別為陽性。二元分類器$D(I, C, s)$由一屬性搭配一個分類切點(cutoff point)以及分類方向所組成。其中，

I　為屬性編號，$I \in \{1,2,3...,A\}$。

C　為分類切點值

s　為分類方向，$s \in \{-1,+1\}$。

當$s=+1$時，若一筆資料中第I個屬性的值大於等於C，此分類器將此筆資料歸類為陽性，反之則陰性。當$s=-1$時，若資料中第I個屬性的值大於等於C，此分類器將此筆資料歸類陰性，反之則歸為陽性。

每一筆資料被二元分類器分類後，皆有四種可能的分類結果。假設有一筆資料實際類別為陽性，預測類別若為陽性，此分類結果稱為真陽性(true positive)；但預測結果若為陰性，此分類結果則稱為假陰性(false negative)。假設此筆資料實際類別為陰性，預測類別若為陰性，此分類結果稱為真陰性(true negative)；但預測結果

若為陽性，此分類結果則稱為假陽性(false positive)。

表 2-1 四種分類結果

		預測類別	
		陽性(Positive)	陰性(Negative)
實際類別	陽性(Positive)	真陽性(True Positive)	假陰性(False Negative)
	陰性(Negative)	假陽性(False Positive)	真陰性(True Negative)

用一個分類器對 N 筆資料做分類，會得到 N 個分類結果。為了方便之後的闡述，我們先定義真陽性的總數為 TP、假陽性的總數為 FP、真陰性的總數為 TN，與假陰性的總數為 FN，接著我們就可以計算這個分類器的真陽性率與假陽性率。

真陽性率(true positive rate)被定義為：

$$TPR = \frac{TP}{TP + FN}$$

假陽性率(false positive rate)被定義為：

$$FPR = \frac{FP}{TN + FP}$$

我們也可以用敏感性與特異性來表示此分類器對 N 筆資料分類的結果。其中，敏感性被定義為：

$$Sensitivity = TPR$$

特異性被定義為：

$$Specificity = 1 - FPR$$

當我們想要利用一個特定屬性來對 N 筆資料作二元分類時，可以藉由改變切點值 C 與分類方向 s 得到不同的分類器，進而得到不同真陽性率與假陽性率的組合。N 筆資料中所有出現的屬性值都可以當作切點值 C。以每一個切點對應的假陽性率作為 X 座標，真陽性率作為 Y 座標，得到的點座標即為此切點對應在這個屬性的 ROC curve 上的位置，將這些點以直線段相連得到的不平滑曲線即為非參數型 ROC curve(Non-parametric ROC curve，簡稱 NP-ROC)。

以 Figure 2-1 屬性 A 值與類別散佈圖的散佈圖為例，圖中縱軸表示資料類別：0 表示陰性，1 表示陽性；橫軸表示屬性 A 的值。此圖例中共有 10 筆資料，類別 1 和類別 0 各 5 筆。其中每一筆資料的屬性 A 值，都可以當作分類切點 C。在這個例子中，我們令 $s = +1$，接著在我們將分類切點 C 從此屬性的最小值，移動到最

大值的過程中，可以得到 11 組不同敏感性與特異性的組合如表 2 2 屬性 A 之 ROC curve 相關資料，將每一個假陽性率(1-特異性)作為 X 座標，敏感性最為 Y 座標，繪製於 ROC curve 上，可以得到 Figure 2-2 ROC curve 示意圖上的小正方形，將這些小正方形用線段連接起來，就可以得到非參數型的 ROC curve。

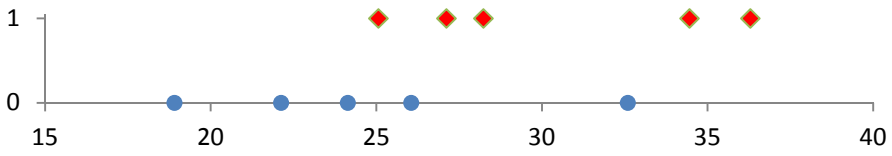


Figure 2-1 屬性 A 值與類別散佈圖

表 2-2 屬性 A 之 ROC curve 相關資料

C	Specificity	Sensitivity	FPR	TPR	C	Specificity	Sensitivity	FPR	TPR
18	0	1	1	1	28	0.8	0.6	0.2	0.6
21	0.2	1	0.8	1	32	0.8	0.4	0.2	0.4
24	0.4	1	0.6	1	34	1	0.4	0	0.4
25	0.6	1	0.4	1	36	1	0.2	0	0.2
26	0.6	0.8	0.4	0.8	∞	1	0	0	0
27	0.8	0.8	0.2	0.8					

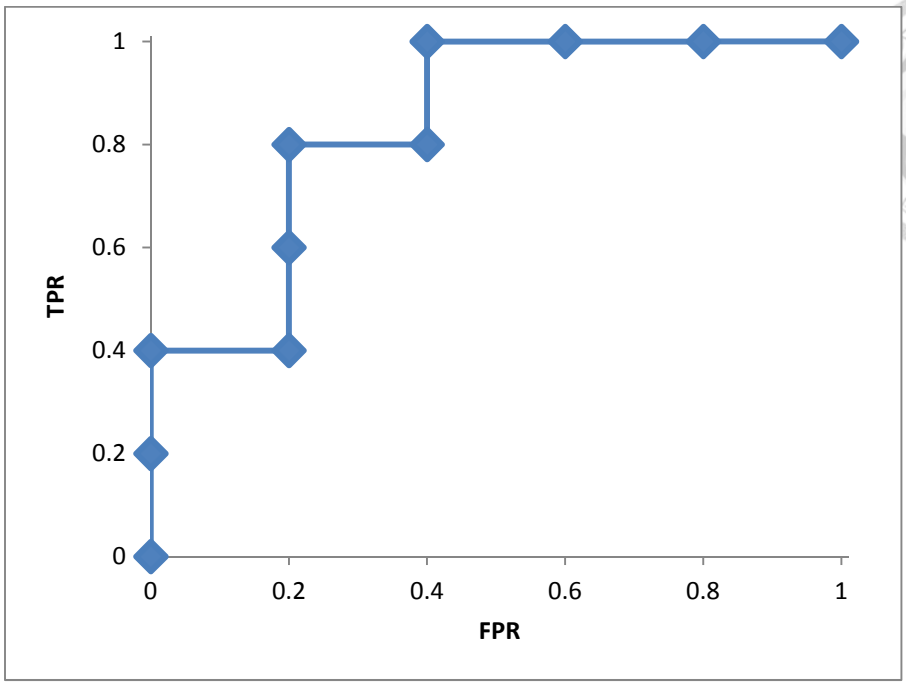


Figure 2-2 ROC curve 示意圖

2.1.2 ROC curve 之線下面積

ROC curve 的線下面積(area under ROC curve，簡稱 AUC)即為整條 ROC curve 以下的面積(如 Figure 2-3 ROC curve 整體線下面積陰影部分面積)，是用來衡量一個屬性分類能力好壞的指標，可以被視為是在 $0 \le FPR \le 1$ 範圍內的平均敏感性或被視為是在 $0 \le TPR \le 1$ 範圍內的平均特異性。一個屬性的分類能力越好，ROC curve 越偏向左上方，因此 AUC 會越大(最完美的分類器其 AUC 為 1)，而完全沒有分類能力時，ROC curve 為 45 度角斜直線（AUC 為 0.5）。當 AUC 小於 0.5 時，我們會改變屬性中分類器的分類方向，也就是改變 s 的正負號，然後重新繪製 ROC curve。在前一小節的例子中，如果我們一開始令 $s = -1$，就會畫出 AUC 小於 0.5 的 ROC curve，遇到這種情況就必須重新改變 s 的正負號，重新繪製 ROC curve 得到 Figure 2-2 ROC curve 示意圖。故 AUC 的範圍會介於 0.5 和 1 之間。

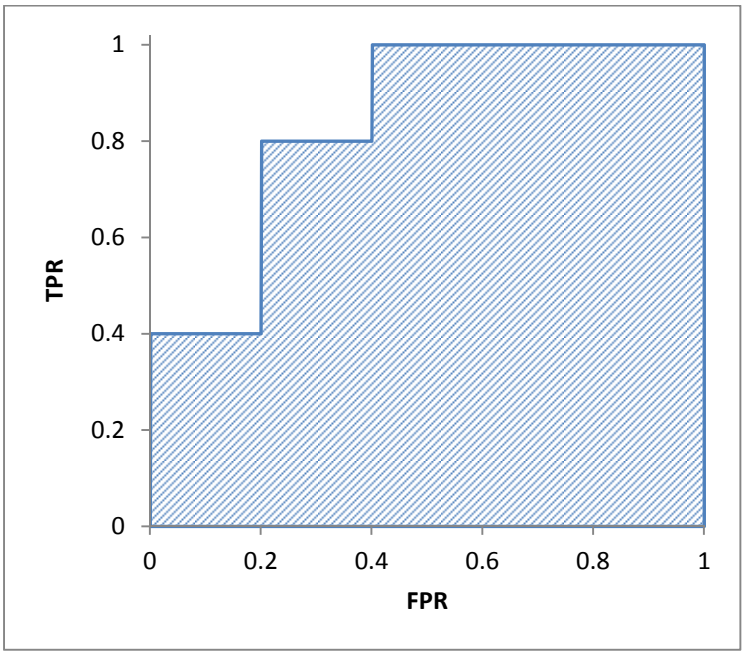


Figure 2-3 ROC curve 整體線下面積

雖然 AUC 確實是衡量一個屬性整體分類能力的有效工具，但是無法反映 ROC curve 上每一個點所傳達的資訊。即使兩條 ROC curve 的 AUC 相同或是相近，它們也可能在不同的特異性範圍，有完全不同的敏感性值。因此，若只使用 AUC 作為衡量屬性分類能力的單一指標，可能依舊無法滿足不同的分類需求。

2.1.3 參數型接收者操作特徵曲線

雖然前面介紹的 ROC 曲線讓我們了解了屬性值與資料類別的關係，但是兩者之間的關係描述卻不夠清楚。因此，為了進一步量化兩者之間的關係，可以使用一條平滑曲線來擬合經驗接收者操作特徵曲線，這條平滑的曲線稱為參數型接收者操作特徵曲線(Pepe,1997)。

最常用來找出擬合曲線的方法，是假設兩個類別的屬性值 T_0^* 和 T_1^* 分別來自不同的常態分佈如 Figure 2-4 常態假設下的兩類別分布。即

$$T_0^* \sim N(\mu_0, \sigma_0^2);\ T_1^* \sim N(\mu_1, \sigma_1^2)$$

其中，μ_0 和 μ_1 分別為兩個常態分布的平均值，σ_0^2 和 σ_1^2 分別為兩個常態分佈的變異數。令

$$a = (\mu_1 - \mu_0)/\sigma_1;\quad b = \sigma_0/\sigma_1$$

在得到 a 和 b 的估計值 $\hat{a}$ 和 $\hat{b}$ 以後，我們可以得到不同切點值 c 對應在特徵曲線上

的座標為

$$[1-\Phi(c),1-\Phi(\hat{b}c-\hat{a})]$$

其中 $-\infty < c < \infty$ ，$\Phi(x)$ 為標準常態分佈的累積機率分布函數。

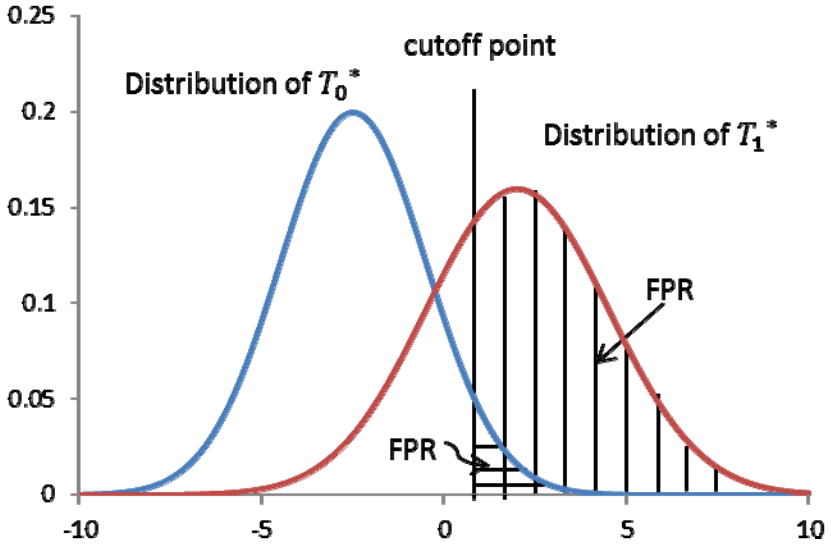


Figure 2-4 常態假設下的兩類別分布

2.2 非參數型接收者操作特徵曲線線下面積之統計檢定

2.2.1 NP-ROC 之線下面積

2.1 節介紹的 ROC 曲線不僅讓我們了解屬性值與資料類別的關係，而且說明 AUC 是衡量一個屬性整體分類能力的重要指標。為了進一步探究 AUC 的數值在分類問題內所代表的意義，本小節將說明 AUC 與 Wilcoxon Statistic 之間的關聯。

Wilcoxon Statistic 通常被用來檢測來自不同母體的量化變數的差異是否明顯。假定存在兩母體(population (A)和 population (N))，其 X_A 為來自 A 母體、X_N 為來自 N 母體。若 $\theta = prob(X_A > X_N) = 0.5$，代表 X 為無效用的區別值，即 X 無法有效地將母體的樣本區分開來。反之，若此機率越接近 1，代表 X 的區別能力越強。令 n_A 為來自母體 A 的樣本數、n_N 為來自母體 N 的樣本數，有 $n_A n_N$ 對可能的比較組合數，定義以下分數函數

$$S(x_A, x_N) = \begin{cases} 1 & if\ x_A > x_N \\ 0.5 & if\ x_A = x_N \\ 0 & if\ x_A < x_N \end{cases} \quad (0.1)$$

將所有可能的比較組合進行評分後，可得平均得分，即為

$$W = \frac{1}{n_A \square n_N} \sum_{1}^{n_A} \sum_{1}^{n_N} S(x_A, x_N) \quad (0.2)$$

於母體 A 中隨機獨立選擇一樣本 a，且於母體 N 中隨機獨立選擇一樣本 n，比較兩樣本的觀測值 X 的大小，即比較 x_a, x_n 的大小，其 $prob(x_a > x_n)$ ，就是 W 代表的意義。

根據 Figure 2-1 屬性A值與類別散佈圖屬性 值與類別散佈圖，我們可以計算屬性 A 的 Wilcoxon Statistic 值為 0.84，而根據 Figure 2-3 ROC curve 整體線下面積 ROC curve 整體線下面積也正好為 0.84。NP-ROC 之 AUC 計算方式為將該屬性的整體線下面積拆成數個矩形之後再相加，而 Wilcoxon Statistic 的計算方式為將所有可能的比較組合進行評分後所得到的平均得分，因而兩者的計算方式恰好一致，即意味著 $AUC = prob(X_A > X_N)$。

2.2.2 AUC 之無母數統計檢定

我們在2.1.1小節闡釋了NP-ROC之AUC意義，其意義即為Wilcoxon Statistic。Sen(1960)提供了一個比較兩條不同 ROC 之 AUC 差異程度的方法。對於其中一條 ROC，假設 X 為從母體C_1獨立觀察而得的區別值，Y 為從母體C_2獨立觀察而得的區別值，其中$X_i, i=1,...,m$ 為母體C_1抽取出的 m 個樣本，Y_j, j = 1,..., n 為母體C_2抽取出的 n 個樣本。2.1.1 說明其 AUC 之計算方式如下：

$$\hat{\theta}=\frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\psi\left(X_i,Y_j\right) \tag{0.3}$$

其中

$$\psi\left(X,\mathrm{Y}\right)=\begin{cases}1 & \mathrm{Y}<\mathrm{X}\\ 0.5 & \mathrm{Y}=\mathrm{X}\\ 0 & \mathrm{Y}>\mathrm{X}\end{cases} \tag{0.4}$$

接下來定義以下不同 $X-components$ 和 $Y-components$

$$\mathrm{V}_{10}(\mathrm{X}_i)=\frac{1}{n}\sum_{j=1}^{n}\psi(X_i,Y_j)\quad (i=1,2,...,m) \tag{0.5}$$

和

$$\mathrm{V}_{01}(\mathrm{Y}_j)=\frac{1}{m}\sum_{i=1}^{m}\psi(X_i,Y_j)\quad (\mathrm{j}=1,2,...,\mathrm{n}) \tag{0.6}$$

V_{10} V_{01} 分別代表每個 X_i Y_j 對於 $\hat{\theta}$ 的平均貢獻。

$$S_{10}=\frac{1}{m-1}\sum_{i=1}^{m}\left[V_{10}(X_i)-\hat{\theta}\right]\left[V_{10}(X_i)-\hat{\theta}\right] \tag{0.7}$$

$$S_{01}=\frac{1}{n-1}\sum_{j=1}^{n}\left[V_{01}(\mathrm{Y}_j)-\hat{\theta}\right]\left[V_{01}(\mathrm{Y}_j)-\hat{\theta}\right] \tag{0.8}$$

S_{10} S_{01} 分別代表 X Y 對於 $\hat{\theta}$ 的平均平方差。

對於 $\hat{\theta}$，其變異數的估計值為

$$S=\frac{1}{m}S_{10}+\frac{1}{n}S_{01} \tag{0.9}$$

結合 Sen(1960)和 Arveson(1969, Theorem 16)的結果，當 $\lim_{N\to\infty}\frac{m}{n}$ 不為零或無窮大時$(N=m+n)$ ，其 $N^{\frac{1}{2}}\left(\hat{\theta}-\theta\right)$ 將會是一漸進常態分配，其分配的平均數為 0，

變異數為σ^2，$\sigma^2 = N\left(\frac{1}{m}S_{10} + \frac{1}{n}S_{01}\right)$。

藉由V_{10}、V_{01}、S_{10}、S_{01}以及(2.9)，可以得到$Var(\hat{\theta})$的估計值，接著就可以檢定線下面積是否等於任意特定值A_0，其虛無假設與對立假設為

$$H_0 : \hat{\theta} = A_0$$

$$H_1 : \hat{\theta} \neq A_0$$

檢定值

$$z = \frac{\hat{\theta} - A_0}{\sqrt{\mathrm{Var}(\hat{\theta})}} \tag{0.10}$$

藉由V_{10}、V_{01}、S_{10}、S_{01}以及(2.9)，可以得到$Var(\hat{\theta})$的估計值，接著就可以檢定線下面積是否等於任意特定值A_0，其虛無假設與對立假設為

$$H_0 : \hat{\theta} = A_0$$

$$H_1 : \hat{\theta} \neq A_0$$

檢定值

$$z = \frac{\hat{\theta} - A_0}{\sqrt{\mathrm{Var}(\hat{\theta})}} \tag{0.11}$$

2.1.2 小節提到，ROC curve 在屬性沒有分類能力時為 45 度角斜直線，因此在做統計檢定時，通常會令A_0為 45 度角斜直線下的面積。如 Figure 2-5 線下面積檢定之示意圖，當我們希望檢定$\hat{\theta}$ 時，先會令A_0為虛線下的陰影面積，即$A_0 = 0.5$。然後，透過檢定值z，我們可以算得線下面積檢定的$p-value$值，並從$p-value$值了解該屬性線下面積的顯著程度。

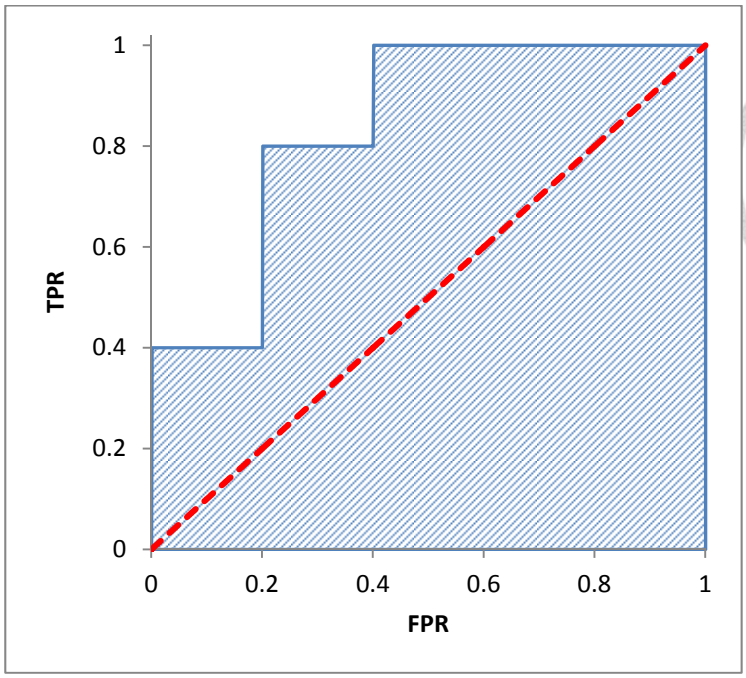


Figure 2-5 線下面積檢定之示意圖

2.3 分類樹

2.3.1 C&ART

C&ART 是透過一個序列式的二元分割所構成，其目標是讓分類樹的子節點的同質性越高越好。一個 C&ART 的模型的建構可以分成兩個階段。第一個階段是先建立一個完整的分類樹，此階段會不斷的進行二元分割直到無法再找到有用的分割為止。其方法為在母節點搜尋一個最好的屬性和切點，然後用此屬性跟切點把這個母節點分割成兩個子節點，只有子節點的樣本數太少或是子節點裡已經只剩一個類別，此子節點才停止再繼續分割，否則的話此子節點就要繼續往下進行二元分割。由於此階段在建立分類樹時會針對訓練樣本建立一個完整結構的樹，把訓練樣本完全分乾淨，但通常會造成過度配適(over-fitting)的情形，所以第二階段需要修剪掉一些分支來避免過度配適。修剪的方法是利用交叉驗證把表現不好的分支刪掉，交叉驗證會利用成本複雜(cost-complexity)函數來平衡分錯的機率跟樹的大小。

由於每次在進行分割時要搜尋一個最好的屬性跟切點，所以需要有一個分割的準則來評估屬性與切點的效能，其中較常見的準則為 Gini index。Gini index 是一種衡量節點內資料類別分布不純度(impurity)的準則，所以一個節點的 Gini index 越小，表示資料的同質性越高。當對一個節點作分割時，必須選擇一個屬性配上

一個對應的切點，利用此切點產生兩個子節點。藉由計算兩個子節點的 Gini Index 對節點樣本數的加權平均，就能得到此次分割的 Gini Index，所以每個屬性皆可搜尋一個最佳的對應切點。在進行變數選擇時，只要比較每個屬性搭配上其對應的最佳切點後的 Gini Index 即可選出此節點最佳分割的屬性與切點之搭配。

當給定一節點 t，在二元分類問題中，此節點的 Gini index 為

$$Gini(t) = 1 - P(0|t)^2 - P(1|t)^2 \tag{0.12}$$

其中 $P(k|t)$ 為類別 k 在節點 t 所佔的比例。若是給定一節點 t 搭配切點 C_i，將節點 t 分割成 t_L 和 t_R，可以計算此分割的不純度為

$$I(t, C_i) = \frac{n_L}{N} Gini(t_L) + \frac{n_R}{N} Gini(t_R) \tag{0.13}$$

其中 n_L 為節點 t_L 的樣本數；n_R 為節點 t_R 的樣本數；$N = n_L + n_R$ 為所有樣本總和。

為用 C&ART 對 Figure 2-6 模擬案例散布圖左圖的模擬資料作分類的示意圖，在原始資料進入根節點(root)時，類別 0 和類別 1 各有 50 筆，以 $(50,50)$ 表示，其中前項表示類別 0 的個數，後項表示類別 1 的個數。藉由 Gini index，C&ART 一開始選擇 X_1 作為分割的屬性：X_1 值小於 30 的 30 筆資料被分到左邊子節點，剩下 70 筆資料被分到右邊子節點。由於左邊子節點的資料類全部是 0，C&ART 會停止分割該節點。右邊的子節點則再用 X_2 分割：X_2 值小於 40 的 50 筆資料被分到左邊子節點，剩下 20 筆資料被分到右邊子節點。由於新的兩個子節點內部都只剩下一個類別，C&ART 會停止分割這兩個子節點。這些不須再被分割的節點，在樹狀結構中稱為葉子(leaf)。當樹狀結構中沒有任何未分割且非葉子的節點時，C&ART 演算法就會終止。Figure 2-6 模擬案例散布圖右圖為分割結果的示意圖。

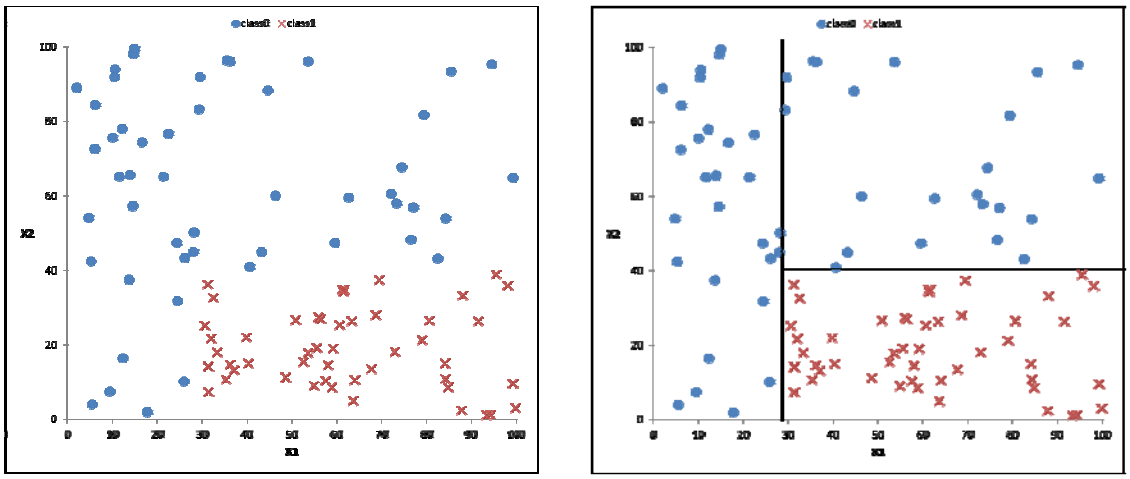


Figure 2-6 模擬案例散布圖

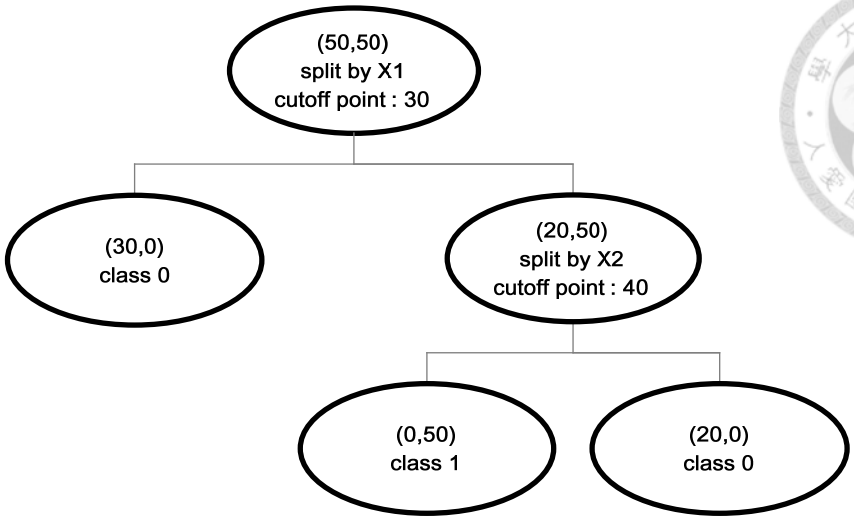


Figure 2-7 C&ART 示意圖

2.3.2 多層判別分析

在研究甲狀腺腫瘤良惡性判別時，有一些屬性只能判別某個類別，例如腫瘤裡囊腫(cyst)所占的面積只能拿來判別良性腫瘤，若腫瘤裡大部分都是囊腫，則很有可能為良性腫瘤，但若腫瘤裡面無囊腫無法代表它為惡性腫瘤(Pepe,1997)。如 Figure 2-8 內囊腫所占比例之良惡性散布圖(巫信融，2009)所示，x 軸為腫瘤內囊腫面積所占的比例，為了使散布圖方便觀察，令 y 軸為任意產生的一組大於 0 小於 1 的隨機變數。由 Figure 2-9 瘤內囊腫所占比例之良惡性散布圖(巫信融，2009)可知，腫瘤內囊腫面積所占比例超過 0.35 的腫瘤裡，有 22 個為良性，3 個為惡性，但在腫瘤內囊腫面積所占比例較小的那邊則很混雜。而腫瘤內的鈣化點所占的比例(CI)則是只能判別惡性腫瘤，若腫瘤內的鈣化點所占比例很高，則為惡性腫瘤的機率很高，但若鈣化點所占比例很低，其可能惡性腫瘤也可能為良性腫瘤，如 Figure 2-10 腫瘤周圍環狀區域血管多寡之良惡性散布圖(巫信融，2009)。還有一些屬性是能拿來判別良性腫瘤也能判別惡性腫瘤的，例如腫瘤周圍環狀區域的血管多寡(ringPDVImax)，若腫瘤周圍環狀區域的血管很多，則很有可能為良性腫瘤，若血管很少，則很有可能為惡性腫瘤。如 Figure 2-10 腫瘤周圍環狀區域血管多寡之良惡性散布圖(巫信融，2009)。

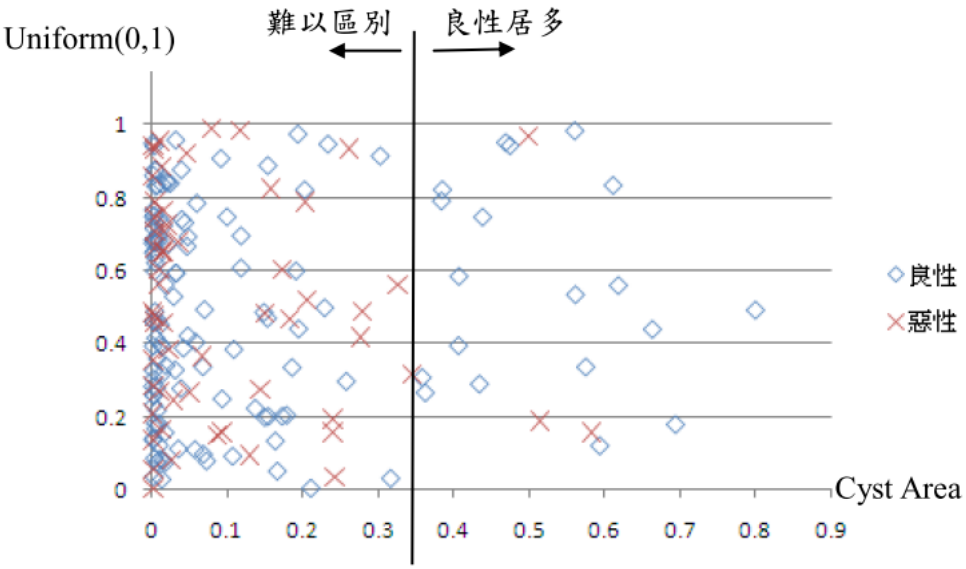


Figure 2-8 腫瘤內囊腫所占比例之良惡性散布圖(巫信融，2009)

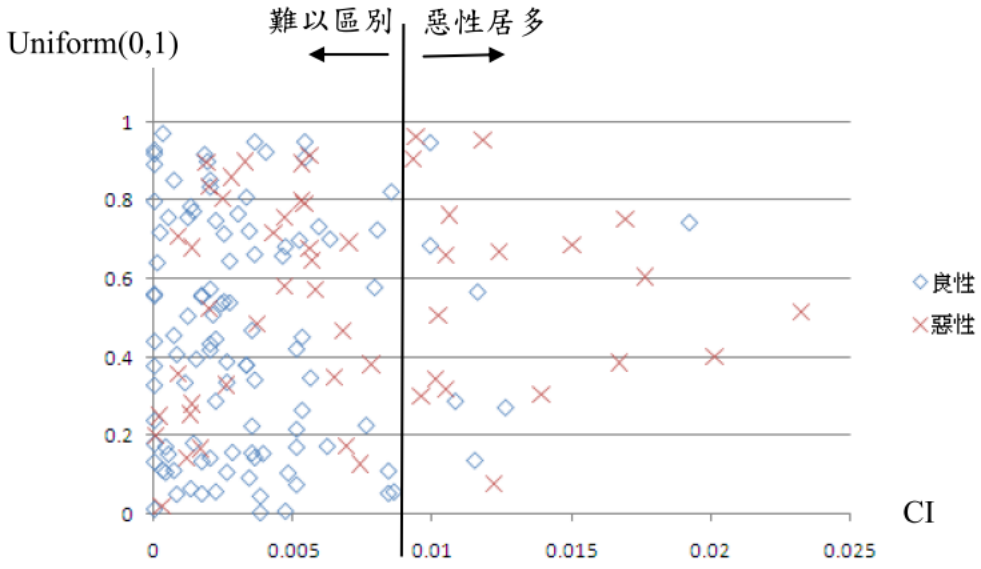


Figure 2-9 腫瘤內鈣化點所占比例之良惡性散布圖(巫信融，2009)

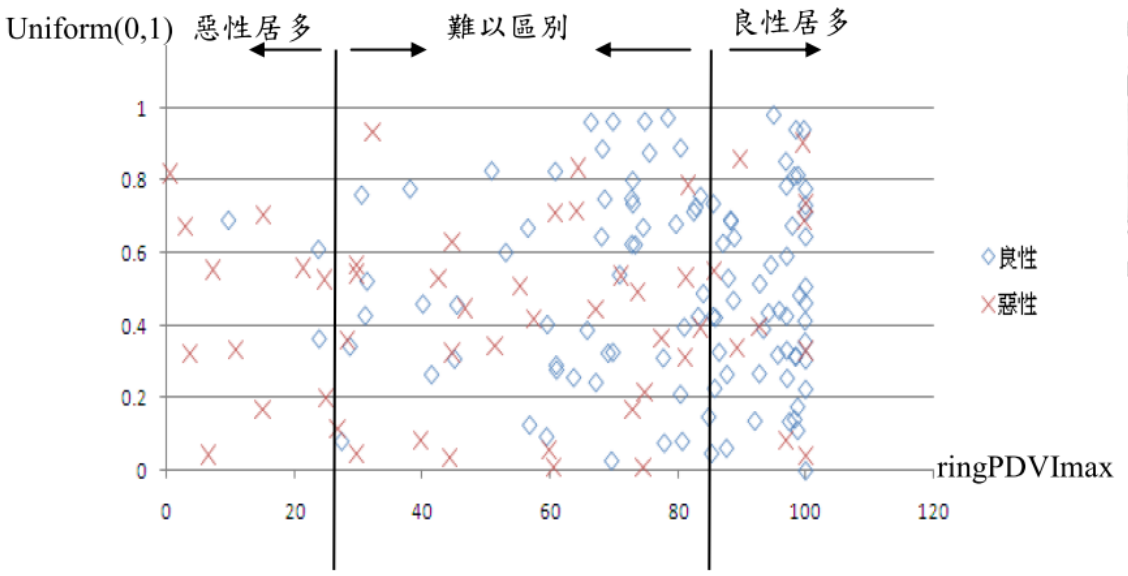


Figure 2-10 腫瘤周圍環狀區域血管多寡之良惡性散布圖(巫信融，2009)

因此在建構多層判別分析分類樹時，每一層可能只判別出惡性或良性中的一個類別或是同時判別出良性惡性，若是無法在這一層中區別出類別的樣本則會留到下一層用下一層的模型來判斷它的類別，且每一層的模型不一定只有單一屬性，可以透過費雪判別分析(Fisher Linear Discriminant，簡稱 FLD)來結合多個屬性，使其分類的效能更優於只選擇單一屬性去分類。

多層判別分析使用 Gini index 或 Wilks' lambda 選擇屬性，選擇完屬性後，必須選出二個切點將資料節分割為 t_A 、t_M ，和 t_B 三節點，因此不純度的計算調整為

$$I = \frac{N_A \times Gini(t_A) + N_M \times Gini(t_M) + N_B \times Gini(t_B)}{N_A + N_M + N_B} \tag{0.14}$$

由於我們需要同時選出兩個切點，當資料中有 N 個樣本時，則有 $N(N-1)/2$ 種可能的切點組合，當樣本數很大時，試過所有可能的切點組合會相當緩慢，因此發展出一個快速搜尋二個切點的方法。

該方法主要分為兩個階段，第一階段，先搜尋一個能將所有資料分成兩群後不純度最小的切點 C_0，然後利用 C_0 可將資料分成兩群：$Node_A$ 和 $Node_B$。第二階段，分別在此二節點中尋找能將資料分成兩群後不純度最小的切點 C_1 和 C_2 ，C_1 將 $Node_A$ 分割成 $Node_{AA}$ 和 $Node_{AB}$ ；C_2 將 $Node_B$ 分割成 $Node_{BA}$ 和 $Node_{BB}$ 。

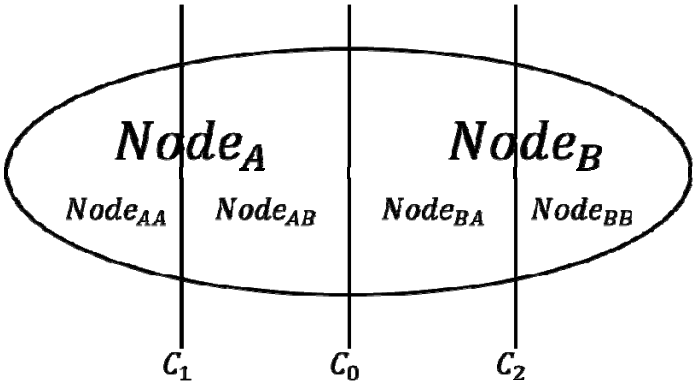


Figure 2-11 多層判別分析候選切點示意圖

得到三個候選切點C_0、C_1和C_2，利用此三候選切點可以組成(C_0,C_1)、(C_0,C_2)、(C_1,C_2)三組切點組合，用(0.14)式比較此三種切點組合將資料分成三群後的不純度，選出一組切點使不純度最小者。在尋找最佳切點組合時，希望將同性質高的樣本放在左右兩側，所以在搜尋C_1時會設下限制，用 C_1切出來的兩群資料裡，比較遠離C_0的那群資料$Node_{AA}$的不純度要比另一群資料$Node_{AB}$小，即

$$Gini(Node_{AA}) < Gini(Node_{AB}) \quad ;$$

同樣地，在搜尋C_2時也要設下相同限制，

$$Gini(Node_{BB}) < Gini(Node_{BA})$$

以此搜尋方法，只需 2N 次來尋找三個候選切點，再比較三組切點組合即可。在多層判別分析中，每次要加一個新屬性進入模型時，必須考慮四種方案：

方案一： 加入新屬性到原有資料節點中，利用 FLD 將新屬性與原有屬性做結合。

方案二：以中間未分類資料繼續分類。

方案三：以中間未分類資料結合判別 Class A 資料繼續分類。

方案四：以中間未分類資料結合判別 Class B 資料繼續分類。

分別如 Figure 2-12 多層判別分析四種分割方案所示。

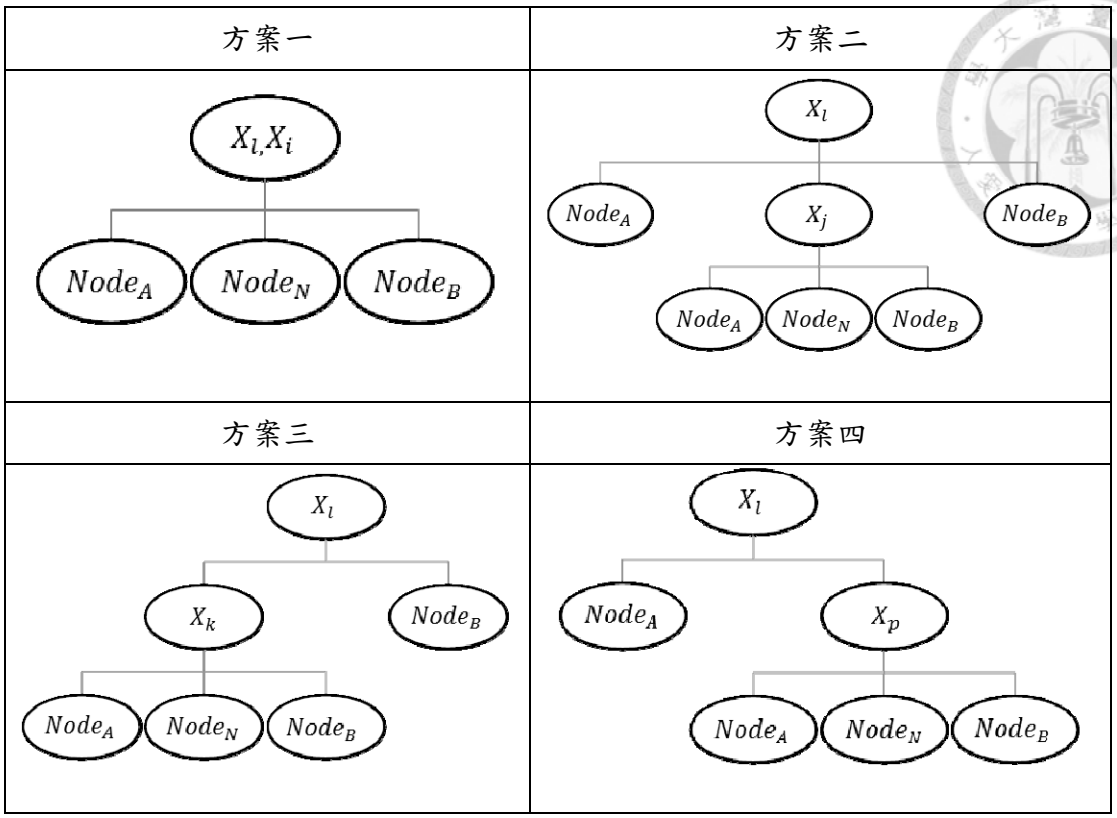


Figure 2-12 多層判別分析四種分割方案

將每種方案同類別的節點合併在一起後，就可以用(0.14)式比較四種方案的效能，選出不純度最低的方案進行分割。

為了避免模型過度拓展，在傳統方式上使用修剪方法（Pruning Procedure）解決此問題。多層判別分析中，在決定是否繼續向下分割節點時，利用選變數時提到的 Wilks'lambda 做區別檢定，以避免多元分類樹不必要的拓展，若不拒絕虛無假設，代表在剩餘的樣本中，無法找到能把不同類別顯著區分開來的屬性，因此該節點停止繼續分割。

若加入之新屬性夠顯著，再用評估效能模型中提到的方法比較整體模型效能，如無法提升效能，則停止加入新屬性。在模型的最後一層，則根據節點中的不同類別樣本比重進行強迫分類。Figure 2-13 為多層判別分析的範例。

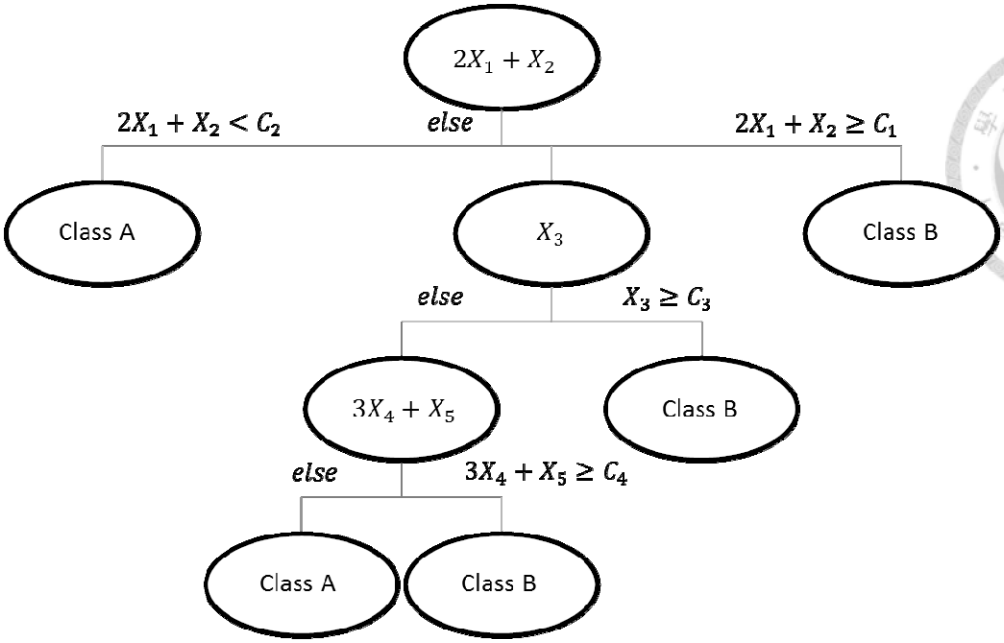


Figure 2-13 多層判別分析範例

2.3.3 C&ART 分類樹與多層判別分析分類能力之說明與比較

透過 2.3.1 和 2.3.2 小節中的介紹，我們了解了 C&ART 和多層判別分析的原理和建構過程。然而，面對實際的案例資料，C&ART 和多層判別分析互有優劣，在一定程度上還存在互補關係（賴淑俐，2010）。因而，本節將針對標竿資料，通過分別說明 C&ART 分類樹與多層判別分析之分類能力，從側面證實 C&ART 分類樹最有效率的模型與多層判別分析最有效率的模型恰好互補，即 C&ART 分類樹最有效率的模型可以補足多層判別分析無法將資料判別無誤的狀況。

本小節使用的資料類型是二類別資料，類別 A 或類別 B，假設類別由 2 個屬性 x_1 與 x_2 決定，$x_1, x_2 \overset{i.i.d}{\sim}$ 落在 0~1 的均勻分配。如 Figure 2-1，單位正方形劃分為 2 個區域Ⅰ與Ⅱ。區域Ⅰ為多邊形，由 6 個端點 $(0,0)$、$(0,1)$、$(a,1)$、$(a,0.5)$、$(1-a,0.5)$ 與 $(1-a,0)$ 所組成，類別 A 的實例(instances)均勻分布在區域Ⅰ內，其中 $0 \leq a \leq 0.5$；區域Ⅱ亦為多邊形，由 6 個端點 $(a,0.5)$、$(a,1)$、$(1,1)$、$(1,0)$、$(1-a,0)$ 與 $(1-a,0.5)$ 所組成，類別 B 的實例隨機均勻分布在區域Ⅱ內。類別 A 與類別 B 的樣本大小相同，各有 n 個實例，以 C_i 表示第 i 個實例的類別，其中 $C_i = A \forall_i = 1 \square n$；$C_i = B \forall_i = n+1 \square 2n$，亦即第 1 個到第 n 個實力分佈於區域Ⅰ內，而第 $n+1$ 個到第

$2n$ 個實例則分布於區域Ⅱ內。以 z_{ij} 表示第 i 個實例的屬性 x_j 的值，由於 x_j 的值落在 0□1，$z_{ij} \in [0,1]$；$\forall_i = 1$□ $2n$，其中 $j = 1,2$。例如 z_{11} 為第 1 個實例屬性 x_1 的值。探討的資料共有 $2n$ 個實例，以 T 表示所有實例的集合，$T = \{i : i = 1,2,\ldots,2n\}$，稱之為標竿資料。

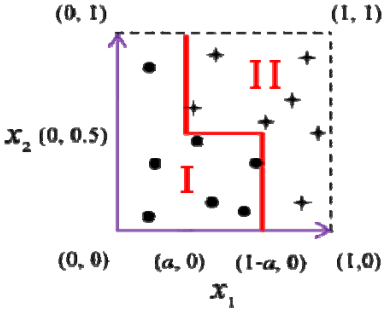


Figure 2-14 探討的資料型態分佈圖

一般認為使用 C&ART 傳統分類樹可以有效率的分類標竿資料。若 C&ART 先使用 x_2 判別資料再使用 x_1，只用兩層的分割便可將所有類別區分完成。C&ART 分類樹類別 A、B 的樣本大小相同，但是根據 a 的大小會改變標竿資料的分佈狀態，進而影響分類樹選擇的第一個屬性去做分割。因此賴淑俐學者（2010）推導並證明了性質一，具體內容如下，

性質一、 假定如標竿資料分佈之下使用分類樹，當 $0 < a \le 0.22815$ 時，會建構出如 Figure 2-17 的模型，而 a=0 時則建構出如 Figure 2-16 的模型；當 $0.22815 < a < 0.5$ 時，會建構出如 Figure 2-19 的模型， 而 a=0.5 時則建構出如 Figure 2-18 的模型。

由性質一，可以得知 C&ART 分類樹方法並非總是能有效率地分類標竿資料，如 Figure 2-15 所示，因為參數 a 的改變導致資料分佈情況不同，進而在選擇模型上產生錯誤，如此一來將影響到 C&ART 分類樹模型的成長。

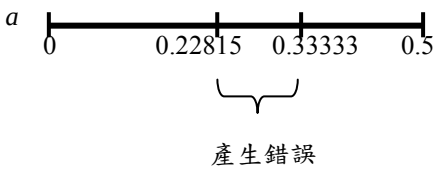


Figure 2-15 變數 a 的範圍影響模型的選擇示意圖

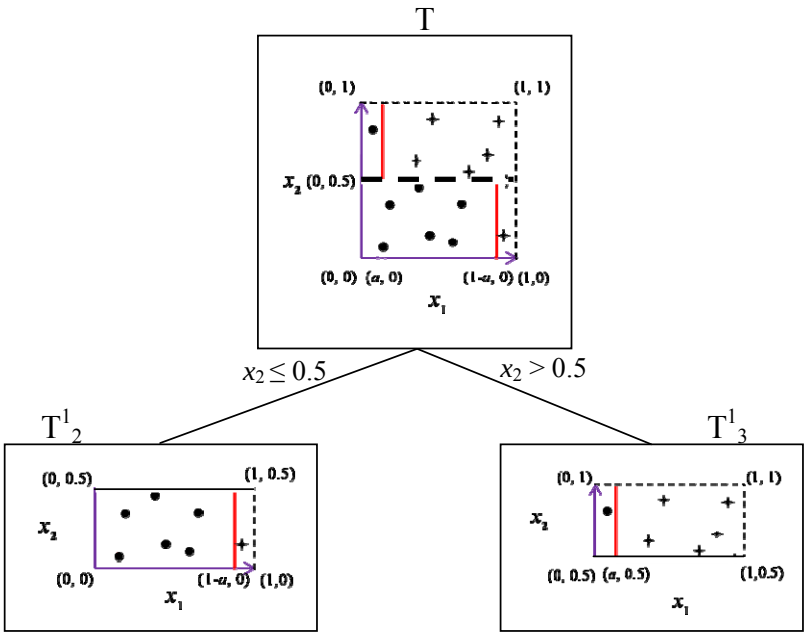


Figure 2-16 模型一，一層分類樹的建構過程，粗黑虛線為切割線

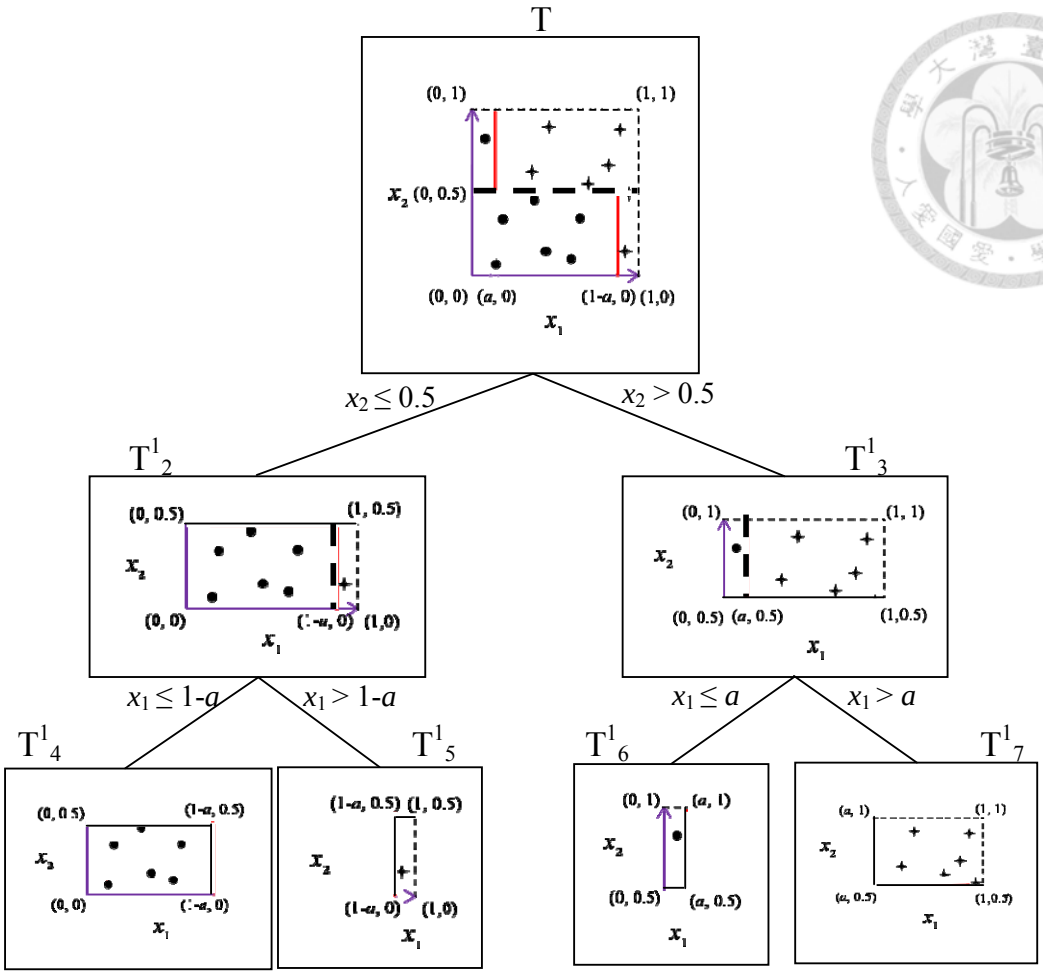


Figure 2-17 模型一，兩層的分類樹

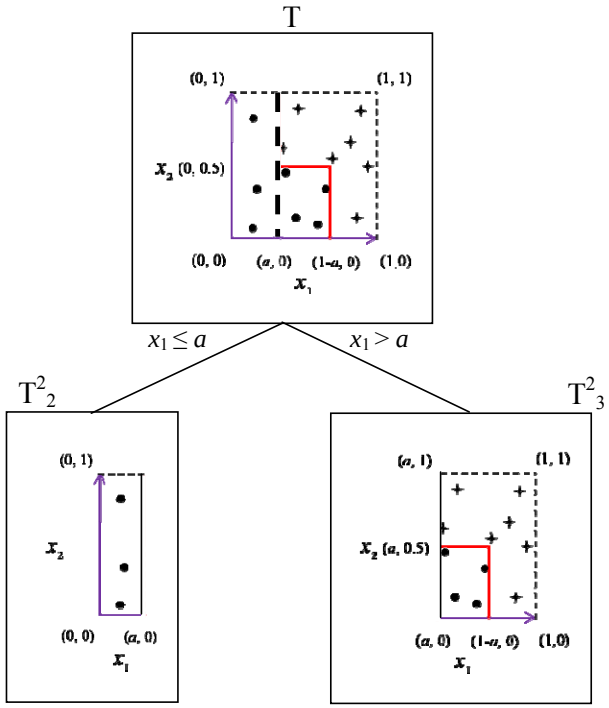


Figure 2-18 模型二，一層分類樹的建構過程，粗黑虛線為切割線

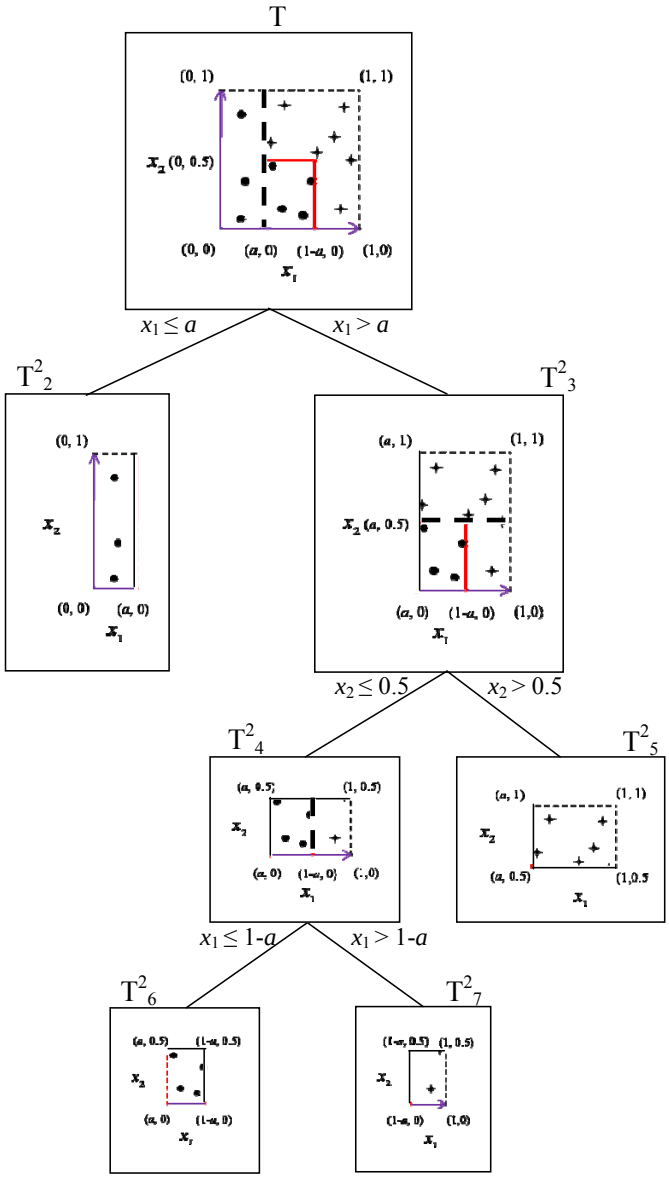


Figure 2-19 模型二，三層的分類樹

為了確保標竿資料被有效率的分類，多層判別分析必須先使用 x_1 判別資料再使用 x_2 判別。如果先使用 x_2 判別資料再使用 x_1 則會有誤判的情況發生，且沒有特定的建構模式。雖然資料類別 A、類別 B 的實例樣本大小相同，但是根據 a 的大小會改變標竿資料的分佈狀態，影響到多層判別分析選擇的第一個屬性去做分割，進而影響分類狀況。因此賴淑俐學者（2010）推導並證明了性質二，具體內容如下，

性質二、 假定標竿資料分佈之下使用多層判別分析，當 a=0 時將建構出如 Figure 2-16 的模型，而 $0 < a \le 0.19098$ 時，建構出來的模型仍存有誤判的資料；

當 0.19098< a < 0.5 時，會建構出如 Figure 2-20 模型三，a=0.5 時則建構出如 Figure 2-18 的模型。

由性質二可知，多層判別分析在 a 的影響下，將有可能選擇到 x_2 做為第一個進入演算法程序中的屬性，而無法將標竿資料分類無誤，關於此不足之處，分類樹恰好在此是選擇到效率最高的模型。

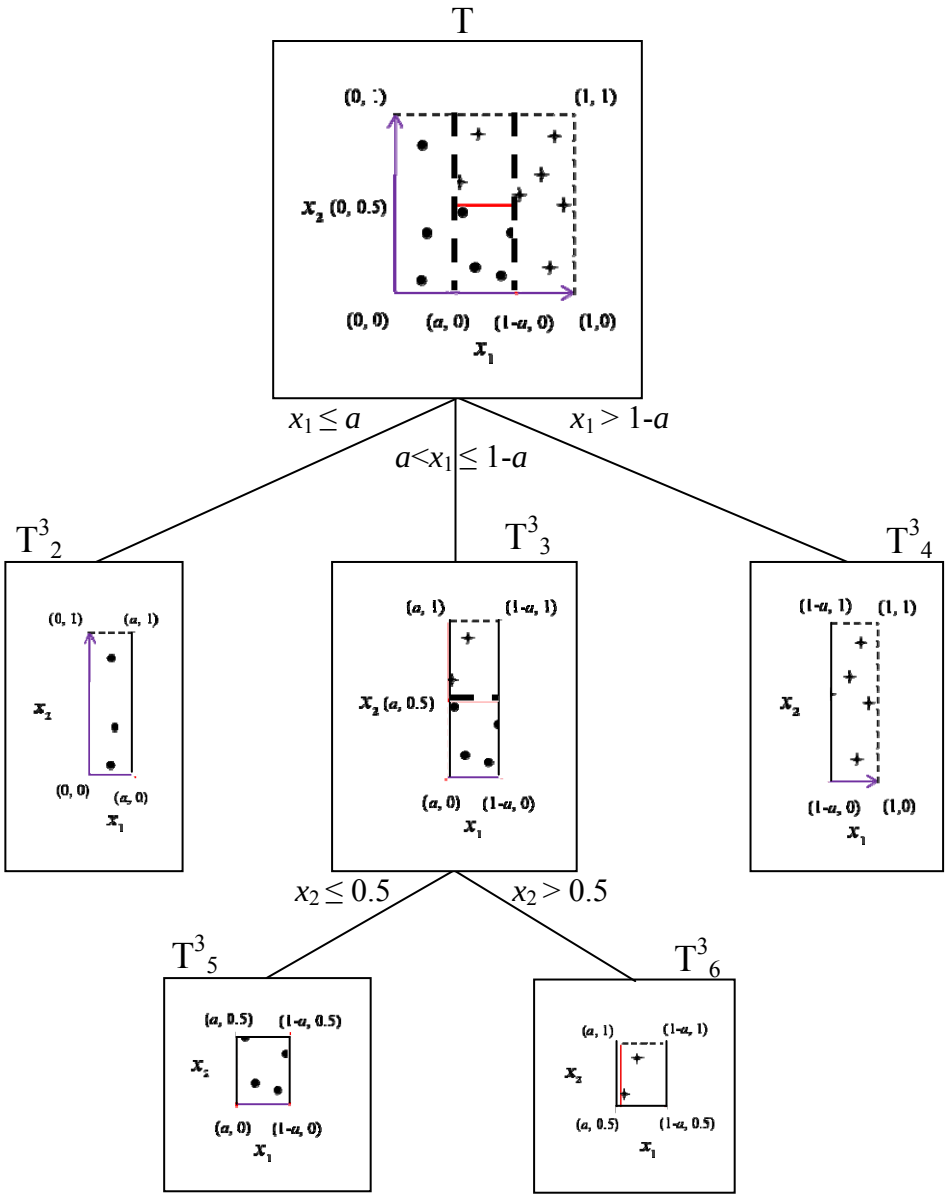


Figure 2-20 模型三，兩層多層判別分析的建構過程

2.4 費雪線性判別

2.4.1 費雪線性判別分析

費雪線性判別分析是由 R.A. Fisher 所發展的方法，其目標是尋找出一區別函數 (discriminant function)，同時最小化組內變異和最大化組間變異。幾何上的意義，FLD 把資料從原本的空間投影到一個最小化組內變異和最大化組間變異的一個方向，再使用投影過後的資料進行分類。

假設現有 N 個樣本，g 個類別，p 個屬性，每個類別有 n_k 個樣本，$N=\sum_{k=1}^{g} n_k$ 。為了要同時最小化組內變異和最大化組間變異，FLD 的做法為最大化以下此目標函數，

$$\max_{w_1} \quad J(w_1)=\frac{w_1^T S_B w_1}{w_1^T S_W w_1} \tag{0.15}$$

其中 $S_B=\sum_{k=1}^{g} n_k(\bar{x}_k-\bar{x})(\bar{x}_k-\bar{x})^T$ 為組間變異矩陣，$S_W=\sum_{k=1}^{g}\sum_{s=1}^{n_k}(\mathrm{x}_s^k-\bar{x}_k)(\mathrm{x}_s^k-\bar{x}_k)^T$ 為組內變異矩陣，x_s^k 是類別 k 的第 s 個樣本，$\bar{x}_k=\frac{1}{n_k}\sum_{l=1}^{n_k} x_l^k$ 是類別 k 的樣本平均值所組成的向量，$\bar{x}=\frac{1}{N}\sum_{i=1}^{N} x_i$ 是所有 N 個樣本的平均值所組成的向量，我們可以從(2.14)推得(2.15)

$$\begin{aligned}
\max_{w_1} \quad J(w_1) &= \frac{w_1^T S_B w_1}{w_1^T S_W w_1} \\
&= \frac{w_1^T(\sum_{k=1}^{g} n_k(\bar{x}_k-\bar{x})(\bar{x}_k-\bar{x})^T)w_1}{w_1^T(\sum_{k=1}^{g}\sum_{l=1}^{n_k}(\mathrm{x}_l^k-\bar{x}_k)(\mathrm{x}_l^k-\bar{x}_k)^T)w_1} \\
&= \frac{\sum_{k=1}^{g} n_k(w_1^T\bar{x}_k-w_1^T\bar{x})(\bar{x}_k^T w_1-\bar{x}^T w_1)}{\sum_{k=1}^{g}\sum_{l=1}^{n_k}(w_1^T\mathrm{x}_l^k-w_1^T\bar{x}_k)(\mathrm{x}_l^{kT} w_1-\bar{x}_k^T w_1)} \\
&= \frac{\sum_{k=1}^{g} n_k(\bar{d}_{1k}-\bar{d}_1)(\bar{d}_{1k}-\bar{d}_1)^T}{\sum_{k=1}^{g}\sum_{l=1}^{n_k}(\mathrm{d}_{1l}^k-\bar{d}_{1k})(\mathrm{d}_{1l}^k-\bar{d}_{1k})^T}=\frac{S_{D_1,B}}{S_{D_1,\mathrm{W}}}
\end{aligned} \tag{0.16}$$

其中 $d_{1i}=w_1^T x_i=w_{11}x_{i1}+\ldots+w_{1p}x_{ip}$ 為第 i 個樣本的第一區別分數。$d_{1s}{}^k=w_1^T x_s^k$

為第 s 個樣本在類別 k 的第一區別分數，

$\bar{d}_{1k} = \frac{1}{n_k}\sum_{l=1}^{n_k} d_l^k = \frac{1}{n_k}\sum_{l=1}^{n_k} w_1^T x_l^k = w_1^T\left[\frac{1}{n_k}\sum_{l=1}^{n_k} x_l^k\right] = w_1^T \bar{x}_k$ 為類別 k 裡第一區別分數的平均值，而 $\bar{d}_1 = \frac{1}{N}\sum_{i=1}^{N} d_{1i} = \frac{1}{N}\sum_{i=1}^{N} w_1^T x_i = w_1^T\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = w_1^T \bar{x}$

為所有 N 個樣本的第一區別分數

的平均值，$d_1 = [d_{11}, d_{12}, ..., d_{1N}]^T$ 為第一區別分數向量，

$S_{D_1,B} = \sum_{k=1}^{g} n_k(\bar{d}_{1k} - \bar{d}_1)(\bar{d}_{1k} - \bar{d}_1)^T$ 為第一區別分數的組內變異，

$S_{D_1,W} = \sum_{k=1}^{g}\sum_{l=1}^{n_k}(d_{1l}^k - \bar{d}_{1k})(d_{1l}^k - \bar{d}_{1k})^T$ 為第一區別分數的組間變異，

從(2.14)可以很明顯的看到，FLD 在找一個區別函數可以同時最大化第一區別分數的組間變異和最小化第一區別分數的組內變異。 若有三個類別以上，我們可以藉由最大化 $S_{D_2,B} / S_{D_2,W} = w_2^T S_B w_2 / w_2^T S_W w_2$ 得到第二區別函數的係數 $w_2 = [w_{21}, w_{22}, ..., w_{2p}]^T$ ，其中 w_2 必須和 w_1 正交(orthogonal)，也就是 $w_2^T w_1 = 0$ ，不同的區別函數係數都是正交的，$w_i^T w_j = 0, i \neq j$ 。

我們可以藉由解 $S_W S_B^{-1}$ 的特徵值(eigenvalue) $\lambda_1 > \lambda_2 > ... > \lambda_r > 0$ 和其對應的特徵向量(eigenvector) $w_1, w_2, ... w_r$ 來解最大化 FLD 目標函數的問題，最大特徵值對應到的特徵向量 w_1 即為第一區別函數的係數，第二大特徵值對應到的特徵向量即為第二區別函數的係數，一共會有 r 個區別函數，其中 $r = Min(p, g-1)$ 。第 j 個區別函數解釋變異的百分比為 $\frac{\lambda_j}{\sum_{i=1}^{r}\lambda_i}$ 。

2.4.2 相對重要性指標

複迴歸分析是常被使用的統計方法，在複迴歸分析中常會遇到的問題就是共線性(collinearity)，共線性是來自於變數之間的高度相關所引起。當變數間共線性程度小時，標準化迴歸係數即可代表各變數間的相對重要性(relative importance)，但當共線性程度大時，標準化迴歸係數就無法代表各變數間的相對重要性，因此文獻上使用 Dominance Index(Budescu, 1993)和 Relative Weight(Jonhson, 2000)來處理共線性的問題，以找出變數之間的相對重要性，兩種方法所得到的結果十分相似，但是 Dominance Index 比 Relative Weight 的計算負擔還要多很多。

傳統對於相對重要性的研究，著重在複迴歸分析上，但是對於其他線性模型，如：費雪線性區別分析，較少著墨。但概括性相對重要性指標(王彥龍, 2013)是把在羅吉斯迴歸分析中尋找相對重要性的方法推廣，並應用於費雪線性區別分析上，能夠找出費雪線性區別分析上各變數的相對重要性，當資料變數較多時，提供了一個指標來解決如何選擇變數的方法。

假設原始資料有 p 個屬性，n 筆資料，則屬性以矩陣表示為 $\mathbf{X}_{n*p}$，資料類別有兩類，標示為 0 或 1，以矩陣表示為 $\mathbf{Y}_{n*1}$，費雪線性區別函數為：

$$D_i = \beta_i x_{i1} + ... + \beta_j x_{ij} + ... \beta_p x_{ip},\ i = 1,...,n\ ,\ j = 1,...,p \qquad (0.17)$$

D 為費雪線性區別函數值，也可成為區別計分(Discriminant Score)，也就是各個觀察值投影到一維空間中的數值大小，類似複迴歸中的相依變數估計值 $\hat{y}$，以矩陣表示為 $\mathbf{D}_{n*1}$，$\boldsymbol{\beta}_j$ 為區別函數的係數值，$j = 1,...,\mathrm{p}$，以矩陣表示為 $\boldsymbol{\beta}_{p*1}$。

將費雪線性區別中求取概括性相對重要性指標的步驟如下：

1. 將 $\mathbf{X}_{n*p}$ 轉成 unit norm 形式
2. 將 $\mathbf{X}_{n*p}$ 做 S.V.D.，得出正交變數 $\mathbf{Z}_{n*p}$。($\mathbf{X} = \mathbf{USV}^T$ ，$\mathbf{Z} = \mathbf{UV}^T$)
3. 得出 **Z** 與 **X** 之間的迴歸係數矩陣。($\boldsymbol{\Lambda}^* = \mathbf{USU}^T$)
4. 將 **Z** 內 p 個變數皆進行標準化，使平均數 $\overline{Z_j} = 0$，變異數 $S_{Z_j}{}^2 = 1$，$j = 1,..., p$
5. 得出 **Z** 與 **Y** 之間的迴歸係數矩陣($\boldsymbol{\beta}_z$)。
6. 求出 R^2。($R^2 = cor^2(\mathrm{y},\mathrm{p})$)
7. 從 5 得到費雪線性區別計分的標準差 S_D。

8. 從 4、5、6、7 得到標準化迴歸係數矩陣$\boldsymbol{\beta}^*$。($\boldsymbol{\beta}^* = \frac{\beta_z S_z}{(S_D / R)}$)

9. 得出概括性相對重要性指標矩陣。($\boldsymbol{\varepsilon} = \boldsymbol{\Lambda}^{*2}\boldsymbol{\beta}^{*2}$)

2.5 BI-RADS 腫瘤分級系統

乳房影像報告暨資料分析系統(Breast Imaging Reporting and Data System,簡稱 BI-RADS)為美國放射醫學會(American College of Radiology,簡稱 ACR)所發展出的一套書寫報告的方式，目的在於降低乳房檢查的報告書寫的差異性，便於不同可別醫師之間進行有效的溝通。

完整的 BI-RADS 分級系統將乳房腫瘤分為九個等級，排除代表檢驗資料不齊全之 BI-RADS 0、無任何腫瘤顯像之 BI-RADS 1 以及典型良性腫瘤之 BI-RADS 2、確診為惡性腫瘤之 BI-RADS 6 以外的五個有或高或低惡性可能的級別(即 BI-RADS 3、4A、4B、4C、5)為本研究所探討的分級範圍。BI-RADS 分級系統的詳細描述如表 2-3 BI-RADS 級別與惡性機率對照表所示。

表 2-3 BI-RADS 級別與惡性機率對照表

BI-RADS 級別	3	4A	4B	4C	5
惡性機率	0~0.02	0.02~0.25	0.25~0.50	0.50~0.89	0.89~1

Chapter 3　利用 NP-ROC 建構多層混合分類樹

本章 3.1 節說明部分線下面積統計檢定方法之選擇； 3.2 節介紹建構模型的流程；3.3 節介紹模型架構；3.4 節介紹屬性評估方案的建構流程；3.5 節說明參數之功能與影響；3.6 節介紹參數之設定。

3.1　部分線下面積統計檢定方法之選擇

在實際的甲狀腺腫瘤的案例研究中，張富皓學者（2014）發現 NP-ROC 在兩點上會優於 Parametric-ROC：1. 若資料分布不為常態分布時，則擬合出的 ROC 曲線可能跟實際的情況有所出入，如 Figure 3-1。2. 當資料中有離群值存在使得 $\hat{a}$ 的估計失準時，其 AUC 也無法忠實反應屬性的分類能力，如 Figure 3-2。因而本研究將使用 NP-ROC 來建構多層混合分類樹。

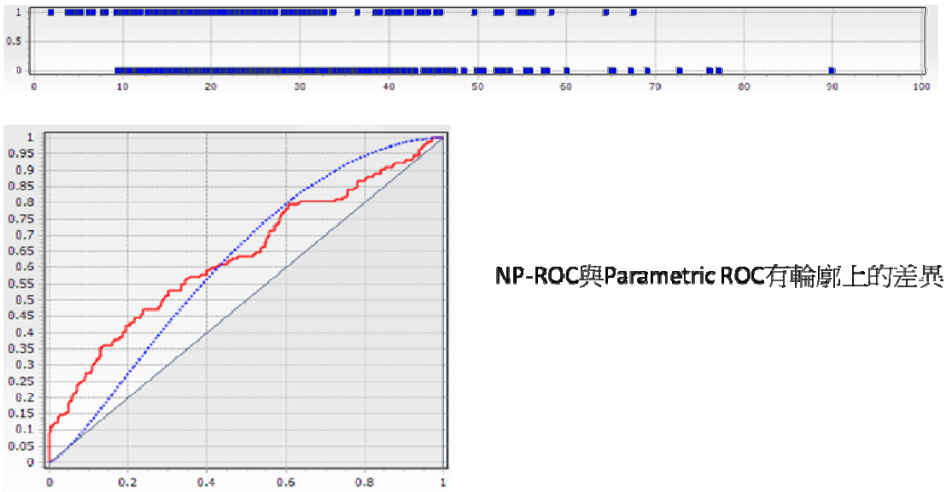


Figure 3-1 MI 的資料分布圖以及 ROC 圖

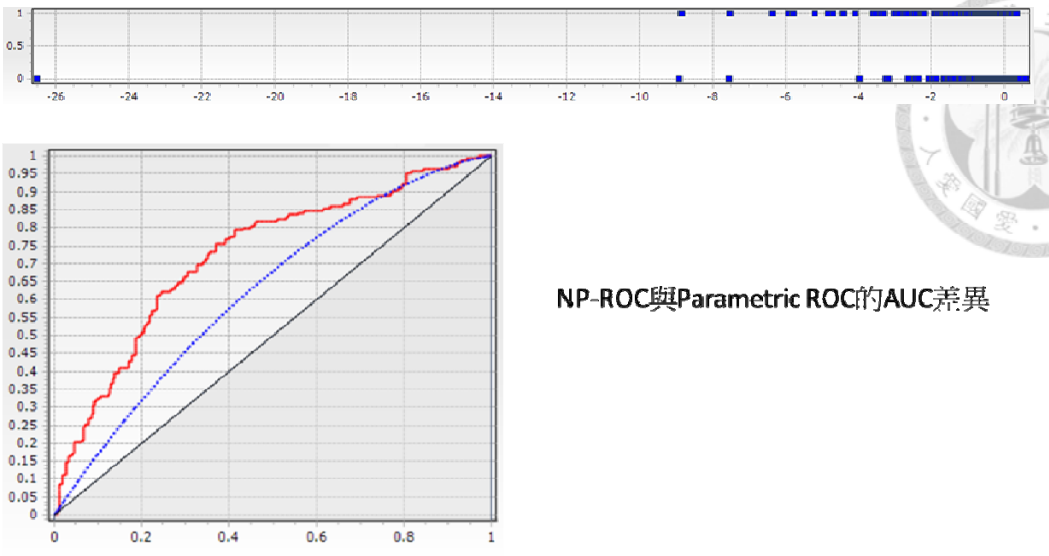


Figure 3-2 EI 的資料分布圖以及 ROC 圖

3.2 建構模型之流程

本節將分別介紹基於單一屬性和結合費雪線性判別兩種多層混合分類樹模型的建構流程。其中，3.2.1 小節將介紹基於單一屬性的多層混合分類樹模型的建構流程，而在 3.2.2 小節中將說明結合費雪線性判別的多層混合分類樹模型的建構流程。

3.2.1 基於單一屬性之建構流程

本小節所介紹的基於單一屬性的多層混合分類樹的建構流程將會以 Figure 3-3 流程圖敘述演算法的過程。

在建構多層混合分類樹模型時，演算法會判斷每一層中所有節點是否都經過「屬性評估」，當同一層中的所有節點都經過屬性評估過後，演算法才會再對下一層的節點進行「屬性評估」；在單一屬性的情況下，所謂的「屬性評估」主要是判斷該節點是否符合執行多層判別分析的條件。

在一開始進入演算法時，設定 current Layer=0，代表目前演算法判斷的層數，同時也代表目前進行「屬性評估」的目標節點就是根節點(Root)。節點一旦滿足繼續切割的條件，將會進入「屬性評估」環節，由三個部分區域線下面積檢定 p-value 的大小關係及 $P_{critical}$ 來決定該節點應該產生多層判別分析方案或 C&ART 方案。在

產生完方案之後將根據其所執行的方案決定所產生的一個或者兩個子節點需要繼續進行切割。然後，我們將會繼續判斷 current Layer 中的節點是否都已進行「屬性評估」，若都已經進行「屬性評估」，則演算法將會對下一層的節點進行「屬性評估」。當所有節點已經無法繼續進行切割或是沒有切割的必要時，基於單一屬性的多層混合分類樹模型就建構完畢。

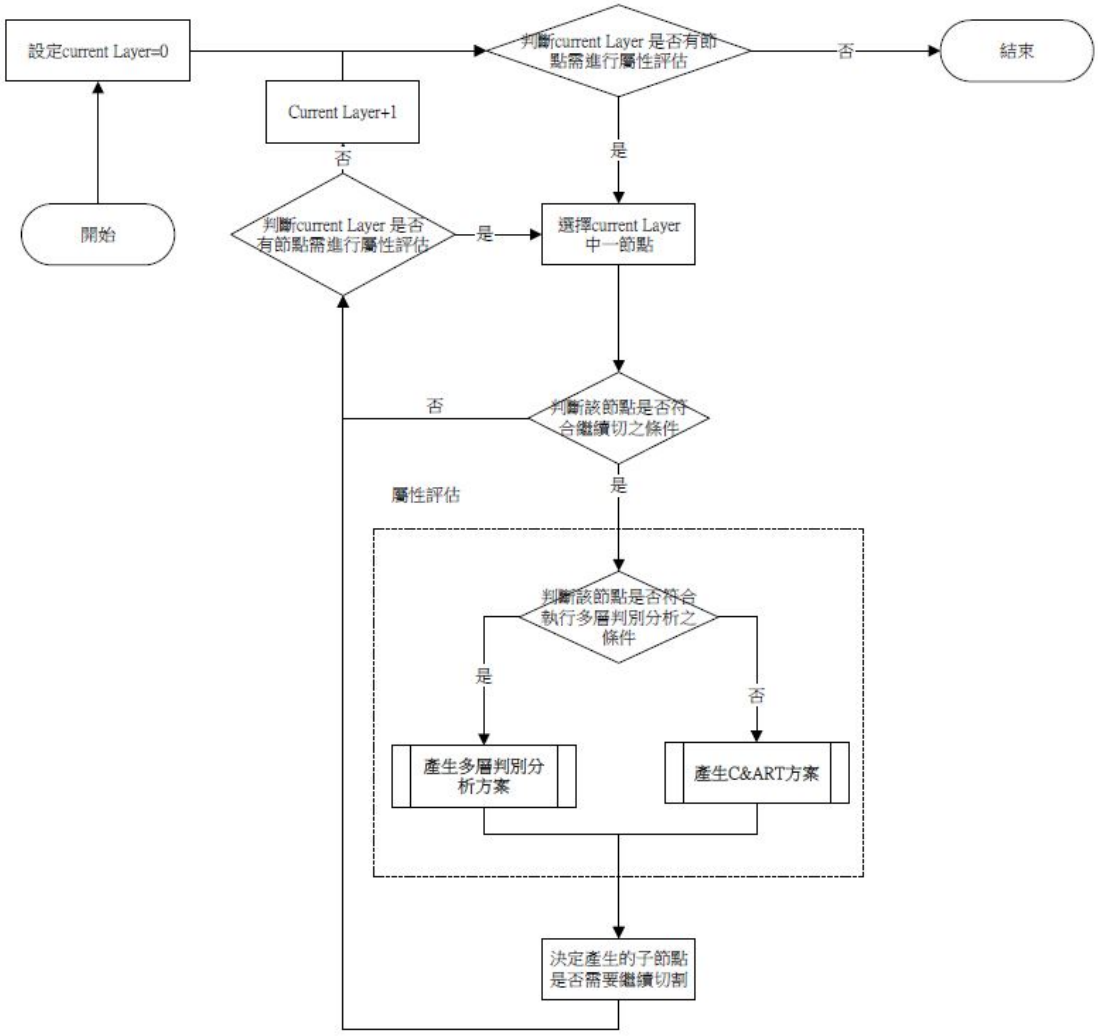


Figure 3-3 基於單一屬性的多層混合分類樹之建構流程

3.2.2 結合費雪線性判別之建構流程

小節 3.2.1 介紹了基於單一屬性的多層混合分類樹的建構流程，然而為了能同時考慮多個屬性且有較好的正確性度，本小節將介紹結合費雪線性判別的多層混合分類樹的建構流程，演算法的過程如 Figure 3-4 流程圖所示。

建構結合費雪線性判別的多層混合分類樹模型的流程跟上一小節所敘述的基於單一屬性的多層混合分類樹大致相同。只有在「屬性評估」中，我們會在單一屬性方案之外多考慮費雪線性組合屬性方案。然而，能否產生費雪線性組合屬性方案由兩個因素來決定：一是參數 P_{FLD} 的值，若 P_{FLD}=0，則多層混合分類樹就不會產生費雪線性組合屬性方案；二是費雪線性組合屬性方案的分類效果是否優於使用相同數量屬性的多層組合屬性方案。

具體執行時，結合費雪線性判別的多層混合分類樹每次對節點進行「屬性評估」前會判斷該節點是否做過切割。若是未做過切割，則先產生單一屬性方案、將單一屬性方案作為原方案。若是已做過切割，則會一直添加新的屬性進入模型，產生費雪線性組合屬性方案，直到費雪線性組合屬性方案的效能遜於在原方案基礎上建構的多層組合屬性方案時才會停止。在停止添加新的屬性之後，本演算法將會判斷選定之特徵數個數，若只有一個，則產生單一屬性方案，若有兩個或以上，則產生費雪線性組合屬性方案。本研究所建構的模型每次產生新的費雪線性組合屬性方案時，均需要對做完線性組合之後的判別分數做統計檢定並得到 p-value，然後判斷 p-value 是否小於 P_{FLD}。只有 p-value 確實小於 P_{FLD}，該方案才會順利產生。因此，P_{FLD} 的大小會影響到組合屬性方案中的費雪線性組合屬性方案的數量。

在目標節點停止考慮「屬性評估」之後，我們將會繼續判斷 current Layer 中的節點是否都已進行「屬性評估」，若都已經進行「屬性評估」，則演算法將會對下一層的節點進行「屬性評估」。當所有節點皆無法繼續進行切割或是沒有切割的必要時，多層混合分類樹模型就建構完畢。

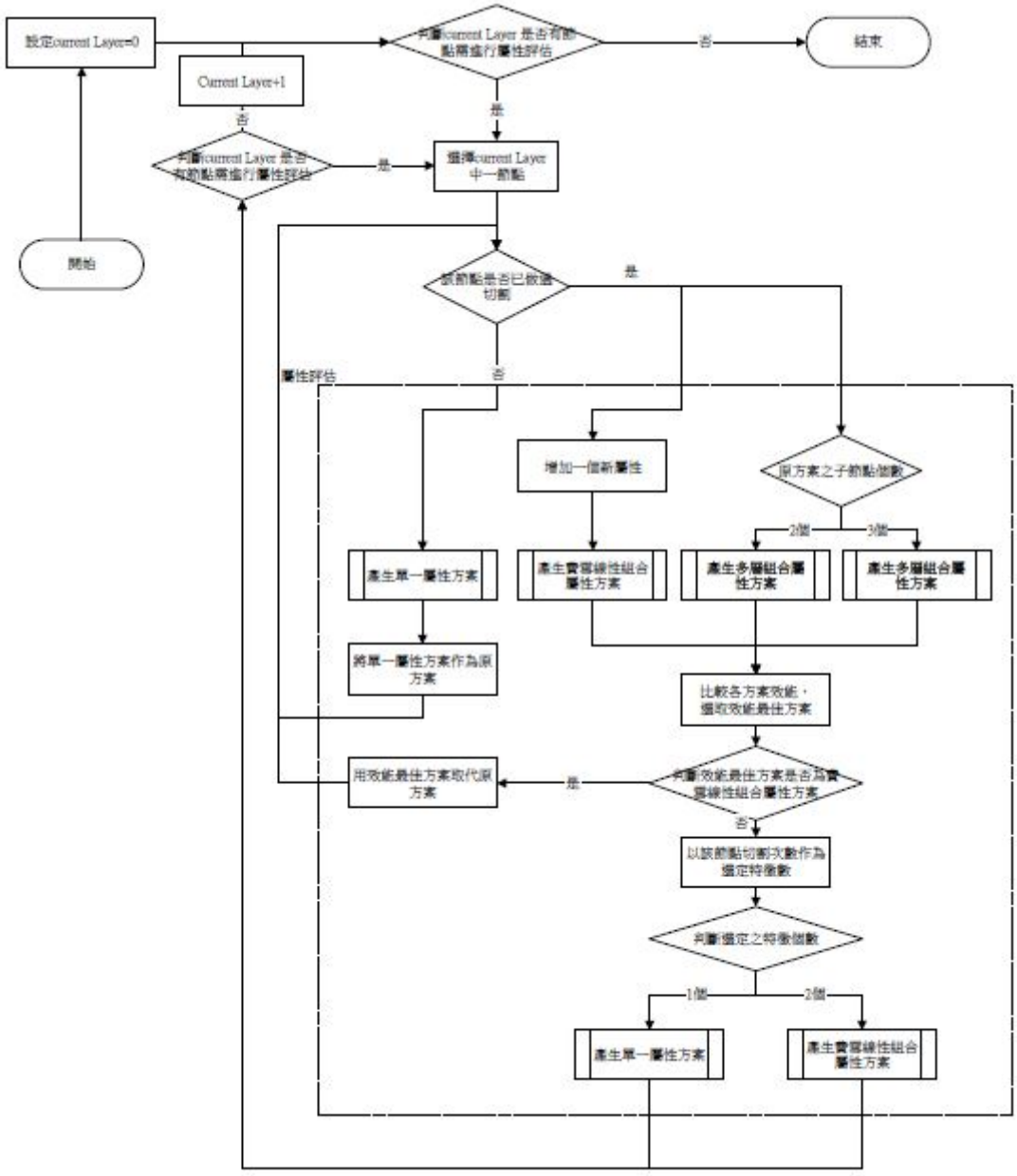


Figure 3-4 結合費雪線性判別的多層混合分類樹之建構流程

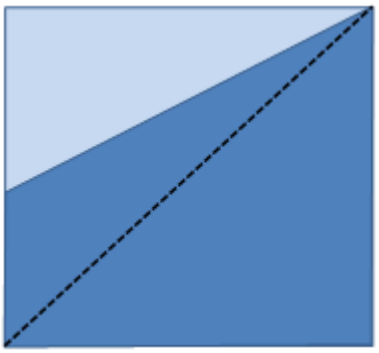
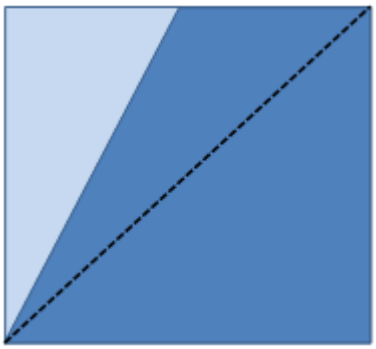
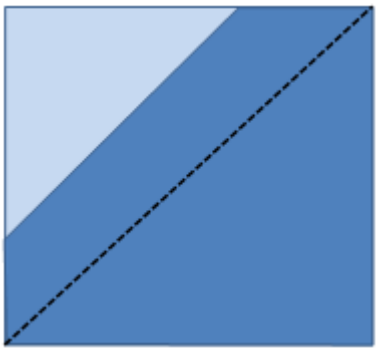
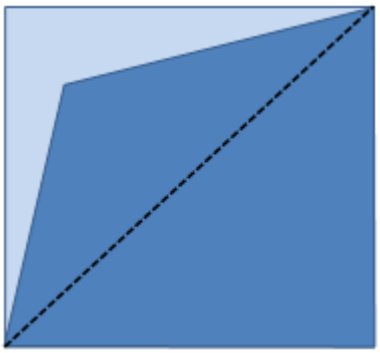
3.3 模型架構

多層混合分類樹的模型架構有點類似多層判別分析，先依據 NP-ROC 尋找到的三個候選切點將節點中的資料分割成三個不同的子資料群並分別進行AUC的統計檢定，分別得到 $P^L(i^*)$、$P^M(i^*)$ 和 $P^R(i^*)$，再根據三個線下面積檢定值的大小順序將資料形態定義為四種不同的情況，即 Left、Right、Both 和 Middle，最後依據判定的情況選擇不同的切點並決定子節點是否需要繼續切割。然而，由表 3-1 可知，其不同於多層判別分析之處在於界定 Left、Right 和 Both 三種情況時，同時要

求切點所對應的 p-value 要小於 $P_{critical}$，若是條件得不到滿足，則自動歸類為 Middle 情況並執行 C&ART 機制，即將母節點分割成一個類別 0 較多的子節點和一個類別 1 較多的子節點，兩個子節點皆可留置下一層繼續做切割。因此，多層混合分類樹的每一層不僅可以同時執行 C&ART 分類樹和多層判別分析，而且可以透過參數 $P_{critical}$ 的設置來調整兩者的相對比重。

同多層判別分析一樣，多層混合分類樹每次要加入一個新的屬性時，會考慮整體模型的效能來決定要在原有的層內結合新的屬性讓原本那一層的判斷力更佳還是要加一個新的屬性來對那些尚未分類出來的樣本進行分類、不斷的在此模型中加入新的屬性直到此模型達到停止條件。

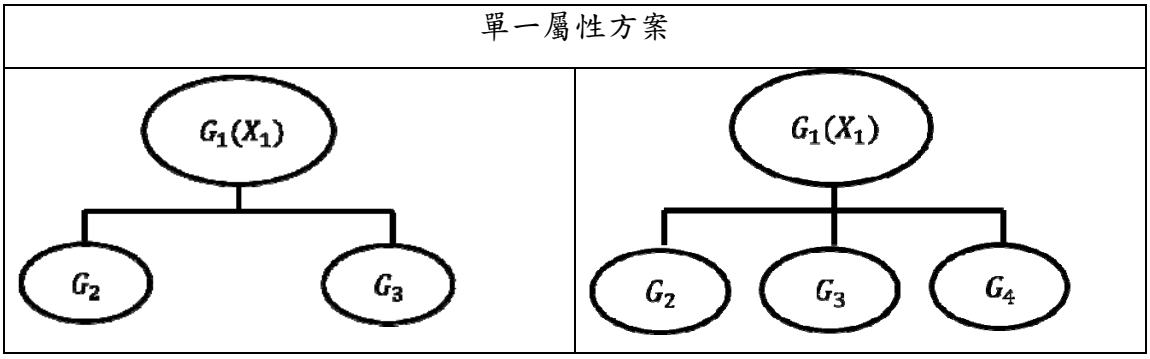
表 3-1 不同的資料形態類型所對應的理想 ROC curve、分割方案條件式和對應分類樹機制

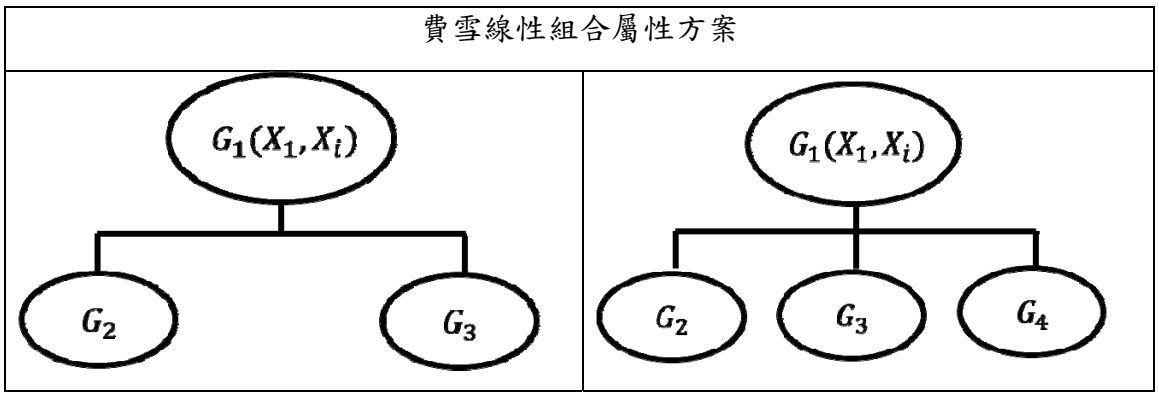
資料形態	理想的 ROC curve	部分線下面積檢定的關係	分類樹機制
Left		條件式：$P^L(i^*) < P^M(i^*) \leq P^R(i^*)$ && $P^L(i^*)$小於$P_{critical}$	多層判別分析
Right		條件式：$P^R(i^*) < P^M(i^*) \leq P^L(i^*)$ && $P^R(i^*)$小於$P_{critical}$	多層判別分析
Both		條件式：$P^L(i^*) < P^M(i^*)$ && $P^R(i^*) < P^M(i^*)$ && $P^L(i^*) < P_{critical}$ && $P^R(i^*) < P_{critical}$	多層判別分析
Middle		非以上三種方案的任一種	C&ART 機制

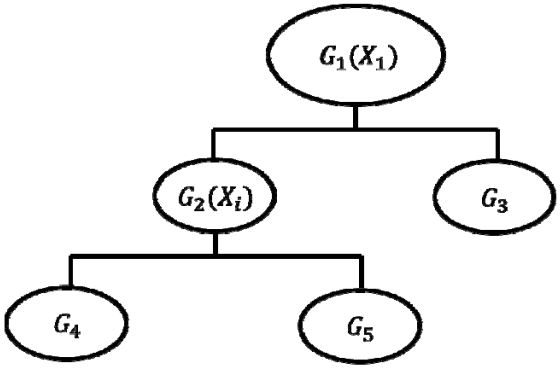
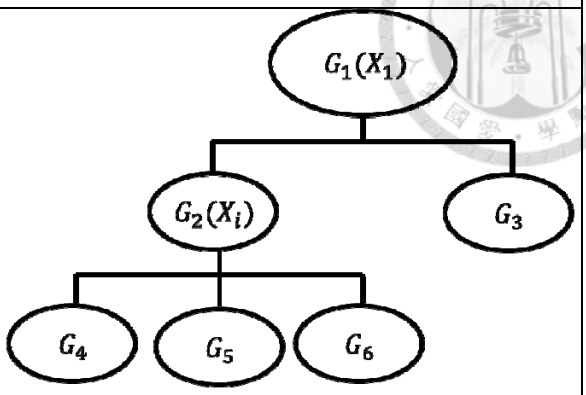
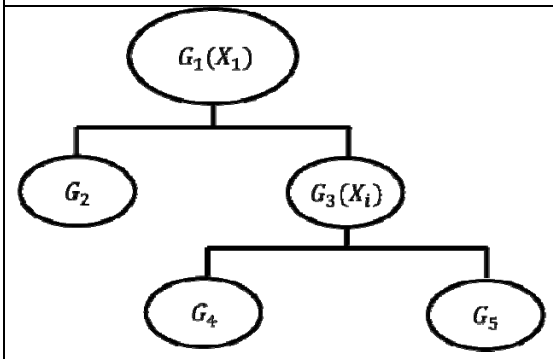
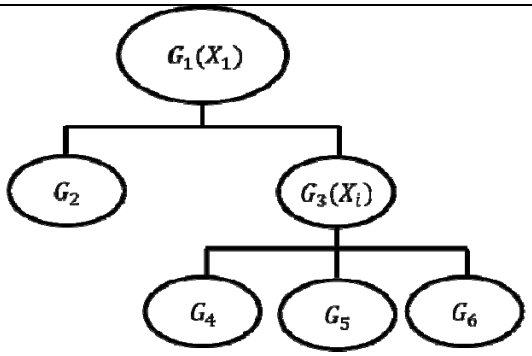
3.4 主要屬性評估方案建構流程

結合費雪線性判別的多層混合分類樹的屬性評估方案包含單一屬性方案以及組合屬性方案兩種不同的類型，而基於單一屬性的多層混合分類樹中僅包含單一屬性方案。其中，單一屬性方案是指一次僅使用一個屬性進行「屬性分類能力評估」以及「尋找屬性值切點」，而組合屬性方案通常使用兩種或兩種以上屬性進行「屬性分類能力評估」以及「尋找屬性值切點」。組合屬性方案可繼續分為費雪線性組合屬性方案以及多層組合屬性方案，其差異在於費雪線性組合屬性方案在進行「屬性分類能力評估」、「尋找屬性值切點」前會使用 FLD 對多屬性進行線性組合，而多層組合屬性方案是在現有模型中，針對子節點進行單一屬性的「屬性分類能力評估」、「尋找屬性值切點」。

Figure 3-5 為各屬性評估方案的示意圖，$G_i, i=1,2,...$ 代表節點名稱，$X_j, j=1,2,...$ 代表不同屬性，而 $G_i(X_j)$ 代表第 i 個節點使用第 j 個屬性進行切割(這裡為了方便說明而先省略屬性值切點)，$G_i(X_j, X_k)$ 代表第 i 個節點使用第 j 個和第 k 個屬性的線性組合進行切割。





多層組合屬性方案	
 $G_1(X_1)$ $G_2(X_i)$ G_3 G_4 G_5	 $G_1(X_1)$ $G_2(X_i)$ G_3 G_4 G_5 G_6
 $G_1(X_1)$ G_2 $G_3(X_i)$ G_4 G_5	 $G_1(X_1)$ G_2 $G_3(X_i)$ G_4 G_5 G_6

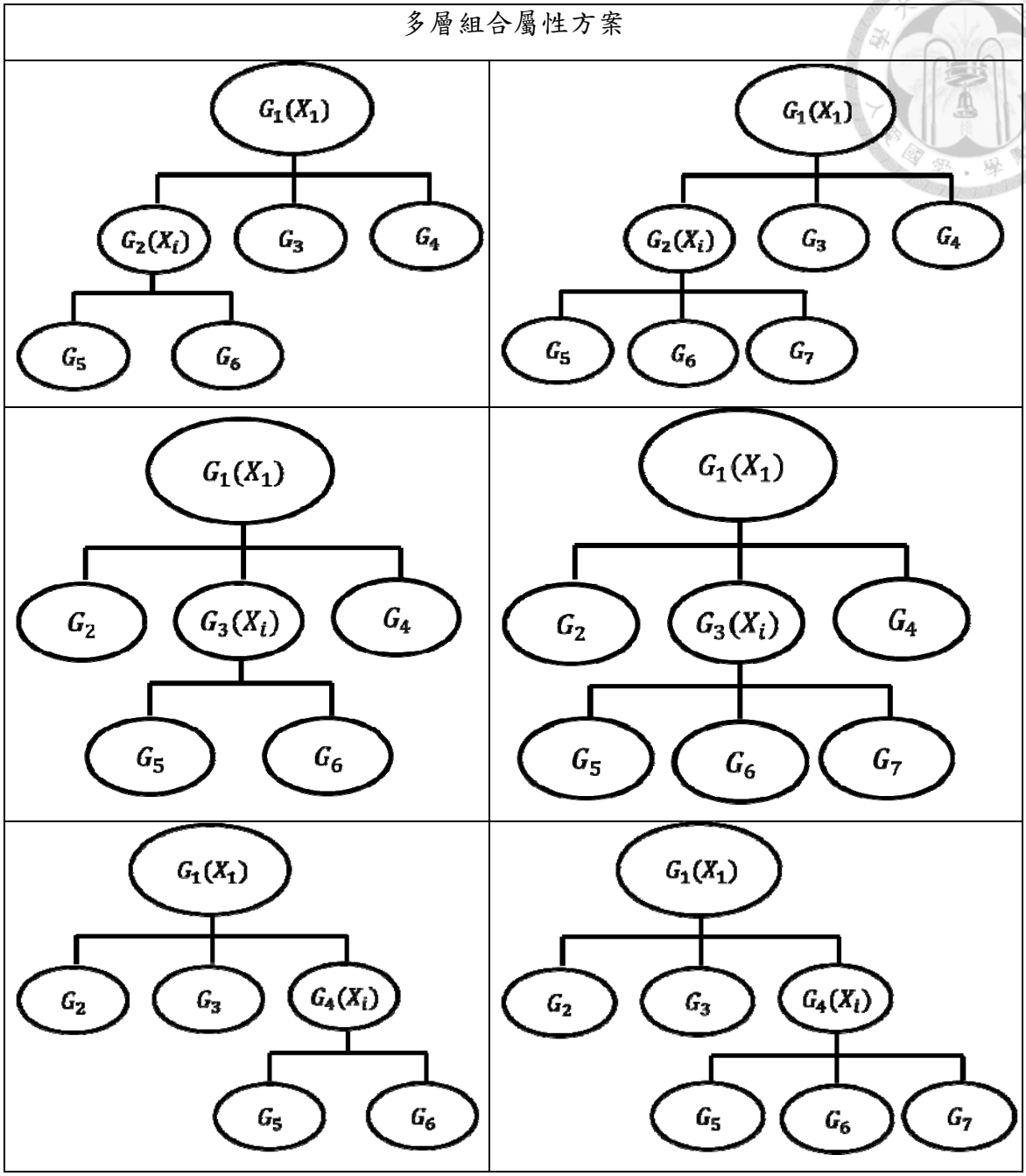


Figure 3-5 各種方案示意圖

每種方案皆會經歷「屬性分類能力評估」以及「尋找屬性值切點」。3.4.1 節主要介紹單一屬性方案的建構流程，而需要結合相對重要指標來選擇組合屬性的費雪線性組合屬性方案則會在 3.4.2 節中介紹。

3.4.1 單一屬性及多層組合屬性方案建構流程

3.4 節提及了許多屬性評估方案，然而無論是單一屬性方案還是組合屬性方案皆需要進行「屬性分類能力評估」以及「決定屬性切點值」，不同之處僅在於多層組合屬性方案在決定分割節點時不受 $P_{critical}$ 的影響。本節將會介紹建構屬性評估方案的流程，其中包含利用 NP-ROC 尋找候選切點、屬性分類能力評估、決定屬性切點值和停止建構條件。Figure 3-6 為建構屬性評估方案的流程圖，此流程圖主要適用於「單一屬性方案」。

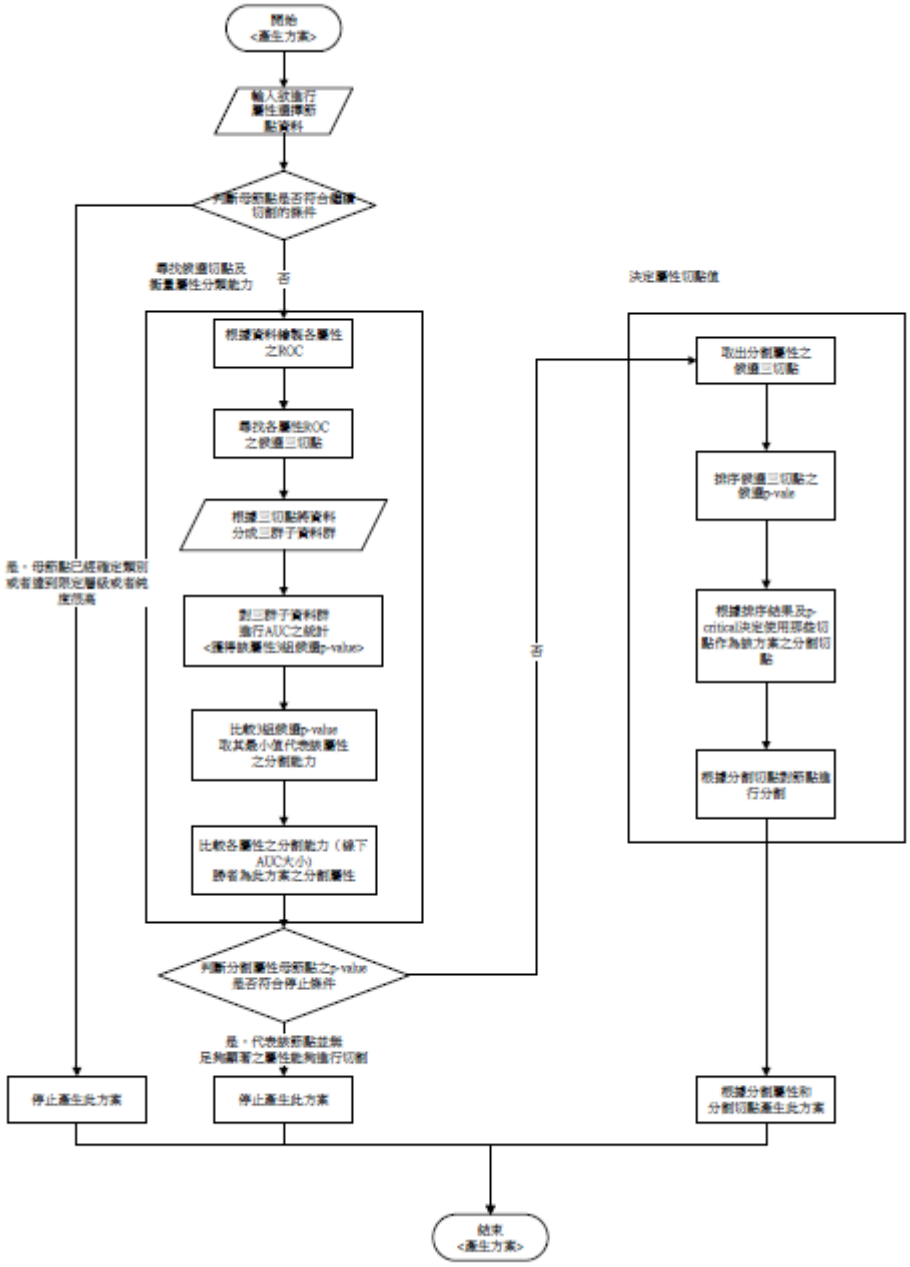


Figure 3-6 屬性評估方案建構流程圖

（1） 利用 NP-ROC 決定候選切點

本研究所提出的多層混合分類樹在分割節點時，先利用 NP-ROC 和 Youden's index(Youden, 1950)，尋找左切點、右切點和中切點三個候選切點，再利用 NP-ROC

的線下面積檢定值及$P_{critical}$從多層判別分析和 C&ART 中選擇所應執行的分類樹機制，最後再利用相應的切點對資料進行分割。

令一屬性共有 N 筆資料，其中N_0筆資料為類別 0，N_1筆資料為類別 1。定義C_i為屬性中由小到大排序後，第i小的屬性值，$i=1,2,...,N$，$C_1 \le C_2 \le ... \le C_N$。又$D_i$為屬性值由小到大排序後，第$i$小的屬性值真實類別。若$C_i$的類別為 0，則$D_i=0$，若$C_i$的類別為 1，$D_i=1$。當切點值為$C_i$時，

其敏感性

$$sensitivity_i = \frac{\sum_{n=i}^{N} D_n}{\sum_{n=1}^{N} D_n} \tag{0.18}$$

特異性

$$specificity_i = \frac{\sum_{n=1}^{i-1} (1-D_n)}{\sum_{n=1}^{N} (1-D_n)} \tag{0.19}$$

之後計算當切點值為屬性中第 i 小的值時，也就是切點值為C_i情況下的 Youden's index

$$YD_i = sensitivity_i + specificity_i - 1 \tag{0.20}$$

當一個特定切點分類的結果越好，其敏感性與特異性會越大，因而 Youden's index 也會越大。Youden's index 最大可以達到 1，此時類別 0 跟類別 1 將被完全區分開。

本研究引用「利用參數型接收者操作特徵曲線建構分類樹之研究與應用」(馬康恆，2013)中提到的找尋候選切點的 Initial cutoff point method(以後簡稱 ICP)：

ICP 之步驟：

1. 尋找 ROC 上 Youden's index 最大值，令其為候選切點R_M。
2. 計算(0,0)至R_M之斜率r_1；計算R_M至 (1,1)之斜率r_2。
3. 尋找一斜率為r_1，通過 NP-ROC，且截距項最大之直線，該直線與 NP-ROC 之交點為R_L。

4. 尋找一斜率為 r_2，通過 NP-ROC，且截距項最大之直線，該直線與 NP-ROC 之交點為 R_R。

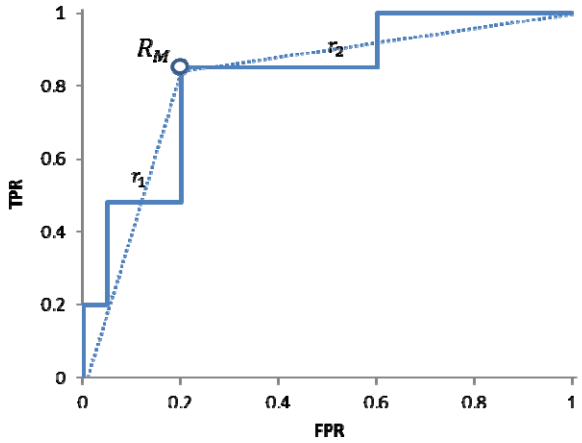


Figure 3-7 ICP 步驟 2 的示意圖

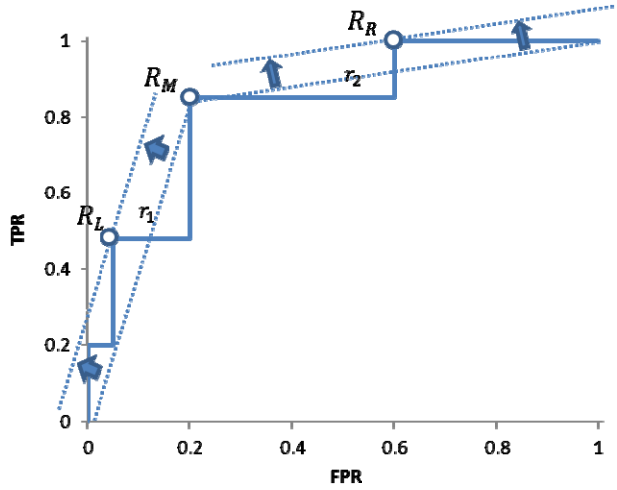


Figure 3-8 ICP 步驟 3、4 的示意圖

ICP 在實際的資料上意義為， 尋找 R_M 所對應到的切點值 C_M 只考慮屬性值內大於 C_M 的子資料群，以此子資料群繪製之 ROC 的最大 Youden's index 所對應到的切點值為 R_L；考慮屬性值內小於 C_M 的子資料群，以此子資料群繪製之 ROC 的最大 Youden's index 所對應到的切點值為 R_R。

（2） 根據 AUC 之統計檢定結果衡量屬性分類能力

在本小節利用 NP-ROC 決定候選切點的內容中，我們介紹了 ICP 方法，接下來我們會利用這三候選切點來衡量一個屬性的分類能力。通過 2.1.2 小節的介紹，我們了解到一個屬性的整體分類能力越好則 AUC 將會越大，然而由 Figure 3-9 可知單一類別分類能力越強，該 ROC 將會偏向 FPR=0 或是 TPR=1。因此，我們將會利用上述兩個特性來選取不同的資料群進行統計檢定。

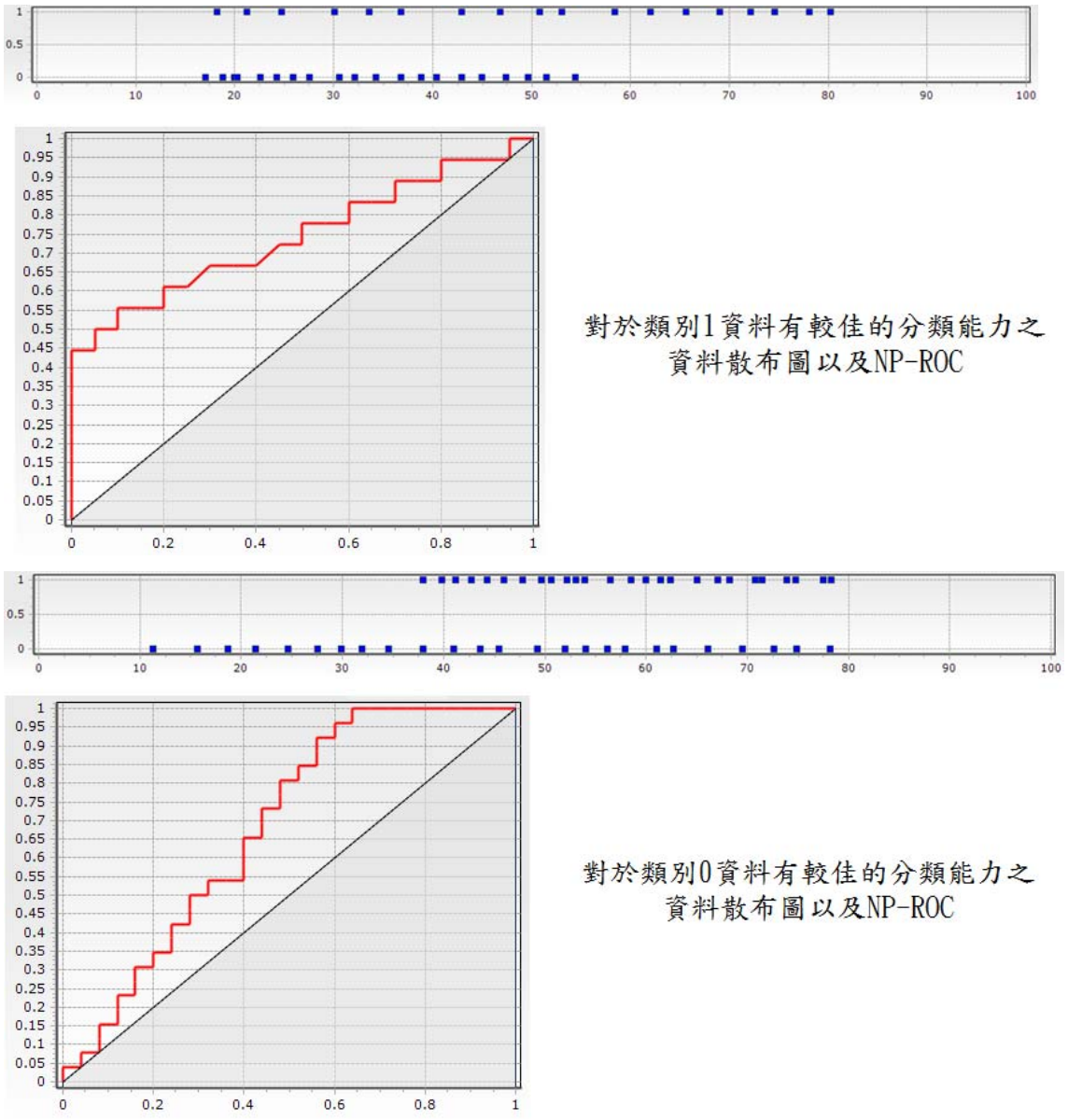


Figure 3-9 對某一類別資料具有較佳分類能力之資料散布圖以及 NP-ROC

令一屬性共有 N 筆資料，其中 N_0 筆資料為類別 0，N_1 筆資料為類別 1。利用 ICP 方法所尋找的三個候選切點 R_M、R_L、R_R，我們從 N 筆資料中挑選出 ROC curve 上 R_M 以左的資料群作為代表該屬性使用 R_L 作為切點的分類能力依據；挑選出 ROC curve 上 R_M 以右的資料群作為代表該屬性使用 R_M 作為切點的分類能力依據；挑選出 ROC curve 上介於 R_L 和 R_M 之間的資料群作為代表該屬性使用 R_M 作為切點的分類能力依據。 (如 Figure 3-10 所示)

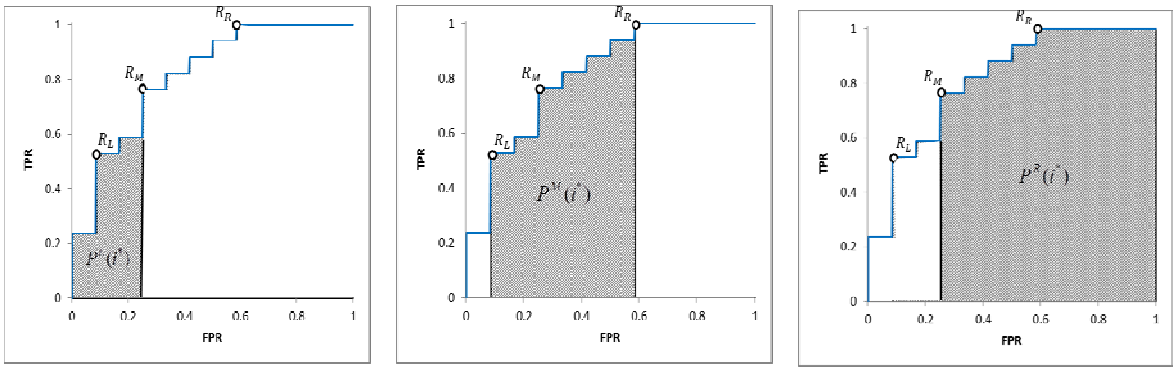


Figure 3-10 以三候選切點選擇不同資料群示意圖

基於以上提及的三子資料群，我們可以分別繪製三條新的 NP-ROC，並按照 2.2.2 小節提到的流程對其 AUC 分別進行統計檢定，將檢定結果之 p-value 分別定義為 $P^L(i)$、$P^M(i)$、$P^R(i)$，i 代表為第 i 個屬性，我們可以求得 $P^*(i)=Min(P^L(i),P^M(i),P^R(i))$，$P^*(i)$ 為第 i 個代表屬性的代表指標，接下來比較每個屬性的 P^*，選擇 P^* 最小的屬性作為該方案的分割屬性 $i^*=\arg\min\{P^*(i)\}$。

（3） 決定方案屬性值切點

本小節根據 AUC 之統計檢定結果衡量屬性分類能力中決定了方案的分割屬性，接著說明如何選擇應該執行什麼分類樹機制和決定對應的切點。在多層混合分類樹中，切割資料主要有四種切割方式，其中前三種切割方式屬於多層判別分析，而最後一種分割方式屬於 C&ART 分類樹，以下為四種切割方式的具體介紹：

A. 選擇多層判別分析，使用一切點，將資料切割成類別 0 和難以判別的子節點。

B. 選擇多層判別分析，使用一切點，將資料切割成類別 1 和難以判別的子節點。

C. 選擇多層判別分析，使用兩切點，將資料切割成類別 0、類別 1 和難以判別的子節點。

D. 選擇 C&ART 分類樹，使用一切點，將資料切割成類別 0 樣本數較多和類別 1 樣本數較多的子節點。

而我們可以根據 $P^L(i^*)$ 、 $P^M(i^*)$ 、 $P^R(i^*)$ 的大小關係及 $P_{critical}$ 來決定使用上述何種切割方式，而大小關係有以下六種：

(1) $P^L(i^*) < P^M(i^*) < P^R(i^*)$

(2) $P^R(i^*) < P^M(i^*) < P^L(i^*)$

(3) $P^M(i^*) < P^L(i^*) < P^R(i^*)$

(4) $P^M(i^*) < P^R(i^*) < P^L(i^*)$

(5) $P^L(i^*) < P^R(i^*) < P^M(i^*)$

(6) $P^R(i^*) < P^L(i^*) < P^M(i^*)$

由於 $P^R(i^*)$ 、 $P^L(i^*)$ ，和 $P^M(i^*)$ 的值有可能彼此一樣大，所以除了上述六種大小關係外，還有其他的可能存在。演算法實際執行時，會先後根據條件式判斷應該以表 3-2 中的四種方案的哪一種做分割。如果符合第一個方案的條件式，就直接使用第一種方案做分割，如果不符合，則會接著判斷是否符合下一個方案的條件式，之後以此類推。因此若 $P_{critical}$ 設置為 0 時，前三個方案皆不符合，就只能使用第四種方案做分割，代表分類樹完全不會執行多層判別分析，多層混合分類樹則轉化為 C&ART。

表 3-2 方案屬性值切點選擇列表

方案 A	條件式： $P^L(i^*) < P^M(i^*) \le P^R(i^*)$ && $P^L(i^*) < P_{critical}$
	切點選擇： R_L
方案 B	條件式： $P^R(i^*) < P^M(i^*) \le P^L(i^*)$ && $P^R(i^*) < P_{critical}$
	切點選擇： R_R
方案 C	條件式： $P^L(i^*) < P^M(i^*)$ && $P^R(i^*) < P^M(i^*)$ && $P^L(i^*) < P_{critical}$ && $P^R(i^*) < P_{critical}$
	切點選擇： R_L 和 R_R
方案 D	條件式：非以上三方案的任一種
	切點選擇： R_M

（4） 方案停止建構條件

Figure 3-6 屬性評估方案建構流程圖中，在「決定候選切點以及衡量屬性分類能力」、「決定屬性切點值」這兩大流程之前都必須判斷方案是否能夠繼續建構。

在「決定候選切點以及衡量屬性分類能力」執行之前，我們需要判斷待分割節點層數是否已經達到限定層級及節點中的單一類別個數是否小於或者等於 1（節點純度已經很高，沒有必要繼續切割）。若是該方案是費雪線性組合屬性方案，則需要判斷節點中的資料個數小於屬性個數+2，以避免 RW 出現 Single 的情況。在「決定方案切割方式」執行之前，我們也需要比較待分割節點的檢定值$P_{total}(i^*)$和$P_{threshold}$的大小。若$P_{total}(i^*) < P_{threshold}$代表在該方案中找不到足夠顯著的屬性來進行分割，則停止建構方案。

3.4.2 費雪線性組合屬性方案建構之流程

本小節將會說明如何利用 FLD-RW 來建構費雪線性組合屬性方案。

費雪線性組合屬性方案與在 3.4.1 小節中所介紹的單一屬性方案的建構流程有所不同。在一開始進行建構時，我們使用多種屬性來進行線性組合，使用幾種屬性是根據目標節點目前已使用的屬性數量+1。決定完使用幾種屬性後(假設使用 n 種)，要利用相對重要指標(王彥龍，2013)來選取屬性。我們只選取排名前 n 名的屬性，并根據這些被選擇的屬性，使用 FLD 來找出其各屬性的係數。最後，根據這些被選擇的屬性以及其係數可以計算出一新的線性組合屬性。

當節點進行費雪線性組合屬性方案評估時，關鍵在於決定應該用幾個屬性來進行分割。我們依次添加新屬性進入到節點中，直到費雪線性組合屬性方案效能遜於多層組合屬性方案時，才停止屬性評估。因此，演算法並非在一開始就決定該目標節點需要使用的屬性數。

3.5 多層混合分類樹參數之功能與影響

建構一棵多層混合分類樹，一共有三個參數需要調整和訓練，分別為方案停止建構條件之一的$P_{threshold}$、用來調整費雪線性組合方案比例的P_{FLD}和調節 C&ART 和多層判別分析相對比重的$P_{critical}$。

為了驗證這些參數的實際功能與影響，本研究利用專門用於參數訓練和選擇的 266 筆乳房腫瘤案例來做實例研究。本研究會建構 1000 棵多層混合分類樹，並收集相應的統計量，每次建構時會隨機選取 80%案例作為訓練樣本，而其餘的 20%

作為測試樣本。

3.5.1 $P_{threshold}$ 在樹群大小調整中的影響及作用

本章 3.4.1 小節方案停止建構條件中提及在「決定方案切割方式」執行之前，需要比較待分割節點的檢定值 $P_{total}(i^*)$ 和 $P_{threshold}$ 的大小。若 $P_{total} < P_{threshold}$，代表在該方案中找不到足夠顯著的屬性來進行分割，則停止建構方案。$P_{threshold}$ 的大小會影響到分類樹的規模，具體表現為最高層級（MaxLayer）和分割結點數(Split Nodes)。如 Figure 3-4 所示，隨著 $P_{threshold}$ 的增大，多層混合分類樹的最高層級和分割結點數均呈現上升的趨勢，其中隨著 $P_{threshold}$ 從 0.15 提高至 0.5，最高層級從平均 3.03 層升到 5.02 層，而分割結點數從平均 3.12 個升至 6.3 個。

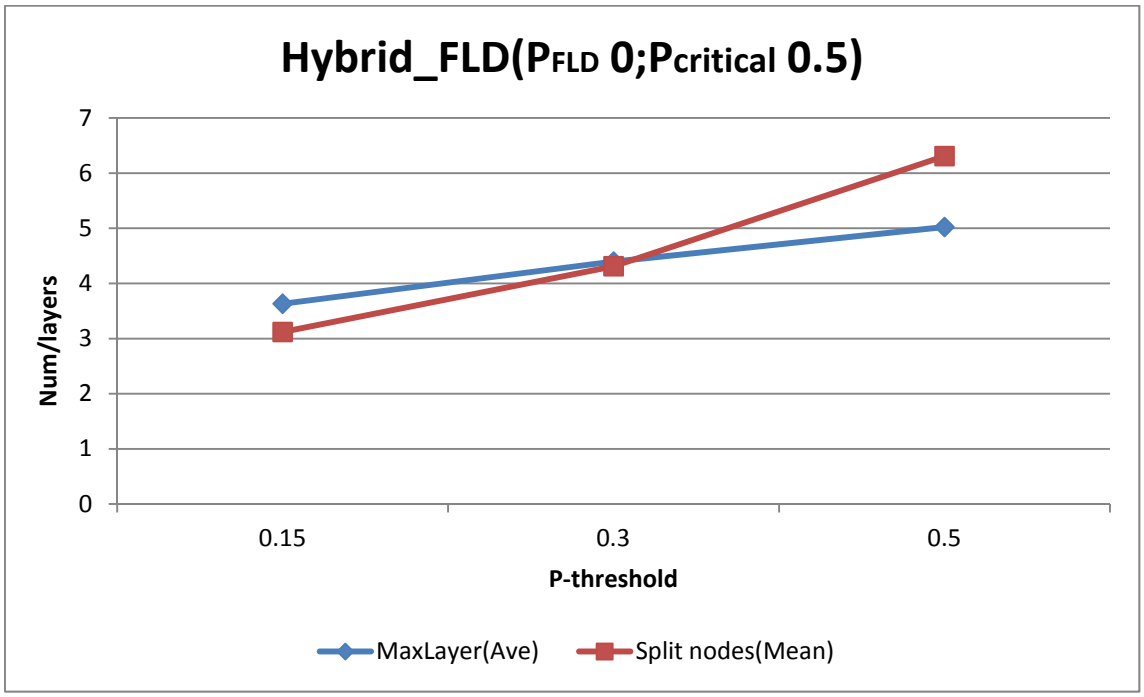


Figure 3-11 多層混合分類樹平均層級及分割節點數隨著 $P_{threshold}$ 的變化圖

3.5.2 P_{FLD} 在線性組合方案比重調整中的作用及影響

本章 3.2.2 小節中提及多層混合分類樹的每一層不僅可以執行單一屬性方案，而且也可以執行費雪線性組合屬性方案。費雪線性組合屬性方案利用 FLD 結合多個屬性，即可計算出節點內的資料經過線性組合之後所得到的區別分數，利用區別分數結合 NP-ROC 線下面積檢定對結點做分類。費雪線性組合屬性方案區別于單一屬性方案之處就在於同時考慮了多個屬性且有較好的正確度(Yildiz 與

Alpaydin, 2001)。

然而，在本小節 266 筆乳房腫瘤案例研究中，費雪線性組合屬性方案的使用不一定能夠有效提升分類樹的分類效能。由表 3-3 可見，結合了費雪線性判別（FLD）的多層判別分析在相同 $P_{threshold}$ 設定下，其 Youden_Ave 的平均值皆更高且波動更大。因此，本研究在多層混合分類樹中發展了 P_{FLD} 來調節費雪線性組合屬性方案在分類樹中的比例，從而進一步增加多層混合分類樹的靈活性，有效處理不同類型的數據。為了視覺化理解 P_{FLD} 的作用，我們統計了 3000 次多層混合分類樹中在不同的 P_{FLD} 下費雪線性組合方案（FLD）出現的平均次數及在總次數中的比重，如 Figure 3-12 副坐標和主坐標所標示。由 Figure 3-12，我們可以發現費雪線性組合方案出現的次數及其占總次數的比重皆隨著 P_{FLD} 的增加而不斷提高，從平均 0 個(0%)逐步提高至 12.68 個(13%)，因此調節 P_{FLD} 確實可以影響費雪線性組合屬性方案出現的次數。

表 3-3 基於單一屬性和結合費雪線性判別的多層判別分析效能比較圖

	ML-ROC					
	noFLD			FLD		
$P_{threshold}$	0.15	0.3	0.5	0.15	0.3	0.5
Youden_Ave(Mean)	0.5822	0.5829	0.5811	0.5621	0.5676	0.5635
Youden_Ave(STD)	0.0693	0.0689	0.0744	0.0759	0.0760	0.0801

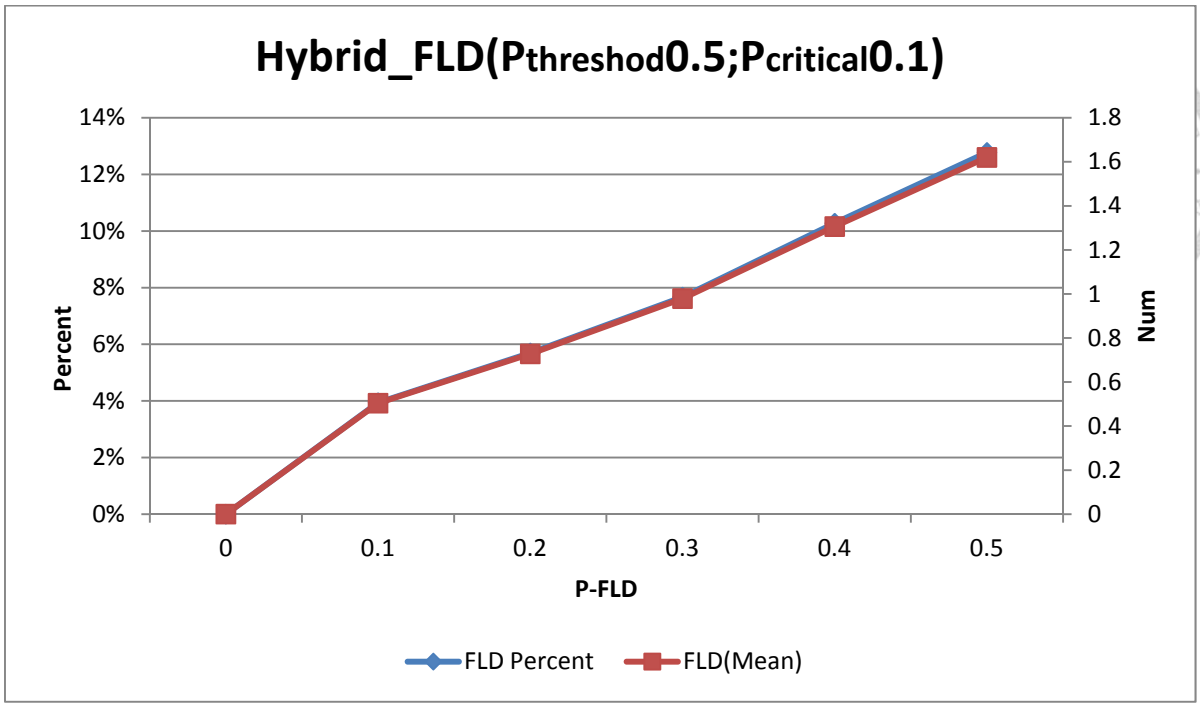


Figure 3-12 多層混合分類樹節點各方案類型數量隨 P_{FLD} 變化圖

3.5.3 $P_{critical}$ 在 C&ART 和多層判別分析比重調整中的作用及影響

本章 3.3 小節中提及多層混合分類樹的每一層不僅可以同時執行 C&ART 分類樹和多層判別分析，而且可以透過參數 $P_{critical}$ 的設置來調整兩者的相對比重。為了能夠視覺化理解 $P_{critical}$ 在多層混合分類樹中的作用和影響，本研究將統計基於乳房腫瘤實例所構建的樹狀結構中 Left、Right、Both 和 Middle 出現的次數，相關結果可見 Figure 3-13。Figure 3-13 的主坐標標示 Left、Right 和 Both 的個數，而副坐標標示 Middle 和 SUM 的個數。由 Figure 3-13 的曲線可以看出，$P_{critical}$ 的值越小，Middle 出現的次數越多，當 $P_{critical}$ 為 0 時，Left、Right 和 Both 的出現次數將變為零，僅剩 Middle 一種情況，多層混合分類樹就轉換為 C&ART 分類樹；而隨著 $P_{critical}$ 的值增大，Middle 出現的次數越變越少，最後趨近於零，代表多層混合分類樹中多層判別分析的比重不斷增加，而 C&ART 的比重不斷下降。

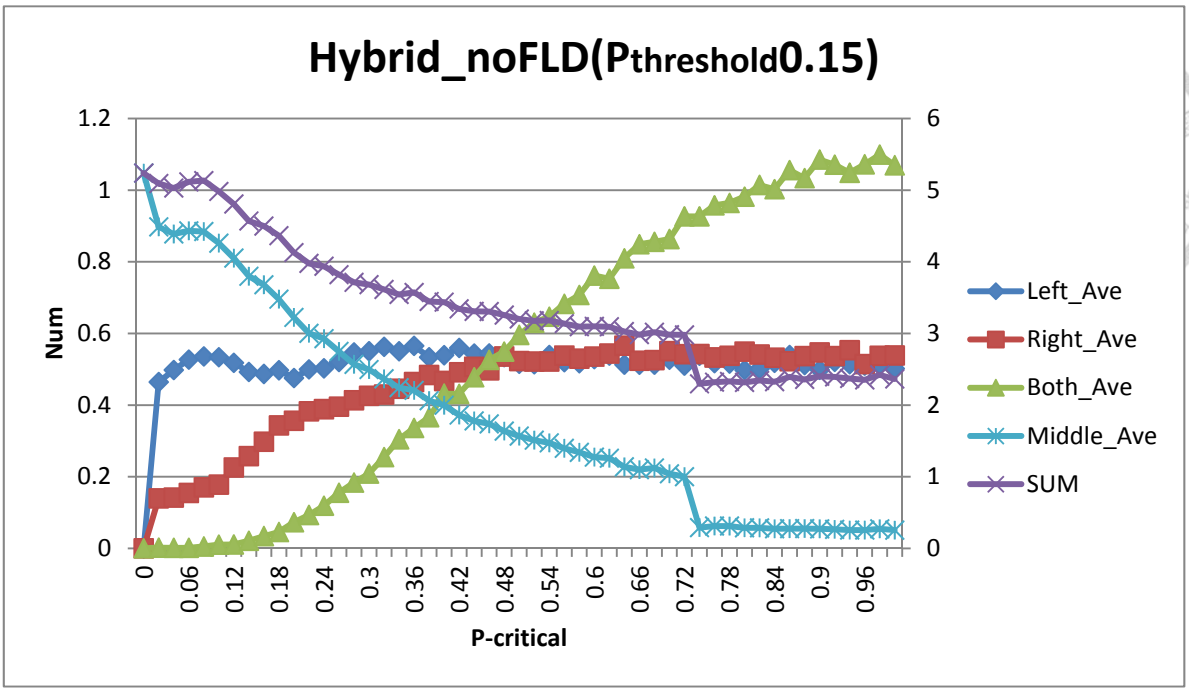


Figure 3-13 多層混合分類樹不同資料形態出現的次數彙整圖

3.6 多層混合分類樹參數之設定

多層混合分類樹存在三個需要調節的參數，分別為 $P_{threshold}$、P_{FLD} 和 $P_{critical}$，其中后兩項為多層混合分類樹所特有。為了設定合適的參數組合，建構單一分類樹和多階段調適樹群（莊曙詮，2012），本研究採用多次隨機抽樣測試的平均結果來決定最後的參數組合。每次測試時會從專門用於參數訓練和選擇的 266 筆乳房腫瘤案例，隨機抽取 80%的樣本作為訓練樣本構建樹狀結構，然後用剩餘 20%的樣本來測試，重複進行 1000 次測試，取其平均結果來決定最佳的參數組合。然後，單一建樹篩選得到的最佳參數組合將直接用於建構多階段調適樹群。本研究參數的調整範圍及間隔如表 3-4 所示。

表 3-4 多層混合分類樹參數的調節範圍和間隔彙整表

參數	範圍及間隔
$P_{threshold}$	0.15, 0.3, 0.5
P_{FLD}	範圍 0~0.5；間隔 0.1
$P_{critical}$	範圍 0~1；間隔 0.02

Chapter 4　實例驗證

4.1　資料說明

本研究使用的乳房腫瘤實例皆帶有 9 項屬性，其中 6 項由灰階超音波顯影片所得，為 B-mode 屬性，另 3 項由彩色顯影片所得之彈性度相關資訊為 Elastography 相關屬性(以下略稱為 EI 屬性)。266 筆（良性案例 193 筆，惡性案例 73 筆）將作為訓練樣本，用於參數選擇和樹狀模型的建構。為了客觀公正地衡量和比較不同分類方法的效能，本研究將另外使用 100 筆案例(良性案例 61 筆，惡性案例 39 筆)作為獨立測試樣本，獨立測試樣本僅在多階段調適樹群的測試環節使用。訓練樣本、獨立測試樣本的良惡性個數及惡性案例佔比可見表 4-1 。

表 4-1 訓練樣本及獨立測試樣本的良惡性佔比

	Benign	Malign	Malign Percent	SUM
Training	193	73	27.44%	266
Independent Test	61	39	39.00%	100

獨立測試樣本的 100 筆案例皆有資深臨床人員所做的 BI-RADS 評估(如 Figure 4-1 所示)，可用以比較參數化三階段樹群模型所做的 BI-RADS 分級與臨床人員所做的分級間之差異。

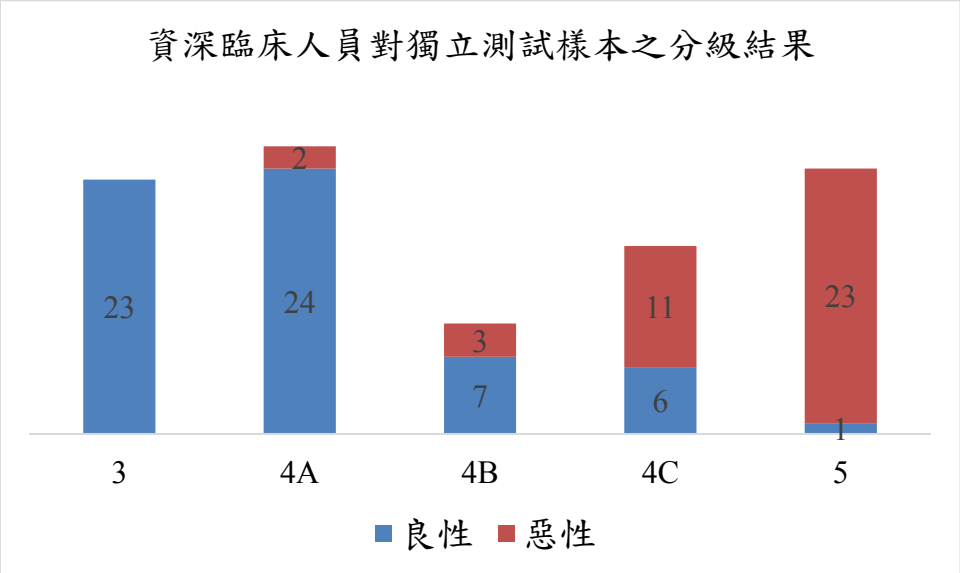


Figure 4-1 獨立測試樣本 BI-RADS 分級結果（資深臨床人員評級）

4.2 單一建樹結果彙整及多階段調適樹群模型最佳參數之決定

3.6 節提及本研究採用多次隨機抽樣測試的平均結果來決定最後的參數組合。每次建構分類樹，我們會從 266 筆訓練樣本中隨機抽取 213 筆案例（占總案例比重為 80.1%）作為訓練資料，剩餘 53 筆案例（占總案例比重為 19.9%）作為測試樣本，重複進行 1000 次測試，取其平均結果來決定最佳的參數組合。然後，單一建樹篩選得到的最佳參數組合將直接用於建構多階段調適樹群。

為了更好地比較和了解多層混合分類樹的分類效能和特點，本研究將同時比較六種分類樹方法的效能，並選擇它們最佳的參數組合來建構三階段調試樹群模型。六種分類樹分別為 Hybrid-noFLD、Hybrid-FLD、C&ART-Gini、ML-ROC、ML-FLD-ROC 和 Enhanced-ML-ROC。在六種分類樹中需要設置的參數主要有三個，分別是 $P_{threshold}$、$P_{critical}$ 和 P_{FLD}，後兩個參數為 Hybrid 所特有，最高層級統一設置為 7 層（含 Root）。不同參數的調節範圍如表 3-4 所示，其中 Hybrid 的 $P_{threshold}$ 的調節間隔為 0.02。調節原則皆是選擇 Youden_Ave 平均值最大的參數或參數組合。各分類樹最佳的參數設置組合如表 4-2 所示，而測試樣本的分類結果則如表 4-3 所示。

表 4-2 六種不同分類樹測試樣本最佳的參數設置組合彙整表

	C&ART-Gini	ML-ROC	ML-FLD-ROC	Enhanced-ML-ROC	Hybrid-noFLD	Hybrid-FLD
$P_{threshold}$	--	0.3	0.3	0.15	0.5	0.5
$P_{critical}$	--	--	--	--	0.9	0.9
P_{FLD}	--	--	--	--	--	0

表 4-3 六種不同分類樹測試樣本分類結果彙整表

	C&ART-Gini	ML-ROC	ML-FLD-ROC	Enhanced-ML-ROC	Hybrid-noFLD	Hybrid-FLD
Layer(Ave)	7.0000	3.6990	4.5593	4.9207	6.3610	6.3770
Layer(Std)	0.0000	1.4491	1.2286	1.0935	0.9527	0.9782
Sensi(Ave)	0.4435	0.5524	0.4653	0.4413	0.4648	0.4604
Sensi(Std)	0.1403	0.3225	0.3038	0.2105	0.1773	0.1839
Speci(Ave)	0.7459	0.6135	0.6698	0.7461	0.7746	0.7779
Speci(Std)	0.0952	0.2876	0.2944	0.1873	0.1177	0.1176
Youden_Ave(Ave)	0.5947	0.5829	0.5676	0.5937	0.6197	0.6192
Youden_Ave(Std)	0.0740	0.0689	0.0760	0.0745	0.0748	0.0744

由表 4-3 可知，Hybrid-noFLD 和 Hybrid-FLD 的 Youden_Ave 平均值分別位於前兩位，其值分別為 0.6197 和 0.6192，C&ART-Gini 和 Enhanced-ML-ROC 則分別位於第 4 和第 5，分別為 0.5947 和 0.5937，且四者的 Youden_Ave 的波動非常接近。仔細查看四個方法 Sensi(Ave)和 Speci(Ave)，可以發現 Hybrid-noFLD 和 Hybrid-FLD 兩個方法在 Speci(Ave)和 Sensi(Ave)均較后兩者高，因而導致其 Youden_Ave 的值較高。

由表 4-2 可知，Hybrid-FLD 最佳的參數組合中的P_{FLD}為 0，代表不採用 FLD 的表現最佳。因此，接下來只比較含 Hybrid-noFLD 在內的五種分類樹多階段調適樹群（莊曙詮，2012）的 BI-RADS 分級結果。

4.3 乳癌腫瘤實例驗證

本節說明之乳房腫瘤實例驗證除了採用六項由灰階超音波顯影片萃取的數據外，再加入三項與腫瘤彈性(EI)相關的彩色影像數據做為屬性，分別以五種不同的分類樹建構三階段調適樹群模型，相關參數則依照表 4-2 來設置。

4.3.1 模型建構結果彙整與比較（方法一）

本小節會呈現 10 次獨立測試的 AUC 和 BIRADS 3 的分級結果，如表 4-4 所示。不同分類樹的 BIRADS 的箱型分佈圖和條形圖可見附錄。由表 4-4 可知，ML-FLD-ROC 的 AUC 平均值最高為 0.7681，但其波動也最大為 0.0436，

C&ART-Gini 次之（0.7605）但波動最小（0.025），然而 Hybrid-noFLD 的 AUC 平均值緊跟其後（0.7603）。不過 Hybrid-noFLD 的 BIRADS 3 中良性平均個數最多（10.7）且惡性比例為 3%，雖然略高於理想惡性比例 2%，但是小於臨床實踐上可以接受的比例 4%。

表 4-4 五種不同分類樹三階段調試樹群（方法一）獨立測試結果彙整表

		C&ART-Gini	ML-ROC	ML-FLD-ROC	Enhanced-ML-ROC	Hybrid-noFLD
AUC	Mean	0.7605	0.7461	0.7681	0.7490	0.7603
	STD	0.0250	0.0245	0.0436	0.0372	0.0350
BIRADS 3(Benign)	Mean	9.5000	5.4000	6.3000	8.7000	10.7000
	STD	2.4608	2.2706	2.7101	3.1640	4.0838
BIRADS 3(Malign)	Mean	0.1000	0.0000	0.2000	0.4000	0.3000
	STD	0.3162	0.0000	0.4216	0.6992	0.4830
Malign Percent		1%	0%	3%	4%	3%
Ideal Malign Percent		2%	2%	2%	2%	2%

4.3.2 模型建構結果彙整與比較（方法二）

本小節會呈現 10 次獨立測試的 AUC 和 BIRADS 3 的分級結果，如表 4-5 所示。不同分類樹的 BIRADS 的箱型分佈圖和條形圖可見附錄。由表 4-5 可知，Enhanced-ML-FLD 的 AUC 平均值最高為 0.7947，但其波動也最大為 0.0545，ML-FLD-ROC 次之（0.7904）且波動也次大（0.0484），然而 Hybrid-noFLD 的 AUC 平均值位於第三位（0.7750）但波動較小（0.0364）。不過 Hybrid-noFLD 的 BIRADS 3 中良性平均個數最多（10.7）且惡性比例為 3%，雖然略高於理性惡性比例 2%，但是小於臨床實踐上可以接受的比例 4%。

比較表 4-4 和表 4-5，我們可以發現在五種分類樹中方法二的 AUC 平均值皆高於方法一，由此可見方法二更適合本研究所使用的乳房腫瘤實例。

表 4-5 五種不同分類樹三階段調試樹群（方法二）獨立測試結果彙整表

		C&ART-Gini	ML-ROC	ML-FLD-ROC	Enhanced-ML-ROC	Hybrid-noFLD
AUC	Mean	0.7811	0.7692	0.7904	0.7947	0.7848
	STD	0.0329	0.0277	0.0484	0.0545	0.0364
BIRADS 3(Benign)	Mean	9.5000	5.4000	6.3000	9.0000	10.7000
	STD	2.4608	2.2706	2.7101	2.0548	4.0838
BIRADS 3(Malign)	Mean	0.1000	0.0000	0.2000	0.0000	0.3000
	STD	0.3162	0.0000	0.4216	0.0000	0.4830
Malign Percent		1%	0%	3%	0%	3%
Ideal Malign Percent		2%	2%	2%	2%	2%

Chapter 5　結論與未來研究建議

在模型構造中，本研究通過引入參數 P_{FLD} 和參數 $P_{critical}$ 分別調整費雪線性組合屬性方案的比例及多層判別分析和 C&ART 分類樹的相對比重，從而增加多層混合分類樹的靈活性，以便有效處理不同形態的數據。

在建立模型的過程中，當每一個節點進入演算法中時，先通過 P_{FLD} 和多層組合屬性方案決定是否需要採用費雪線性組合屬性方案及相應的特徵數，再通過 $P_{critical}$ 和非參數型接受者操作特徵（NP-ROC）來決定節點和切割方案，即決定是否分割成 C&ART 的兩個節點或多層判別分析的兩個節點或三個節點，然而一旦節點確定類別，則不進行再次分割。

多層混合分類樹存在三個需要調節的參數，分別為 $P_{threshold}$ 、P_{FLD} 和 $P_{critical}$ ，本研究通過三千次隨機獨立測試來選擇最佳的參數組合。

在本研究第三章 3.5 節，利用乳房超音波資料實際驗證了三個參數 $P_{threshold}$ 、P_{FLD} 和 $P_{critical}$ 的調節效果。進而，在第四章將多層混合分類樹應用在乳房超音波資料的診斷上，可以發現它不僅在單一建樹環節的整體分類效能優於 C&ART、多層判別分析和強化多層判別分析，且在 Speci(Ave)和 Sensi(Ave)均較 C&ART 和強化多層判別分析高，而且在三階段調適樹群中的 BIRADS 3 的良性個數較其他方法更多且惡性比例維持在可接受的範圍內。

對於未來相關研究提出可供學者繼續研究的方向：

(1) 在實例驗證環節，我們皆是使用一千次單一建樹的測試平均結果所篩選出的最佳參數作為三階段調適樹群的最佳參數。然而，三階段調適樹群模型本身中加入了 Boosting 的機制，會基於不同分類樹的分類效能給予相應的權重，因此單一建樹最佳的參數組合不一定是三階段調適樹群模型的最佳參數。因此未來可以進一步探討單一建樹和三階段調適樹群模型參數設置之間的關係，從而進一步提高三階段調適樹群的分級效能。

(2) 在多層混合分類樹中，在決定進行線性組合的屬性數時需要對費雪線性組合屬性方案和多層組合屬性方案進行比較。多層組合屬性方案的建構仍然延續賴淑俐學者(2010)所提出的想法，即每個子節點皆有可能繼續切割下

去，然而在多層混合分類樹中，一旦子節點的類別被判斷清楚了，它應該無法繼續切割下去。因此，未來可以進一步探討該部分的機制設計。

(3) 在具體的實例驗證研究中，我們發現若是能夠在最初的幾層中增加使用兩個切點的機率，可以有效地增加正確度和可行度，然而在現行的機制設計中暫時還沒有探討到這一部分。因此可以考慮在未來的機制設計進行適當考量，從而進一步提高多層混合分類樹的分類效能。

REFERENCE

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984).*Classification and regression trees*. CRC press.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*(3), 542.

Chang, K. J., Chen, W. H., Chen, A., Chen, C. N., Ho, M. C., Tai, H. C., ... & Wu, H. J. (2013). *U.S. Patent No. 8,572,006*. Washington, DC: U.S. Patent and Trademark Office.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

Fisher, R. A. (1950). The use of multiple measurements in taxonomic problems, Annual Eugenics, 7, Part II, 179-188 (1936); also in Contributions to Mathematical Statistics.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, *35*(1), 1-19.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, *9*(3), 190-195.

Pepe, M. S. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, *84*(3), 595-608.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32-35.

張富皓，2014，利用非參數型接收者操作特徵曲線建構統計分類樹之研究與應用，國立台灣大學工業工程學研究所碩士論文。

莊曙詮，2010，多階段調適樹群模型建構方法及其於腫瘤分級之應用，國立台灣大學工業工程學研究所碩士論文。

賴淑俐，2010，多層判別分析理論與方法擴張及其於腫瘤診斷上的應用，國立台灣大學工業工程學研究所碩士論文。

馬康恆，2013，利用接收者操作特徵曲線建構分類樹之研究與應用，國立台灣大學工業工程學研究所碩士論文。

巫信融，2009，多層判別分析及其應用，國立台灣大學工業工程學研究所碩士論文。

王彥龍，2013，概括性相對重要指標及變數選擇之研究及其於費雪線性區別分析於 Cox 比例風險迴歸之應用，國立台灣大學工業工程學研究所碩士論文。

附錄：五種分類樹方法獨立測試 BIRADS 分級結果

1.1 C&ART-Gini 樹群（方法一）

根據 10 次 C&ART-Gini 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-2 所示，而條形圖如 Figure 4-3 所示。

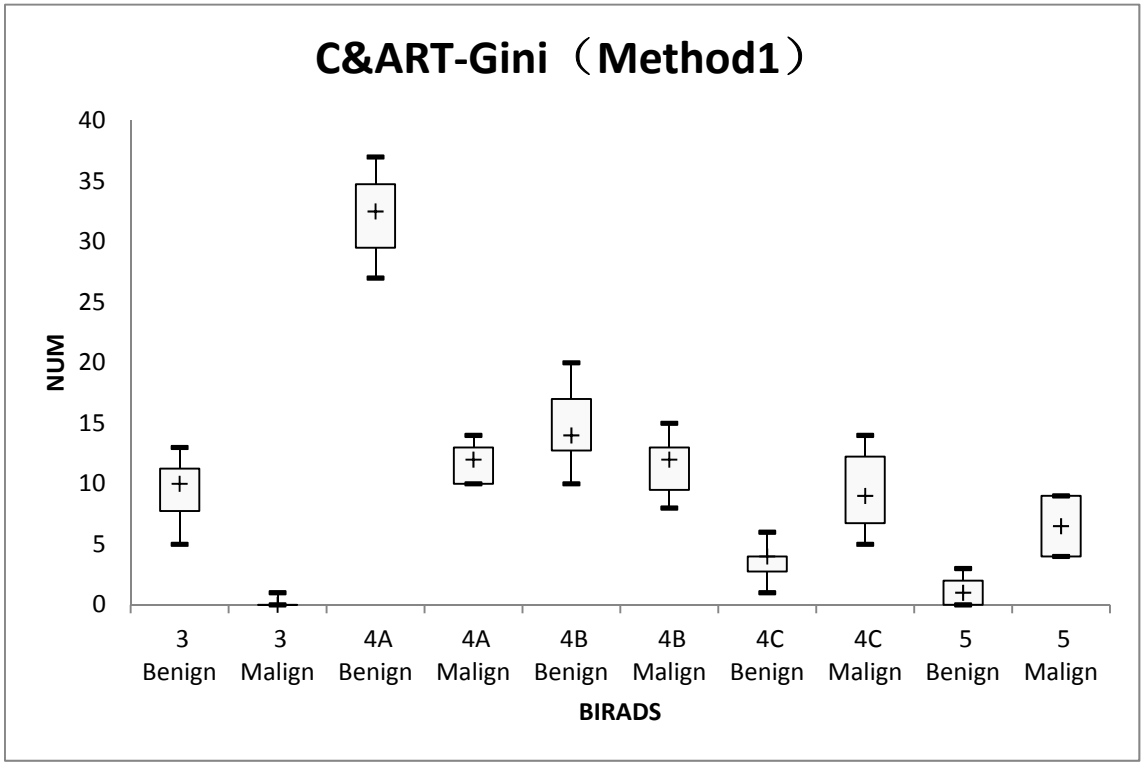


Figure 5-1 C&ART-Gini 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果箱型圖

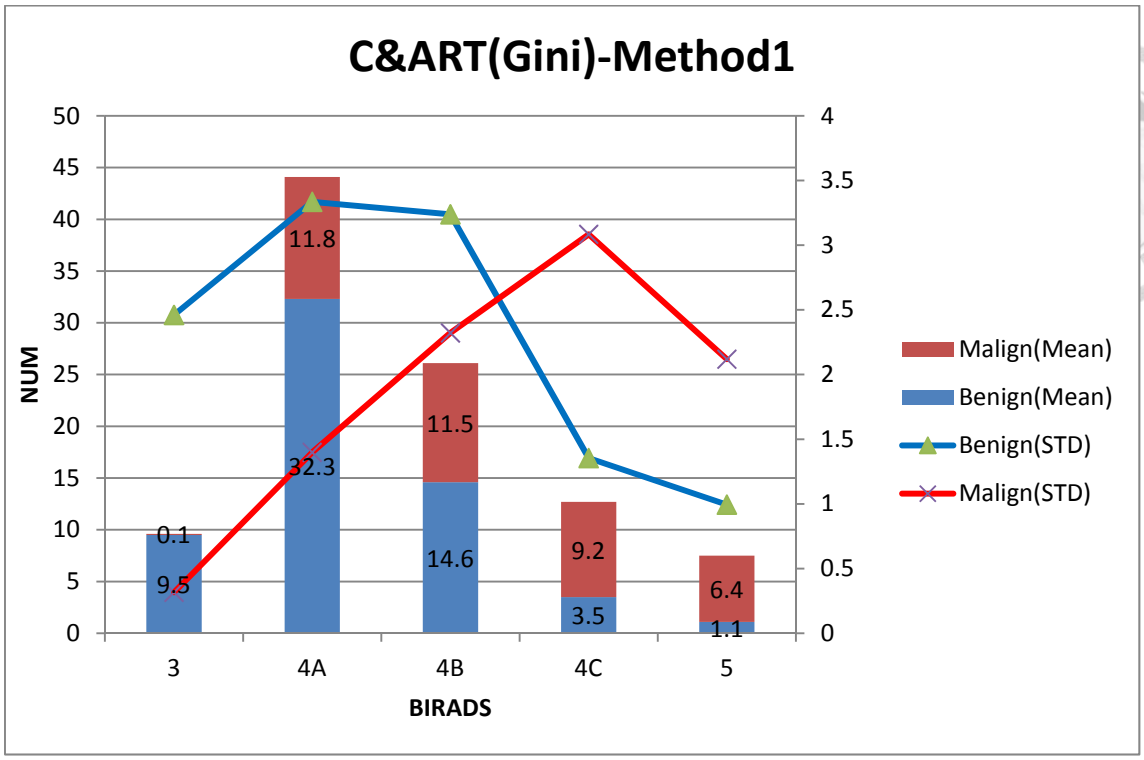


Figure 5-2 C&ART-Gini 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果條形圖

1.2 ML-ROC 樹群（方法一）

根據 10 次 ML-ROC 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-6 所示，而條形圖如 Figure 4-7 所示。

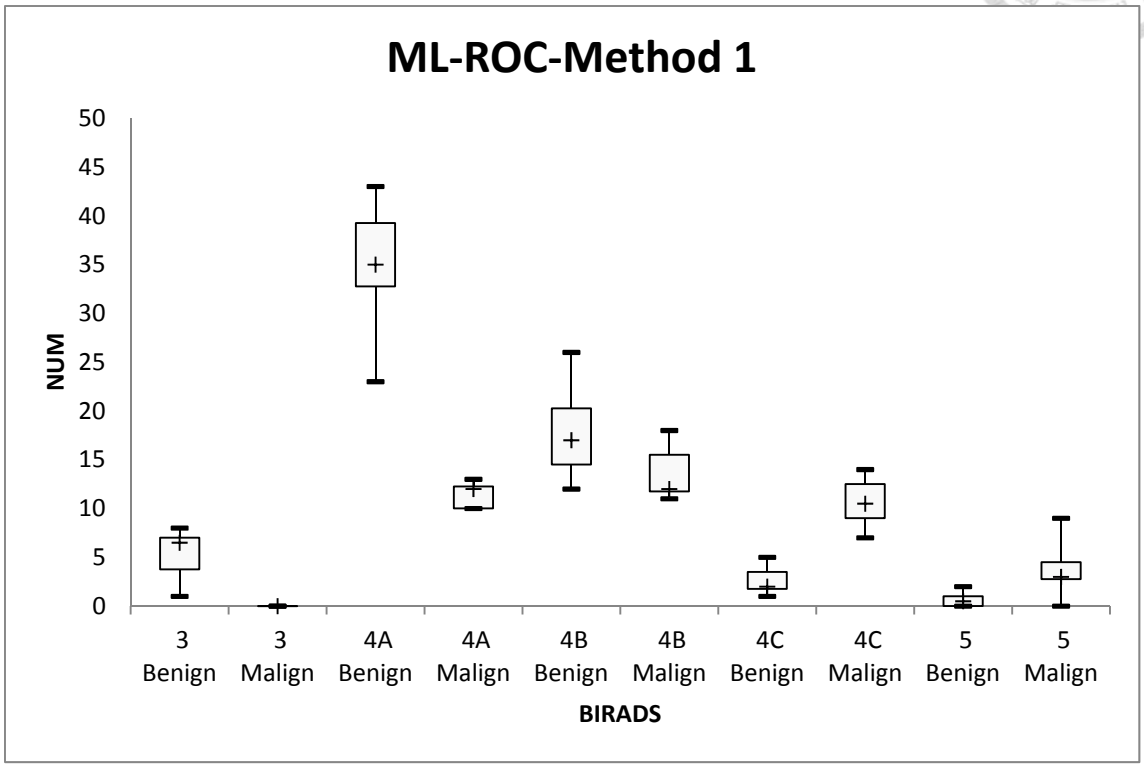


Figure 5-3 ML-ROC 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果箱型圖

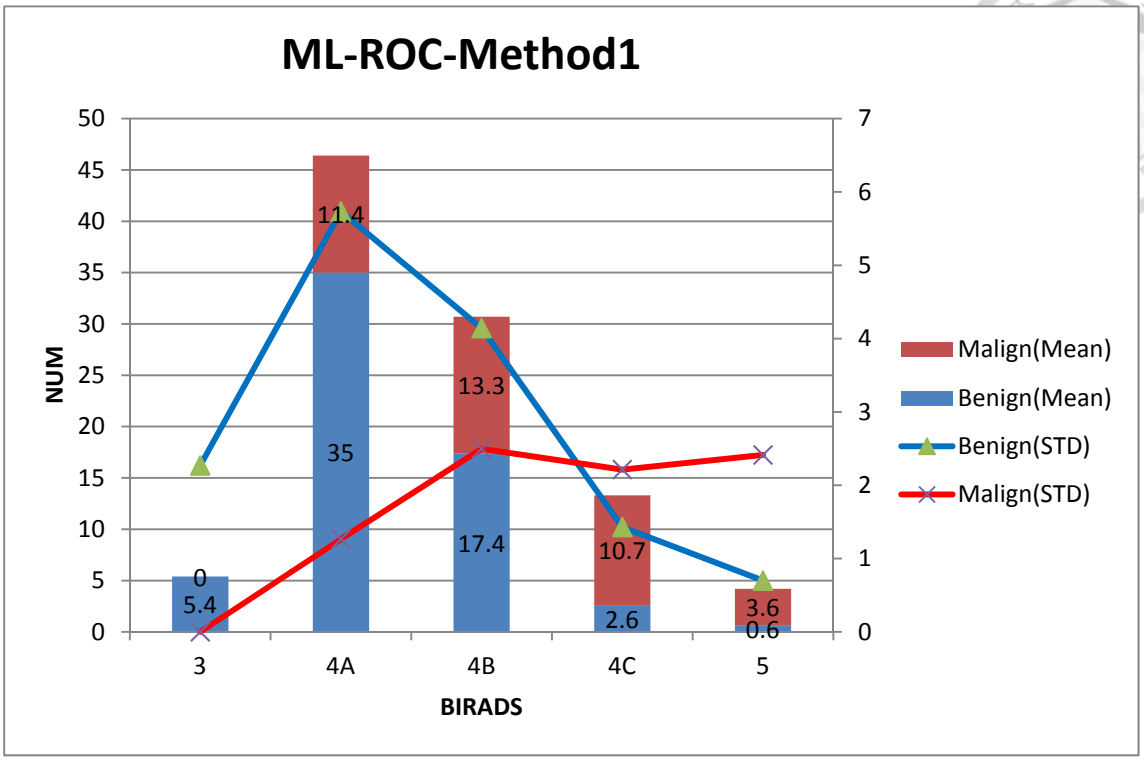


Figure 5-4 ML-ROC 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果條形圖

1.3 ML-FLD-ROC 樹群（方法一）

根據 10 次 ML-FLD-ROC 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-12 所示，而條形圖如 Figure 4-13 所示。

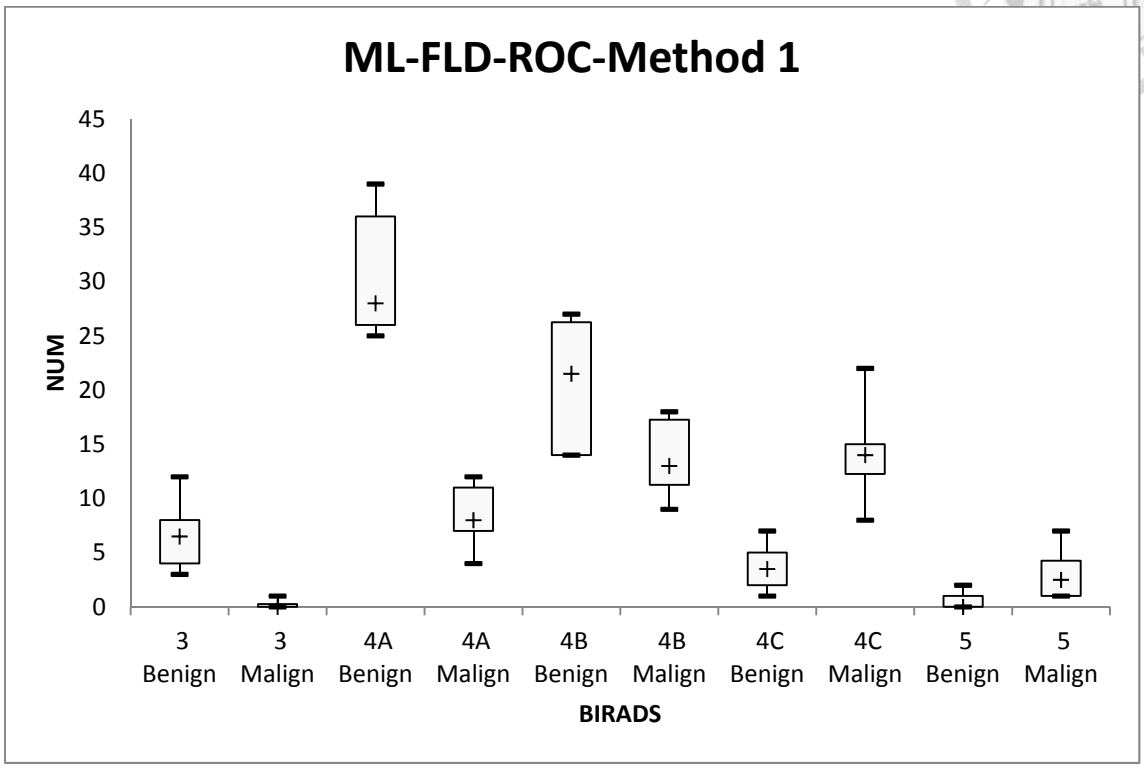


Figure 5-5 ML-FLD-ROC 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果箱型圖

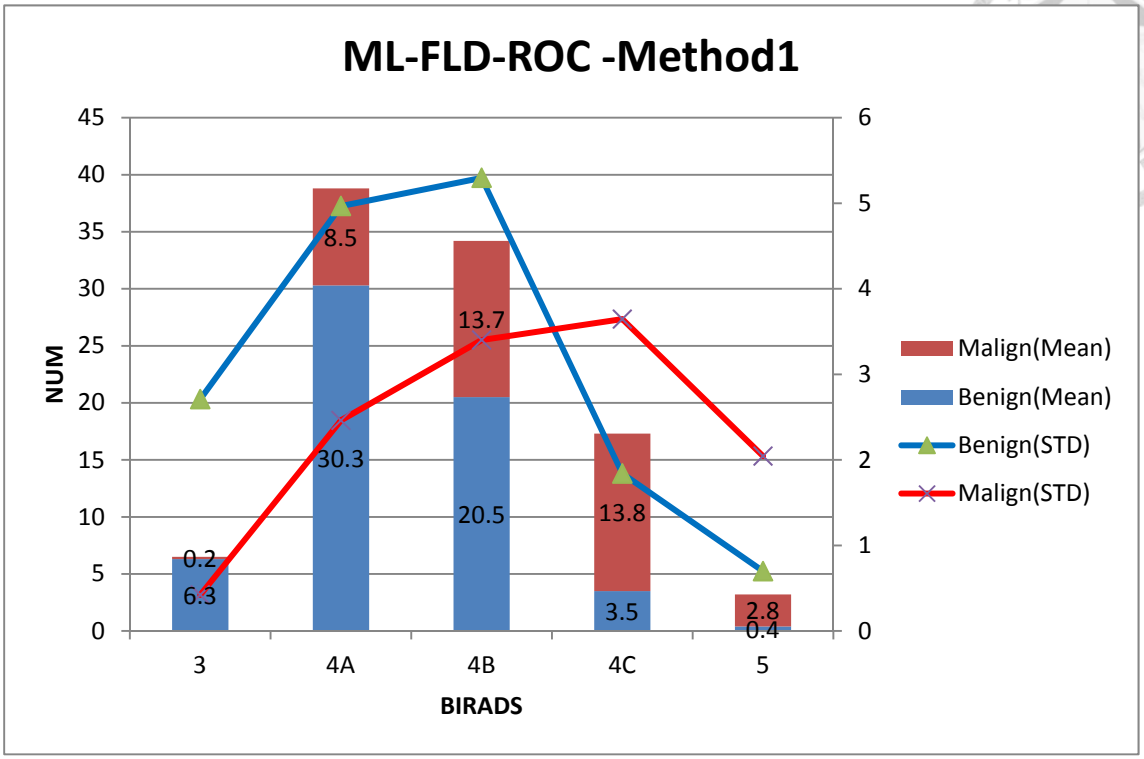


Figure 5-6 ML-FLD-ROC 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果條形圖

1.4 Enhanced -ML-ROC 樹群（方法一）

根據 10 次 Enhanced -ML-FLD-ROC 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-16 所示，而條形圖如 Figure 4-17 所示。

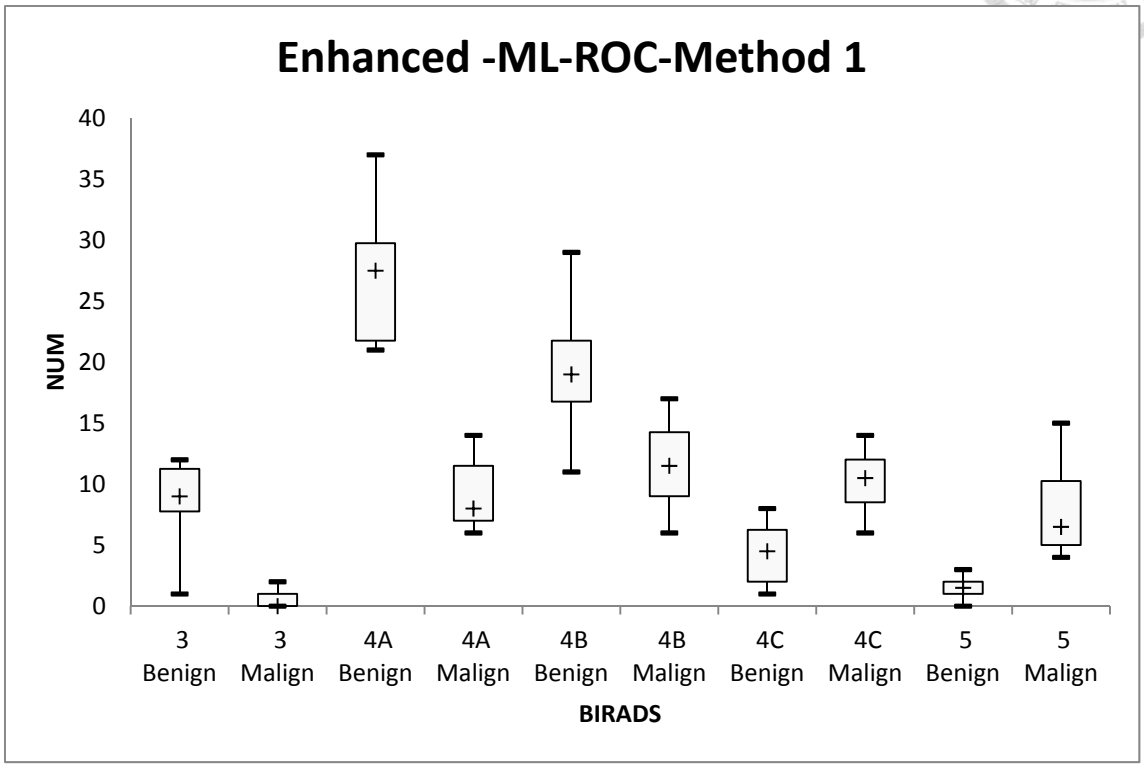


Figure 5-7 Enhanced -ML-FLD-ROC 三階段調試樹群(方法一)獨立測試的 BIRADS 分級結果箱型圖

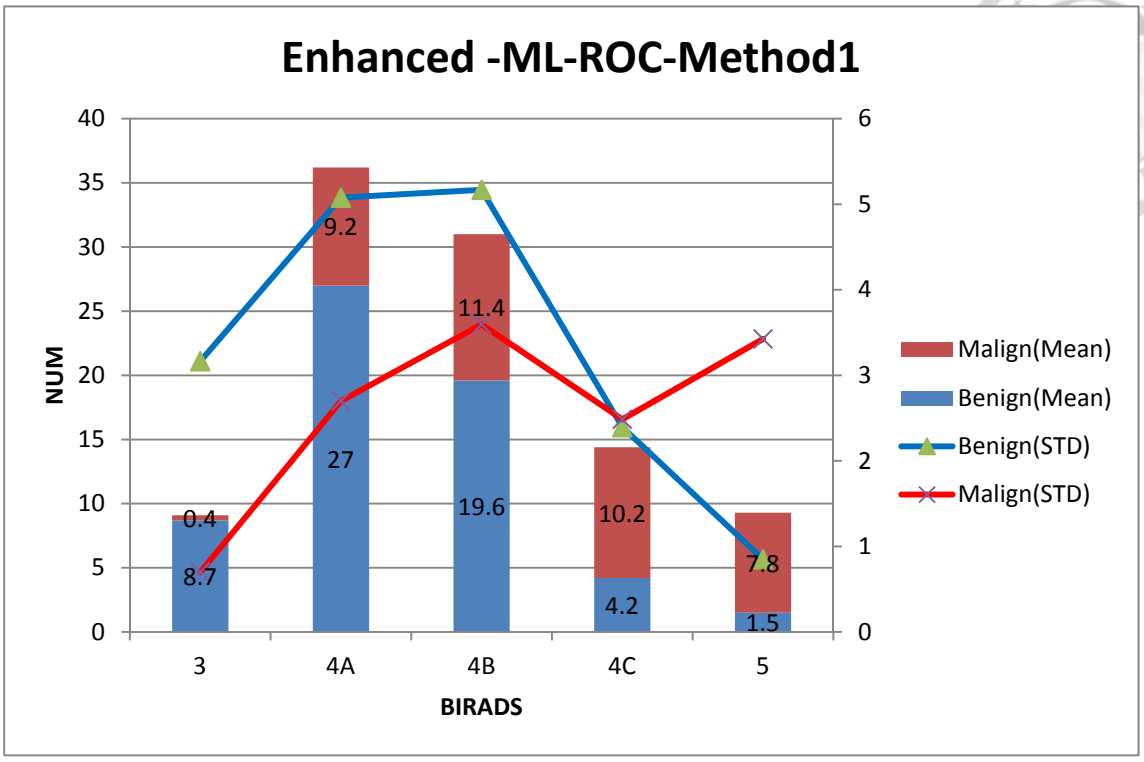


Figure 5-8 Enhanced -ML-FLD-ROC 三階段調試樹群(方法一)獨立測試的 BIRADS 分級結果條形圖

1.5 Hybrid-noFLD 樹群（方法一）

根據 10 次 Hybrid-noFLD 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-8 所示，而條形圖如 Figure 4-9 所示。

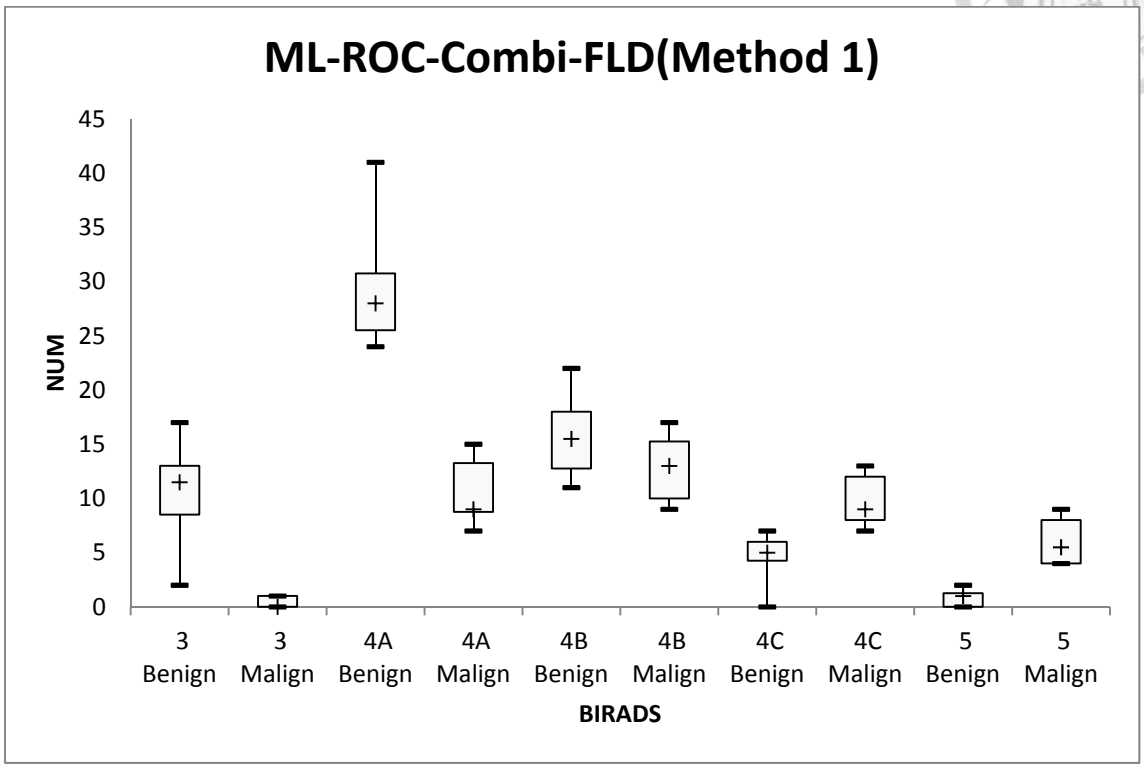


Figure 5-9 Hybrid-noFLD 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果箱型圖

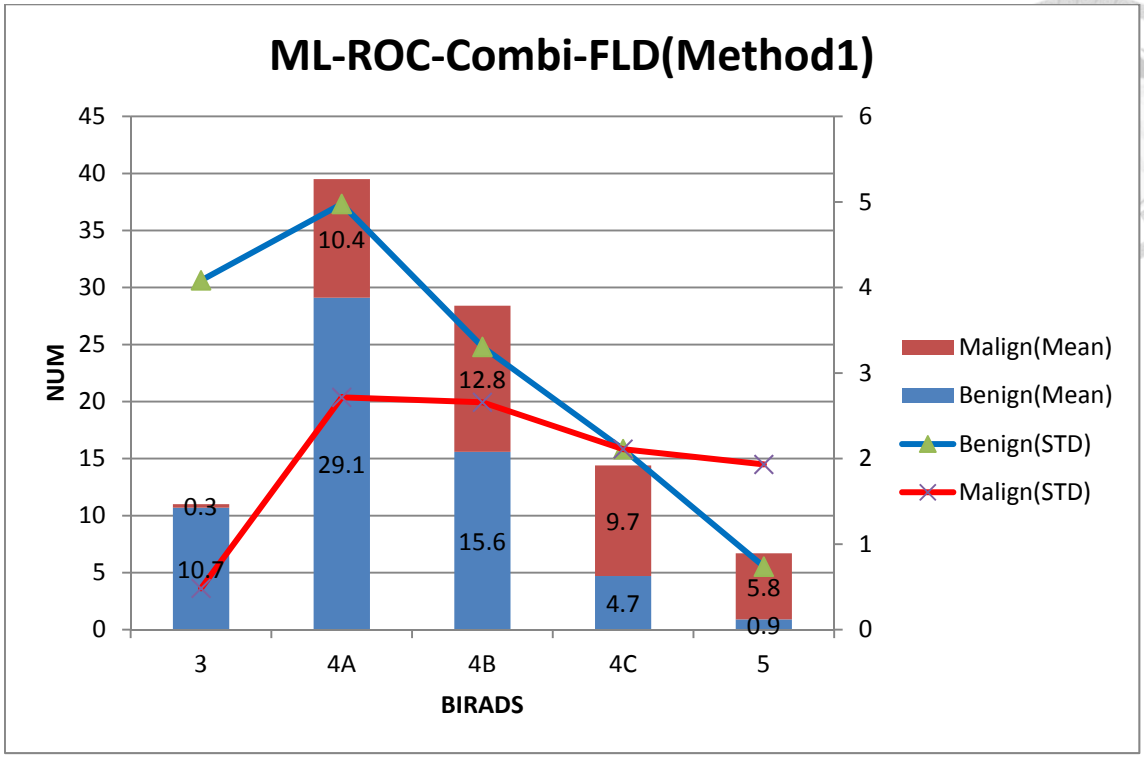


Figure 5-10 Hybrid-noFLD 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果條形圖

1.6 C&ART-Gini 樹群（方法二）

根據 10 次 C&ART-Gini 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-18 所示，而條形圖如 Figure 4-19 所示。

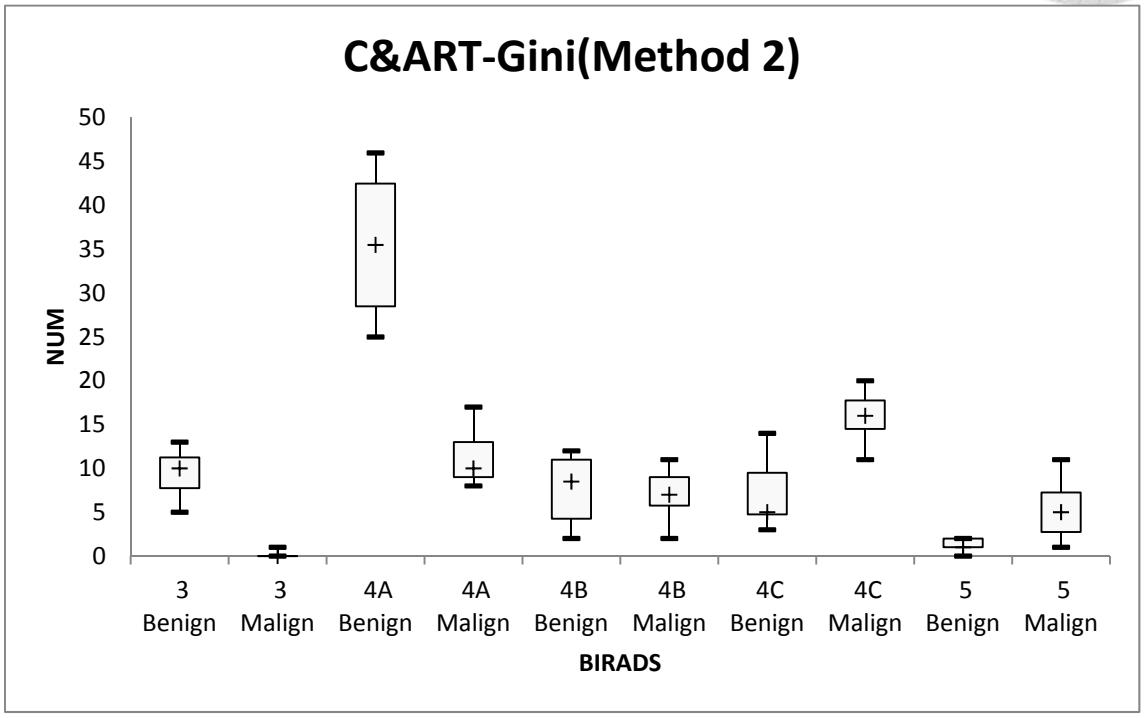


Figure 5-11 C&ART-Gini 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果箱型圖

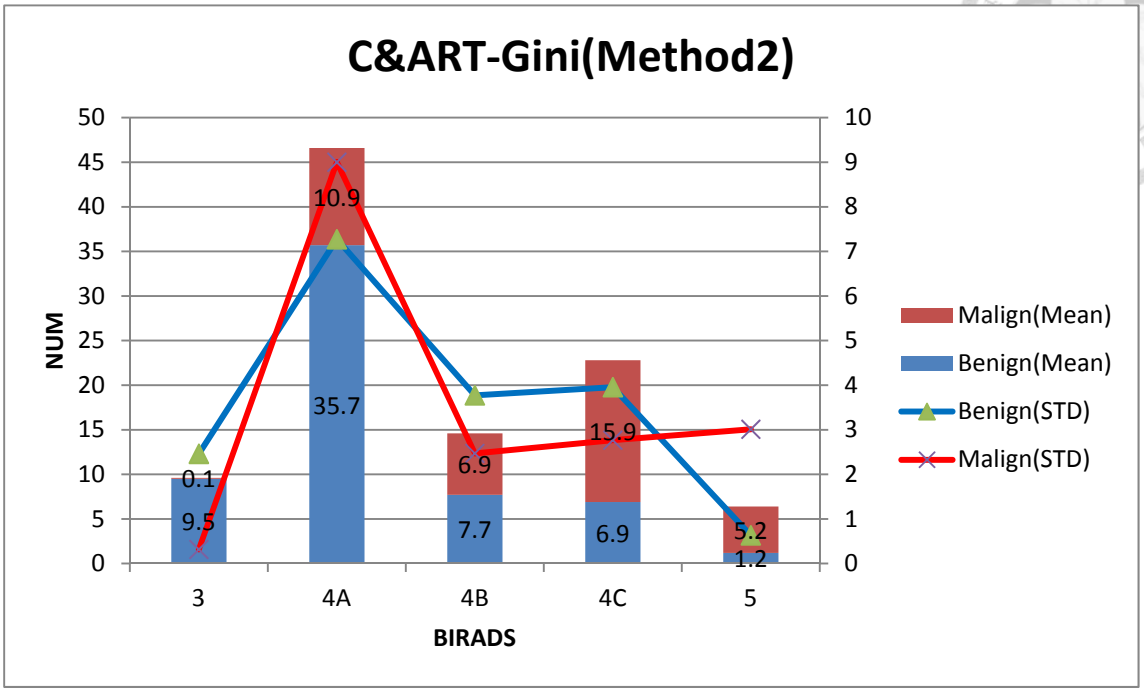


Figure 5-12 C&ART-Gini 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果條形圖

1.7 ML-ROC 樹群（方法二）

根據 10 次 ML-ROC 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-22 所示，而條形圖如 Figure 4-23 所示。

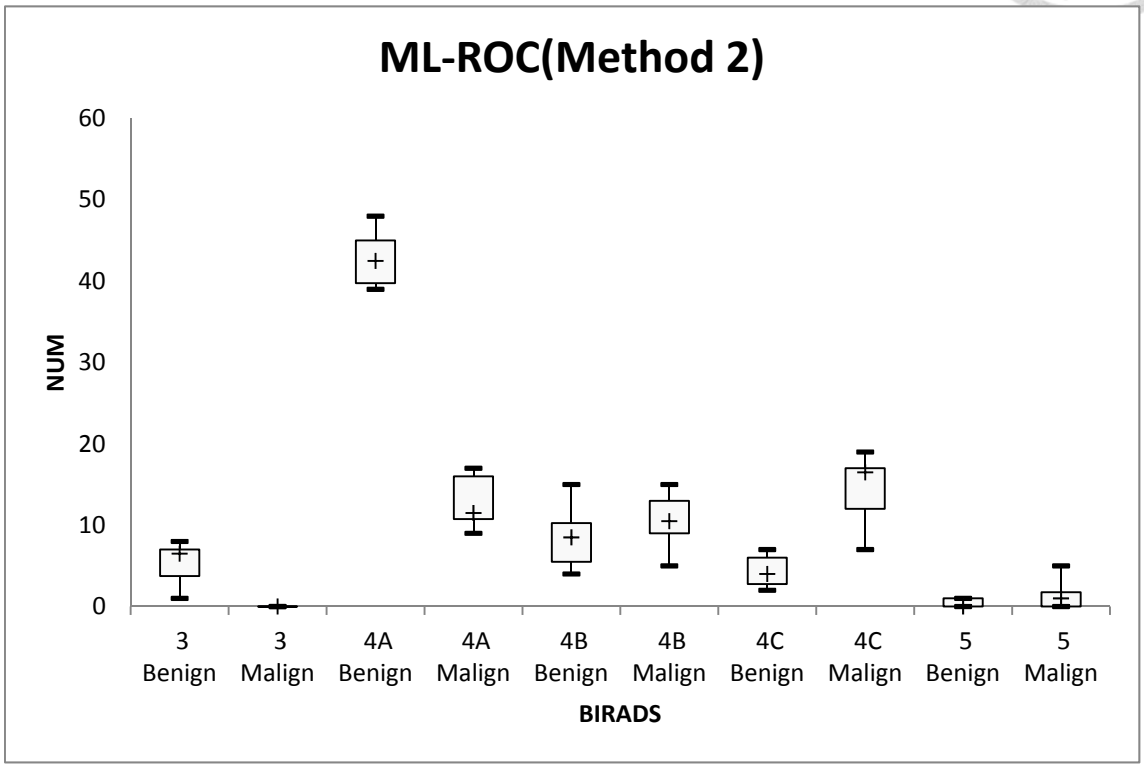


Figure 5-13 ML-ROC 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果箱型圖

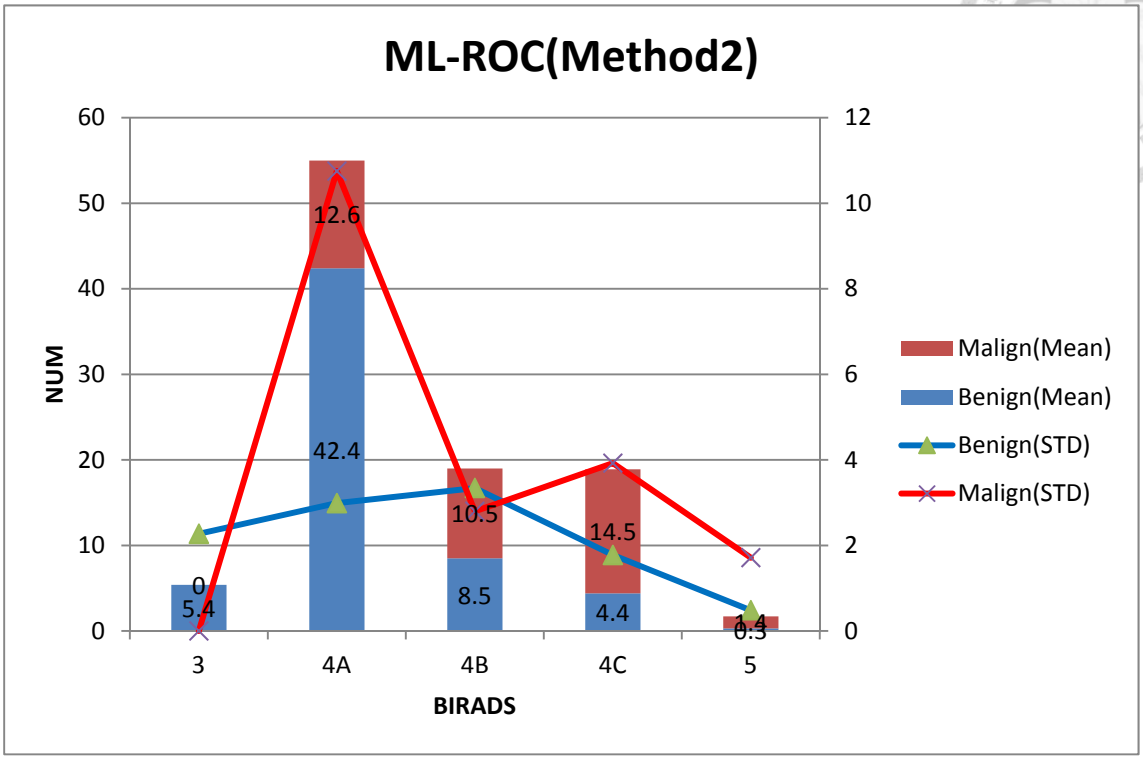


Figure 5-14 ML-ROC 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果條形圖

1.8 ML-FLD-ROC 樹群（方法二）

根據 10 次 ML-FLD-ROC 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-28 所示，而條形圖如 Figure 4-29 所示。

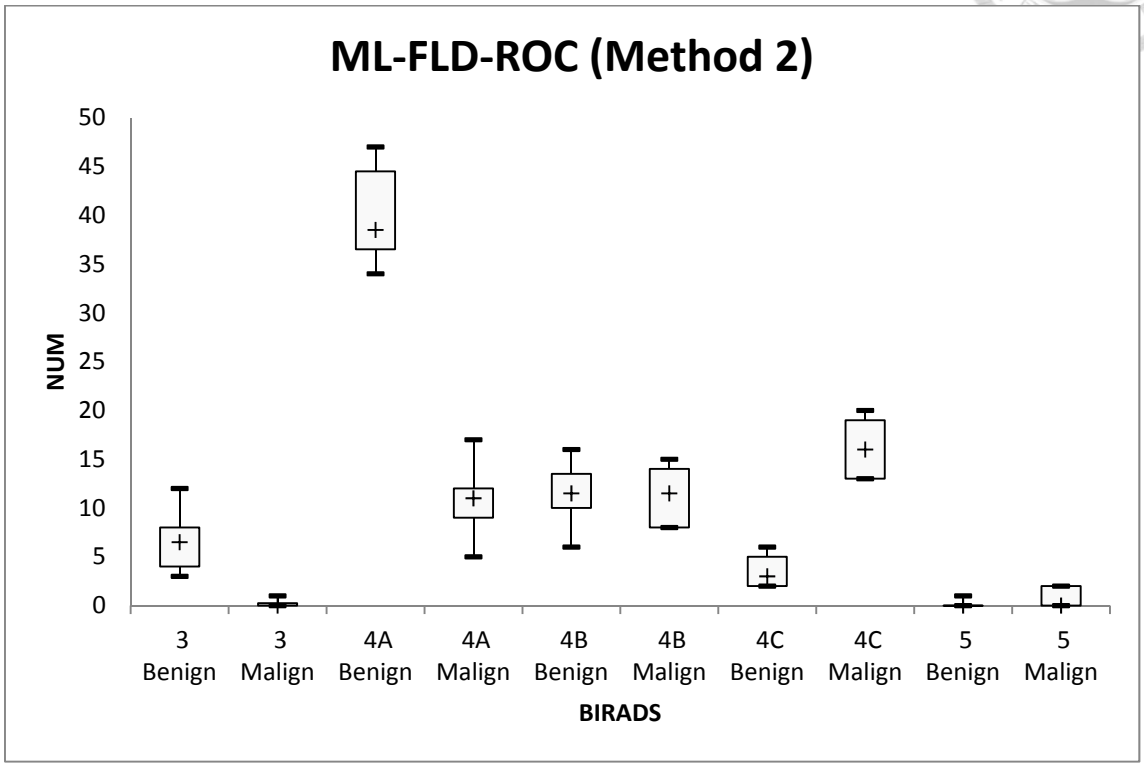


Figure 5-15 ML-FLD-ROC 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果箱型圖

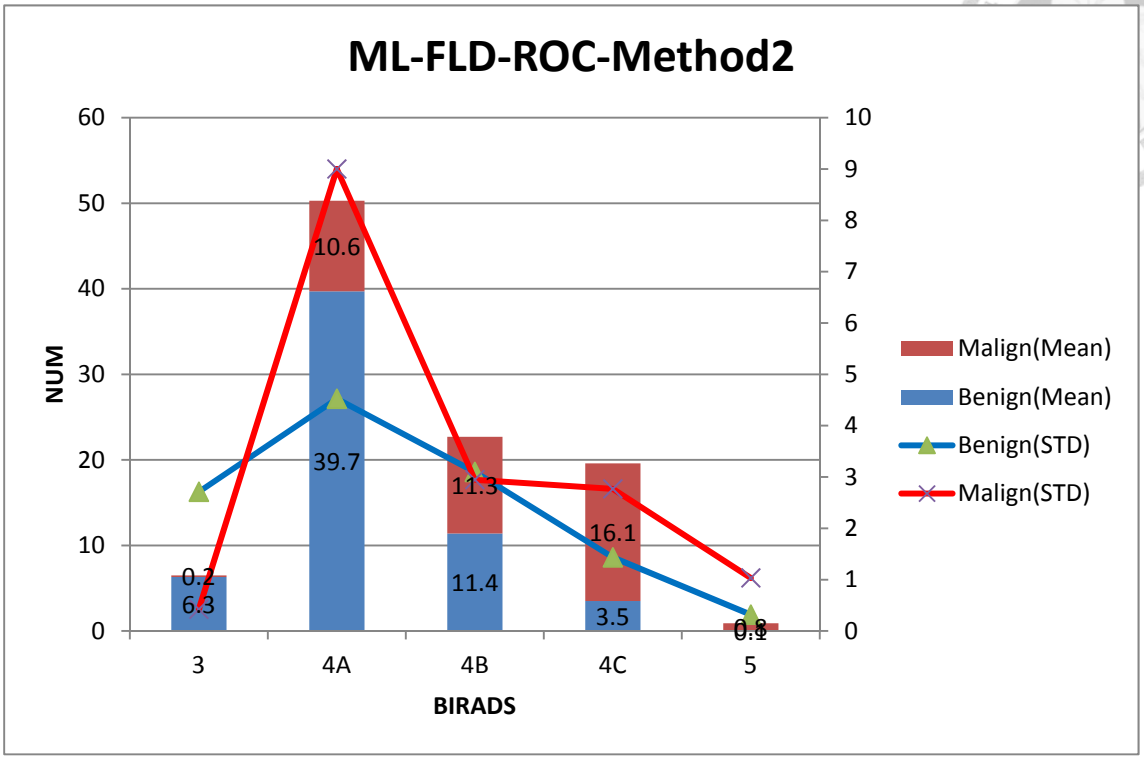


Figure 5-16 ML-FLD-ROC 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果條形圖

1.9 Enhanced -ML-FLD-ROC 樹群（方法二）

根據 10 次 Enhanced -ML-FLD-ROC 三階段調試樹群（方法一）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-32 所示，而條形圖如 Figure 4-33 所示。

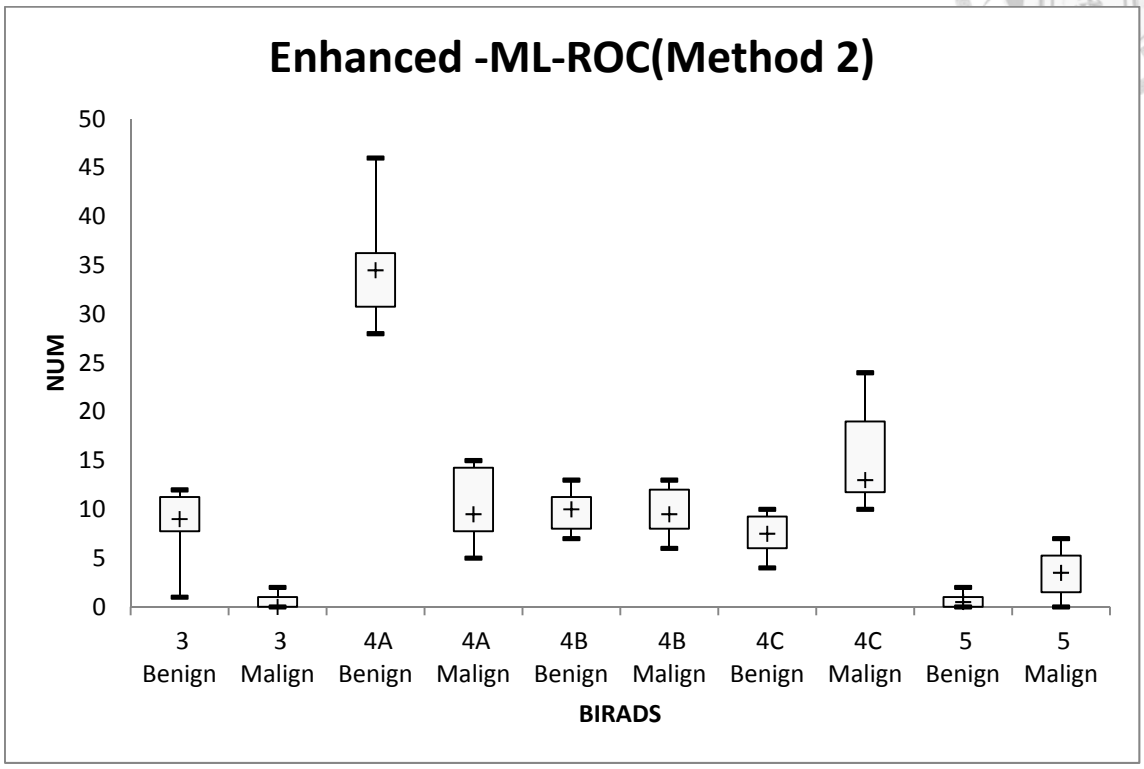


Figure 5-17 Enhanced -ML-FLD-ROC 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果箱型圖

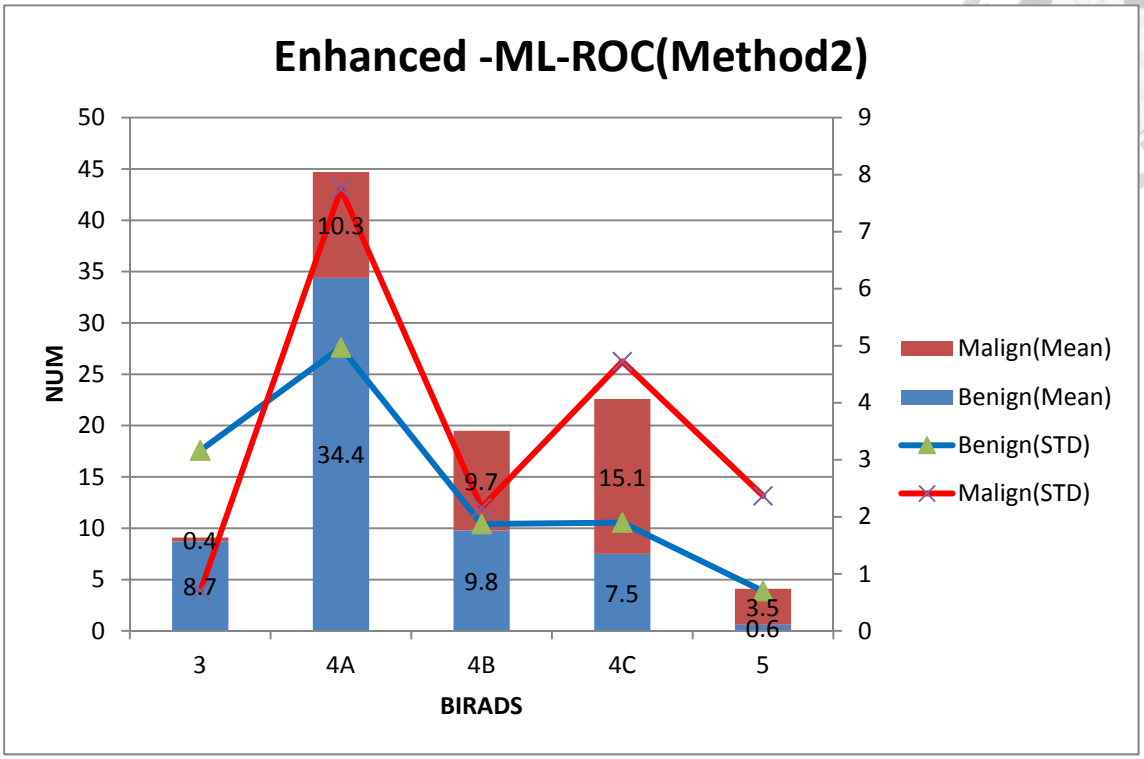


Figure 5-18 Enhanced -ML-FLD-ROC 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果條形圖

1.10 Hybrid-noFLD 樹群（方法二）

根據 10 次 Hybrid-noFLD 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果繪製的箱型圖如 Figure 4-24 所示，而條形圖如 Figure 4-25 所示。

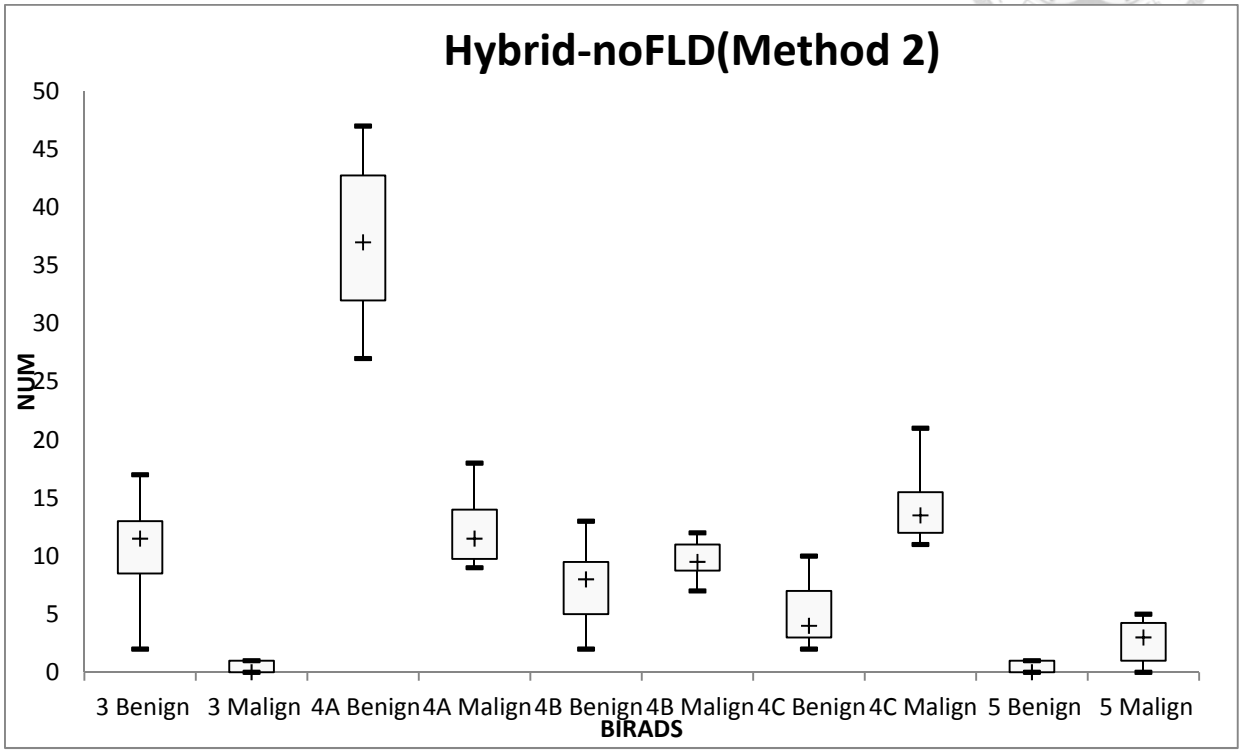


Figure 5-19 Hybrid-noFLD 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果箱型圖

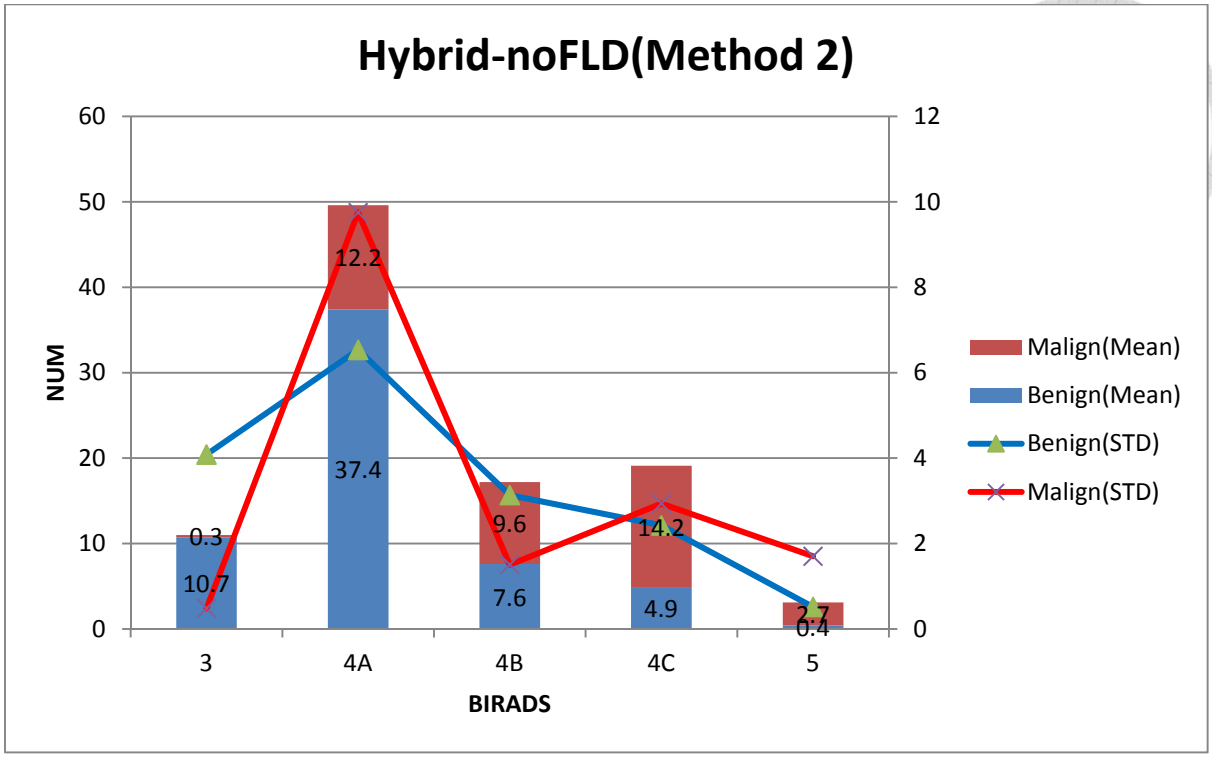


Figure 5-20 Hybrid-noFLD 三階段調試樹群（方法二）獨立測試的 BIRADS 分級結果條形圖