科技部補助

大專學生研究計畫研究成果報告

＊ ********* ＊＊ ＊
＊ 計 畫 ＊
＊ ： 整合資料探勘技術與集成學習架構預測卵巢癌復發 ＊
＊ 名 稱 ＊
＊ ********* ＊＊ ＊

執行計畫學生： 鄧婷
學生計畫編號： MOST 103-2815-C-040-028-E
研 究 期 間 ： 103 年 07 月 01 日至 104 年 02 月 28 日止，計 8 個月
指 導 教 授 ： 張啟昌

處理方式： 本計畫可公開查詢

執 行 單 位： 中山醫學大學醫學資訊學系

中華民國 104 年 03 月 31 日

整合資料探勘技術與集成學習架構預測卵巢癌復發

摘要

卵巢癌在臨床上通常是依據疾病的發展提供適合的進程治療。因此，對於癌症復發徵候的偵測及其後續無症狀復發事件的觀察而言，是與個體的存活率密切相關。過去很多研究將變因的觀察以全民健保資料庫抽樣檔的門診處方及治療明細檔作為資料分析，缺乏實際觀察個別病患深入特定臨床路徑的移轉、復發和治療的時序關聯樣式，以提供臨床醫師對可能的病情發展有更多資訊可參考。因此，為了提高治癒率與存活率，從實際診療紀錄中找出預測復發因子提供臨床醫師治療的資訊是非常關鍵且重要。本研究規劃使用支援向量機、快速學習器、C5.0 決策樹、MARS 以及隨機森林五種資料探勘演算法，並利用集成學習策略提高整體的準確度，改善一般學習演算法的缺點。本研究所需的病歷記錄和病理資料的來源為中山醫學大學附設醫院癌症防治中心。初步經由三位資深臨床醫師討論的復發的危險因子有：年齡(Age)、組織型態(Histology)、分化(Grade)、病理 T(Pathologic T)、病理 N(Pathologic N)、病理 M(Pathologic M)、病理期別(Pathologic Stage)、FIGO 期別(The International Federation of Gynecology and Obstetrics)、手術邊緣(Surgical Margins)、體能狀態(Performance status)、CA125、適當減積(Operation Optimal Debulking)、化療指引(Chemotherapy Guideline)，資料清理後共計有效個案 987 筆。有鑑於過去單一學習方法在分類預測準確性的缺點：統計問題、計算問題和代表性問題。本研究除了使用支援向量機、快速學習器、C5.0 決策樹、MARS、隨機森林資料探勘演算法分析外；針對卵巢癌復發的特性，我們加以考量納入集成學習架構，亦即第一階段採用集成學習投票機制，先行篩選出重要變數後，再進行第二階段支援向量機、快速學習器、C5.0 決策樹、MARS、隨機森林資料探勘演算法分析。結果顯示，各種方法預敏感度與特異度各有不錯的預測結果，進一步整合資料探勘技術與集成學習架構，經由變數篩選後各種方法能夠經由集成學習機制突顯各種方法的分類準確率，可以有效改善單獨資料探勘技術方法的預測結果。

關鍵字：卵巢癌復發、資料探勘、集成學習

Abstract

This study applied advanced machine learning techniques and combined with ensemble learning, widely considered as the most successful method to produce objective to an inferential problem of recurrent ovarian cancer. In this study, five machine learning approaches including SVM(support vector machine), C5.0, ELM(extreme learning machine), MARS(Multivariate Adaptive Regression Splines) and RF(Random Forests) were considered to find important risk factors and to predict the recurrence-proneness for ovarian cancer. Furthermore, we use ensemble learning to improve the defect of classification accuracy used normal machine learning: first, selecting important risk factors by ensemble learning, then, using the five machine learning approaches to analyze again. The medical records and pathology were accessible by the Chung Shan Medical University Hospital Tumor Registry. The existing literature on recurrent ovarian cancer reveals that factors include Age, Histology, Grade, Pathologic T, Pathologic N, Pathologic M, Pathologic Stage, The International Federation of Gynecology and Obstetrics (FIGO), Surgical Margins, Performance status, CA125, Operation Optimal Debulking, Chemotherapy Guideline. There are totally 987 patients in the data set. In our study, C5.0 is the superior approach in predicting recurrence of ovarian cancer. Moreover, the classification accuracy of C5.0, MARS, RF and SVM indeed increases after using ensemble learning. Particularly the classification accuracy of C5.0 obviously improves by using ensemble learning.

(一) 研究動機與研究問題

卵巢癌是台灣婦科癌症中死亡率占第三位，僅次於乳癌與子宮頸癌 (衛生福利部, 2013)。對於女性來說卵巢癌是重要的醫療問題，因卵巢癌的發生率與死亡率雖然比子宮頸癌低，但致死率卻高居婦科癌症之首 (台灣婦癌醫學會,2006)，晚期的卵巢癌復發率很高，一旦復發，能存活的機率很低 (李耀泰等人, 2007)。卵巢癌是婦科常見的癌症之一，屬於實體性癌症。當正常細胞轉變成癌細胞時，會出現異常(不受控制的)分裂情形，並形成腫瘤，腫瘤會對卵巢附近的器官造成壓迫，此外，癌細胞也可能會從腫瘤擴散到身體的其他部位，產生新的腫瘤(轉移)(台灣癌症防治網, 2013)。大部份的卵巢癌早期無明顯的症狀，大都是因為在做子宮頸檢查或大腸直腸檢查時發現，卵巢癌不像子宮頸癌或大腸直腸癌有比較明確且快速的方法可以被篩檢出來，一旦被發現時通常都已經擴散，造成相當高的死亡率，像是最常見的上皮性卵巢癌 (epithelial ovarian cancer) (黃仁治等人, 2012)，上皮性卵巢癌是婦科癌症中死亡率最高的一種，主因發現時約 75%已為晚期病灶(3、4 期)，即使以進步的手術加上新而更有效的藥物，這些晚期病灶大於 60%仍會復發 (李耀泰等人, 2007)，所以若能早期發現及治療，就能提高治癒率及減少死亡率。由於治療卵巢癌復發然是一項臨床挑戰，許多研究試圖找出影響卵巢癌復發因素，更可提高臨床的管理。研究顯示復發因素包括 13 項：(1)年齡；(2)組織型態；(3)分化；(4)病理 T；(5)病理 N；(6)病理 M；(7)病理期別；(8)FIGO 期別；(9)手術邊緣；(10)體能狀態；(11)CA125；(12)適當減積；(13)化療指引。

隨著資訊技術的發展，資料探勘(Data Mining)技術逐漸成為臨床診療指引及教學研究上最有價值的工具。所謂的資料探勘又稱之為機器學習(Machine Learning)就是從儲存於資料庫中的資料表、資料記錄及資料欄位內容裡的大量資料中分析出我們所感興趣而隱藏於資料集內的重要資訊。利用資料探勘方法的分類技術也已經成為國內外熱門的研究領域，在此種情況下，使用現代的資料探勘方法可找出卵巢癌復發重要因子之間的關聯(Ho,2004 & Thangavel, 2006)。

一般而言，輸出結果只產生一個假說的分類演算法普遍都會遭遇三個問題：統計問題、計算問題和代表性問題。然而，這些問題通常是可以透過集成學習的方法加以解決的。首先對於統計問題部分，當分類演算法的訓練資料數量過於龐大假說時，就會產生所謂的統計問題。因此，若能夠針對所有分類器所進行投票機制，將可有效的降低這種風險。其次是計算問題，常因為在分類演算法不能保證可以從假說找到一個最好可能發生狀態時，就像統計問題一樣，如果可以使用加權方式，是可以有效降低選擇錯誤而陷入本地最佳化的危險。最後關於代表性的問題，當假說不包含任何真實函式 f 時，代表性問題就會產生。若能提供不同權重的投票方法賦予假說條件，整體分類演算法是可以找到一個非常接近真實函式 f 的近似值。因此，過去的研究報告證實集成學習架構能降低學習演算法的偏差和變異。

有鑑於過去單一學習方法在分類預測準確性的缺點：統計問題、計算問題和代表性問題。本研究除了使用支援向量機、快速學習器、C5.0 決策樹、MARS 資料探勘演算法分析外；針對卵巢癌復發的特性，我們加以考量納入集成學習架構，亦即第一階段採用隨機森林演算法，先行篩選出重要變數後，再進行第二階段支援向量機、快速學習器、C5.0 決策樹、MARS 資料探勘演算法分析。本研究工作有 ：

- 採用五種不同的資料探勘方法預測準確度，希望透過各方法分析出來的結果，找出準確度較高的預測模型。
- 利用集成學習策略提高整體資料及的預測準確度、敏感度及特異度，改善一般學習方法的缺點。

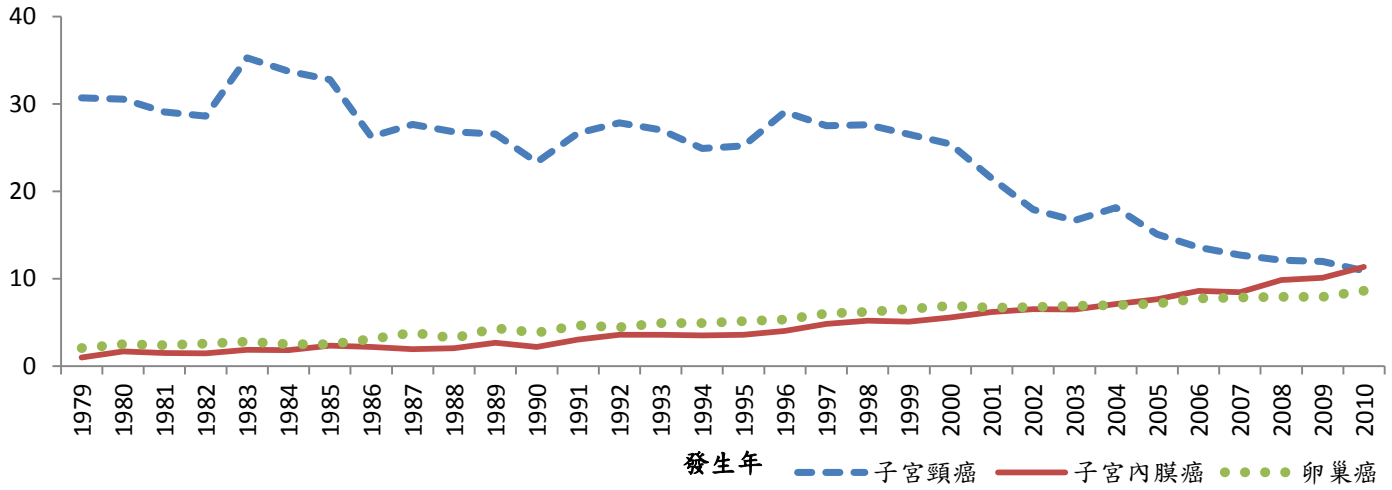
(二) 文獻回顧與探討

本研究文獻探討內容分為兩個部分討論:卵巢盛行率、資料探勘方法與集成學習。

一、卵巢癌盛行率

卵巢癌是婦科腫瘤導致死亡的主要原因。在 2010 年，美國估計大約有 21,880 人被診斷為卵巢癌，

有 13,850 人因此而死亡。卵巢癌罹患率會隨著年齡增長而增加，其中以 50-59 歲最高；有卵巢癌家族史是最高的危險因素(黃仁治等人,2012)。台灣地區婦女卵巢癌在 101 年死亡數為每十萬名女性 528 例，而年齡化標準化死亡率是每十萬例 3.2 人，平均年齡約在 50-59 歲，在婦科癌症死亡率中占第三位，僅次於女性乳癌及子宮頸癌，且發生率有逐年增高的趨勢。



圖一、1979-2010 台灣子宮頸癌、子宮內膜癌、卵巢癌發生率(每 10 萬人口)
(資料來源:衛生福利部國民健康署, 2013)

二、資料探勘方法

在醫療領域中，資料探勘的應用可以被用來預測疾病模式，並可預測不同群體之間的重要因子。本研究將應用以下五種不同的資料探勘方法來預測卵巢癌復發的重要因子：

(1) Support Vector Machine(SVM)：支援向量機是由 Vladimir Vapnik 從 1995 年開始發展的一種分類方法，被視為最具成效的監督式學習方法之一。現在成為資料探勘標準工具之一(Shutao et al., 2003)。SVM 的特性是將輸入空間(Input Space)先使用非線性的對應 Mapping 轉換到高維度的特徵空間(Feature Space)再做分類。其中 SVM 所使用的 Mapping 在選擇所對應的核心函數上有很大的彈性，且需為非線性的函數，之後將 Mapping 到高維度的特徵空間中的資料建構線性分類式子，選擇能使分類錯誤降到最小的權重，得到最大化邊界超平面(Maximal Margin Hyperplane)，以完成分類(Mao et al., 2005)。以下簡述 SVM 相關研究及運用：David 研究一個支援向量機對醫學實際資料做分類，利用量測位於螢光上交雜(Fluorescence In-Situ Hybridization, FISH)影像細胞發生的訊號，去診斷發生的併發症狀，研究中突出測試圖樣距離的門檻值，從 SVM 分割超平面去拒絕錯誤分類的圖樣，因此可減少誤差的發生，研究結果與其他先進儀器比對，指出基於 SVM 發展診斷系統的潛力(David and Lerner, 2004)。

(2) C5.0：C5.0 演算法又稱為規則推理模型(rule-based reasoning model)，是 C4.5 演算法的修訂版，屬於監督式學習的一種，適用在處理大資料集，採用 Boosting 方式提高模型準確率，又稱為 Boosting Trees，在軟體上的計算速度比較快，佔用的記憶體資源較少，主要能解析連續型變數與類別型變數，結果可產生決策樹(decision tree)或規則集(rule sets)。張惟智(2009)使用 C5.0 決策樹及類神經網路找出腹主動脈瘤手術併發症三大類併發症的分類規則及重要因子，再利用貝氏網路，找出重要因子間的因果關係並計算出其聯合條件機率。

(3) Extreme learning machine(ELM):快速學習器(Extreme learning machine, ELM)是一種新型態的類神經網路架構，「快速學習的理論與應用」一文於 2006 年由 G.B. Huang、Q.Y. Zhu 和 C.K. Siew 共同發表於「Neurocomputing」上。有別於其他類神經網路，快速學習器採用截然不同的演算規則，屬於單一隱藏層的前饋式類神經網路模式(Single hidden Layer Feed-forword neural Network, SLFN)，其輸入層到隱藏層間的權重稱之為輸入權重，是隨機產生的，而隱藏層到輸出層之間的權重則稱為輸出權重，

是由 MP 轉置矩陣(Moore-Penrose inverse)分析後得到，ELM 的學習速度相較於傳統的陡坡降法(gradient-based)明顯快速許多，許多文獻皆已證實此一特點(歐宗殷， 2010)。Vani 等人(2010)使用 ELM 方法應用於乳房 X 光檢查異常的分類，結果表明 ELM 方法在分類乳房 X 光檢查異常的效能優於其他演算法。

(4) Multivariate Adaptive Regression Splines(MARS)：多元適應性雲形回歸(Multivariate Adaptive Regression Splines, MARS)是由 Friedman 等人提出來處理多元複雜資料問題的新方法。MARS 目前較被廣泛運用的領域，大部份是資料探勘的分類問題，以及預測。Falk 等人(2006)比較了 CART、MARS 和 GMR 方法在預測經歷癌症切除患者乳腺癌復發時間。結果表明，GMR 算法證明相較於 CART 和 MARS 有較好的效率。

(5) Random Forests (RF)：隨機森林是 Breiman(2001)提出的一個新式決策樹演算法。是一整合多決策樹進行分類預測與重要變數(variable importance)，(Breiman, 2001 ; Liaw and Wiener, 2002; Svetnik et al., 2003) 。採用分類迴歸樹(Classification and Regression Trees, CART)作為元分類器，將變數隨機投入，以 Gini 方式進行子節點分裂，Bagging 方式得出整合分類結果。隨機森林不同於傳統決策樹是，傳統決策樹，僅以單一決策樹為單位作出決策，隨機森林則以多個決策樹整合得出分類結果。對於分類與規則上相較於舊有的 CART、CHAID 與 C5.0…等擁有精確的分類預測能力(許智宇,2010)。李放歌(2011)等人認為隨機森林方法已經被用來研究乳腺癌和哮喘, 研究顯示交互作用對疾病發生有影響。

三、集成學習

集成學習(ensemble learning)指的是透過多個分類工具加以整合成為一個新的綜合分類器，它的優點是能提供給預測模型不錯的泛化能力，進而成為一個強學習器。整體學習演算法的運作是透過多次執行基礎學習演算法，並且針對每次產生的假說進行投票，最後整合投票的結果構成一致同意的假說。一般而言，設計整體學習演算法的技巧有兩種主要的方法。第一種方法是「使用獨立的模式去建造每一個假說」，每一單獨的假說對於新資料點的預測，具有某一個合理低的出錯率，但是假說和假說彼此之間，在大多數預測裡常常是不一致的。如果能夠統合單獨假說的預測，並建立一個具有整體性時，會比起任何一個單獨或個別的分類器更具有高準確度的預測；第二種設計整體學習的方法是「採用連接模式來建造假說」。此一連接模式是把權重高的票投給和實際資料誤差小的假說，然後把權重低的票投給和實際資料誤差大的假說，藉由不同權重的投票方式結合所有的假說，並產生一個比任何單獨假說都逼近實際資料的整體假說。因應卵巢癌復發的臨床反應，本研究將採取第二種設計整體學習的方法：採用連接模式來建造假說，迫使學習演算法產生多樣化特性的目的是在每次呼叫學習演算法時，都採用一個具有不同輸入特徵的子集合。利用隨機森林整合多決策樹去選取輸入特徵的復發因素，最後形成一個群體性特徵的重要變數，最後導致的整體性的分類結果比任何個別資料探勘方法處理的結果更為準確。在此情況下更能針對復發因素：(1)年齡；(2)組織型態；(3)分化；(4)病理 T；(5)病理 N；(6)病理 M；(7)病理期別；(8)FIGO 期別；(9)手術邊緣；(10)體能狀態；(11)CA125；(12)適當減積；(13)化療指引，進行病患特性更深入的臨床解釋。

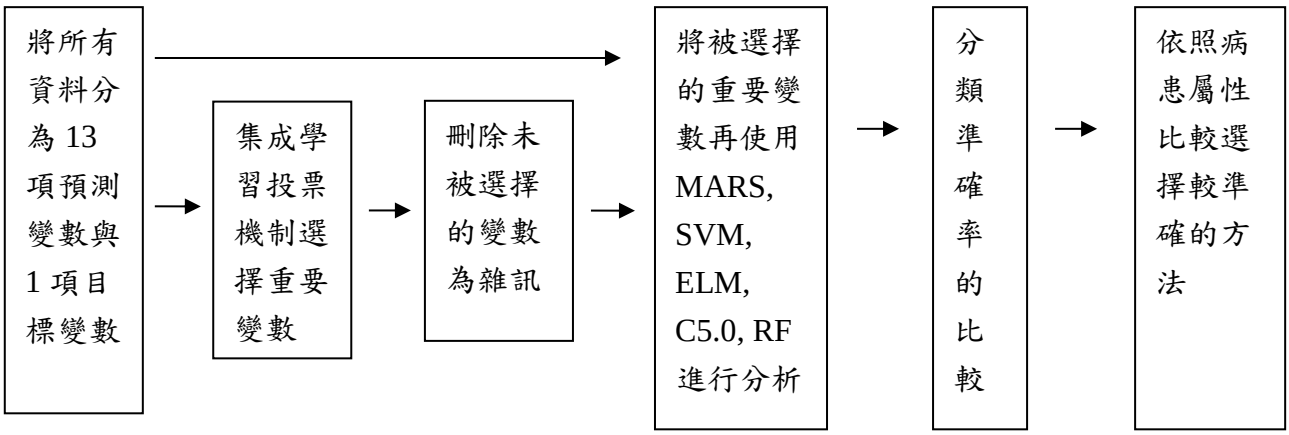
國際機器學習界權威 T.G. Dietterich 指出當前常見的三種集成學習策略，分別是 Bagging, boosting 和 stacking。Bagging 針對相同的演算法，去訓練出多個分類器，使用非加權的方法進行投票，即採用多數決的方法作為最後集成模型的決策。而 Boosting 利用類似 bagging 的作法，皆選用相同的演算法去訓練出多個分類器，兩者差別在於 Boosting 是採用各分類器的預測結果作加權投票準確率也較 bagging 高。Stacking 和前兩種策略最主要的不同在於可以使用不同的演算法去得到多元的分類器，在決策結果上，則可使用加權或不加權投票的處理方式(洪智力、陳勁宏， 2007)。從另一個角度來看，整體學習也可以是一種附加模型(additive model)。所謂的附加模型通常是指一個新增的資料點，最後所指定的類別標籤，是由部分或所有的附屬模型(component model)經由賦予不等的權重後，再加總所得到的結果。Freund 和 Schapire(1996, 1997)提出的 Adaboost 演算法，可以說是建造附加模型極有效的方法。透過學習演算法，極盡可能地將分類錯誤減少到最小的方式去產生一個假說，每次增加一個假說到整體學習之中，分類錯誤就相對的降低。在多數的研究實驗中(Freund & Schapire, 1996;

Dietterich, 2000)都說明了 Adaboost 確實可以提供大部分數據資料最好的表現結果。若針對包含較多貼錯標籤(mislabeled)的訓練資料來說，Adaboost 把非常高的權重放在雜訊的資料點上，然後生成一個非常差的整體分類器。目前確實有許多的研究工作，著重在如何延伸 Adaboost 的功能，使之能夠在處理較高雜訊的訓練資料(莊永裕，2006)。

(三) 研究方法與步驟

研究架構:

為了比較重要變數篩選的差異，研究設計架構如圖二所示：在圖二中，首先依據文獻查證與臨床醫師討論後決定 13 項預測變數((1)年齡；(2)組織型態；(3)分化；(4)病理 T；(5)病理 N；(6)病理 M；(7)病理期別；(8)FIGO 期別；(9)手術邊緣；(10)體能狀態；(11)CA125；(12)適當減積；(13)化療指引)進行復發的預測。在圖二上方研究流程中未經變數篩選直接以 SVM、C5.0 決策樹、ELM、MARS 方法進行預測；在圖二下方研究流程中則是藉由集成方法，經過變數篩選後再以 SVM、C5.0 決策樹、ELM、MARS 方法進行預測。在進一步比較兩個流程所分析分類準確率；最後針對所分析的變數結果，依照病患屬性完成臨床後續預測卵巢癌復發重要因子的建議。



圖二、研究流程圖

研究方法：

在醫學衛生領域中，資料探勘應用已大幅成長，因為可以被用來直接取得預測不同群體之間患者的相關資訊。我們最好的資料探勘方法分類技術尚未被利用於分析卵巢癌復發。因此，本研究我們試圖利用五種資料探勘方法由卵巢癌的資料庫中進行分類並進一步分析。

研究工具:

一、支援向量機(Support Vector Machine, SVM)

支援向量機廣泛被使用來處理統計分類及回歸分析，適合應用於解決具有較小範圍、非線性及高維度等特性的問題。從有限的訓練樣本中學習得到決策規則，對獨立的測試集合仍能夠得到較小的預測誤差。支援向量機將資料映射至高維空間當中，希望從映射過後的結果找出一個可將資料分隔成兩組不同集合的超平面(hyperplane)。透過此超平面分類方法對資料進行分類，區分出互不重疊的分類集合。支援向量機從二維空間中找出一條分隔線區分兩種類型資料，且此分隔線與兩集合之距離越大越好，藉由此分隔線對資料進行分類。以分隔線將資料分隔成兩組不互相重疊之集合，並可找出集合中最鄰近分隔線且各自平行於分隔線的兩條平行線。SVM 算法如下：假設$\{(x_i, y_i)\}_{I=1}^{N}, x_i \in R^d, y_i \in \{-1,1\}$, 資料集合為可輸入向量之訓練組，N 為樣本數量，而 d 為每一觀測值之維度。y_i是已知的目標。此算法為了求超平面(hyperplane)$w \cdot x_i + b = 0$其中 w 為超平面向量，b 為偏移量，區分兩超平面的最大寬度為$2 / \|w\|^2$，所有在範圍內的點皆稱為支援向量(Vapnik, 2000)。

$$\text{Min}\Phi(x) = \frac{1}{2}\|w\|^2 \quad (1)$$

$$\text{S.t.}\, y_i(w^T x_i + b) \geq 1\,, i = 1,2,\dots,N$$

由於(1)式較難解，需透過拉格朗乘數法(Lagrange method)將理想化問題轉換成對偶問題。拉格朗乘數法的數值為非負實係數，(1)式被轉換為以下形式：

$$\text{Max}\Phi(w,b,\xi,\alpha,\beta) = \sum_{i=1}^{N}\alpha_I - \frac{1}{2}\sum_{i=1,j=1}^{N}\alpha_i\alpha_j y_i y_j {x_i}^T x_j \quad (2)$$
$$\text{S.t.}\sum_{j=1}^{N}\alpha_j y_j = 0, 0 \leq \alpha_i \leq C, i = 1,2,\dots,N$$

在(2)式中C為懲罰因子並決定懲罰的權重，被視為可調整參數，用於控制最大極限與分類誤差之間的交換。一般情況下，在所有可應用的數據無法找到線性分離的超平面，最佳的解決方法為將原始非線性數據轉換為更高線性分離的維度。最常見的核心函數為線性、多項式、半徑式函數(RBF)。雖然核心函數具多種選擇且可被利用的，但RBF仍較被廣泛使用。其定義為：

$K(x_i, x_j) = exp(-\gamma\|x_i - x_j\|^2), \gamma \geq 0$, (Vapnik, 2000)其γ表示RBF寬度。因此RBF被用於研究上。最原先的SVM設計為二元分類，構建多類SVM仍然是一個正在研究的問題。本研究我們將使用二元分類SVM方法(Hsu and Lin, 2003)。

二、C5.0 決策樹

C5.0分類器是將一龐大數據分類與分析出隱藏資料的方法，亦可用決策樹呈現出有用的資料(Larose, 2005)。此算法採用決策樹，由循環式劃分與採用選擇的方式在訓練組主要部分中取得方法。C5.0由C4.5改善了一些問題。如：變得更快、記憶效率更高、透過更小的決策樹區分較相似的結果、準確度更高、權重不同與分類錯誤的型態、降低干擾(Larose, 2005)。C4.5中Quinlan(1993)利用具信息熵概念的ID3演算法(Iterative Dichotomiser 3)由一組已分類的訓練組建立決策樹，訓練組資料擴大由每個樣本所屬類別包括屬性向量，每個資料屬性可以用來做決策。C4.5在決策樹的每個節點上使用資訊獲取量(Information Gain)來選擇測試屬性，選擇最高資訊獲取量的屬性作為節點的測試屬性。該屬性使得對產生之劃分中的樣本分類所需的資訊量最小，能反應劃分的最小隨機性與不純性(impurity)(Han and Micheline, 2001)。以計算A的屬性為例，計算資訊獲取率GainRatio(A)，S表一資料樣本集，p_i為屬於B_i的任意樣本概率。假設有*n*個不同類B_i的值,其中*(i = 1, … ,n)*，假設S_i類別*B*的樣本數，*Info(S)*表示在現有樣本內的信息熵，計算過程如下：

$$Info(S) = \sum_{i=1}^{n} p_i \log(p_i) \quad (3)$$

假設 A 屬性有 *n* 個不同值$\{A_1, A_2, \dots, A_n\}$，使用 A 將 S 劃分為 *n* 個子集合$\{S_1, S_2, \dots, S_n\}$，S_j為A_j在子集合中的樣本數，S_{ij}為S_j子集合中B_i類別的樣本數，*Info(S, A)*為要計算的信息熵。計算過程如下：

$$Info(S,A) = \sum_{j=1}^{n}\frac{S_{1j}+S_{2j}+\cdots+S_{nj}}{S} Info(A) \quad (4)$$

以分割的信息 *SplitInfo (A)*是 *S* 裡每個屬性 A 的熵值，用來消除有大量屬性值誤差。計算過程如下：

$$SplitInfo(A) = -\sum_{i=1}^{n}\frac{|S_j|}{|S|} log\left(\frac{|S_j|}{|S|}\right) \quad (5)$$

$$Gain(A) = Info(S) - Info(S,A) \quad (6)$$

$$GainRatio(A) = Gain(A)/SplitInfo(A) \quad (7)$$

三、Extreme learning machine(ELM)

快速學習器是由Huang於2004年提出的單隱藏前饋式類神經網路(SLFNs)演算法(Huang et al., 2006)，可隨機輸入權重與分析輸出權重。本節將介紹單一隱藏層網路的矩陣數學描述，並說明快速學習器演算法。給定 N 個任意的輸入輸出樣本(x_i, t_i)，i =1,...,N，其中：$[x_{i1}, x_{i2}, \dots, x_{in}]^T \in R^n$以及$t_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in R^m$，標準的單一隱藏層網路$\tilde{N}$個隱藏節點以及激活函數(Activation function)$g(x)$可以近似 N 個樣本達到平均零誤差。數學模型為以下式子：

$$H\beta = T \text{ ，} \tag{8}$$

其中

$$H(w_1, \dots, w_{\tilde{N}}, b_1 \dots, b_{\tilde{N}}, x_1 \dots, x_N) = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}};$$

$$\beta_{\tilde{N} \times m} = \left(\beta_1^T, \dots, \beta_{\tilde{N}}^T\right)^t; T_{N \times m} = \left(T_1^T, \dots, T_N^T\right)^t$$

其中$w_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$，$i$ =1,2,...,$\tilde{N}$,為權重向量連接第 i 連接第 i 個隱藏節點和輸入節點 $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$為權重向量連接第 i 個隱藏節點和輸出節點，b_i為第 i 個隱藏節點的開端，$w_i \cdot x_j$ 表示 w_i 和 x_j 的內積。H被稱作網路隱藏層輸出矩陣(Hidden layer output matrix of neural network)；H的 i 行是 i 個隱藏節點的輸出向量跟輸入樣本$x_1, x_2, \dots, x_N$之間的關係，而H的 j 列是隱藏層輸出向量跟輸入樣本 x_j 之間的關係。因此，測定輸出權重(連結隱藏層到輸出層)與找到最小平方解法得到線性系統一樣簡易。透過最低標準LS解法得到線性系統需利用以下式子：

$$\hat{\beta} = H^{\Psi} T \tag{9}$$

是根據 Rao(1971)和 Serre(2002)的 Moore-Penrose 廣義逆矩陣 H，而具有最低的標準的 LS 解法是獨一無二的。快速學習器算法步驟如下：給一訓練樣本集合 $\aleph = \{(x_i, t_i) | x_i \in R^n, t_i \in R^m, i = 1, \dots, N\}$、激活函數 $g(x)$，以及隱藏節點數 $\tilde{N}$。

步驟1. 隨機給一輸入權重 w_i 以及閥值 b_i，$i = 1, \dots, \tilde{N}$。

步驟2. 計算隱藏層輸出矩陣H。

步驟3. 計算輸出權重$\hat{\beta}$。$\hat{\beta} = H^{\Psi} T$其中T=$[t_{1,\dots,} t_n]^T$。

四、Multivariate Adaptive Regression Splines(MARS)

MARS是一個新興的多元適應性雲形回歸數迴歸程序技術，是藉由採用數個線段規則—也就是BF的累加模型解釋非線性狀態的工具(Friedman, 1990)。

$$\hat{f}(x) = a_0 + \sum_{k=1}^{M} a_m \prod_{k=1}^{Km} [S_{km} \cdot (x_{y(k,m)} - t_{km})] \tag{10}$$

上面的方程式是通用的MARS模型，其中BF則是後段累乘的部分(如下所示)，主要是根據需求而變化。

$$B_m(x) = \prod_{k=1}^{Km} [S_{km} \cdot (x_{y(k,m)} - t_{km})] \tag{11}$$

其中a_0與a_m皆為參數值，主要是給予類似迴歸係數的功能；M 為 BF 的個數，經由評估準則決定；K_m為切割的折點個數；S_{km}之值為+1or-1，其作用為顯示方向；$v(k,m)$是對變數的標示；最後t_{km}則是各節點的分界點(數值)。MARS1.0 應用軟體中的 basis function 是以下列的形式表現:

$$max(0, X - c) \quad or \quad max(0, c - X) \tag{12}$$

而在給定目標變數和一個可選擇預測變數的集合下，MARS 令所有模型建立及調度自動化(Steinberg

etal., 1999)，其中包括：將有意義的變數與較不恰當的變數分開、對與目標值呈現非線性關係的預測變數做轉換、決定預測變數間的交互作用、採用新的變數群聚技術來處理遺失值問題、採用大量的自我測試來避免過度配適。

我們可以將 BF 視為每一段規則中所屬的解釋方程式，而每個 BF 則是經由評估其損適性(Loss of Fit, LOF)之判斷標準決定所包含之影響變數的個數，並同時經由前推式及後推式演算法尋找較適當的折點數以及交互作用以解決高維度資料的各種問題(Friedman, 1990)。根據 LOF 決定 BF 個數時，主要是參酌各個 BF 在加入後，是否在主要模型中具有貢獻性，評估其表現是否在可接受的範圍內，以降低模型的複雜度，加速其對資料的處理及判斷。而其權衡的概念如上述，是以 LOF 的觀念加以判斷，而所採用的方法為 GCV(generalized cross-validation)，是由 spline 的研究先趨(Craven and Wahba, 1979)所提出的判斷準則，下列為 GCV 的處理方法：

$$LOF(\hat{f}_M) = GCV(M) = \frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}_M(X_i)]^2 / [1 - \frac{C(M)}{N}]^2 \tag{13}$$

其中C(M)為採用m個BF所需付出的成本，而最主要的概念是來自下列方程式：

$$\Delta[\hat{f}(x), f(x)] = [\hat{f}(x) - f(x)]^2 \tag{14}$$

即針對推估出來的值與實際值的比較，所不同的是增加 BF 的成本觀念，以節省無謂的運算時間。相關研究步驟如下：1. 取得卵巢癌資料庫數據為研究對象。為確保資料的完整性、一致性，將進行資料編碼，不同的數值型態與臨床醫師討論進行轉換，並做分類，最後刪除缺失欄位過多的資料。2. 將資料分為 13 項預測變數與 1 項目標變數。3. 利用 C5.0 選擇重要變數。4. 將選擇出的重要變數，利用 MARS,SVM,ELM,C5.0 再次分析，其他未被選擇的重要變數則被視為影響分析的雜訊從資料中刪除。5. 比較各種方法的預測準確率。6. 最後，與臨床醫師討論並證實重要變數的可信度，並且可以依照病患的屬性決定只用何種方法預測最為準確。

五、 Random Forests (RF)

隨機森林演算法是將多數類樣本劃分為數個獨立的子集合；然後將每一個獨立子集合進行交叉組合以構成不同的訓練樣本集，並針對不同的訓練樣本集利用決策樹分類器加以學習；最後根據平均加權法產成隨機森林，進而獲得決策規則(吳華芹，2013)。計算方法為給定 K 個分類器以及隨機向量 x、y，定義邊際函數如下：(張華偉等人，2006)

$$mg(x,y) = av_k I(h_k(x) = y) - max_{j \neq y} av_k I(h_k(x) = j) \tag{15}$$

其中，I(•)是可能性函數，邊際函數顯示向量X所得到正確分類Y的平均得票數超過其它任何類平均得票數的程度。由此可知邊際越大分類的可信度就越高。分類器誤差定義：

$$PE^* = P_{x,y}(mg(x,y) < 0)$$

將上面的結論推廣到隨機森林函數： $h_k(X) = h(X, \theta_k)$，邊際函數如下：

$$mr(x,y) - P_\theta(h(x,\theta) = y) - max_{j \neq Y} P_\theta(h(x,\theta) = j) \tag{16}$$

隨著樹的數目增加，PE* 就會趨向於

$$P_{x,y}(p_\theta(h(x,\theta) = y) - max_{j \neq Y} p_\theta(h(x,\theta) = j) < 0) \tag{17}$$

而分類器{h(X,θ)}的強度可以表示為

$$s = E_{X,Y} mr(x,y) \tag{18}$$

假設$s \geq 0$，根據契比雪夫不等式，(16),(17)兩式可以得到：

$$PE^* \leq var(mr)/s^2 \tag{19}$$

根據Breiman(2001)可推導出

$$\begin{aligned} var(mr) &= \bar{\rho}\left(E_\theta sd(\theta)\right)^2 \\ &\leq \bar{\rho} E_\theta var(\theta) \\ &\leq 1 - s^2 \end{aligned} \tag{20}$$

隨機森林的目標誤差上界是$PE^* \leq \bar{\rho}(1 - S^2)/S^2$

研究步驟:

1. 取得卵巢癌資料庫數據為研究對象。為確保資料的完整性、一致性，將進行資料編碼，不同的數值型態與臨床醫師討論進行轉換，並做分類，最後刪除缺失欄位過多的資料。
2. 將資料分為 13 項預測變數與 1 項目標變數
3. 利用機器學習法 C5.0, RF, MARS, ELM, SVM 進行預測，並選出貢獻率高的變數
4. 利用集成學習策略篩選並剔除低貢獻率之變數除，再次利用五種方法分析分類準確率之變化。
5. 比較各種方法的預測準確率，以及比較集成學習方法與一般機器學習。
6. 最後，與臨床醫師討論並證實重要變數的可信度，並且可以依照病患的屬性決定只用何種方法預測最為準確。

(四) 實證研究

在研究中，我們由中山醫學大學附設醫院癌症防治中心癌症登記資料庫提供的卵巢癌數據集，使用C5.0、MARS、RF、SVM、ELM驗證其敏感度與特異度，並預測卵巢癌復發之重要因子。數據集中共包含13個預測變數，分別為年齡(Age)、組織型態(Histology)、分化(Grade)、病理T(Pathologic T)、病理N(Pathologic N)、病理M(Pathologic M)、病理期別(Pathologic Stage)、FIGO期別(The International Federation of Gynecology and Obstetrics)、手術邊緣(Surgical Margins)、體能狀態(Performance status)、CA125、適當減積(Operation Optimal Debulking)、化療指引(Chemotherapy Guideline)，以及1個目標變數為復發型態(Type of Recurrence)，共987筆資料，隨機選取300筆資料為測試樣本，其餘687筆資料為訓練樣本，進行重複取樣十次。

以{1}代表復發；{2}則代表未復發。因此{1-1}代表：原始的判定為復發，而經由模式判定後亦為復發；而{2-2}則表示：原始判定為沒有復發，經由模式判定亦為沒有復發。由表一可知C5.0的整體正確判別率為75.67%，而個別的判別正確率{1-1}的比率為84.04%：即原始群體為第1類的樣本正確的被判別到第1類的比率為84.04%。其中有30個原本群體為第1類的樣本，被錯分為第2類的群體中；而有43個原本群體為第2類的樣本，被錯分為第1類的群體中。

表一、C5.0之預測結果

實際類別	預測類別	
	1 (有復發)	2 (未復發)
1 (有復發)	158 (84.04%)	30 (15.96%)
2 (未復發)	43 (38.39%)	69 (61.61%)
平均準確率	75.67 %	

由表二可知MARS的整體正確判別率為59.67%，而個別的判別正確率{1-1}的比率為87.23%：即原始群體為第1類的樣本正確的被判別到第1類的比率為87.23%；而{2-2}的判別正確率為13.39%。其中有24個原本群體為第1類的樣本，被錯分為第2類的群體中；有97個原本群體為第2類的樣本，被錯分為第1類的群體中。

表二、MARS之預測結果

實際類別	預測類別	
	1 (有復發)	2 (未復發)
1 (有復發)	164 (87.23%)	24 (12.77%)
2 (未復發)	97 (86.61%)	15 (13.39%)
平均準確率	59.67 %	

由表三可知RF的整體正確判別率為44.33%，而個別的判別正確率{1-1}的比率為27.13%：即原始群體為第1類的樣本正確的被判別到第1類的比率為27.13%；而{2-2}的判別正確率為73.21%。其中有137個原本群體為第1類的樣本，被錯分為第2類的群體中；而有30個原本群體為第2類的樣本，被錯分為第1類的群體中。

表三、RF之預測結果

實際類別	預測類別	
	1 (有復發)	2 (未復發)
1 (有復發)	51 (27.13%)	137 (72.87%)
2 (未復發)	30 (26.79%)	82 (73.21%)
平均準確率	44.33 %	

由表四可知SVM的整體正確判別率為61.33%，而個別的判別正確率{1-1}的比率為96.28%：即原始群體為第1類的樣本正確的被判別到第1類的比率為96.28%；而{2-2}的判別正確率為2.68%。其中有7個原本群體為第1類的樣本，被錯分為第2類的群體中；而有109個原本群體為第2類的樣本，被錯分為第1類的群體中。

表四、SVM之預測結果

實際類別	預測類別	
	1 (有復發)	2 (未復發)
1 (有復發)	181 (96.28%)	7 (3.72%)
2 (未復發)	109 (97.32%)	3 (2.68%)
平均準確率	61.33 %	

由表五可知ELM的整體正確判別率為63.67%，而個別的判別正確率以{1-1}的比率最高，為99.47%：即原始群體為第1類的樣本正確的被判別到第1類的比率為99.47%；而{2-2}的判別正確率為3.57%。其中有1個原本群體為第1類的樣本，被錯分為第2類的群體中；而有108個原本群體為第2類的樣本，被錯分為第1類的群體中。

表五、ELM之預測結果

實際類別	預測類別	
	1 (有復發)	2 (未復發)
1 (有復發)	187 (99.47%)	1 (0.53%)
2 (未復發)	108 (96.43%)	4 (3.57%)
平均準確率	63.67 %	

根據分析結果表示，ELM 模型在{1-1}最高的平均準確率為99.95% (實際有復發並會被預測為有復發)，而RF模型在{2-2}(實際未復發病患預測為未復發)中擁有最高的平均準確率90.80%。而整體而言，最高的平均準確率是由C5.0產生的75.87%。C5.0在整體準確率優於其他四種方法，表示C5.0確實比另外四個方法提供更佳的分類準確率。因此，C5.0是對於卵巢癌分類最有效的方法。

本研究第二階段使用集成學習投票策略，相關貢獻率之排序如表六所示，在C5.0，依被選取次數將重要變數排名依序為FIGO期別、病理T、病理M、化療指引、分化、病理期別、CA125、病理N、手術邊緣、體能狀態、適當減積、組織型態、年齡；MARS依貢獻率排名依序為FIGO期別、病理N、病理M、手術邊緣、體能狀態、CA125、適當減積、化療指引、病理T、病理期別、分化、組織型態、年齡；RF依貢獻率排名為年齡、組織型態、分化、病理期別、FIGO期別、病理T、病理M、體能狀態、病理N、手術邊緣、化療指引、適當減積、CA125；ELM依貢獻率排名為年齡、病理M、FIGO期別、手術邊緣、體能狀態、化療指引、CA125、分化、組織型態、病理期別、病理T、適當減積、病理N；SVM依貢獻率排名為年齡、病理T、組織型態、病理期別、病理N、FIGO期別、分化、病理M、手術邊緣、化療指引、體能狀態、化療指引、CA125。表七為經過集成學習投票機制選擇重要變數，根據結果依名次排序為FIGO期別、病理M、年齡、病理T、化療指引、病理期別、手術邊緣、體能狀態、分化、病理N、組織型態、CA125、適當減積。所有機器學習法經集成學習投票後FIGO期別為貢獻率最高之變數，因此可以得知FIGO期別是最重要的卵巢癌復發因子。另外，CA125、適當減積為貢獻率最低的兩個變數，因此我們將這2項變數從資料中刪除，且再次以五種機器學習法進行分析。

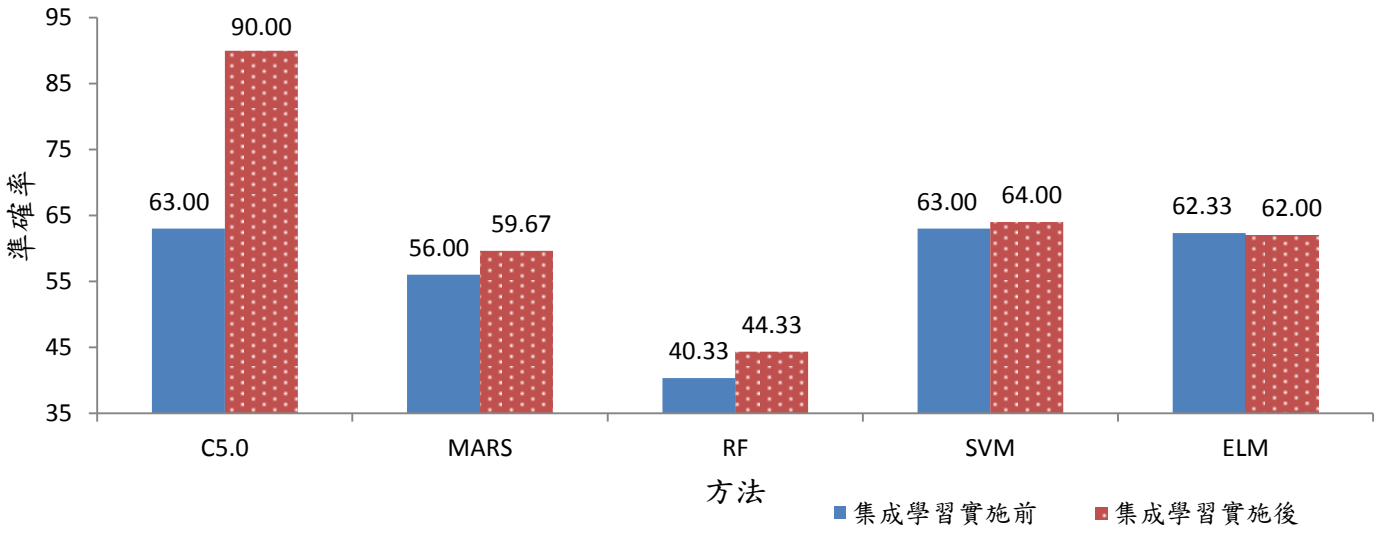
表六、各機器學習法之貢獻率比較

C5.0		MARS		RF		ELM		SVM	
變數	貢獻率%	變數	貢獻率%	變數	貢獻率%	變數	貢獻率%	變數	貢獻率%
FIGO期別	11.4754	FIGO期別	0.0017	年齡	24.1560	年齡	0.0360	年齡	3.0845
病理 T	9.8361	病理 N	-0.0004	組織型態	14.3140	病理M	0.0360	病理T	1.7030
病理 M	9.8361	病理 M	-0.0004	分化	11.2200	FIGO期別	0.0360	組織型態	1.4240
化療指引	9.8361	手術邊緣	-0.0004	病理期別	11.1800	手術邊緣	0.0030	病理期別	1.2905
分化	8.1967	體能狀態	-0.0004	FIGO期別	7.8080	體能狀態	0.0030	病理N	0.6114
病理期別	8.1967	CA125	-0.0004	病理T	7.8050	化療指引	0.0030	FIGO期別	0.5450
CA125	8.1967	適當減積	-0.0004	病理M	5.4860	CA125	0.0020	分化	0.5100
病理 N	6.5574	化療指引	-0.0004	體能狀態	5.0820	分化	-0.0240	病理M	0.4085
手術邊緣	6.5574	病理 T	-0.0010	病理N	4.5800	組織型態	-0.0310	手術邊緣	-0.0349
體能狀態	6.5574	病理期別	-0.0017	手術邊緣	3.5660	病理期別	-0.0310	化療指引	-0.1349
適當減積	6.5574	分化	-0.0048	化療指引	1.7320	病理T	-0.0320	體能狀態	-0.1689
組織型態	4.9180	組織型態	-0.0112	適當減積	1.5400	適當減積	-0.0320	化療指引	-0.2027
年齡	3.2787	年齡	-0.0336	CA125	1.5250	病理N	-0.0650	CA125	-0.2364

表七、集成學習重要變數之篩選結果

變數	貢獻率排名
FIGO期別	1
病理M	2
年齡	3
病理T	4
化療指引	5
病理期別	6
手術邊緣	7
體能狀態	7
分化	9
病理N	10
組織型態	11
CA125	12
適當減積	13

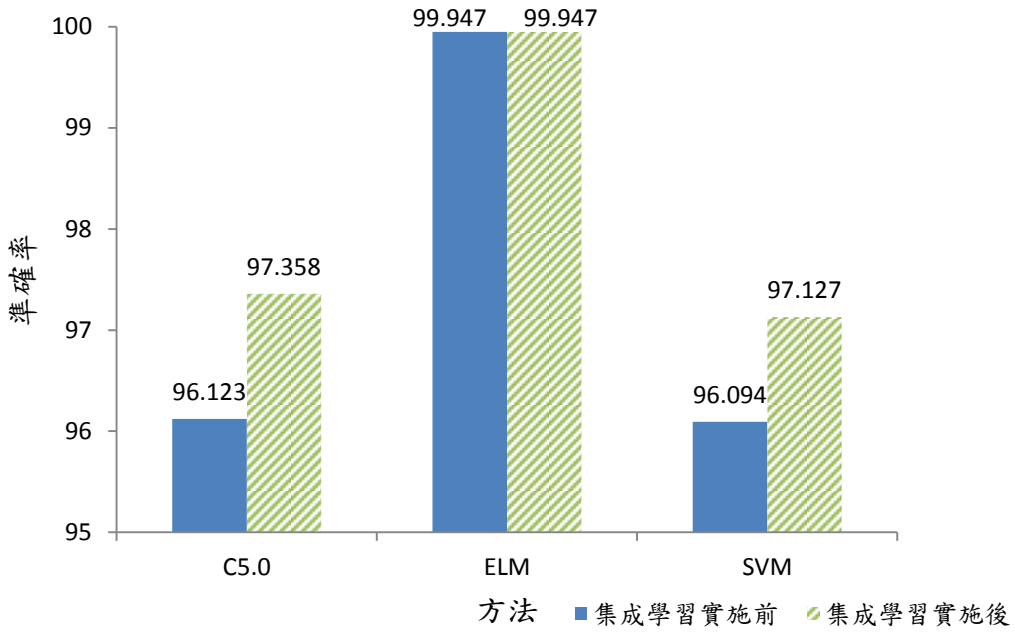
經過集成學習後分類準確率結果如圖三所示，C5.0、MARS、RF、SVM與ELM分類準確率分別為63.00%、56.00%、40.33%、63.00%與62.00%，經過集成學習分析後，五種方法之分類準確率變為90.00%、59.67%、44.33%、64.00%與62.00%。可以發現C5.0經過集成學習透票機制變數篩選後，其分類準確率明顯提高，MARS之分類準確率由56.00%提高至59.67%，RF之分類準確率由40.33%提高至44.33%，SVM之分類準確率由63.00%提高至64.00%，唯ELM經過集成學習投票機制篩選變數後分類準確率下降，由62.33%降為62.00%。



圖三、各機器學習篩選變數前後之比較

敏感度分析結果見圖四，C5.0、ELM、SVM之敏感度分別為96.12%、99.95%、96.09%，經過集成學習篩選變數後，敏感度變為97.36%、99.95%、97.13%。其中ELM篩選變數前後皆維持99.95%的極高敏感度，證明ELM在敏感度的預測是非常穩定的。而C5.0與SVM經過集成學習變數篩選後敏感度皆有提高。RF篩選變數前(90.80%)後(92.10%)的良好特異度，證明RF在特異度的預測是非常穩定

的。



圖四、集成學習策略前後之敏感度結果比較

(五) 結論

卵巢癌的致死率高居婦科癌症之首，晚期的卵巢癌復發率很高，一旦復發，能存活的機率很低，過去很多研究將變因的觀察以全民健保資料庫抽樣檔的門診處方及治療明細檔作為資料分析，至今仍缺乏以資料探勘方法分析之相關研究。因此本研究使用了資料探勘分析，且更進一步加入集成學習投票機制來改善一般機器學習法的缺點。經由結果比較後可以驗證集成學習架構用於卵巢癌復發預測之成效。本研究結果顯示：1.使用 C5.0、MARS、RF、SVM 和 ELM 五種方法預測卵巢癌復發的準確度，我們發現 C5.0 為準確度最高之機器學習方法; 2.經過機器學習預測卵巢癌復發之風險因子，並針對風險因子，使用集成學習策略來改善一般機器學習之缺點。此方法是有效的，大多數之機器學習法再經過集成學習改善，其分類準確率皆提高，尤其以 C5.0 之效果最明顯。本研究臨床實務建議：針對病患個案需鑑別遭遇復發之預測可以使用 C5.0 方法進行敏感度分析；相對地，對於針對病患個案需識別無復發之預測可以使用 RF 方法進行特異度分析。相關文獻顯示預測卵巢癌復發因子中的確診年齡與 Gadducci 等人(2013)的研究有一致的結果；另外對於年齡(Previset al., 2014; Gadducciet al., 2013)、期別(Previset al., 2014)、分化(Paulsen et al., 2011; Previset al., 2014)、組織型態(Paulsen et al., 2011)、體能狀態(Previset al., 2014)是否扮演復發重要的預後因子，建議未來可以深入分析。最後，重要的考量是個案資料不完整可能造成的數據缺失問題，但是若能提高樣本數，相信也能具體反應卵巢癌復發之重要變數。

(六) 文獻參考

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L. “Bagging predictors”, Machine Learning, 1996, Vol. 24, No. 2, pp.123-140.

Chang, Hsiu-Ping, Lo, Chuan-Wei, Chang, Ting-Chang and Chou, Hung-Hsueh “Usefulness of ^{18}F-FDG PET in the Management of Recurrent Ovarian Cancer Patients with Unexplained Tumor Marker CA 125 Elevation: A Preliminary Report” , Ann Nucl Med Sci 2009;22:135-143 Vol. 22 No. 3 September 2009.

Craven, P., Wahba, G. “ Smoothing Noisy Data with Spline Functions. Estimating the Correct Degree of

Smoothing by the Method of Generalized Cross-Validation", Craven, P. , Wahba, G. " Smoothing Noisy Data with Spline Functions. Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation", Numberische Mathematik Vol.31, 1979, pp.317-403.

David, A. and Lerner, L., "Pattern classification using a support vector machine for genetic disease diagnosis", Electrical and Electronics Engineers in Israel, 23rd IEEE Convention of Prceedings, 2004, pp. 289-292.

Dietterich, T.G., "Ensemble methods in machine learning," Proceedings of the First International Workshop on Multiple Classifier Systems (MCS00), 2000, pp 1-15.

Freund, Y., Schapire, R. E., "A decision-theoretic generaliation of on-line learning and an application to boosting". Journal of Computer and System Sciences, 1997, Vol. 55, No. 1, pp. 119-139.

Friedman, J. H. "Multivariate Adaptive Regression Splines". Department of Statistics, Stanford University, Technical Report 102 Rev, August 1990.

Gadducci, A., Cosio, S., Zola, P., Sostegni, B., Fuso, L., & Sartori, E. (2013). Prognostic factors and clinical outcome of patients with recurrent early-stage epithelial ovarian cancer: an Italian multicenter retrospective study. International Journal of Gynecological Cancer, 23(3), 461-468.

Han, J.W. and Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann, New York, 2001.

Ho, S.H., Jee, S.H., Lee, J.E., Park, J.S. (2004) Analysis on risk factors for cervical cancer using induction technique. Expert Systems with Applications 27(1):97-105

Hsu, C. W., Chang, C. C., Lin, C. J.: A practical guide to support vector classification. Taipei, Taiwan: Department of Computer Science and Information Engineering, National Taiwan University (2003)

Huang, G. B., Zhu, Q.Y. and Siew, C. K., "Extreme learning machine: a new learning scheme of feedforward neural networks," School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Vol.2, July 2004, pp. 985 - 990.

Huang, G.R., Zhu, Q.Y., Siew, C.X. (2006) Extreme learning machine: theory and applications. Neurocomputing 70: 489-501

Kim, H.S., Park, N.H., Kang, S.B. (2008) Rare Metastases of Recurrent Cervical Cancer to the Pericardium and Abdominal Muscle. Archives of Gynecology and Obstetrics 278: 479-482

Larose, D.T. (2005) Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey: John Wiley & Sons, Inc.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R news, 2(3), 18-22.

Li, S., Kwok, J. T., Zhu, H., & Wang, Y. (2003). Texture classification using the support vector machines. Pattern recognition, 36(12), 2883-2893.

Louie, K.S., de Sanjose, S, Mayaud, P. (2009) Epidemiology and prevention of human papillomavirus and cervical cancer in sub-Saharan Africa: a comprehensive review. Tropical Medicine & International Health 14(10):1287-1302

Mao, Y., Zhou, X., Pi, D., Sun, Y., & Wong, S. T. (2005). Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. BioMed Research International, 2005(2), 160-171.

Nizar, A.H., Dong, Z.Y., Wang, Y. (2008) Power utility nontechnical loss analysis with extreme learning machine method. IEEE Transactions on Power Systems 23(3):946-955

Parkin, D.M., Bray, F.I., Devesa, S.S. (2001) Cancer burden in the year 2000: the global picture. European Journal of Cancer 37:S4-S66

Paulsen, T., Kærn, J., & Tropé, C. (2011). Improved 5-year disease-free survival for FIGO stage I epithelial ovarian cancer patients without tumor rupture during surgery. Gynecologic oncology, 122(1), 83-88.

Previs, R. A., Bevis, K. S., Huh, W., Tillmanns, T., Perry, L., Moore, K., ... & Secord, A. A. (2014). A prognostic nomogram to predict overall survival in women with recurrent ovarian cancer treated with bevacizumab and chemotherapy. Gynecologic oncology, 132(3), 531-536.

Quinlan J.R. (1993) C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufinann.

Rao C.R. and Mitra S.K., Generalized inverse of matrices and its applications, New York: Wiley, January 1971.

Schapire, R. E., "The Strength of weak learnability". Machine Learning, 1990,Vol.5, No.2, pp.197-227.

See5: An Informal Tutorial <http://www.rulequest.com/see5-win.html>, (Accessed May 10, 2007).

Serre, D., Matrices: Theory and applications, New York: Springer, September 2002.

Steinberg, D., Bernard B., Phillip C., Kerry M. MARS User Guide. San Diego, CA: Salford Systems, 1999.

Sun, Z.L., Choi, T.M. (2008) Sales forecasting using extreme learning machine with applications in fashion

retailing. Decision Support Systems 46:411-419
Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 43(6), 1947-1958.
Thangavel, K., Jaganathan, P.P., Easmi, P.O. (2006) Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Technique. Asian Journal of Information Technology 5(4):413-417
Falk, T. H., Shatkay, H., & Chan, W. Y. (2006, May). Breast cancer prognosis via Gaussian mixture regression. In Electrical and Computer Engineering, 2006. CCECE'06. Canadian Conference on (pp. 987-990). IEEE.
Vani G., Savitha, R. and Sundararajan, N.. Classification of Abnormalities in Digitized Mammograms using Extreme Learning Machine. Automation, Robotics and Vision Singapore, 7-10th December 2010.
Vapnik, V.N. (2000) The Nature of Statistical Learning Theory. Springer, Berlin
Wald, N., Cuckle, H. Reporting the assessment of screening and diagnostic tests. Br J Obstet Gynaecol 1989; 96: 389-96.
Witten, I.H. and Frank, E., Data Mining, (second edition), Elsevier, Morgan Kaufmann Publishers, 2005.
中央健保局(2006)。全民健保預防保健服務。取自http://www.nhi.gov.tw/.
衛生福利部統計處(2013)。民國101年主要死因分析。取自http://www.mohw.gov.tw/
衛生署福利部國民健康署(2013)。1979-2010台灣子宮頸癌、子宮內膜癌、卵巢癌發生率。取自http://www.hpa.gov.tw/
台灣婦癌醫學會(2006)。卵巢癌。取自http://www.tago.org.tw/
台灣癌症防治網(2013)。認識卵巢癌。取自http://www.tccf.org.tw/
台灣癌症登記中心(2013)。卵巢癌1979 – 2010年年齡標準化發生率長期趨勢。取自http://tcr.cph.ntu.edu.tw/
莊永裕(2006)。整體學習(Ensemble Learning)入門。取自http://www.csie.ntu.edu.tw/~cyy/
張惟智(2009)。運用資料探勘分類模型對腹主動脈瘤術後併發症之探討與研究。國立台北護理學院資管系研究所碩士論文。
歐宗殷(2010)。資料探勘為基礎之零售業銷售預測模式以連鎖超商鮮食商品為例。國立清華大學工業工程與工程管理研究所博士論文。
李耀泰，陳福民，趙德讓，柯天龍，郭宗正。復發性卵巢癌的第二次減積術。中華民國婦癌醫學雜誌;1:24-28,2007。
李放歌，王志鵬，戶國，& 李輝。(2011). 全基因組關聯研究中的交互作用研究現狀. HEREDITAS (Beijing), 33(9), 901-910.
黃仁治，林淑慧，林秋杏，李俐瑤，餘政穎。淺談卵巢癌。藥學雜誌第112冊第28卷第3期,2012年9月。
張華偉，王明文 & 甘麗新。(2006)。基於隨機森林的文本分類模型研究. 山東大學學報: 理學版, 41(3), 139-143.
吳華芹。(2013)。基於訓練集劃分的隨機森林演算法. 科技通報, 29(10), 124-126.
洪智力，陳勁宏(2007) 。以選擇性集成為基礎的破產預測模型, 2007 人工智慧與應用研討會，台灣，雲林，2007 年11 月16 日。
洪智力，陳勁宏(2007)。『破產預測選擇性集成模型比較』，全國計算機會議，第 12-20~21頁。
許智宇(2010)。整合KMV模型、約略集合及隨機森林應用於企業信用評等之研究。國立臺北科技大學商業自動化與管理研究所碩士學位論文。