科技部補助

大專學生參與專題研究計畫研究成果報告

＊ ********* ** ＊
＊ 計 畫 ＊
＊ ： 應用機器學習預測子宮頸癌復發 ＊
＊ 名 稱 ＊
＊ ********* ** ＊

執行計畫學生： 鄧婷
學生計畫編號： NSC 102-2815-C-040-005-H
研 究 期 間 ： 102 年 07 月 01 日至 103 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 103 年 03 月 31 日

應用機器學習預測子宮頸癌復發

Application of machine learning to predict the recurrence-proneness for cervical cancer

摘要

子宮頸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，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、快速學習器以及 C5.0 決策樹三種資料探勘演算法，探討子宮頸癌復發的危險因子，並深入探討三種方法預測的準確度。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：(1)年齡；(2)組織型態；(3)分化；(4)腫瘤大小；(5)病理 T；(6)病理期別；(7)手術邊緣；(8)淋巴結轉移；(9)其他放射治療臨床標靶體積治療次數；(10)放射治療臨床標靶體積摘要；(11)區域治療與全身性治療順序；(12) 淋巴血管侵犯，資料清理後共計有效個案 168 筆。實驗結果表示，整體而言，C5.0, SVM, ELM 的平均準確率為 92.44%, 74.44%, 91.56%，C5.0 為最佳的預測模型。我們的研究表明四個最重要的復發因子為病理期別、病理 T、組織型態、放射治療臨床標靶體積摘要。特別是：病理期別、病理 T 是重要與獨立的預後因子，而組織型態與放射治療臨床標靶體積摘要對於復發有顯著的關聯性。為了研究輔助療法的利益，臨床試驗應隨機將病患透過這些預後因子進行分層，並提高治療後的監控，可以提早發現復發與改善預後。

關鍵字：子宮頸癌復發、Support vector machine、Extreme learning machine、C5.0

Abstract

This study applied advanced machine learning techniques, widely considered as the most successful method to produce objective to an inferential problem of recurrent cervical cancer. Traditionally, clinical diagnosis of recurrent cervical cancer was based on physician’s clinical experience with various risk factors. In this study, three machine learning approaches including SVM(support vector machine), C5.0 and ELM(extreme learning machine) were considered to find important risk factors and to predict the recurrence-proneness for cervical cancer. The medical records and pathology were accessible by the Chung Shan Medical University Hospital Tumor Registry. The existing literature on recurrent cervical cancer reveals that factors include (1) Age, (2) Cell Type, (3) Tumor Grade, (4) Tumor Size, (5) Pathologic T, (6) Pathologic Stage, (7) Surgical Margin Involvement, (8) Lymph Node Metastases (LNM), (9) Number of Fractions of Other RT, (10) RT target Summary, (11) Sequence of Locoregional Therapy and Systemic Therapy and (12) Lympho-Vascular Space Involvement (LVSI). There are totally 168 patients in the data set. The overall correct rate of classification were 92.44 %, 74.44%, 91.56% which is provided by the C5.0, SVM, ELM model, respectively. Consequently, based on the results of this study, we can conclude that the C5.0 model is the most effective classification model. Further, after 10 runs, the selected important independent variables were Pathologic Stage, Pathologic T, Cell Type and RT target Summary. Our findings support that Pathologic Stage and Pathologic T were important and independent prognostic factor. Particularly, Cell Type and RT target Summary were significantly related to the recurrence. For clinical interpretation, however, a further clinical cooperation with physician was necessary.

Keywords: Recurrent cervical cancer, Support vector machine, Extreme learning machine, C5.0

一、 研究動機與問題

子宮頸癌仍是全球婦女癌症死亡的主要原因之一 (Parkin et al., 2001 ; Goldie et al,2001)，即使近年來發病率及死亡率已減少。在台灣子宮頸癌是第二常見的且佔所有女性癌症的四分之一。根據歷史記載子宮頸癌發生於正常的上皮細胞因階段性病變導致最後化生不良，癌前上皮內瘤病變(CIN)分為三個階段：CIN 1、CIN 2、CIN 3，最終子宮頸癌細胞侵犯(ICC)。從癌前病變到子宮頸癌細胞侵犯間隔很長一段時間，最重要的治療方法就是在癌細胞侵犯之前定期檢查且盡早阻止上皮內瘤病變 (Delgado et al., 1990)。如果盡早發現，子宮頸癌的治癒機率相當高。根據國際婦產科聯盟分期系統（FIGO 系統），癌症第一期 B2 至第四期大約 30％的患者最終仍會復發(Lai et al., 1999； Waggoner, 2003)。第一次主要治療若失敗，第二次治療成功的機率幾乎是微乎其微。子宮頸癌復發或是骨盆轉移，一年內預後的生存率僅約 15%到 20% (Berek and Hacker, 2005)。由於治療子宮頸癌復發是一項臨床挑戰，許多研究試圖找出影響子宮頸癌復發因素，以提高臨床的管理。過去研究顯示復發因素包括：(1)年齡；(2)組織型態；(3)分化；(4)腫瘤大小；(5)病理 T；(6)病理期別；(7)手術邊緣；(8)淋巴結轉移；(9)其他放射治療臨床標靶體積治療次數；(10) 放射治療臨床標靶體積摘要；(11) 區域治療與全身性治療順序；(12) 淋巴血管侵犯(Kamura et al., 1992 ; Grisaru et al., 2003)。

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。資料探勘又稱之為機器學習（Machine Learning）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 (Ho et al., 2004 ; Thangavel et al., 2006)。本研究目的為使用現代的資料探勘方法找出子宮頸癌復發的重要因子。Brinton(1992)與 Kim 等人(2008)利用流行病學分析與探討患者生存的關鍵因子。然而大部分的預測技術結合多種潛在的致病原因，因此無法完整解釋子宮頸癌復發預測的因素。本研究試圖於治療後發現子宮頸癌復發的危險因子以提供臨床重要參考，可使估計更加精確以及研究結果達到最佳化。

二、文獻探討

本研究文獻探討內容分為兩個部分:子宮頸癌症死亡率、資料探勘方法。

2.1 子宮頸癌症死亡率

根據世界衛生組織的資料統計顯示，子宮頸癌是全球婦女第二常見的癌症，每年大概有三十八萬左右的新病例產生，中南美洲、非洲、印度半島等地是發生率較高的地區，台灣在全球的排名約在二十幾名，也算是高發生率的國家之一。雖然近年來子宮頸癌死亡率下降，但根據行政院衛生署的統計，台灣地區婦女子宮頸癌的年齡化標準化死亡率是每十萬例 3.9 人，在全國女性主要癌症死亡原因中占第七位，也是婦科癌症死亡率占第二位，僅次於乳癌，表一為全國女性主要癌症死亡原因民國 101 年與 100 年之比較，在民國 101 年死亡人數為 669 人，相較於民國 100 年降低了 12 人，但整體降幅並不明顯。因此，子宮頸癌對於女性仍然還是重要且必須被重視的癌症。

2.2 資料探勘方法

在醫療領域，資料探勘的應用可以被用來預測疾病以及可預測不同群體之間的危險預測因子。本研究將應用以下三種不同的資料探勘方法來預測子宮頸癌復發的重要因子：

- 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（Li 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）。

- 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）。近年有關 C5.0 的相關研究如下：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。

- 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 Huang、Zhu 和 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。

表一、全國女性主要癌症死亡原因

癌症死亡原因	民國 101 年			民國 100 年			標準化死亡率增減數
	順位	每十萬人口 死亡率	標準死亡率	順位	每十萬人口 死亡率	標準死亡率	
惡性腫瘤		141.2	95.1		134.3	93.4	1.7
氣管/支氣管/肺癌	1	25.5	17.0	1	24.2	16.5	0.4
肝和肝內膽管癌	2	21.7	14.4	2	20.7	14.3	0.1
結腸/直腸/肛門癌	3	18.7	12.1	3	17.7	11.9	0.1
女性乳房癌	4	16.5	11.6	4	16.0	11.6	0.0
胃癌	5	7.6	5.0	5	7.0	4.7	0.3
胰臟癌	6	6.0	4.0	6	6.1	4.1	-0.2
子宮頸	7	5.8	3.9	7	5.9	4.1	-0.2
卵巢癌	8	4.5	3.2	8	3.9	2.8	0.4
非何杰金氏淋巴瘤	9	3.5	2.4	9	3.1	2.1	0.2
白血病	10	3.1	2.4	10	2.9	2.2	0.2
膽囊和其他膽道癌	11	2.6	1.7	12	2.2	1.5	0.2
膀胱癌	12	2.4	1.5	11	2.3	1.5	-0.0
腦癌	13	1.8	1.5	13	1.9	1.6	-0.0
口腔癌	14	1.8	1.2	17	1.3	0.9	0.3
腎臟癌	15	1.7	1.1	14	1.9	1.3	-0.1

（資料來源：國民健康署，2014）

三、研究方法

本研究以 SVM、C5.0 決策樹、ELM 三種模型等相關研究基礎，建立預測子宮頸癌復發的重要因子，並探討三種資料探勘方法預測之準確度。

在醫學衛生領域中，資料探勘應用已大幅成長，因為具備可以被用來直接取得預測不同群體之間患者相關資訊的優點。因此，本研究我們試圖利用三種資料探勘方法由子宮頸癌的資料庫中進行分類並進一步分析。

3.1 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。支援向量機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，如手寫辨識及建立預測分析模型。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 $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$， $\mathbf{x}_i \in R^d$， $y_i \in \{-1,1\}$ 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) $\mathbf{w}\cdot\mathbf{x}_i + b = 0$ 其中 $\mathbf{w}$ 為超平面向量，b 為偏移量，區分兩超平面的最大寬度為 $2/\|\mathbf{w}\|^2$，所有在範圍內的點皆稱為支援向量(Vapnik, 2000)。

$$\text{Min} \quad \Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{1}$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, 2, ..., N$$

(1)式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：

$$\text{Max} \quad \Phi(\mathbf{w}, b, \xi, \alpha, \beta) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i{}^T\mathbf{x}_j \tag{2}$$
$$\text{s.t.} \quad \sum_{j=1}^{N}\alpha_j y_j = 0, \quad 0 \leq \alpha_i \leq C, i = 1, 2, ..., N$$

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： $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2), \gamma \geq 0$ ，(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。本研究我們使用多類SVM方法（Hsu et al., 2003）。有關詳細內容，請參閱（Hsu et al., 2003）。

3.2 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 為類別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：

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

假設 A 屬性有 n 個不同值$\{A_1, A_2, ..., A_n\}$，使用 A 將 S 劃分為 n 個子集合$\{S_1, S_2, ..., S_n\}$，S_j 為 A_j 在 A 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：

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

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：

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

$$\text{Gain(A)=Info(S)-Info(S,A)} \tag{6}$$

$$\text{GainRatio(A)=Gain(A)/ SplitInfo(A)} \tag{7}$$

3.3 Extreme learning machine（ELM）

快速學習器是由Huang等人於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。快速學習器已成功地被運用(Sun and Choi, 2008; Nizar et al, 2008)，不僅能在許多的例子上表現出好的效能且學習比傳統的前饋網路快上數千倍，如反向傳播(BP)時有好的泛化能力，而且排除困難提出了陡坡降法，如停止標準、學習率、學習時期、局部最小值、過度分割等。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：$x_i = [x_{i1}, x_{i2}, ..., x_{in}]^T \in R^n$ 以及 $t_i = [t_{i1}, t_{i2}, ..., t_{im}]^T \in R^m$，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：$H\beta = T$，

$$H(w_1, ..., w_{\tilde{N}}, b_1, ..., b_{\tilde{N}}, x_1, ..., x_N) = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}} ; \tag{8}$$

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

其中 $w_i = [w_{i1}, w_{i2}, ..., w_{in}]^T$，$i = 1,2,..., \tilde{N}$, 為權重向量連接第$i$連接第$i$個隱藏節點和輸入節點

$\beta_i = [\beta_{i1}, \beta_{i2}, ..., \beta_{im}]^T$ 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 $x_1, x_2, ..., x_N$ 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：

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

H^{Ψ} 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

$X=\{(x_i,t_i)|\ x_i\in R^n, t_i\in R^m, i-1,...,N\}$、激活函數$g(x)$，以及隱藏節點數$\tilde{N}$。順序分別為：

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

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$T=[t_1,...,t_N]^T$ 。

相關研究步驟為:

1. 取得子宮頸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. 利用 SVM、C5.0 決策樹、ELM 三種資料探勘方法進行預測，分析出各資料探勘方法之敏感度、特異度和準確度。
3. 分析重要變數，分別抽離 12 個預測變數，若抽離後準確度下降則為重要變數。
4. 最後，與臨床醫師討論並證實重要變數的可信度。

四、實證研究

本研究為了驗證 C5.0、SVM、ELM 的有效性與可行性，所需子宮頸癌資料庫由中山醫學大學附設醫院癌症登記中心提供。在數據集中每位病人包括 12 個預測變數，即年齡、組織型態、分化、腫瘤大小、病理 T、病理期別、手術邊緣、淋巴結轉移、其他放射治療臨床標靶體積治療次數、放射治療臨床標靶體積摘要、區域治療與全身性治療順序、淋巴血管侵犯，而結果的變數則是復發或是未復發。癌登資料庫中有效筆總共有 168 例病患，其中 118 例隨機選取為訓練樣本，而其餘 50 例病患做為測試樣本。

在 C5.0 分類模型的建構，被選擇最多次的預測變數即是最重要的。最後 C5.0 模型選擇最多次包括兩個顯著獨立變量，即病理期別及病理 T。C5.0 的測試樣本分類結果如下表二。

表二、C5.0 模型分類結果

實際類別	預測類別	
	1（有復發）	2（未復發）
1（有復發）	34 (95.00%)	0 (0.00%)
2（未復發）	2 (12.50%)	14 (87.50%)
平均準確率	96.00 %	

[34(實際有復發且預測有復發)+14(實際未復發且預測未復發)]/50(測試樣本總數)*100%

從表一我們可以觀察，平均準確率是 96.0%，有 0 位病患在第 1 類被錯誤分類為第 2 類(第 1 類為復發，第 2 類是未復發)；另外有 2 位病患在第 2 類被錯誤分類為第 1 類。

建構 SVM 分類模型，首先將全部 12 個預測變數被輸入。其中，支援向量機的性能主要設定兩個參數(C 和 γ)，因為在這項研究中採用了 RBF 核心函數，γ 表示 RBF 的寬度。在這項研究提出了 Hsu 等人的研究使用網格搜索參數設定，使用網格搜索之後，參數集(C=2^{13}, γ =2^{-5})是 SVM 最佳的參數集。測試樣本的分類結果如表三所示

表三、SVM 模型分類結果

實際類別	預測類別	
	1（有復發）	2（未復發）
1（有復發）	32 (94.12%)	2 (5.88%)
2（未復發）	14 (87.50%)	2 (12.50%)
平均準確率	68.00 %	

[32(實際有復發且預測有復發)+2(實際未復發且預測未復發)]/50(測試樣本總數)*100%

表三顯示的平均準確率為 68.00%，有 2 位病患在第 1 類被錯誤分類為第 2 類(第 1 類為復發，第 2 類是未復發)；另外有 14 位病患在第 2 類被錯誤分類為第 1 類。

建構 ELM 模型，首先 12 個獨立變量被使用時作為輸入層的 12 個節點。ELM 最重要與最關鍵的參數是隱藏節點的數量和觀察 ELM 在單次分類不穩定狀態，因此，ELM 模型從 1 至 30 建構出不同的隱藏節點，對於每個節點數量重複 30 次並選擇隱藏節點數量最小測試均方根誤差值。在這項研究中，有 23 個隱藏節點具有較小的均方根誤差。

表四、ELM 模型分類結果

實際類別	預測類別	
	1（有復發）	2（未復發）
1（有復發）	31 (91.18%)	3 (8.82%)
2（未復發）	0 (0.00%)	16 (100.00%)
平均準確率	94.00 %	

[31(實際有復發且預測有復發)+16(實際未復發且預測未復發)]/50(測試樣本總數)*100%

表四顯示了 ELM 模型的分類結果。平均分類準確率為 94.00%，有 3 位病患在第 1 類被錯誤分類為第 2 類(第 1 類為復發，第 2 類是未復發)；另外有 0 位病患在第 2 類被錯誤分類為第 1 類。綜合表二、三、四的結果，C5.0, SVM., ELM 的平均分類準確率分別為 96.00%, 68.00%, 94.00%。其中 C5.0 有最佳的分類能力且對子宮頸癌的復發提供更有效率的預測。

根據表五結果,ELM 模型在{1-1}最高的平均準確率為 93.14%(實際有復發並會被預測為有復發)，而 C5.0 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 91.27%。整體而言，最高的平均準確率是由 C5.0 產生的 92.44%。由於 C5.0 在{2-2}和整體皆優於 SVM 與 ELM，表示 C5.0 確實比另外兩個方法提供更佳的分類準確率。因此，C5.0 是對於子宮頸癌分類最有效的方法。在這項研究中，C5.0 不僅產生最佳的分類結果，且更可以用來選擇子宮頸癌分類中重要的變數。被選擇的重要獨立變數可用來提供子宮頸癌治療更有用的資訊。在這項研究中，經過 10 次測試後，被選擇的重要變數有：病理期別、病理 T、組織型態、放射治療臨床標靶體積摘要。

表五、C5.0、SVM、ELM 準確率比較

組別	{1-1} (實際有復發且預測有復發)			{2-2} (實際未復發且預測未復發)			整體		
	C5.0	SVM	ELM	C5.0	SVM	ELM	C5.0	SVM	ELM
1	100.00	94.12	91.18	87.50	12.50	100.00	96.00	68.00	94.00
2	91.67	94.44	94.44	100.00	7.14	100.00	94.00	70.00	96.00
3	95.00	92.50	92.50	80.00	10.00	70.00	96.00	76.00	88.00
4	89.47	89.47	94.74	100.00	16.67	83.33	92.00	72.00	92.00
5	91.89	91.89	91.89	92.31	7.69	92.31	92.00	76.00	92.00
6	94.87	89.74	92.31	81.82	18.18	90.10	92.00	94.00	92.00
7	84.38	96.88	96.88	94.44	11.11	83.33	88.00	68.00	92.00
8	97.14	94.29	91.43	80.00	0.00	86.67	92.00	66.00	90.00
9	95.12	92.68	95.15	100.00	0.00	77.78	96.00	76.00	92.00
10	88.89	88.89	88.89	92.86	0.00	92.86	90.00	72.00	90.00
平均	92.05	92.31	93.14	91.27	7.87	86.26	92.44	74.44	91.56

五、結論

子宮頸癌在確診到臨床復發會歷經一段潛伏期，對於臨床醫師而言掌握復發的危險因子是非常重要的。為了有更好的預後結果，許多研究人員試圖找出危險的復發因素。雖然多年來的臨床研究與經驗可以選擇出重要的因子，但仍然有可能選擇錯誤。事實上，包括腫瘤大小，淋巴血管侵犯，腫瘤侵

犯深度，淋巴節轉移，這些因素也是有相關的，但這些因素並不能反映的預後結果並不理想。在本研究中，病理期別是侵入腫瘤最危險的因素，而病理 T 相對其他類似的分析是獨立的危險因素。根據結果， C5.0 決策樹是最佳的預測模型，其中四個最重要的復發因子為：病理期別、病理 T、組織型態、放射治療臨床標靶體積摘要。特別是，病理期別、病理 T 是重要與獨立的預後因子。而組織型態與放射治療臨床標靶體積摘要對於復發也具備顯著的關聯性。因樣本數量限制，有淋巴血管侵犯與未有淋巴轉移侵入式腫瘤的患者使用不同療法，無法深入分析臨床預後。為了分析輔助療法的差異，建議後續研究可再深入探討。

六、參考文獻

Berek, J.S., Hacker, N.F. (2005) Practical gynaecologic oncology. New York: Lippincott Williams & Wilkins.

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.

Delgado G., Bundy B., Zaino R., Sevin B.U., Creasman W.T., Major F.. (1990) Prospective surgical – pathological study of disease-free interval in patients with stage Ib squamous cell carcinoma of the cervix: a gynecologic oncology group study. Gynecologic Oncology 38:352-357.

Goldie, S.J., Kuhn, L., Denny, L., Pollack, A., Wright, T. (2001) Policy analysis of cervical cancer screening strategies in low-resource setting: clinical benefits and cost effectiveness. The Journal of the American Medical Association 2001(285):3107-3115.

Grisaru D.A., Covens A., Fransen E., Chapman W., Shaw P, Colgan T (2003) Histopathologic score predicts recurrence free survival after radical surgery in patients with stage IA2-IB1–2 cervical carcinoma. Cancer 97:1904-1908.

Han, J. and Micheline, K., 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, 2004(2):985-990.

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

Kamura T., Tsukamoto N., Tsuruchi N., Saito T., Matsuyama T., Akazawa K. (1992) Multivariate analysis of the histopathologic prognostic factors of cervical cancer in patients undergoing radical hysterectomy. Cancer 1992(69):181-186.

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 2008(278):479-482.

Lai, C.H., Hong, J.H., Hsueh S. (1999) Preoperative prognostic variables and the impact of postoperative adjuvant therapy on the outcomes of stage IB or II cervical carcinoma patients with or without pelvic lymph node metastases. Cancer 1999(85):1537-1546.

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

Li S., James T. Kwok, Zhu H., and Wang Y., “ Texture classification using the support vector machines”, Pattern Recognition, 2003(36):2883 – 2893.

Mao Y., Zhou X., Pi D., Sun Y., and Stephen T. C. Wong, “ Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection”, Journal of Biomedicine and Biotechnology, 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 2001(37):S4-S66.

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

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

Sun, Z.L., Choi, T.M. (2008) Sales forecasting using extreme learning machine with applications in fashion retailing. Decision Support Systems 2008(46):411-419.

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.

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, VN (2000) The Nature of Statistical Learning Theory. Springer, Berlin.

Waggoner, SE (2003) Cervical cancer. Lancet 361:2217-2225.

中央健保局(2006)。全民健保預防保健服務。取自http://www.nhi.gov.tw/. (2014/03/01)

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。