「數據分析」
2010 年 6 月，5 卷 3 期，頁 72-83

高維度資料特徵選取之探討–應用於分類蛋白質質譜儀資料

郭訓志[1] 黃仁澤[2] 薛慧敏[3]

摘 要

一般健檢的腫瘤指標的靈敏度和特異性皆不高，也無法偵測較小的腫瘤，因此通常無法及早診斷出腫瘤。本研究的資料為應用蛋白質晶片與表面強化雷射解吸電離飛行質譜技術（SELDI）的血清蛋白質質譜資料，血清樣本來自健康的正常人以及三組不同時期的攝護腺癌症病人。研究目的在選取有助於區分不同時期攝護腺癌症的蛋白質特徵，利用重複隨機抽樣的交叉驗證和支援向量機（Support Vector Machine），先以 t 檢定的平均 p 值、Kruskal-Wallis 檢定的平均 p 值、或平均分錯率對於所有蛋白質特徵進行排序，再利用向前選取方式找出最小分錯率模型之特徵變數。為了精簡模型，本研究同時考慮佐以相關係數與判定係數萃取後的特徵變數之分類結果。在各個方法比較上，使用 Kruskal-Wallis 檢定之最小 p 值特徵選取法的分類效果較好，而輔助的萃取方法以最大相關係數萃取法最能有效縮減特徵個數，同時又保持分類效果。

關鍵詞：特徵選取；蛋白質質譜儀資料；支援向量機；交叉驗證

[1]政治大學統計系助理教授
[2]政治大學統計所碩士
[3]政治大學統計系副教授
收件日期：2011.5.17 ；修改日期：2011.5.28 ；接受日期：

Journal of Data Analysis
June 2010, Vol.5 No.3, pp.72-83

On Feature Selection of High Dimensional Data - Application on Classifying Proteomic Spectra Data

Hsun-Chih Kuo[1] Jen-Tse Hunag[2] Huey-miin Hsueh[3]

Abstract

Often the time the tumor marker of regular health evaluation is low in sensitivity and specificity so that it could not detect tumor of small size in time. This research aims to develop a classification tool for early diagnosis of tumor by studying proteomic mass spectra of prostate cancer data at different stages. The prostate cancer data studied are the Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (SELDI-TOF-MS) generated from 327 serum samples. Of the 327 serum samples, 81 are from unaffected healthy men (HM), 78 are from patients diagnosed with benign prostatic hyperplasia (BPH), 84 are from patients with organ-confined PCA (T1/T2), and 84 are from patients with non-organ-confined PCA (T3/T4). The goal of this research is to select features (peaks) of the mass spectra that are useful for classifying different stages of prostate cancer via repeated random subsampling cross-validation. The forward minimum-p_value method (derived from t test or Kruskal-Wallis test) and the forward minimum-classification-error method incorporated with SVM are proposed in this study. In addition, maximum-correlation method and maximum-R2 method are considered for further feature selection. In comparison, the forward minimum-p_value method derived from Kruskal-Wallis test often outperforms other methods in terms of classification rate. Moreover, the maximum-correlation method not only can reduce the number of features effectively but also can preserve the classification rate at the same time.

Keywords: feature selection; proteomic spectra; SVM; cross-validation

[1] Department of Statistics, National Chengchi University
[2]
[3]

Manuscript received: 2011.5.17 ; Revised: 2011.5.28 ; Accepted:

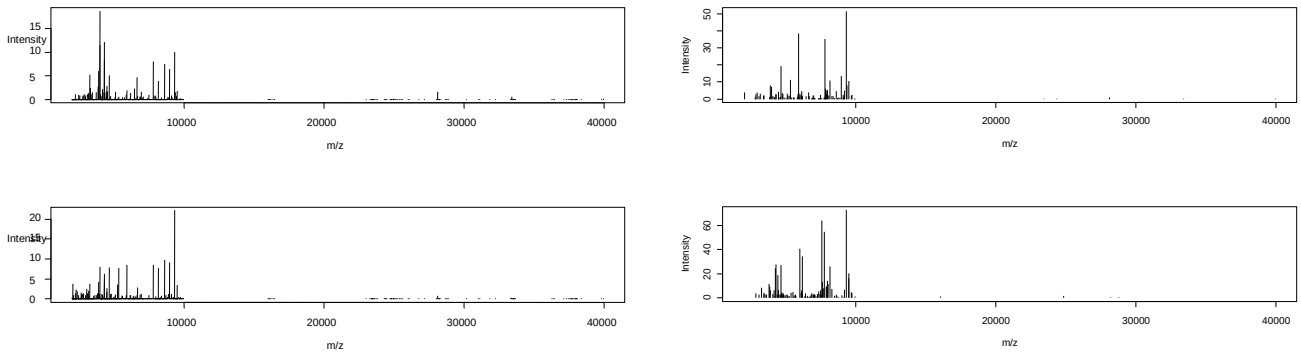
壹、 前言

根據行政院衛生署歷年進行的國人死因調查統計資料，癌症自民國 73 年起即成為國人第一大死因，大部份的研究皆指出癌症早期的發現可增加治療方式的選擇及改善存活率，因此，如何能夠及早發現癌症是一個重要的議題。然而一般健檢的腫瘤指標不僅靈敏度、特異性不高，也無法偵測出小於 1 公分的腫瘤，使得當腫瘤被診斷出來時通常為時已晚。本研究使用美國東維吉尼亞醫學院（Eastern Virginia Medical School）提供的攝護腺癌蛋白質質譜資料，藉由分析癌症不同時期的蛋白質表現質譜圖之變化，利用重複隨機抽樣方式的交叉驗證（Cross-Validation）與支援向量機（Support Vector Machine）的分類方法來挑選對於分類效果有助益的縮氨酸（Peptide）特徵，期望能利用這些挑選出來的縮氨酸特徵來發展較為準確的腫瘤預測分類機制，幫助早期診斷發現腫瘤以增加治療方式的選擇及改善存活率，避免醫療資源浪費。

本文第二節為此攝護腺癌蛋白質質譜資料的描述；第三節為本文所使用特徵選取方法與架構，包括最小 p 值特徵選取法與最小分錯率特徵選取法；第四節為以相關係數與判定係數的粹取法；第五節為總結與建議。

貳、 資料描述

本研究使用的資料來自於美國東維吉尼亞醫學院，收集 327 位受測者的血清樣本所產生的蛋白質表面強化雷射解吸電離飛行質譜(SELDI-TOF-MS)資料，其中包含 81 位沒有罹患癌症的正常人、78 位良性腫瘤病人、84 位攝護腺癌症早期病人以及 84 位攝護腺癌症晚期病人。此蛋白質質譜資料已經經過事前處理（Pre-Processing），亦即是將原始質譜經過扣除基線（Baseline Subtraction）、校準(Calibration）、正規化（Normalization）等手續。經事前處理之後，質譜範圍為 2000～40000 質量／電荷比（mass-to-charge，m/z）,特徵變數總個



圖一、四類病況中各一位病例的質譜（左上:正常、左下:良性腫瘤、右上:癌症早期、右下:癌症晚期）

數為 779 個。圖一為正常、良性腫瘤、癌症早期、癌症晚期各取一位案例的質譜，由圖中可發現在 10000 質量／電荷比之後幾乎沒有明顯的縮氨酸表現。

參、 特徵變數選取法

本研究主旨在於從大量的特徵中選取部份具有代表性的縮氨酸特徵來預測受測者的病況，事前處理的過程並不是我們研究的主要內容，因此不再進一步去探討事前處理的方式。研究中除了將四種病狀類別（正常、良性腫瘤、癌症早期與癌症晚期）分別做兩兩之間的比較，另外也同時對四種病狀進行特徵選取，並做分類效果的驗證。本文中的質譜資料將隨機抽取 2/3 為訓練資料（Training Set）與 1/3 為驗證資料(Validation)，為降低樣本選取的偏差(Selection Bias)，本研究採取重複抽樣(Resampling) 100 次之方式。

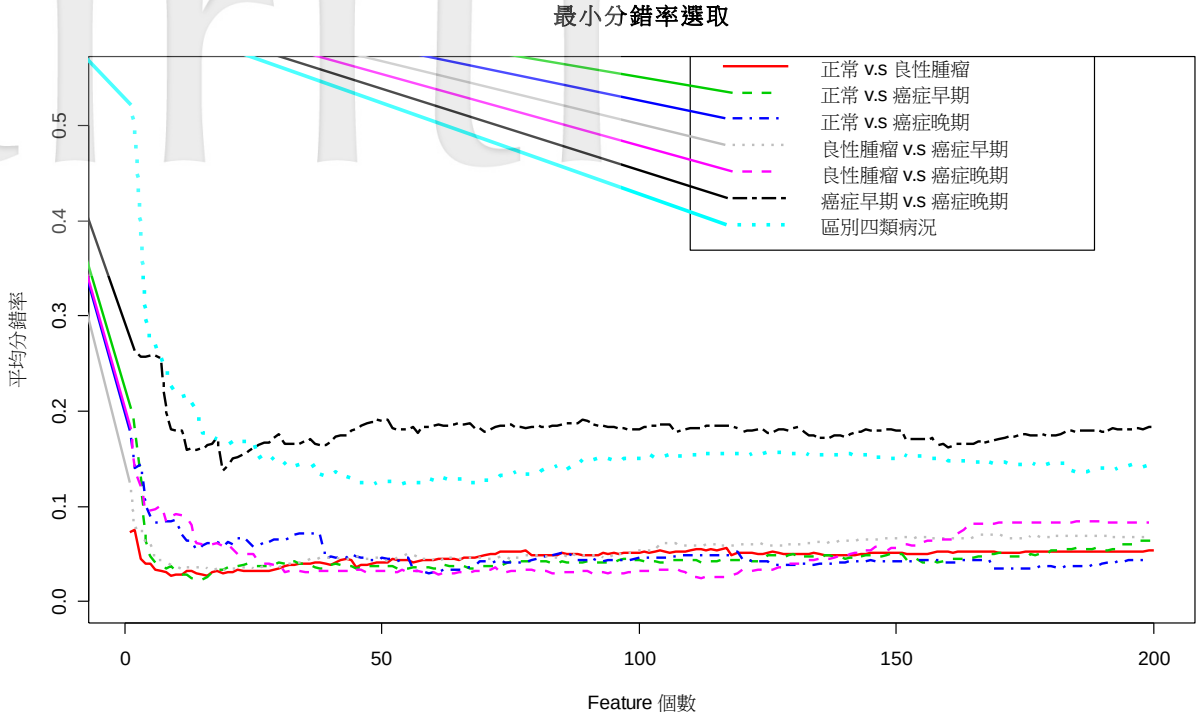
在選取特徵前，我們考慮兩種評估特徵變數對病狀的相關程度的評估準則，所有的特徵變數將會依評估準則排序。第一種使用的特徵評估準則為辨識程度，其定義為支援向量機分類對訓練資料組 100 次之平均分錯率，本研究中我們使用的支援向量機的核函數（Kernel Function）是線性核函數 $K(x_i, x_j) = x_i^T x_j$ 。第二種使用的特徵評估準則為統計假設檢定之 p 值，針對每個特徵變數在不同病況下之平均差異檢定所得之 p 值來決定其特徵變數在辨識上的重要程度，再按照所得 p 值將特徵變數排序。

理想的做法是類似所有可能迴歸法(All Possible Regression)或逐步迴歸(Stepwise Regression)方式選取重要特徵，但計算量將非常龐大，在時間成本考量下，改以對排序過的特徵中逐一遞增地選取特徵變數，藉此找出最佳的特徵組合。

一、 最小分錯率特徵選取法

我們使用個別特徵變數運用支援向量機方法將訓練組資料做分類並計算其分錯率，藉由 100 組訓練資料可得到每個特徵的平均分錯率，再依照個別特徵的平均分錯率由小到大排列，便可將縮氨酸特徵依照辨識程度排序。接著第一次挑選第一名的特徵，第二次挑選前兩名特徵……。在本研究中僅考慮前 200 名特徵變數，每次選取特徵之後即代入支援向量機辨別系統，並計算其在驗證資料之平均分錯率，以期在 200 個特徵組合中求得一最佳特徵組合，使得有最低平均分錯率。

圖二為最小分錯率特徵選取法在七種分類下的分錯率走勢圖，以「癌症早期 vs.癌症晚期」的分類效果最差，甚至低於「區別四類病況」，而其他的五種分類結果差異不大，都可以有不錯的分類效果。



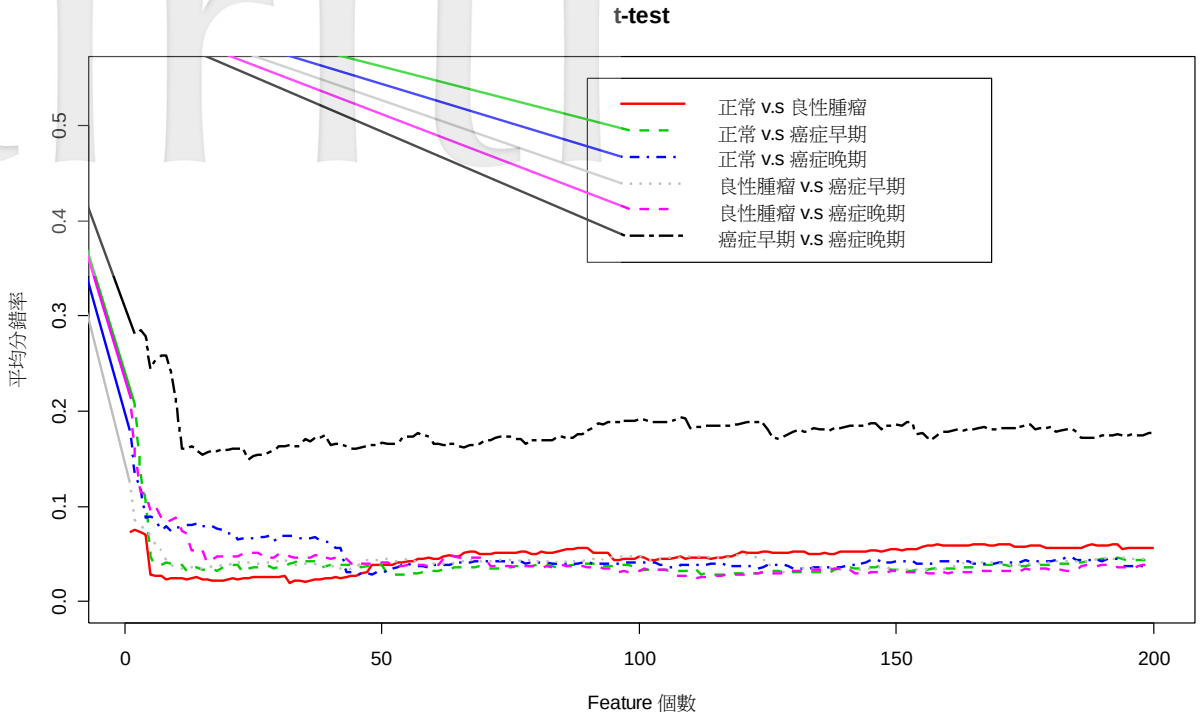
圖二、最小分錯率選取在各種分類下的平均分錯率

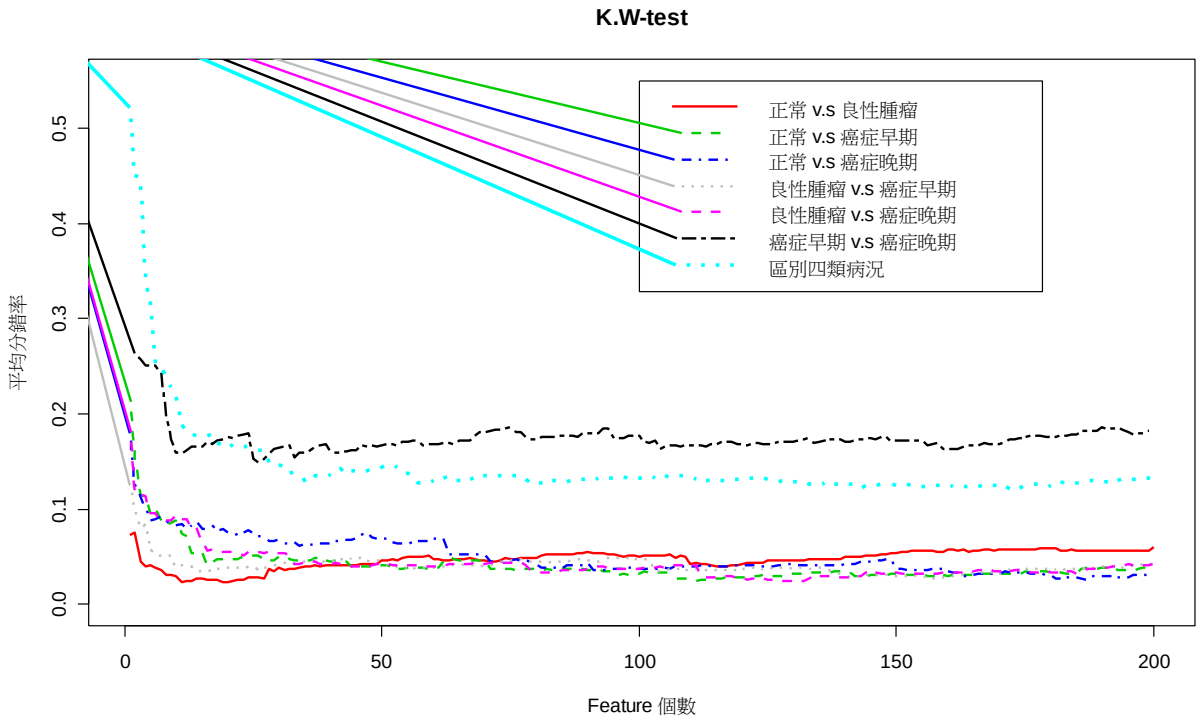
二、 最小p值特徵選取法

我們提出的第二個特徵評估準則為統計假設檢定中之 p 值，針對每個特徵變數以 100 組訓練資料下之平均差異檢定所得之 100 個 p 值的平均來排序。本研究考慮 t 檢定與 Kruskal-Wallis 檢定(為簡化起見，以下簡稱 KW 檢定)所得之 p 值，在時間成本的考量下，如同最小分錯率選取法，在利用平均 p 值排序過的特徵中逐一遞增地選取特徵變數，透過支援向量機分類器決定分類準則，並以 100 次抽樣的驗證資料代入分類準則可獲得 100 次分類結果的平均分錯率，以決定最適的特徵集合。

由圖三的 t 檢定之最小 p 值特徵選取法之分類走勢圖可發現，「癌症早期 vs.癌症晚期」的分類效果依然是最差，其他五種分類的效果仍很接近且分得不錯。總體而言，以 t 檢定之最小 p 值特徵選取法所得到的結果與最小分錯率特徵選取法的效果是一致的。而以 KW 檢定之最小 p 值選取法所得到的分類效果也與最小分錯率選取與 t 檢定之最小 p 值選取法都有類似的效果，兩個方法所得之分類效果還算一致。

表一為三種選取方法在各個分類的最小平均分錯率與對應的特徵個數，括號內為分錯率的標準差。由表中結果可發現此三種選取方法在七種分類中大部份的分類結果是差不多的，就整體而言，最小 p 值選取法之 KW 檢定選取所得之分類準確率稍微高一些，但是大部分的特徵個數仍為太多，因此將再進一步改善，使得所選取的特徵更加精簡化。圖四為三種方法分類結果之平均分錯率趨勢圖。

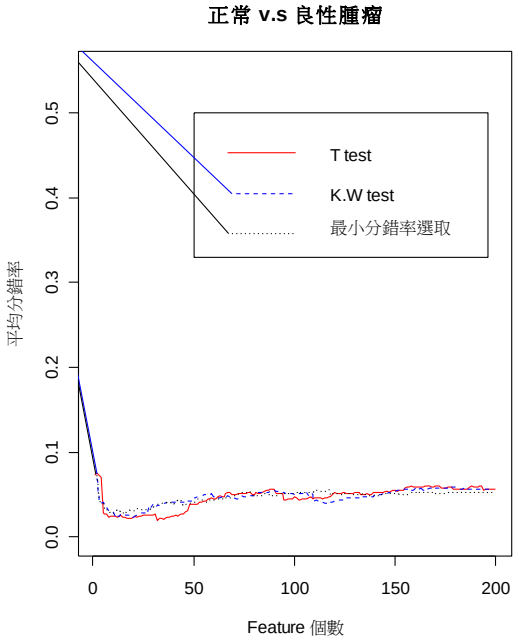


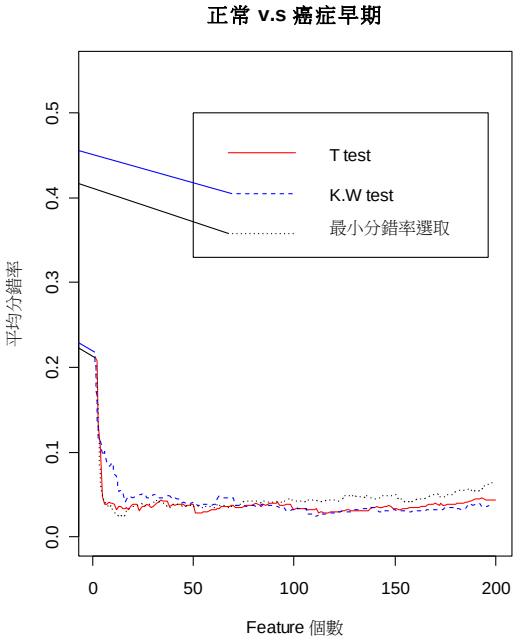


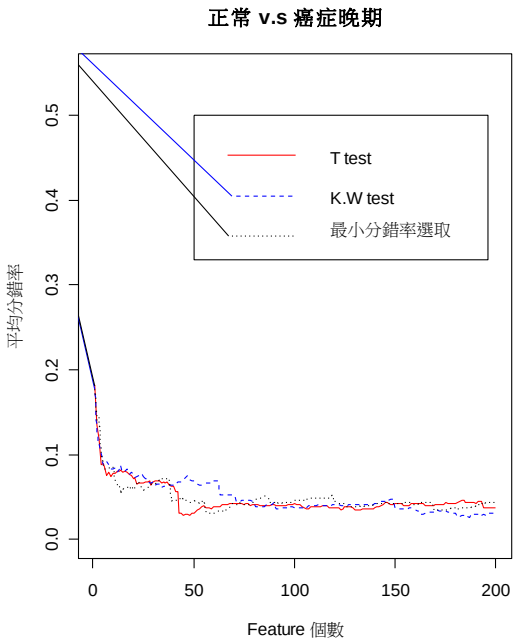
圖三、最小 p 值特徵選取法在各種分類下的平均分錯率

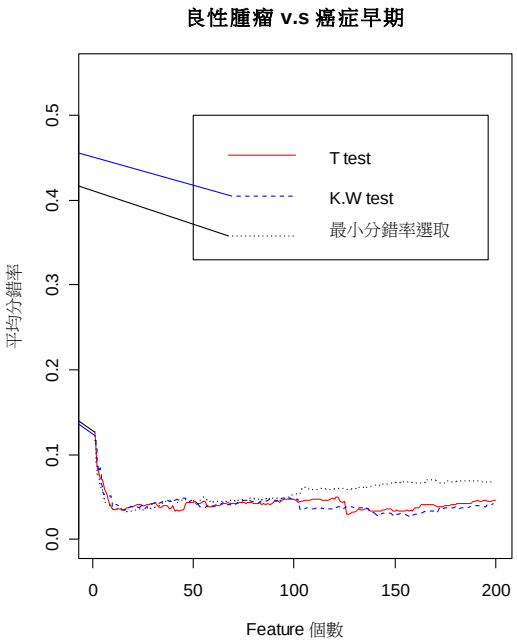
表一、最小平均分錯率與特徵個數

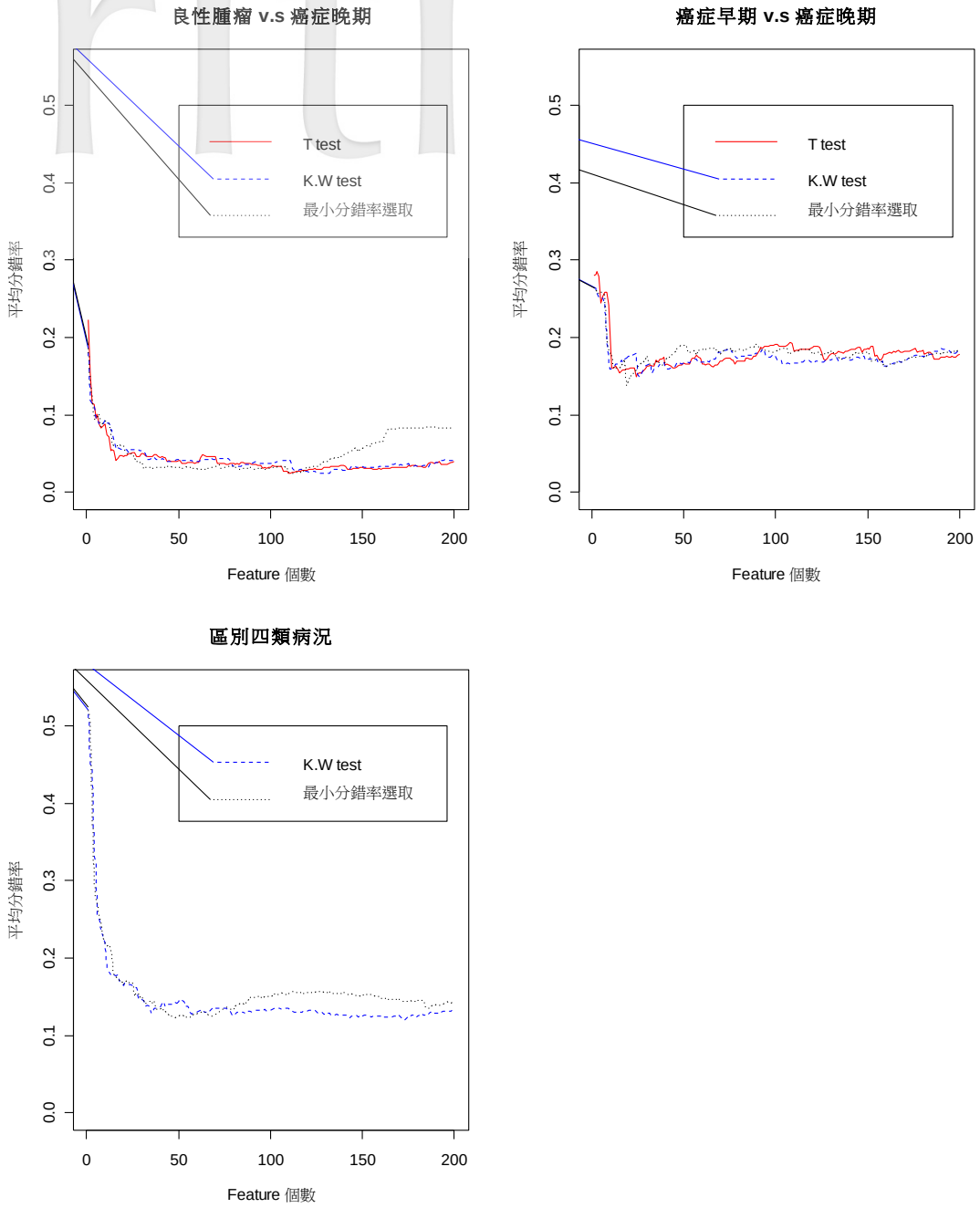
	t 檢定		KW 檢定		最小分錯率選取法	
	分錯率(%)	個數	分錯率(%)	個數	分錯率(%)	個數
正常/良性腫瘤	2.02 (1.27)	32	2.27 (1.41)	20	2.69 (1.56)	16
正常/癌症早期	2.78 (1.67)	117	2.52 (1.79)	110	2.32 (1.43)	15
正常/癌症晚期	2.82 (1.76)	48	2.65 (1.55)	187	3.02 (1.86)	59
良性腫瘤/癌症早期	2.93 (1.97)	127	2.71 (2.25)	157	3.28 (1.80)	17
良性腫瘤/癌症晚期	2.52 (1.79)	110	2.47 (1.80)	128	2.49 (1.89)	112
癌症早期/癌症晚期	14.95 (3.23)	24	14.86 (3.60)	26	13.79 (3.67)	19
區別四類病況	—	—	12.04 (2.48)	173	12.28 (2.28)	48











圖四、最小分錯率選取法與最小 p 值選取法的比較

肆、 特徵篩選法之改進

比較最小分錯率特徵選取法與最小 p 值特徵選取法，我們發現利用 KW 檢定之最小 p 值選取法通常可以獲得較佳的分類準確性，但是特徵個數不如最小分錯率特徵選取法來的精簡。在先前的方法中，我們使用前 200 名顯著特徵下，以遞增選取方式找到最低平均分錯率以決定最佳的特徵組合，以下我們進一步利用相關係數萃取法與判定係數萃取法來萃取更精簡的特徵變數。

一、 最大相關係數特徵萃取法

當已知有多個特徵入選時，挑選與其相關程度較低的其他特徵有許多方法，本文以最大相關係數當作衡量相關性的指標。已知特徵總數 m＝779，將特徵變數依照支援向量機平均分錯率或 KW 檢定之平均 p 值做排序可得 $f_{(1)}, f_{(2)}, \ldots\ldots, f_{(m)}$ ，其中 $f_{(i)}$ 表示第 i 名的特徵之觀測向量，$i = 1,2,\ldots\ldots,m$ 。定義 $\gamma_{ij} = Corr(f_{(i)}, f_{(j)})$ 為 $f_{(i)}$ 與 $f_{(j)}$ 之樣本相關係數。若已知入選特徵之指標集合為 A，則令

$$|\rho_{(i)}| = \max\left\{ |\gamma_{ij}|, j \in A \right\}, i \notin A ,$$

即$|\rho_{(i)}|$表示第 i 名特徵與已選入特徵間的最大絕對值相關係數。

我們的做法為先根據支援向量機平均分錯率或Kruskal-Wallis檢定之平均p值將所有特徵變數做排列，第一步先將第一名的特徵選入，此時 $A = \{1\}$ 。而第二名特徵選入與否將取決於其與第一名特徵之相關係數絕對值的大小，若此相關係數絕對值太高，亦即$|\rho_{(2)}| = |\gamma_{21}| \geq \lambda$ (λ 為門檻值)，則剔除此特徵變數。之後繼續考慮第三名特徵變數，若其與已入選的第一名特徵之相關係數絕對值低於門檻值，亦即$|\rho_{(3)}| = |\gamma_{31}| < \lambda$ ，即入選；故此時入選的特徵變數則為第一名與第三名特徵變數，此時 $A = \{1, 3\}$ 。接下來的第四名特徵是否入選則視其與第一名與第三名已入選特徵變數之最大絕對值相關係數，亦即$|\rho_{(4)}| = \max\left\{|\gamma_{4j}|, j \in \{1, 3\}\right\}$，當此最大絕對值相關係數$|\rho_{(4)}|$低於門檻值 λ 時才會選進此特徵，之後依序進行萃取動作。經過此階段萃取後，再以遞增選取方式逐一選取特徵代入驗證資料，以支援向量機方法分類所得之最低平均分錯率的特徵便是最佳特徵組合。

表二、經最大相關係數萃取法篩選後的特徵數目

相關係數選取法	SVM 分錯率排序特徵 特徵個數	KW *p* 值排序特徵 特徵個數
正常/良性腫瘤	150	199
正常/癌症早期	168	223
正常/癌症晚期	174	242
良性腫瘤/癌症早期	134	223
良性腫瘤/癌症晚期	120	235
癌症早期/癌症晚期	260	253
區別四類病況	318	308

表二列出將所有特徵做排列經過最大相關係數特徵萃取法所選入的特徵個數，本文所使用的門檻值為 0.8。由表四中可得知以最大相關係數做特徵選取的修正，可以減少特徵變數的個數同時維持相當的分類準確性。但是有些分類中經過最大相關係數篩選過的特徵數量仍比預期的還多，在部份分類下，仍選取超過一百個特徵變數，因此，下一節另外改以判定係數篩選法逐一篩選特徵變數來修正。

二、 判定係數特徵萃取法

由於特徵維度過高，使用前述之相關係數萃取法篩選特徵可能仍不足以解決共線性的問題，因為其只考慮兩兩特徵間的相關係數來作篩檢，但卻忽略整體特徵間的共線性，導致篩選後的特徵仍不夠精簡。為了從大量的有序特徵中更嚴格地剔除共線性的特徵，我們採用迴歸分析之判定係數進行特徵篩選，有點類似逐步迴歸的過程。在以支援向量機平均分錯率或 Kruskal-Wallis 檢定之平均 p 值所排序過的特徵變數中，依序每次選入一個新特徵當作反應變數，而先前已選入的舊特徵則當做解釋變數，以判定係數決定新特徵是否入選。判定係數可用來衡量解釋變數對模型的共線性程度，若判定係數太高代表新的特徵變數可由已經選入的特徵變數取代，故可剔除之。

若經過支援向量機平均分錯率或 Kruskal-Wallis 檢定之平均 p 值所排序過的特徵變數為：$f_{(1)}, f_{(2)}, \cdots\cdots, f_{(m)}$，首先以 $f_{(1)}$ 當作解釋變數，$f_{(2)}$ 當作反應變數，配適迴歸模型並得其判定係數，若判定係數高於臨界值 λ，則剔除 $f_{(2)}$，再考慮下一個特徵 $f_{(3)}$ 當作反應變數，以 $f_{(1)}$ 當作解釋變數來配適迴歸模型，若迴歸分析中之判定係數低於臨界值 λ，則選取該特徵 $f_{(2)}$，接著考慮特徵 $f_{(4)}$，以 $f_{(4)}$ 當作反應變數，以已入選之 $f_{(1)}$ 與 $f_{(3)}$ 當作解釋變數配適一複迴歸模型，若此迴歸模型之判定係數低於 λ 則選取特徵 $f_{(4)}$，重複相同步驟依次將所有特徵變數做過篩選。

我們將七種分類皆利用此篩檢方式，在所有有序特徵中，萃取出彼此相關程度低的特徵，我們嘗試 λ 在 0.1－0.9 範圍的臨界值之後發現，λ =0.7 可以使七種分類下的平均分類準確率達到最高，結果整理如表四。表三為以 λ =0.7 經過判定係數特徵萃取法所選入的特徵個數，與表二比較可知經判定係數萃取法所篩選的特徵個數較精簡。

表三、經判定係數萃取法篩選後的特徵數目

判定係數選取法	SVM 分錯率排序特徵 特徵個數	K.W p 值排序特徵 特徵個數
正常/良性腫瘤	115	138
正常/癌症早期	132	155
正常/癌症晚期	126	153
良性腫瘤/癌症早期	104	142
良性腫瘤/癌症晚期	96	142
癌症早期/癌症晚期	162	158
區別四類病況	228	228

三、 萃取法的結果比較

在表四中，經最小分錯率特徵選取法所選取的特徵個數原本就不多(除了「良性腫瘤 v.s 癌症晚期」之外) ，因此最大相關係數或判定係數萃取法縮減的特徵個數並不多，比較上來說，判定係數萃取法

縮減的幅度稍大。然而以分類效果而言，經最大相關係數萃取之後的分類結果不只優於經判定係數萃取之後的分類結果，同時也和未經萃取的最小分錯率特徵選取法分類結果相當。

而經 Kruskal-Wallis 檢定之最小 p 值特徵選取法所選取的特徵個數，不論經由最大相關係數或判定係數萃取皆可有效縮減特徵個數，其中還是以判定係數萃取法縮減的幅度較大。而以分類效果而言，經最大相關係數萃取之後的分類結果仍皆優於經判定係數萃取之後的分類結果，同時也和未經萃取的分類結果相近。

表四、事前處理資料經過特徵萃取後最佳特徵集合結果的比較

SVM最小分錯率選取法	未經萃取		最大相關係數萃取法		判定係數萃取法	
	分錯率(%)	特徵個數	分錯率(%)	特徵個數	分錯率(%)	特徵個數
正常/良性腫瘤	2.69 (1.56)	16	3.02 (1.39)	10	3.01 (1.39)	9
正常/癌症早期	2.32 (1.43)	15	2.32 (1.43)	16	2.33 (1.45)	15
正常/癌症晚期	3.02 (1.86)	59	3.07 (1.67)	49	3.55 (1.70)	40
良性腫瘤/癌症早期	3.28 (1.80)	17	3.19 (1.86)	18	3.25 (1.89)	17
良性腫瘤/癌症晚期	2.49 (1.89)	112	2.71 (1.78)	77	3.15 (1.87)	65
癌症早期/癌症晚期	13.79 (3.67)	19	12.9 (3.31)	14	13.42 (3.41)	16
區別四類病況	12.28 (2.28)	48	12.47 (2.24)	43	13.02 (2.47)	36
最小*p*值選取法(K.W)	未經萃取		最大相關係數萃取法		判定係數萃取法	
	分錯率(%)	特徵個數	分錯率(%)	特徵個數	分錯率(%)	特徵個數
正常/良性腫瘤	2.27 (1.41)	20	2.49 (1.50)	20	2.49 (1.50)	20
正常/癌症早期	2.52 (1.79)	110	2.61 (1.40)	65	3.05 (1.40)	55
正常/癌症晚期	2.65 (1.55)	187	2.89 (1.55)	133	4.21 (1.84)	97
良性腫瘤/癌症早期	2.71 (2.25)	157	3.04 (1.74)	120	3.47 (2.13)	42
良性腫瘤/癌症晚期	2.47 (1.80)	128	2.49 (2.08)	83	2.94 (1.73)	64
癌症早期/癌症晚期	14.86 (3.60)	26	13.15 (3.23)	15	14.15 (3.39)	19
區別四類病況	12.04 (2.48)	173	12.63 (2.53)	101	13.60 (2.55)	83

伍、 結論與建議

本論文之研究目標在於透過事前處理的 SELDI 質譜資料進行特徵選取，選取有助於攝護腺癌症分類的特徵變數。我們初步提出以最小分錯率特徵選取法與最小 p 值特徵選取法先將所有特徵變數先依照其辨識程度排序，再來利用遞增選取方式挑選特徵，以找出使得支援向量機分類表現最佳之特徵變數集合。因特徵變數間存在共線性，我們因此更進一步發展出兩個特徵萃取方法(最大相關係數萃取法與判定係數萃取法)，最終便以原有的特徵選取法輔以特徵萃取法來進行特徵篩選。在各方法的分類效果的比較上，我們發現 Kruskal-Wallis 檢定之最小 p 值特徵選取法的分類結果較佳，而經最大相關係數萃取之後不只可以有效縮減特徵個數，同時也和未經萃取的分類結果相近。

雖然已知 SELDI 資料中存有相當程度的測量誤差，但本論文中並未對 SELDI 實驗的測量誤差問題作深入探討。另外本研究中提出的最大相關係數萃取法與判定係數萃取法皆取決於門檻值的設定，如何客觀、並且有效率的決定門檻值將是另一重要的研究課題

參考文獻

1. 西滿正，「癌的最新診斷與治療」，台北：建宏，1996 年。
2. 長庚大學台灣蛋白質體學簡介（2002）。取自http://memo.cgu.edu.tw/inscorelab/corelab/Intro.htm。
3. 黃建榮，「使用支援向量機分類變異特徵之影像查詢」，朝陽科技大學資訊管理系碩士論文，2004 年。
4. 衛生署民國 93 年死因統計結果摘要（2004）。取自http://www.doh.gov.tw/statistic/index.htm。
5. 賴基銘，「癌症篩檢未來的展望：SELDI 血清蛋白指紋圖譜的應用」，國家衛生研究院電子報，第 52 期，2004 年。取自 http://enews.nhri.org.tw/enews_list_new3.php?volume_indx=52&enews_dt=2004-06-25
6. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GL Jr.,"Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men", Cancer Research ,(62), 2002, pp.3609-3614.
7. Alpaydm, E., "Introduction To Machine Learning. Combridge", MA：MIT Press, 2004.
8. Conover, W. J., "Practical Nonparametric Statistics", 3rd ed, New York：Wiley,1999.
9. Fung, E. T. and Enderwick, C., "ProteinChip Clinical Proteomics: Computational Challenges and Solutions", Computational Proteomics Supplement,(32):S34-S41, Ciphergen Biosystems, Fremont, CA, USA,2002.
10. Qu, Y., Adam, B. l., Thornquist, M., Potter, J. D., Thompson, M. L., Yasui, Y., Davis, J., Schellhammer, P., Cazares, L., Clements, M., Jr., Wright, G.L. and Feng, Z., "Data Reduction Using a Discrete Wavelet Transform in Discriminant Analaysis of Very High Dimensionality Data", Biometrics,(59),2003, pp.143-151.
11. Reddy, G. and Dalmasso E. A., "SELDI ProteinChip Array Technology: Protein-Based Predictive Medicine and Drug Discovery Applications", Journal of Biomedicine and Biotechnology ,(4),2003, pp.237-241.
12. Sauve, A. C. and Speed, T. P., "Normalization, Baseline Correction and Alignment of High-Throughput Mass Spectrometry Data", Proceedings Gensips ,2004.
13. Wagner, M. , Naik, D. and Pothen, A., "Protocols for Disease Classification from Mass Spectrometry Data", Proteomics ,(3),2004, pp.1692–1698.