國立政治大學統計研究所

碩士論文

分類蛋白質質譜資料變數選取的探討

On Variable Selection of Classifying Proteomic Spectra Data

指導教授：郭訓志　博士

研究生：林婷婷　撰

中華民國一零一年一月

謝辭

至今，論文的完成最感謝的就是我的指導老師-郭訓志博士。記得第一次與老師會面聊天時，老師對我說道「我覺得你選擇跟我很有勇氣。」，而當下我的心中其實也早已決定要為我所作的選擇負責到底。在撰寫論文的過程中，感謝老師先將分析的流程講解的很詳細，利於發展我之後的研究方向。此外除了感謝老師不時的在過程中提供在學術界常用的統計方法、相關期刊、書籍以及其專業知識外，尤其感謝老師訓練我研究生該有的研究精神、遇到困難的解決方法以及作事情該具備的程序和態度，相信對於之後出社會的我會有很大的幫助。

兩年半的碩士求學生涯終於要完成了。想當初從私立大學考進國立的統研所就讀，一開始還擔心同學們會因為我是私校畢業的而疏遠我或不想跟我作朋友，不過還好這些情形後來都沒有發生。感謝同學們在我撰寫論文的期間提供我程式上問題的協助、對於期刊內容的看法以及翻譯的幫忙，此外大家一起為彼此打氣、加油的感覺真的很好。

最後，要感謝我的爸爸和媽媽。在撰寫論文的過程中，對於你們我常常出現情緒化的現象，如不耐煩等等…，但你們卻都能給予無限的包容和忍耐，仍然會在我睡前或是去學校前表達對我的關心和鼓勵。此外也感謝你們體諒我常常拒絕在假日時與你們一起出去散步的邀約。這段期間你們真的辛苦了，而未來我一定會用我的孝心好好的報答你們對於我的寬容和體諒。

林婷婷

2012 年 6 月

摘要

本研究所利用的資料是來自美國東維吉尼亞醫學院所提供的攝護腺癌蛋白質質譜資料，其資料有原始資料和另一筆經過事前處理過的資料，而本研究是利用事前處理過的資料來作實証分析。由於此種資料通常都是屬於高維度資料，故變數間具有高度相關的現象也很常見，因此從大量的特徵變數中選取到重要的特徵變數來準確的判斷攝護腺的病變程度成為一個非常普遍且重要的課題。那麼本研究的目的是欲探討各(具有懲罰項)迴歸模型對於分類蛋白質質譜資料之變數選取結果，藉由 LARS、Stagewise、LASSO、Group LASSO 和 Elastic Net 各(具有懲罰項)迴歸模型將變數選入的先後順序當作其排序所產生的判別結果與利用「統計量排序」(t 檢定、ANOVA F 檢定以及 Kruskal-Wallis 檢定)以及 SVM「分錯率排序」的判別結果相比較。而分析的結果顯示，Group LASSO 對於六種兩兩分類的分錯率，其分錯率趨勢的表現都較其他方法穩定，並不會有大起大落的現象發生，且最小分錯率也幾乎較其他方法理想。此外 Group LASSO 在四分類的判別結果在與其他方法相較下也顯出此法可得出最低的分錯率，亦表示若須同時判別四種類別時，相較於其他方法之下 Group LASSO 的判別準確度最優。

關鍵詞：LARS、Forward Stagewise、LASSO、Group LASSO、Elastic Net、支持向量機。

Summary

Our research uses the prostate proteomic spectra data which is offered by Eastern Virginia Medical School. The materials have raw data and preprocessed data. Our research uses the preprocessed data to do the analysis of real example. Because this kind of materials usually have high dimension, so it maybe has highly correlation between variables very common, therefore choose from a large number of characteristic variables to accurately determine the pathological change degree of the Prostate is become a very general and important subject. Then the purpose of our research wants to discuss every (penalized) regression model in variable selection results for classifying the proteomic spectra data. With LARS, Stagewise, LASSO, Group LASSO and Elastic Net, each variable is chosen successively by each (penalized) regression model, and it is regarded as each variable's order then produce discrimination results. After that, we use their results to compare with using statistic order (t-test, ANOVA F-test and Kruskal-Wallis test) and SVM fault rate order. And the result of analyzing reveals Group LASSO to two by two of six kinds of rate by mistake that classify, the mistake rate behavior of trend is more stable than other ways, it doesn't appear big rise or big fall phenomenon. Furthermore, this way's mistake rate is almostly more ideal than other ways. Moreover, using Group LASSO to get the discrimination result of four classifications has the lowest mistake rate under comparing with other methods. In other words, when must distinguish four classifications in the same time, Group LASSO's discrimination accuracy is optimum.

Key words: LARS, Forward Stagewise, LASSO, Group LASSO, Elastic Net, SVM.

目錄

表目錄

圖目錄

第一章 緒論

第一節 研究背景

攝護腺，又稱為前列腺。雖然攝護腺從被發現至今已超過 2300 年，但一直到近代，醫學專家才開始研究其構造、生理作用以及病理變化。

攝護腺是只有男生才有的生殖器官，它位於膀胱下方、直腸前面，外形像核桃，且圍繞著尿道。而攝護腺是由三葉組成，外層覆有被膜，兩側為精囊，精囊為一對囊狀腺體。那麼發育完成的成人之攝護腺腺體可分成邊緣、中央與過渡區三個區塊，其中邊緣區占體積的 70%、中央區占 25%而過渡區在這兩區之間，占 5%的體積。那麼「攝護腺肥大」以及「攝護腺癌」是最常見的攝護腺疾病，其中「攝護腺肥大」其實就是攝護腺增生(良性腫瘤)，最主要是發生在過渡區的腺體，因此攝護腺肥大容易產生連道阻塞的現象。另外，「攝護腺癌」主要發生在邊緣區的腺體，所以攝護腺肥大和攝護腺癌發生病變的的位置是在不同的區塊，兩者的病理變化也完全獨立發展，故這兩者容易同時存在。(蒲永孝和黃昌淵，1997)

雖然亞洲國家的攝護腺癌盛行率低於西方，可是近 20 年來大部分的亞洲國家的發生率和死亡率也都逐年上升。台灣行政院衛生署統計，攝護腺癌新診斷人數從 1993 年的 801 人逐年增加，2007 年更增加到了 3367 人。統計每 10 萬人口死於攝護腺癌的比率，在 1993 年是 2.5 人， 2008 年則升高至 7.7 人。在男性癌症的死亡率排行中，攝護腺癌慢慢往前竄升，從 1995 年的 10 名外到 2001 年也升至第 7 名，且維持至今。之所以有這樣的趨勢是因為”老年人口增加(老年人口越多，診斷出攝護腺癌的機會就越高)”、”診斷率大幅提升(由於攝護腺特異性抗原(Prostate-specific antigen，PSA)的運用)”以及”生活型態改變(西化飲食製造更多肥胖者)”。(簡邦平，2006)

早期在診斷攝護腺癌時，就是驗血清中的 PSA。它是一種由攝護腺產生的蛋白質，當攝護腺發生病變時，PSA 就會升高；數值越高的話，癌的機率也就越高，擴散的程度也越大。若 PSA>20 時，則幾乎就確定是癌症。若 PSA>100，則癌細胞應已擴散至骨骼了。雖然 PSA 的敏感度有達到 90%以上，但特異性卻只有 25%，表示這項指標還是有

缺陷。不過，還好有「表面強化雷射解吸電離飛行質譜技術」(Surface-Enhanced Laser Desorption Ionization- Time of Flight，SELDI)的問世。這種診斷方法，不僅有高的敏感度也有高的特異性。若在未來這種技術能夠被有效推廣的話，勢必能降低台灣攝護腺癌症的發生率和死亡率了。所以我們欲利用這種技術所得出的數據來發覺攝護腺在不同時期中蛋白質的變化，並找出精確的生物標誌。(潘荔錞等人，2003；賴基銘，2004)

第二節 研究動機與目的

本研究是利用美國東維吉尼亞醫學院所提供的攝護腺癌蛋白質質譜資料來作分析。而此資料的自變數是為多個質量電荷比的蛋白質(我們往後將稱它為特徵變數)，此外，樣本數也小於這些特徵變數個數，故我們會將其視為一種「高維度資料」。另外，因為這筆資料的應變數是屬於類別型的，因此會讓我們聯想到分類準確度的問題，故在節省人力和時間花費的前提下，我們會希望由這如此大量且具有高度相關的高維度資料中選取對分類結果有幫助的特徵變數來判別分類結果即可，所以很常見又普遍的作法是先藉由特徵選取這步驟再代入分類器得出其分類結果。(Guyon 等人，2002；Degroeve 等人，2002；Weston 等人，2003；Ma 和 Huang，2005)

那麼所謂的特徵選取其實就是由訓練集中盡可能的發現那些對分類結果沒有用處的變數，並將其刪除的一種過程。而最後剩餘下來的這些變數集合，不僅可降低原資料的維度，且對於我們的分類結果也有所幫助。其實在過去十年中，將特徵選取的技術應用至生物資訊學中對於高維度資料的建模、序列分析、微陣列分析和質譜分析已相當普遍。(Efron 等人，2001；Somorjai 等人，2003；Jiang 等人，2004；Fox 和 Dimmic 等人，2006)

在以往特徵選取的步驟中，利用各特徵變數之統計量的顯著性來排序是很常見的作法。那麼除了此排序方法外，我們其實也可以考慮各特徵變數被選入迴歸模型中的順序當作其排序，舉例來說若以向前選取法來選取變數的話，第一步就被選入模型的變數，由於它與應變數最相關，故我們就可以將它排序為第一；到了第二步被選入模型的變數就將它排序為第二…。然而在這麼多種迴歸模型選取變數的方法中，本研究考慮了近期

在學界很著名的縮減維度的懲罰性迴歸方法–最小絕對值壓縮和選取(Least Absolute Shrinkage and Selection Operator，LASSO)，因為這方法也可同時達成估計迴歸係數和縮減維度的目的，另外由於 LASSO 可以算是屬於最小角度迴歸(Least Angle Regression，LARS)的變形，而向前逐段迴歸(Forward Stagewsie Regression)也可由 LARS 變形而成，所以我們除了想探討上述三種迴歸模型之特徵選取情形之外，另外也將 Zou 和 Hastie (2004)提出在高維度的情況下，選取變數的表現上比 LASSO 更令人滿意的彈性網路(Elastic Net)迴歸模型以及 Yuan 和 Lin (2007)為了改善 LASSO 在高維度資料中的缺點而提出的 Group LASSO 迴歸模型一併加入本研究的探討。

第三節 研究架構

本文一共分為六個章節。第二章為蛋白質質譜資料介紹，其中第一節簡述，表面強化雷射解析電離飛行質譜技術、第二節說明攝護腺癌蛋白質質譜資料之內容。接著第三章是文獻回顧，然後第四章分析方法，共分六節，第一節說明分析流程，第二節說明統計量排序方法、第三節說明 LARS、Stagewise 以及 LASSO 的迴歸模型及其演算法、第四節說明 Group LASSO、第五節是 Elastic Net 及其演算法以及第六節支持向量機 SVM 的原理。再來第五章為實証分析，其中第一節為 R 函數之設定、第二節是探討兩兩分類之分錯率結果以及第三節探討四分類之分錯率結果，而第六章為結論與建議。

第二章 蛋白質質譜資料介紹

第一節 表面強化雷射解析電離飛行質譜技術

「表面強化雷射解吸電離飛行質譜技術」英文名稱為Surface-Enhanced Laser Desorption Ionization- Time of Flight（SELDI-TOF）是美國Ciphergen Biosystems 公司根據基質輔助雷射脫附游離飛行時間質譜儀(Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometer，MALDI-TOF MS)為基礎所改良出來的產品。它結合了生物晶片的概念，而SELDI-TOF技術的流程如圖2.1，首先先在蛋白質晶片上進行樣本前處理，在每個晶片中設計不同的化合物，並使它能和特定的蛋白質結合，這樣一來就可去除非專一性的蛋白質，然後再利用雷射脈衝光使晶片上的分析物解析成特異性蛋白分子。而不同質量電荷比(the ratio of mass to charge，M/Z)的離子在儀器中飛行的長短不同，分子量越大的離子在儀器中飛行的時間越長，反之則越短，因此可藉由離子在儀器中飛行的時間長短不同來繪製出質譜圖。最後亦可利用電腦處理而得到質譜圖，藉此能直接顯示樣品中各種蛋白質的分子量等資訊。(潘荔錞等人，2003；Issaq 等人，2002)

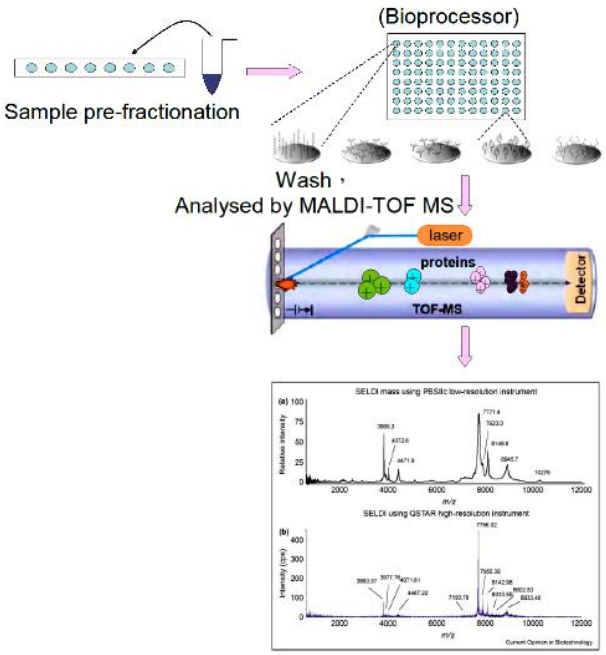

圖 2.1 「表面強化雷射解析電飛行質譜技術」流程圖

資料來源：“The SELDI-TOF MS Approach to Proteomics: Protein Profiling and Biomarker Identification” by Issaq, H. L., Veenstra, T. D., Conrads, T. P. and Felschow, D., 2002, *Biochemical* and *Biophysical Research Communications*, 589.

第二節 攝護腺癌蛋白質質譜資料

本研究所採用的資料是由美國東維吉尼亞醫學院(Eastern Virginia Medical School)利用「表面強化雷射解析電離飛行質譜技術」所建立的攝護腺癌蛋白質質譜資料。而此資料其實有原始資料(raw data)以及另一組事前處理的資料(preprocessed data)。在此，我們是以人工處理的這組資料來作實証分析，那麼所謂的人工處理，就是將原始質譜資料進行扣除基線、正規化和校準等動作。故根據專家的意見，只保留原始資料中其蛋白質的質量電荷比落在 2000 至 40000M/Z 之間的資料，然後進行人工處理後的資料只會擁有 779 個特徵變數。

而這組人工處理的攝護腺癌蛋白質質譜資料收集了 326 位受測者的觀測資料，而且分成四種類別，分別為正常人(Normal，NO)、良性腫瘤(Begin Prostate Hyperplasia，BPH)、癌症早期(Prostate Cancer Stage A and B，CAB)以及癌症晚期(Prostate Cancer Stage C and D，CCD)，此外，每一位受測者都進行兩次的重複實驗，故每位受測者皆會有兩筆樣本筆數，如表 2.1，總計共有 652 筆資料。

表 2.1

四種類別之受測人數和樣本筆數

類別	正常	良性腫瘤	癌症早期	癌症晚期	總計
受測人數	82	77	84	83	326
樣本筆數	164	154	168	166	652

此外，表 2.2 為事前處理的部份資料。其中每一行為受測者編號，在此我們定義以下表格中的 i-1 是表示第 i 位受測者的第一筆資料，i-2 是表示第 i 位受測者的第二筆資料，而第 i 位受測者前頭也會同時標示著所屬類別。此外，每一列則是表示特徵變數(M/Z)，總共會有 779 個，而下表僅列出四種類別的部份資料。在每個細格中的數值是表示某受測者對應至某特徵變數(蛋白質或縮氨酸)所偵測到的強度(intensity)。

表 2.2

事前處理的部份資料

特徵變數	M/Z	NO1-1	NO1-2	NO2-1	NO2-2	NO3-1	NO3-2
1	2009.089071	0	0	0	0	0	0
2	2013.865841	0	0	2.131	0.823	0.813	1.609
3	2020.536925	0	0	0	0	0	0
4	2021.533366	0	0	0	0	0	0
5	2021.909411	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
779	39965.84865	0.026	0.038	0.031	0	0.106	0.067
特徵變數	M/Z	BPH1-1	BPH1-2	BPH2-1	BPH2-2	BPH3-1	BPH3-2
1	2009.089071	0	0	0	0	0	0
2	2013.865841	0	0	6.72	7.198	0	0
3	2020.536925	0	0	0	0	0	0
4	2021.533366	0	0	0	0	0	0
5	2021.909411	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
779	39965.84865	0	0	0	0	0.045	0.076
特徵變數	M/Z	CAB1-1	CAB1-2	CAB2-1	CAB2-2	CAB3-1	CAB3-2
1	2009.089071	0	0	0	0	0	0
2	2013.865841	0	0	0	0	0	0
3	2020.536925	0	0	0	0	0	0
4	2021.533366	0	0	0	0	0	0
5	2021.909411	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
779	39965.84865	0.057	0.051	0	0	0.047	0.053
特徵變數	M/Z	CCD1-1	CCD1-2	CCD2-1	CCD2-2	CCD3-1	CCD3-2
1	2009.089071	0	0	0	0	0	0
2	2013.865841	0	14.987	9.076	4.143	0	0
3	2020.536925	0	0	0	0	0	0
4	2021.533366	0	0	0	0	0	0
5	2021.909411	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
779	39965.84865	0	0	0	0	0	0

第三章 文獻回顧

黃仁澤(2005)提出一篇名為「對於高維度資料進行特徵選取–應用於分類蛋白質質譜儀資料」的論文。作者同時採用原始質譜資料以及事前處理的質譜資料，不過以原始資料為主、事前處理資料為輔來作實証分析。其目的為藉由原始質譜資料來進行特徵選取，除了找出有利於判別兩兩分類以及四分類的特徵選取方法外，亦可得知對於判別分類病況時所組成的特徵變數個數。詳細作法即先將資料分為一百組的訓練集和測試集，然後以最小分錯率特徵選取法和最小p值特徵選取法將所有特徵變數依其對應到的p值由小到大排序，再以遞增選取的方式選取前200個特徵變數，並依序代入支持向量機(Support vector machine，SVM)中建模，藉此找出分錯率最好的特徵變數組合數。此外，因作者發現特徵變數間存在共線性的問題，故又進而發展三種特徵萃取的方法，分別為k-mean分群萃取法、最大相關係數萃取法以及判定係數萃取法。其分析的結果顯示對於分類原始資料時，利用判定係數萃取法搭配最小p值特徵選取法可得最佳的分類結果。而在本研究中，我們將會與該作者所使用之最小分錯率特徵選取方法和最小p值特徵選取方法加上遞增選取方式的分類結果來進行比較。

另外，陳詩佳(2007)也提出將後設學習(Meta-Learning)應用至蛋白質質譜資料的特徵選取方法。其目的為利用後設學習結合分類器搭配逐步選取特徵變數的方法，希望找出能夠利用較少的特徵變數來將資料分類並達到較高正確率的特徵選取方式。其中，後設學習就是把每個分類器融合成一個多元分類器，文章中作者運用到三種分類器，分別為線性判別分析(Linear Discriminant Analysis，LDA)、第K位最接近鄰居(K-th Nearest Neighbor，KNN)以及SVM，並將其結合為一個多元分類器，而作者結合的方法又可分為多數表決法(Majority Vote)、權重投票法(Weighted Vote)以及串聯法(Cascading)。那麼此篇作者先將資料分為一百組的訓練集以及測試集，然後作者利用投票法來分類樣本。其中多數表決法是計算某一個特徵變數在LDA、KNN和SVM下的平均分類正確率，權重投票法來預測每個樣本最後會被預測到的類別。之後，作者又考慮兩種串聯法來結合多種分類器，其中所謂的串聯法就是利用反覆地將分類器結合的過程，串接所有分類器

的預測結果，而每次都會用到前一次之預測結果，然後不斷的更新。一種是多個分類器的串聯方法，而另一種是單一分類器的串聯方法，此處作者是用支持向量機SVM來作單一分類器的串聯。而作者將多個分類器串聯的分類結果、串聯SVM的分類結果與只有用SVM的分類結果比較時，發現就算增加特徵變數也無法提升正確率，故作者就利用Elastic Net加上SVM單一分類器的串聯方法以及判定係數萃取法加上SVM串聯來試圖改善此現象，最後與只用SVM的分類結果比較時，可得出Elastic Net加上SVM單一分類器的串聯方法可稍微改善上述情形，但正確率卻有些許的降低，而判定係數萃取方法確實能夠達到僅用較少的變數來提升正確率。那麼在本研究中，我們將會採用作者所提供的判定係數萃取法搭配SVM串聯法之分類結果來進行比較分析。

而在外國文獻中，Adam 等人(2002)提出藉由計算每個特徵變數在ROC曲線下的面積，並找出合乎其所設門檻值的變數以便加入決策樹中產生分類結果。作者欲利用多種蛋白質來找出更好的生物標誌改善以往攝護腺特異性抗原(PSA)診斷敏感度高、特異性低的缺點，於是利用到事前處理的攝護腺癌蛋白質質譜資料，且將資料分為正常人、良性腫瘤、癌症早期和癌症晚期，然後將癌症早期和晚期合併為癌症病患，接著將正常人、良性腫瘤和癌症病患的所有樣本資料設為訓練集，藉由盲眼測試(blinded test)將15位正常人、15位良性腫瘤和30位癌症病患的資料設為測試集。而其分析方法就是將事前處理的攝護腺癌蛋白質質譜資料中的779個特徵變數先計算其ROC曲線以下的面積(Area under ROC Curve，AUC)，其中ROC曲線的產生方式就是在一個縱軸為敏感度、橫軸為1-特異度的二維平面中，利用診斷工具或診斷方式不斷變動的情形下，所畫出的的一種凹向橫軸的曲線。因此，一旦產生ROC曲線後即可算出AUC，於是作者將AUC≥0.62的那124個特徵變數放入決策樹中來建立分類模型。而兩兩分類之分析結果如圖3.1，一共找出了9個重要的特徵個數，分別是4475、5074、5382、7024、7820、8141、9149、9507和9656這9個特徵變數，然後共產生10個節點，並利用這10個節點來判別受測者是否為正常人(Normal)、良性腫瘤(BPH)和癌症病患(Prostate Cancer，PCA)。

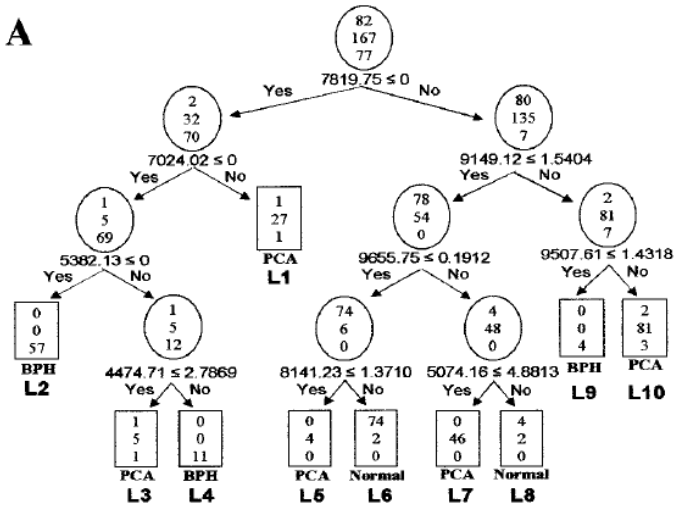

圖3.1 兩兩分類之決策樹

資料來源：“Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men” by Adam, B. L., Qu, Y., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J.,Schellhammer, P. F., Yasui, Y., Feng, Z. and Wright, G. L. Jr., 2002, *Cancer Research, 62*(13), 3611.

最後表3.1列出AUC搭配決策樹之兩兩分類的判別結果，其中包含敏感度和特異度以及分錯率數值。

表3.1

AUC搭配決策樹之兩兩分類的敏感度、特異度以及分錯率

兩兩分類	敏感度	特異度	分錯率
正常人vs.良性腫瘤	93%	100%	3.3%
良性腫瘤vs.癌症早期	93%	80%	13.3%
癌症早期vs.癌症晚期	80%	87%	16.6%
正常人vs.癌症	83%	100%	11%
良性腫瘤vs.癌症	83%	93%	13.3%
正常與良性腫瘤vs.癌症	83%	97%	10%

第四章 分析方法

第一節 分析流程

在事前處理的攝護腺癌蛋白質質譜資料中，我們有四種類別，分別為正常、良性腫瘤、癌症早期以及癌症晚期病人之資料。若想比較兩兩類別或四種類別在各種特徵選取方法下之分類效果，其過程大致可分為三個部份。首先，先將資料分為訓練資料和測試資料，再來排序特徵變數，最後將排名前兩百名的特徵變數依序放入 SVM 中建模，並得出最低的分錯率結果以及其所對應的特徵變數組合數。詳述如下：

第一部份：將欲分析的資料分為一百組的訓練資料和測試資料

如圖 4.1 所示，假使我們想比較正常人和良性腫瘤病人在各特徵選取方法下之分類效果為何，則我們必須先在正常人和良性腫瘤的資料中分別抽百之分六十六點七的樣本資料作為訓練資料以及百分之三十三點三的樣本作為測試資料。此外每次選取樣本(某受測者)時須同時選入(某人)兩次重複的觀測值：

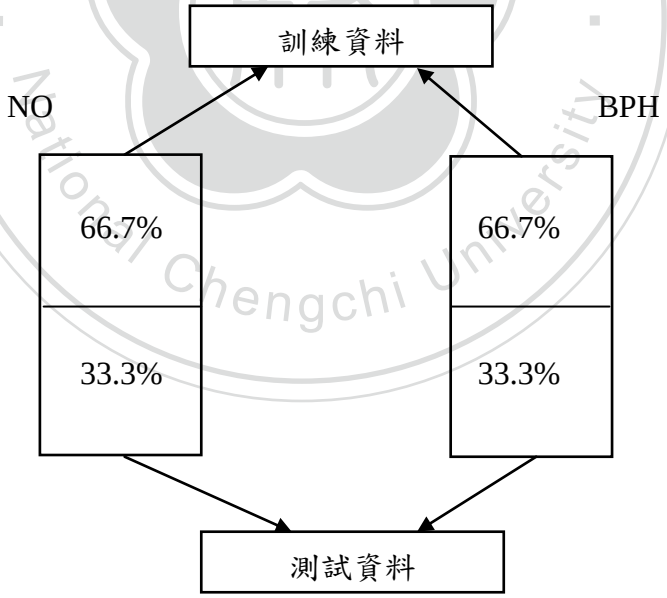

圖 4.1 訓練資和測試資料之抽樣方法

重複上述的程序一百次，即可產生一百組的訓練資料以及一百組的測試資料。

第二部份：排序特徵變數

在排序這個部份，我們對於每組訓練資料中的特徵變數利用其「統計量排序」和被「選入迴歸模型之順序排序」及其個別的「分錯率排序」。其中「統計量排序」我們是利用 t 檢定、ANOVA F 檢定以及 Kruskal-Wallis 檢定來得之變數的統計量。而「選入迴歸模型之順序排序」我們所採用的迴歸估計式分別有 Least Angel Regression、Forward Stagewise regression、LASSO、Group LASSO 以及 Elastic Net，藉此來得到各變數被選入模型的順序。在「分錯率排序」的部份我們是採用支持向量機 SVM 來得出各變數之分錯率。而以上講到的這些方法的理論部份將在此章的第二、三、四、五、六節呈現給大家。

我們先概述如何利用「統計量排序」，如表 4.1 所示，我們以 t 檢定為例來說明，對於每組訓練資料中的每一個特徵變數進行 t 檢定時，由於每個特徵變數在一百組訓練資料中皆可得到一個 t 統計量的值，最後將此一百個統計值取平均可得$\bar{t}_i$，表示第 i 個特徵變數之平均統計值，並利用它來排序每個特徵變數。而平均統計值最小的特徵變數排第一個，而平均統計值最大的特徵變數排最後一個。故 ANOVA F 檢定和 Kruskal-Wallis 檢定也依此作法。

表 4.1

對於每個特徵變數產生的統計量之值取平均過程

特徵變數	訓練資料組別				平均統計值
	1	2	⋯	100	$\bar{t}_i$
X_1	$t_{1,1}$	$t_{1,2}$	⋯	$t_{1,100}$	$\bar{t}_1$
X_2	$t_{2,1}$	$t_{2,2}$	⋯	$t_{2,100}$	$\bar{t}_2$
⋮	⋮	⋮	⋯	⋮	⋮
X_{779}	$t_{779,1}$	$t_{779,2}$	⋯	$t_{779,100}$	$\bar{t}_{779}$

再來我們說明「選入迴歸模型之順序排序」的方法。如表 4.2 所示，對於每組訓練資料中的每個特徵變數配適 Least Angel Regression、Forward Stagewise Regression、LASSO、Group LASSO 以及 Elastic Net 模型時，即可得到一個被選入迴歸模型的「等級」(Rank)，這裡的等級意思即若某個特徵變數的等級為 1，就表示其特徵變數是第一個被選入迴歸模型的；若某個特徵變數的等級為 779，則表示其特徵變數是最後一個被選入迴歸模型的。因為每個特徵變數在一百組訓練資料中皆可得到一個等級，再將這一百個等級取平均即 $\overline{R}_i$，最後利用它來排序每個特徵變數。平均等級最小的特徵變數排第一個、平均等級最大的特徵變數排最後一個。

表 4.2

對於每個特徵變數產生的等級取平均之過程

特徵變數	訓練資料組別				平均等級
	1	2	⋯	100	$\overline{R}_i$
X_1	$R_{1,1}$	$R_{1,2}$	⋯	$R_{1,100}$	$\overline{R}_1$
X_2	$R_{2,1}$	$R_{2,2}$	⋯	$R_{2,100}$	$\overline{R}_2$
⋮	⋮	⋮	⋯	⋮	⋮
X_{779}	$R_{779,1}$	$R_{779,2}$	⋯	$R_{779,100}$	$\overline{R}_{779}$

而「分錯率排序」就是利用支持向量機 SVM 來得出每個特徵變數的分錯率，並利用分錯率由低至高來排序這些特徵變數。如表 4.3，將每組訓練資料中的每個特徵變數分別代入支持向量機中配適模型，然後再計算所建構出來的模型在測試資料下的分錯率。因每個特徵變數在一百組的訓練資料下皆可產生一個分錯率，最後即利用每個特徵變數的平均分錯率($\overline{SVM}_i$)來作為排序這些特徵變數的依據。

表 4.3

對於每個特徵變數產生的 SVM 平均分錯率之流程

特徵變數	訓練資料組別				平均分錯率
	1	2	…	100	$\overline{SVM}_i$
X_1	$SVM_{1,1}$	$SVM_{1,2}$	…	$SVM_{1,100}$	$\overline{SVM}_1$
X_2	$SVM_{2,1}$	$SVM_{2,2}$	…	$SVM_{2,100}$	$\overline{SVM}_2$
⋮	⋮	⋮	…	⋮	⋮
X_{779}	$SVM_{779,1}$	$SVM_{779,2}$	…	$SVM_{779,100}$	$\overline{SVM}_{779}$

第三部分：將排序在前兩百名的特徵變數依序代入 SVM 中建模

由第二個部份所得出的 779 個特徵變數的排名順序後，我們就取前兩百名特徵變數依序代入 SVM 中建模，而所謂依序代入的意思就如表 4.4 中所示，先將順位第一的特徵變數帶入 SVM 中建模得出模型 M_1, 然後由第一和第二順位的變數代入 SVM 中建模組成模型 M_2，將前兩百名變數依循此方式建模即可得兩百組的模型。藉由預測一百組測試資料的結果來分別得知分錯率，最後對其取平均而得各模型之平均分錯率。表中 $\overline{M}_i$ 即表示第 i 個模型的平均分錯率。

表 4.4

對於前兩百名特徵變數依序代入 SVM 建模之流程

特徵變數	模型	測試資料組別				模型的平均分錯率
		1	2	…	100	$\overline{M}_i$
$X_{(1)}$	M_1	$M_{1,1}$	$M_{1,2}$	…	$M_{1,100}$	$\overline{M}_1$
$X_{(1)}$, $X_{(2)}$	M_2	$M_{2,1}$	$M_{2,2}$	…	$M_{2,100}$	$\overline{M}_2$
⋮	⋮	⋮	⋮	…	⋮	⋮
$X_{(1)}\cdots X_{(200)}$	M_{200}	$M_{200,1}$	$M_{200,2}$	…	$M_{200,100}$	$\overline{M}_{200}$

第二節　統計量排序

一、兩獨立母體期望值差 $\mu_1-\mu_2$ 之假設檢定

在比較攝護腺癌蛋白質質譜資料中的兩兩分類時，我們對於每個特徵變數先作兩母體之平均數是否有差異之檢定，因此對於每個特徵變數都要作一次 t 檢定才可得出 T 統計量。

首先令 μ_1 為第一類病人的母體平均數，μ_2 為第二類病人的母體平均數，可得一組假設檢定 $\begin{cases} H_0:\mu_1-\mu_2=0 \\ H_1:\mu_1-\mu_2\neq 0 \end{cases}$，其中虛無假設是表示兩類別的病人於此特徵變數的平均上是沒有差異的，而對立假設為兩類別的病人於此特徵變數的平均上是有差異的。

在此我們假設母體為常態分配且母體變異數未知不相等的情況下，其檢定統計量為 $T=\dfrac{\overline{X}_1-\overline{X}_2}{\sqrt{\dfrac{S_1}{n_1}+\dfrac{S_2}{n_2}}}$。因雙尾檢定故我們將每個特徵變數的 T 統計量皆取絕對值以方便特徵變數的排序。

二、一因子變異數分析完全隨機化設計檢定

變異數分析可以檢定兩個母體以上的平均數檢定，故我們在作兩兩分類和四分類的比較時，對於每個特徵變數亦可用此方法檢定。

一因子變異數分析就是在一因子實驗設計中，隨機分配 n 個實驗單位(總樣本數為 n)至 k 個不同類別的病患，故將各類別的樣本數記為 $n_1,n_2,\cdots,n_k$ ，因此 $n=\sum_{j=1}^{k}n_j$ 。而一因子變異數分析完全隨機化設計模型如下：

$$X_{jm}=\mu+\alpha_j+\varepsilon_{jm}=\mu_j+\varepsilon_{jm} \qquad j=1,2,\cdots,k、m=1,2,\cdots n_j$$

此處
- X_{jm}：第j類病患中之第m個樣本觀測值
- μ：k個類別的混合總平均
- μ_j：第j類病患之母體平均
- $\alpha_j=\mu_j-\mu$：第j類病患之處理效應
- ε_{jm}：第j類病患中第m個樣本之個別誤差效應

其假設檢定為 $\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : \mu_j \text{不全相等} \end{cases}$，其中虛無假設是 k 個類別的母體平均無差異，對立假設是 k 個類別的母體平均不全相等。最後將處理平方和（$SSTR$）除以其自由度 $k-1$ 的 $MSTR$ 與誤差平方和（SSE）除以其自由度 $n-k$ 的 MSE 相除即產生檢定統計量 $F = \frac{MSTR}{MSE}$。

三、多個獨立母體之無母數檢定 Kruskal-Wallis 檢定（KW 檢定）

此法不像上述的一因子變異數分析須要同質變異數或常態分配等假設，只須將 n 筆資料混合排序其值的大小，並利用等級來運算即可。它可以檢定母體的中位數或是母體分配的位置。而在本研究中我們對於各特徵變數分別進行 KW 檢定如下：

$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : \mu_j \text{不全相等} \end{cases}$ 其中虛無假設為 k 個類別的平均皆相等，對立假設則為 k 個類別之之平均不全相等。而檢定統計量為 $H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_k^2}{n_k} - 3(n+1)$，

此處 $\begin{cases} n = n_1 + n_2 + \cdots + n_k : k\text{組獨立樣本之樣本個數總和} \\ n_j : \text{第}j\text{組獨立樣本之樣本個數}, j = 1,2,\cdots,k \\ R_j : \text{第}j\text{組獨立樣本所對應觀測值之等級和} \end{cases}$，且此統計量會服從卡方自由度為 $k-1$ 之分配。

第三節 LARS、Stagewise、LASSO 迴歸模型

此節將 LARS、Stagewise、LASSO 一起討論的原因在於三種方法皆可由 LARS 演算法求得其迴歸係數估計值。此外，在演算法中如何修改使我們得知 Stagewise 和 LASSO 的估計值結果以及各方法運用至高維度資料時的優缺點也會在此節一併說明。

首先，將資料標準化後（$\sum_{i=1}^{n} y_i = 0$、$\sum_{i=1}^{n} x_{ij} = 0$ 和 $\sum_{i=1}^{n} x_{ij}^2 = 1$），我們先定義一個迴歸模型為 $\boldsymbol{y} = X_{n\times m}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$，

其中 $\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X_{n\times m}=[\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_m], \boldsymbol{\beta}=\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}, \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ 且假設 $\boldsymbol{\varepsilon} \sim N(0, I_n\sigma^2)$。

然後藉由最小平方法在最小化殘差平方和(亦即 $S(\hat{\boldsymbol{\beta}})=\left\|\boldsymbol{y}-\hat{\boldsymbol{\mu}}\right\|^2=\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$)之目標下求解迴歸係數後，可得樣本估計式為 $\hat{\boldsymbol{\mu}}=\sum_{i=1}^{n}\boldsymbol{x}_j\hat{\beta}_j=X\hat{\boldsymbol{\beta}}$。

一、Least Angle Regression(LARS)

我們先由圖 4.2 來簡單解釋 LARS 的概念。若現在只考慮兩個變量 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$ 對 $\boldsymbol{y}$ 的影響，先將 $\boldsymbol{y}$ 投影至由 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 生成的線性空間中之一個向量 $\bar{\boldsymbol{y}}_2$ 後，計算個變量與 $\boldsymbol{y}$ 的相關係數哪個較大，由圖中可知 $\bar{\boldsymbol{y}}_2$ 與 $\boldsymbol{x}_1$ 的夾角較小亦即 $\boldsymbol{x}_1$ 與 $\boldsymbol{y}$ 的相關係數較大，則我們就沿著 $\boldsymbol{x}_1$ 的方向前進，前進到 $\hat{\boldsymbol{\mu}}_1$ 處確定可產生一條 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 的角平方線向量(就是此向量與 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 相關程度一樣)且最後能與 $\bar{\boldsymbol{y}}_2$ 相交，然後再沿著 $\boldsymbol{u}_2$ 前進，最終可達到 $\bar{\boldsymbol{y}}_2$ 即配適完 LARS 迴歸估計式。

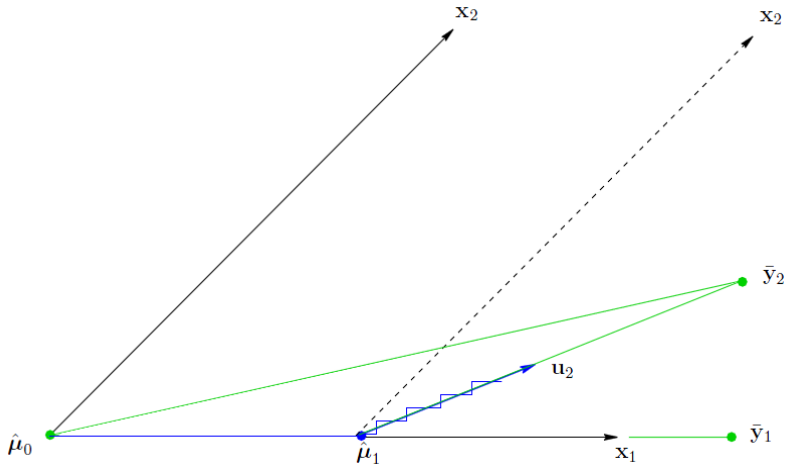

圖 4.2 二維自變數下的 LARS

資料來源：“Least Angle Regression” by Efron, B., Hastie, T., Johnstone, I. and Tibshirani R., 2003, *Annals of Statistics*, *32*(2), 412.

當 $m>2$ 時，我們即以數學式來表示。假設有 m 個變數，n 個觀察值，令 Ω 為 $\{1,2,\cdots,m\}$ 的一個子集合，又可稱為活動集合(active set)，則可定義 $X_\Omega=\left[\cdots s_j\boldsymbol{x}_j\cdots\right]_{j\in\Omega}$，其中 s_j 不是 $+1$ 即為 -1、$A_\Omega=(1_\Omega' G_\Omega^{-1}1_\Omega)^{-1/2}$，其中 $G_\Omega^{-1}=X_\Omega' X_\Omega$ 和

$\boldsymbol{u}_\Omega = X_\Omega w_\Omega$為角平方向量，其中$w_\Omega = A_\Omega G_\Omega^{-1} 1_\Omega$。

若有個當前的 LARS 估計式$\hat{\boldsymbol{\mu}}_\Omega$，則可得到一個當前相關(current correlations)向量$\hat{\boldsymbol{c}} = X'(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_\Omega)$，那麼Ω這個集合就是收集與當前相關取絕對值後最大的第 j 個變數，也就是$\hat{C} = \max_j \left\{ \left| \hat{c}_j \right| \right\}$而$\Omega = \left\{ j : \left| \hat{c}_j \right| = \hat{C} \right\}$，此外$s_j = \text{sgn}\left\{ \hat{c}_j \right\}$。則 LARS 下一步的估計式就是將$\hat{\boldsymbol{\mu}}_\Omega$調整為$\hat{\boldsymbol{\mu}}_{\Omega_+} = \hat{\boldsymbol{\mu}}_\Omega + \hat{\gamma} \boldsymbol{u}_\Omega$，在此定義一個內積向量$\boldsymbol{a} \equiv X' \boldsymbol{u}_\Omega$，並由

$\min^+_{j \in \Omega^c} \left\{ \frac{\hat{C} - \hat{c}_j}{A_\Omega - a_j}, \frac{\hat{C} - \hat{c}_j}{A_\Omega + a_j} \right\}$來決定$\hat{\gamma}$值，最後新的$\Omega_+$就是$\Omega \cup \left\{ \hat{j} \right\}$。

此種迴歸模型選取變數的方法是由 Efron 等人(2003)提出，其優點除了可產生一個完整分段的線性模型並得出各迴歸係數之變化路徑外，也可以很容易的將其修改為其他迴歸模型(如：Stagewise、LASSO)，此外若同時有兩個變數與應變數的相關程度差不多時，則它們的迴歸係數應該會以約略相同的速度增加。不過這方法的缺點是因為利用殘差來疊代，所以若面對高維度資料中變數間容易不為獨立時，會容易影響其模型的選取結果，導致對於變數的排序可能也不太可靠；另外在高維度的資料中，對於模型的配適易受訓練集中樣本數的多寡而受限，也就是說若變數個數(m)大於樣本數(n)的話，LARS 迴歸模型中最多就會有 n 個迴歸係數不為零。(Efron 等人，2003；Leng 等人，2006；Hastie 等人，2009)

二、Forward Stagewise Regression

與向前選取方法類似，差別是在於此方法每一步的移動幅度很小，且較謹慎。簡單來說由第一個模型$\hat{\boldsymbol{\mu}}_0$開始出發，假設$\boldsymbol{x}_j$與$\hat{\boldsymbol{\mu}}_0$的當前相關最大，亦即有最大的$c_j = \boldsymbol{x}_j^T (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)$，接著取一個$\varepsilon$ $(0 < \varepsilon < \left| c_j \right|)$可得第二個模型為$\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_0 + \varepsilon \cdot \text{sgn}(c_j) \boldsymbol{x}_j$，然後重複上述步驟直到模型中所產生的當前殘差與剩餘未被選入模型的變數都不相關即停止。而修正 LARS 演算法來求解 Stagewise 估計值的方法是先考慮 Stagewise 選取變數的每個步驟大小ε趨近於零，假設已作了 N 個 Stagewise 的步驟，就會產生一些估計式$\hat{\boldsymbol{\mu}}$且將N_j定義成將第 j 個變數選入模型中所須之步驟數，其中$j = 1,2,\cdots,m$，當$j \notin \Omega$時，

$N_j = 0$。定義 P 為 $(N_1, N_2, \cdots, N_m)/N$，則對於 $j \in \Omega$ 可產生一個新估計式為

① $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} + N_\varepsilon X_\Omega P_\Omega$，此外我們也已經知道 LARS 演算法是沿著② $\boldsymbol{\mu}_\Omega + \gamma X_\Omega w_\Omega$ 前進。比較①和②可以發現因為 P_Ω 是非負的，故 LARS 估計式在 w_Ω 中有負值時會與 Stagewise 不吻合，這時藉由 $C_\Omega = \left\{ \boldsymbol{v} = \sum_{j \in \Omega} s_j \boldsymbol{x}_j P_j, P_j \geq 0 \right\}$ 集合使得 $\boldsymbol{\mu}_\Omega \in C_\Omega$ 則可以去除這裡的矛盾。

此方法的優點是比向前選取法更謹慎的將變數選入模型中，而且可視為單調版的 LASSO，因為當 LARS 或是 LASSO 對於某個變數之迴歸係數有向零變化之趨勢時，此時的 Stagewise 反而會將其變數之迴歸係數膨脹，並盡可能的將其變數之迴歸係數維持非遞增或非遞減的變化，而此特性會在實証分析的部份有所驗証。而缺點就是比較沒有縮減變數之效果且在高維度資料中，迴歸模型最多會選入與樣本數(n)相同之變數個數。(Efron 等人，2003；Hastie 等人，2009；Leng 等人，2006)

三、LASSO

Tibshirani(1996)提出關於 LASSO 的文章，LASSO 是一種具有懲罰項的迴歸估計式，即限制在 $\sum_{j=1}^{m} \left| \hat{\beta}_j \right| \leq t$ 條件下，能夠最小化殘差平方和的迴歸係數值即為 LASSO 估計值。由於這種限制條件是不等號的情況，因此必須藉由凸函數最佳化問題中的 KKT 條件來求解。另外，LASSO 估計式具有縮減迴歸係數值和變數選取兩種功能，當 t 值小到一定程度的時候，LASSO 估計式能夠使得某些迴歸係數的估計值為零，因此的確可以達到選取變數的作用。當 t 不斷增加時，被選入迴歸模型中的變數也會逐漸增多，且當 t 增大到某個值時，所有的變數都會被選入迴歸模型，而這時迴歸模型中的各迴歸係數值會與利用最小平方法所求出來的迴歸係數值相等。另一方面，也可視它為一種縮減維度的迴歸方法，因為它可以將不顯著的迴歸係數自動估計為零，因此估計迴歸係數和縮減維度可同時被完成。

對於 LASSO 估計式，LARS 演算法之修正方式是先假設已經完成若干 LARS 的步驟，所以這時已經存在一個活動集合 Ω，且可得知一個估計式 $\hat{\boldsymbol{\mu}}_\Omega$。假設 LASSO 的迴歸係數估計值向量為 $\hat{\boldsymbol{\beta}}$，而 $X_\Omega w_\Omega$ 為在 LARS 中的角平分向量。此處定義一個向量 $\hat{d}$, 向量中

的元素是 $j \in \Omega$ 時其值為 $s_j w_j$ ，反之為零。因此 $\hat{\boldsymbol{\mu}}_\Omega = X\boldsymbol{\beta}(\gamma)$ ，其中 $\beta_j(\gamma) = \hat{\beta}_j + \gamma \hat{d}_j$ 。此時若將 LARS 的角平分向量對應至 LASSO 估計值的路徑上會發現 $\beta_j(\gamma)$ 在 $\gamma_j = -\hat{\beta}_j / \hat{d}_j$ 時變號那麼對於我們已經有的 LASSO 估計值 $\boldsymbol{\beta}(\gamma)$ 中的元素會在最小的的那個大於零的 γ_j 處變號且將其記為 $\tilde{\gamma}$ 表示 $\min_{\gamma_j > 0}(\gamma_j)$ ，若沒有 γ_j 大於零的話, 則將 $\tilde{\gamma}$ 記為無窮大。另外，若 $\tilde{\gamma}$ 小於 $\hat{\gamma}$ ，因 $\beta_j(\gamma)$ 之正負號需與 $c_j(\gamma)$ 一致，則對應至 LARS 估計的 $\beta_j(\gamma)$ 就不會成為一個 LASSO 估計值，所以在此狀況下就不能繼續在 LARS 的步驟上繼續前進，而解決辦法就是將與 γ_j 相等的 $\tilde{\gamma}$ 所對應之 $\tilde{j}$ 從 Ω 中刪掉後再繼續進行 LARS 步驟以得出 LASSO 迴歸係數估計值。 (Efron 等人，2003；Leng 等人，2006）

LASSO 這種方法的優點是可同時達到將迴歸係數縮減和變數選取的效果。不過缺點是面對高維度資料變數個數大於樣本數時，除了在對於兩個具有高度相關的變數選入模型的方式是採「任意」的方式將某個變數加入其模型中，因而影響變數的排序；此外最後模型中之非零迴歸係數個數最多並不會超過樣本數(n)。(Hastie 等人，2009)

當面對高維度資料時，以上三種迴歸選模方法相同之處是除了都可產生一連串選取變數模型的過程外，每種方法的步驟數雖不相同，而實際上最後會被選入迴歸模型中的變數個數最多都不會超過樣本數(n)。而在高維度資料中，LARS 選取變數的步驟數目會由樣本數(n)來控制，而 Stagewise 是會根據一個很接近零的 ε 來控制其步驟數目，LASSO 則是依據 t 值的變化來更改其迴歸模型。此外，在選取變數的過程中只有 LARS 必會於每個步驟中將某個變數加入活動集合中(從此再也不會將此變數從活動集合中刪除)，但 Stagewise 以及 LASSO 就可能會出現將某個變數先加入活動集合中，然後經過若干步驟後又將此變數從活動集合中刪除的情形發生。

第四節 Group LASSO 迴歸模型

Yuan 和 Lin(2006)提出 Group LASSO 這種迴歸模型選取方法來改善 LASSO 在面對高維度資料時，由於變數間可能存在著高度相關而導致變數的選取結果不太可性的缺點。其想法是來自眾多的自變數中，也許可分成若干群組，而且在變數選取時，通常是選擇一個群組，而不是一個個別的變數。

接著我們利用數學式來說明此方法，首先假設一基本的迴歸模型如下：

$$Y_{n\times 1} = X_{n\times m}\boldsymbol{\beta}_{m\times 1} + \varepsilon_{n\times 1} \tag{1}$$

再來若我們欲將 m 個變數分成 J 個組，則可將(1)式變為

$$Y_{n\times 1} = \sum_{j=1}^{J} X_j\boldsymbol{\beta}_j + \varepsilon_{n\times 1} \tag{2}$$

其中 $\boldsymbol{\beta}_j$ 的向量長度用 p_j 來表示，也就是對應至第 j 組中的變數個數。接著對於一個向量 $\eta \in R^d$（$d \geq 1$）且定義一 $d\times d$ 的對稱正定矩陣為 κ，並產生一個矩陣為 $\|\eta\|_\kappa = (\eta'\kappa\eta)^{1/2}$，在給定 $\kappa_1,\cdots,\kappa_J$ 後即可得到 Group LASSO 的拉格朗日方程式(Lagrange equation)：

$$\frac{1}{2}\left\|Y - \sum_{j=1}^{J} X_j\boldsymbol{\beta}_j\right\|^2 + \lambda\sum_{j=1}^{J}\left\|\boldsymbol{\beta}_j\right\|_{\kappa_j},\ \lambda \geq 0 \tag{3}$$

最後只要我們找出能夠使(3)式得到最小值的 $\hat{\boldsymbol{\beta}}_j$ 即為 Group LASSO 的迴歸係數估計值。因此 Group LASSO 之迴歸係數估計式如下：

$$\hat{\beta}_j = (1 - \frac{\lambda\sqrt{p_j}}{\|S_j\|})^+ S_j$$

其中 $S_j = X_j'(Y - X\beta_{-j})$ 且 $\beta_{-j} = (\beta_1',\cdots,\beta_{j-1}',0,\beta_{j+1}',\cdots,\beta_J')$。

而在本研究中，我們的作法是將每個變數都視為一個群組，也就是說若共有 417 個變數，則我們就分為 417 個群組，然後來得出各分類之判別結果。(Yuan 和 Lin，2006；Friedman 等人，2010)

第五節 Elastic Net 迴歸模型

這方法的限制式是脊迴歸(Ridge regression)和 LASSO 的融合，由此產生以下 Elastic Net 的拉格朗日方程式：

$$L(\lambda_1,\lambda_2,\boldsymbol{\beta})=\left|Y-X\boldsymbol{\beta}\right|^2+\lambda_2\left|\boldsymbol{\beta}\right|_2^2+\lambda_1\left|\boldsymbol{\beta}\right|_1$$

所以若 $\lambda_1=0$，則為脊迴歸的拉格朗日方程式；若 $\lambda_2=0$，則形成 LASSO 的拉格朗日方程式。(Zou 和 Hastie，2004)

對於此方法的迴歸係數估計值，Park 和 Hastie(2006)提出利用預測−校正(Predictor-corrector)演算法來求得 Elastic Net 之所有迴歸係數的路徑發展。先假設一組具有 n 筆樣本和 p 個變數的資料為 $\left\{(\boldsymbol{x}_i,y_i):\ \boldsymbol{x}_i\in R^p、y_i\in R，i=1,\cdots,n\right\}$，且 Y 為服從指數家族之分配($\mu=E(Y)$ 、 $V=Var(Y)$)。設 $\eta=g(\mu)=\beta_0+\boldsymbol{x}'\beta$ 而 Y 的密度函數為 $L(y;\theta,\phi)=\exp\left\{(y\theta-b(\theta))/a(\phi)+c(y,\phi)\right\}$，其中 $a(.)$、$b(.)$ 和 $c(.)$ 會依 Y 分配不同而改變。

$$l(\beta,\lambda)=-\sum_{i=1}^{n}\left\{y_i\theta(\beta)_i-b(\theta(\beta)_i)\right\}+\lambda\left\|\beta\right\|_1 \tag{4}$$

在此若已知散佈變數 ϕ 為已知，欲找出能夠極大化概似函數的 θ 也就是相當於尋求使第(4)式最小的 $\boldsymbol{\beta}=(\beta_0,\beta')'$。假設 $\boldsymbol{\beta}$ 不為零，在此我們令一函數 H 為：

$$H(\boldsymbol{\beta},\lambda)=\frac{\partial l}{\partial\boldsymbol{\beta}}=-\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu})\frac{\partial\eta}{\partial\boldsymbol{\mu}}+\lambda\,\mathrm{sgn}\begin{bmatrix}0\\ \boldsymbol{\beta}\end{bmatrix} \tag{5}$$

其中

$$\boldsymbol{X}=\begin{bmatrix}1 & x_{11} & \vdots & x_{1p}\\ 1 & x_{21} & \vdots & x_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & \vdots & x_{np}\end{bmatrix}_{n\times(p+1)}，\boldsymbol{W}=\begin{bmatrix}V_1^{-1}(\frac{\partial l}{\partial\eta})_1^2 & 0 & \cdots & 0\\ 0 & V_2^{-1}(\frac{\partial l}{\partial\eta})_2^2 & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & V_n^{-1}(\frac{\partial l}{\partial\eta})_n^2\end{bmatrix}_{n\times n}，$$

$$(\boldsymbol{y}-\boldsymbol{\mu})\frac{\partial l}{\partial\boldsymbol{\mu}}=\begin{bmatrix}(y_1-\mu_1)(\frac{\partial l}{\partial\boldsymbol{\mu}})_1\\ \vdots\\ (y_n-\mu_n)(\frac{\partial l}{\partial\boldsymbol{\mu}})_n\end{bmatrix}_{n\times1}。$$

每給定一個不同的λ值，則第(5)式就需重新計算迴歸係數值。因此預測-校正演算法的目標就是在λ從最大值(λ_{max})往零移動的過程中記錄迴歸係數估計值的變化路徑。而預測校正演算法大致可分為以下四個部份：

1.預測步驟(Predictor step)

假設在第k個預測步驟，則$\boldsymbol{\beta}(\lambda_{k+1})$可藉由以下公式概略計算得來：

$$\hat{\boldsymbol{\beta}}^{k+} = \hat{\boldsymbol{\beta}}^{k} + (\lambda_{k+1} - \lambda_k)\frac{\partial \boldsymbol{\beta}}{\partial \lambda} = \hat{\boldsymbol{\beta}}^{k} - (\lambda_{k+1} - \lambda_k)(\boldsymbol{X}_{\Omega}^{'}\boldsymbol{W}_k\boldsymbol{X}_{\Omega})^{'}\,\mathrm{sgn}\begin{bmatrix} 0 \\ \hat{\boldsymbol{\beta}}^{k} \end{bmatrix}$$

其中$\boldsymbol{W}_k$和$\boldsymbol{X}_{\Omega}$分別表示當前的權重矩陣以及在當前活動集合中的共變量矩陣，且此處$\boldsymbol{\beta}$是只有包含當前迴歸係數被估計成非零的迴歸係數矩陣。

2.校正步驟(Corrector step)

利用$\hat{\boldsymbol{\beta}}^{k+1}$為初始值，來找尋可以最小化$l(\boldsymbol{\beta}, \lambda_{k+1})$ l(β, λ_{k+1})的$\boldsymbol{\beta}$。基本上，$\hat{\boldsymbol{\beta}}^{k+}$通常會很接近$\hat{\boldsymbol{\beta}}^{k+1}$。在此步驟中不僅可找給定一個$\lambda$下的精確解，也提供了之後$\boldsymbol{\beta}$在預測步驟中的方向。

3.活動集合(Active set)

一開始的活動集合中只有截距項，但之後在作每一個修正步驟時，若某個原本不在活動集合中的變數滿足以下第(6)式時，就放入活動集合中。且一直重複校正步驟直到活動集合不再擴大，然後我們再將原本在活動集合中的變數代入第(7)式，若發現某變數之迴歸估計值為零，則再將此變數從活動集合中剔除。

$$\left|\boldsymbol{x}_j^{'}\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\mu})\frac{\partial \eta}{\partial \mu}\right| > \lambda，\ \forall j \in \Omega^c \Rightarrow \Omega \leftarrow \Omega \cup \{j\} \tag{6}$$

$$\left|\hat{\beta}_j\right| < 0，\ \forall j \in \Omega \Rightarrow \Omega \leftarrow \Omega \setminus \{j\} \tag{7}$$

4.步驟長度(Step length)

就是估計前一個步驟與後一個步驟的距離(長度)，也就是$\lambda_k - \lambda_{k+1}$。假設由一個修正步驟中可得到$\boldsymbol{y}$的估計$\hat{\boldsymbol{\mu}}$而第(8)式即表示$\hat{\boldsymbol{\mu}}$所對應的權重相關係數矩陣。

$$\hat{\boldsymbol{c}} = \boldsymbol{X}^{'}\hat{\boldsymbol{W}}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})\frac{\partial \eta}{\partial \mu} \tag{8}$$

然後因在下一個預測步驟中會擴展 $\hat{\boldsymbol{\beta}}$，故當前的相關係數矩陣也會改變。而第(9)式為下降一單位的 λ 時其相關矩陣的變動情形。

$$\boldsymbol{c}(h) = \hat{\boldsymbol{c}} - h\boldsymbol{a} = \hat{\boldsymbol{c}} - h\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}_{\Omega}(\boldsymbol{X}_{\Omega}'\hat{\boldsymbol{W}}\boldsymbol{X}_{\Omega})^{-1}\operatorname{sgn}\begin{bmatrix}0\\ \hat{\boldsymbol{\beta}}\end{bmatrix} \tag{9}$$

再來我們利用第(10)式可解得一個估計步驟長度的估計值即第(11)式。

$$\left|c_j(h)\right| = \left|\hat{c}_j - ha_j\right| = \lambda - h \ ,\ \forall j \in \Omega^c \tag{10}$$

$$h = \min_{j\in\Omega^c}\left\{\frac{\lambda-\hat{c}_j}{1-a_j}, \frac{\lambda+\hat{c}_j}{1+a_j}\right\} \tag{11}$$

那麼將 Elastic Net 運用至預測-校正運算法中來求取迴歸係數的路徑方式就是將第(12)式中的 λ_2 固定成某個很小的正數且 $\lambda_1 \in (0,\infty)$。

$$\hat{\boldsymbol{\beta}}(\lambda_1) = \arg\min_{\boldsymbol{\beta}}\left[-\log L(\boldsymbol{y};\boldsymbol{\beta}) + \lambda_1\|\beta\|_1 + \frac{\lambda_2}{2}\|\beta\|_2^2\right] \tag{12}$$

假設 $\boldsymbol{\beta}$ 皆不為零且 $\boldsymbol{X}$ 沒有行獨立時，則 $\partial H(\boldsymbol{\beta},\lambda)/\partial\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ 將會形成不可逆矩陣，然而再藉由增加一個二次懲罰項的話，則我們可將 H 重新定義為第(13)式。

$$\tilde{H}(\boldsymbol{\beta},\lambda_1,\lambda_2) = -\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu})\frac{\partial\eta}{\partial\boldsymbol{\mu}} + \lambda_1\operatorname{sgn}\begin{bmatrix}0\\ \boldsymbol{\beta}\end{bmatrix} + \lambda_2\begin{bmatrix}0\\ \boldsymbol{\beta}\end{bmatrix} \tag{13}$$

此時對於任何的 $\lambda_2 > 0$，$\dfrac{\partial\tilde{H}}{\partial\boldsymbol{\beta}}$ 就是個可逆矩陣。因此對於自變數間有高度相關的資料之處理方式即固定 λ_2 為一個小的值，然後將 λ_1 由它的最大值往零移動來得出迴歸係數的整個估計路徑。因此，不同於 LASSO 迴歸模型。(Park 和 Hastie，2006)

此迴歸方法除了考慮到變數的縮減之外也考慮到群組效果(grouping effect)，也就是說它具有群組選取(group selection)的能力。也就是說遇到有兩個具有高度相關的變數時，此方法會同時將這兩個變數選入模型中，這樣一來不僅沒有使其模型的預測準確度消失同時也達到縮減變數的效果。因此在處理高維度資料時，Elastic Net 所選取到的變數與 LASSO 相比之下也較可靠，不過其演算法確實也複雜許多。(Zou 和 Hastie，2004)

第六節 支持向量機 SVM

支持向量機的英文全名為 Support Vector Machine，簡稱為 SVM，是一種基於統計學習理論(statistical learning theory)基礎的學習機器。其概念為對於一群資料而言，我們會希望依據資料的某些特性將這群資料分為兩群。以二維的例子來說，如圖 4.3，我們希望能找出一條線將黑點和白點分開，且這條線距離這兩個類別的邊際(margin)越大越好，才能夠很明確的分辨某個點是屬於那個類別，否則在計算上容易因精確度的問題而產生誤差。

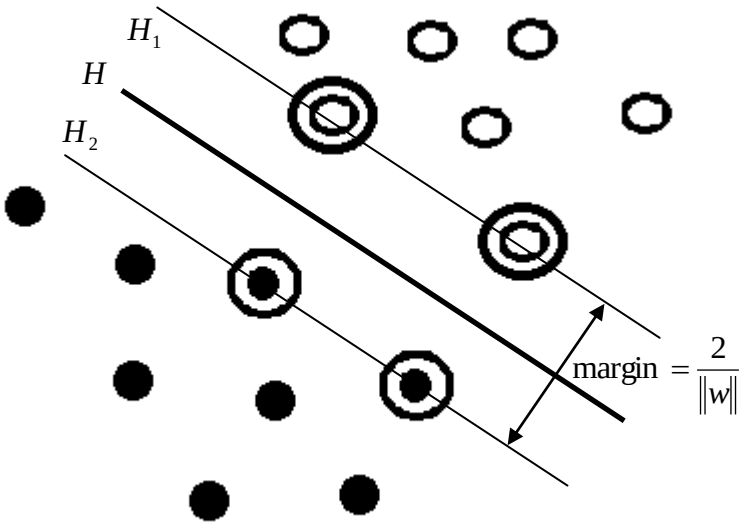

圖 4.3 SVM 原理示意圖

因此以兩種分類的 SVM 來說，假設訓練資料為 $\{\boldsymbol{x}_i, y_i\}$ 其中 $i=1,\cdots,l$、$x_i \in R^d$、$y_i \in \{+1,-1\}$ 且 l 為訓練資料個數、d 是維度。而在此分類超平面上的 $\boldsymbol{x}$ 需滿足 $\boldsymbol{w}^T\boldsymbol{x}+b=0$，其中 $\boldsymbol{w} \in R^l$ 是訓練樣本中係數向量，b 是截距常數。若我們令 $f(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}+b$ 為決策函數，那麼將測試資料代入決策函數後即可根據決策函數之值來加以分類，也就是希望找到一條線 $f(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}+b$ 使所有 $y_i=+1$ 的點落在 $f(\boldsymbol{x})>0$ 的這一邊，且使所有 $y_i=-1$ 的點落在 $f(\boldsymbol{x})<0$ 這一邊，這樣就可根據 $f(\boldsymbol{x})$ 的正負號來區分點是屬於這兩個分類的哪一類，故此超平面稱為分類超平面(separating hyperplane)，而

距離兩邊邊界最大的距離且到各類最近點的距離相等時稱為最佳分類超平面，而此處的距離相等是指此分類超平面到兩邊界的距離皆為$1/\|\boldsymbol{w}\|$，因此為了求取最佳分類超平面就是要求$\boldsymbol{w}$之最小值。

此外，由於一般的分類平面方程式為$\boldsymbol{w}^T\phi(\boldsymbol{x})+b$，而$\phi$是指將樣本空間轉換到另一個高維度空間的對應函數。當$\phi(\boldsymbol{x}_i)=\boldsymbol{x}_i$時，即表示在原本空間中可找到一個平面將資料分類並獲得最小的分類錯誤以及不同類別之間的最大間隔，我們就稱此為「線性可分問題」。

若遇到線性不可分的問題時，則必須將原始向量空間對應到較高維度的空間，來尋找分類的超平面，此時分類平面方程式會變成$\sum_{i=1}^{l}\alpha_i y_i\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_i)+b$，其中的轉換核心$\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_i)$以$K(\boldsymbol{x}_i,\boldsymbol{x})$取代後，分類平面方程式即成為$\sum_{i=1}^{l}\alpha_i y_i K(\boldsymbol{x}_i,\boldsymbol{x})+b$，通常使用的轉換核心有以下三種：

- 線性（Linear）：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{x}_i^T\boldsymbol{x}_j$
- 半徑基底函數（Radial basis function, RBF）：$e^{-\gamma\|\boldsymbol{x}_i-\boldsymbol{x}\|^2}$
- 多項式（Polynomial）：$(\gamma(\boldsymbol{x}_i^T\boldsymbol{x}_j)+\delta)^d$

其中$\gamma,d,\delta\in R$是轉換核心的參數，其分類規則也可以表示成$\text{sgn}(\sum_{i=1}^{l}\alpha_i y_i K(\boldsymbol{x}_i,\boldsymbol{x})+b)$。

第五章 實証分析

本研究是利用事前處理的攝護腺癌之蛋白質質譜資料來進行判別兩兩分類以及四分類的分析，而我們的目的最主要是探討各(具有懲罰項)迴歸模型在各分類之判別結果的表現情形。此部份我們會先利用五種「選入迴歸模型之順序排序」的特徵變數選取結果與三種「統計量排序」和「分錯率排序」的特徵變數選取結果作比較，然後再將五種迴歸模型中表現最突出者與 Adam 等人(2002)以及陳詩佳(2007)之判別結果作比較。

首先分別將四種類別，正常、良性腫瘤、癌症早期以及癌症晚期分為訓練資料和測試資料兩個部份，而表 5.1 是各分類之訓練資料和測試資料的筆數。

表 5.1

訓練資料和測試資料筆數

類別	簡稱	受測者	資料筆數	訓練資料	測試資料
正常	NO	82	164	110	54
良性腫瘤	BPH	77	154	102	52
癌症早期	CAB	84	168	112	56
癌症晚期	CCD	83	166	110	56

再來由於我們要比較兩兩分類以及四分類的判別結果，故我們將正常、良性腫瘤、癌症早期以及癌症晚期四種類別的資料合併為「正常 vs.良性腫瘤」(NO vs. BPH)、「正常 vs.癌症早期」(NO vs. CAB)、「正常 vs.癌症晚期」(NO vs. CCD)、「良性腫瘤 vs.癌症早期」(BPH vs. CAB)、「良性腫瘤 vs.癌症晚期」(BPH vs. CCD)、「癌症早期 vs.癌症晚期」(CAB vs. CCD)以及、「正常 vs.良性腫瘤 vs.癌症早期 vs.癌症晚期」(四分類)這七個類別資料檔。其實我們從這七個資料中會發現某些特徵變數在其資料中的觀測值皆為零，我們會認為此特徵變數在其資料中是無意義的，於是將它刪除再對資料進行分析，因此表 5.2 是分別在七個類別資料檔中，刪除所有受測者測得強度為零的那些特徵變數後所剩餘的特徵變數個數。

表 5.2

七種合併資料刪除零後之特徵變數個數

合併資料	刪除零後之特徵變數個數
NO vs. BPH	470
NO vs. CAB	555
NO vs. CCD	634
BPH vs. CAB	555
BPH vs. CCD	629
CAB vs. CCD	678
四分類	740

第一節 R 函數之設定

本篇研究中當我們在分析兩兩分類以及四分類時，是利用 R package 中的 “LARS” 函數來配適 LARS、Stagewise 以及 LASSO 迴歸模型，因此需在”LARS”函數中作設定，而以下為我們設定的內容：

1. 有截矩項的存在(intercept=TRUE)。
2. 資料為未標準化資料(normalize=FALSE)。
3. 型態的部份分別設定成 type=lar、forward.stagewise 和 lasso。

另外，我們也利用 R package “glmpath”函數來配適 Elastic Net 迴歸模型，其設定內容如下：

1. 將 y 的分配設為二項分配(family=binomial(link = "logit"))
2. 資料為未標準化資料(standardize=FALSE)

上述兩點是對於每組訓練集都採用的設定。但此函數較特別的是，有兩個設定值 “min.lambda“和“relax.lambda“必須依據各組訓練資料而作改變。所謂的“min.lambda“表示我們可以決定第四章第五節中所提到的 λ_1 之下界，由於每一組訓練資料的變數個數皆比觀測個數還要多，因此將此值設為 $1e^{-6}$ 或是其他不為零的值較適當。而 “relax.lambda

“的設定會與某個變數在哪一步被選入迴歸模型有相關，因此若某個變數滿足 $\left|l'(\boldsymbol{\beta})\right| > \lambda_1 \times (1 - \text{ralax.lamb da})$ 就將其加入模型，此處 $l(\boldsymbol{\beta}) = -\sum_{i=1}^{n}\{y_i\theta(\boldsymbol{\beta})_i - b(\theta(\boldsymbol{\beta})_i)\} + \lambda_1\|\boldsymbol{\beta}\|_1$。而“relax.lambda“的預設是為 $1e^{-8}$。若超過 20 個步驟仍然無變數可加入活動集合的話，則可將其值增加到 $1e^{-7}$ 或 $1e^{-6}$。其實有時藉由對於這個值的調整，才可免除在執行 R 軟體時會遇到一些的問題，如：出現無法收斂的警告標語或是 R 軟體呈現無法回應的狀態。

第二節 探討兩兩分類之分錯率結果

在此我們先藉由圖 5.1、圖 5.2、圖 5.3、圖 5.4、圖 5.5 以及圖 5.6 來討論九種特徵選取方法於各兩兩分類下的分錯率表現。

圖 5.1 為判別正常和良性腫瘤的分錯率趨勢圖，可以發現 LARS 和 Stagewise 的分錯率的走向很一致，有同時上升同時下降的現象。另外，當特徵變數組合數為 1 或 2 個時，Elastic Net 所得之分錯率比其他特徵選取方法高很多。此外利用 Group LASSO 所選取的前兩百名特徵變數不論其組合數為何所得之分錯率皆比其他特徵選取方法來得低。

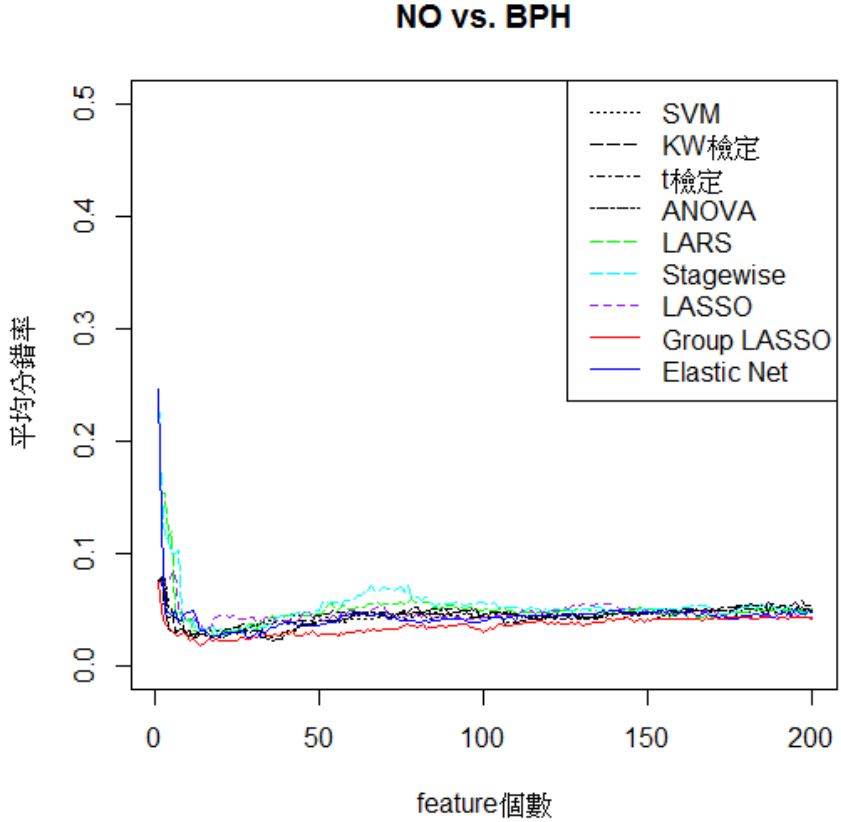

圖 5.1 各特徵選取方法下判別正常與良性腫瘤之分錯率趨勢圖

圖 5.2 為判別正常和癌症早期的分錯率趨勢圖，發現 SVM 之「分錯率排序」方法在組合數超過 84 個以後其分錯率的表現與其他相較之下最不理想。 而 LARS 和 Stagewise 在組合數 1 至 15 個時，其分錯率完全相同。而 Group LASSO 在組合數 21 至 120 個時之分錯率比其他者更優。

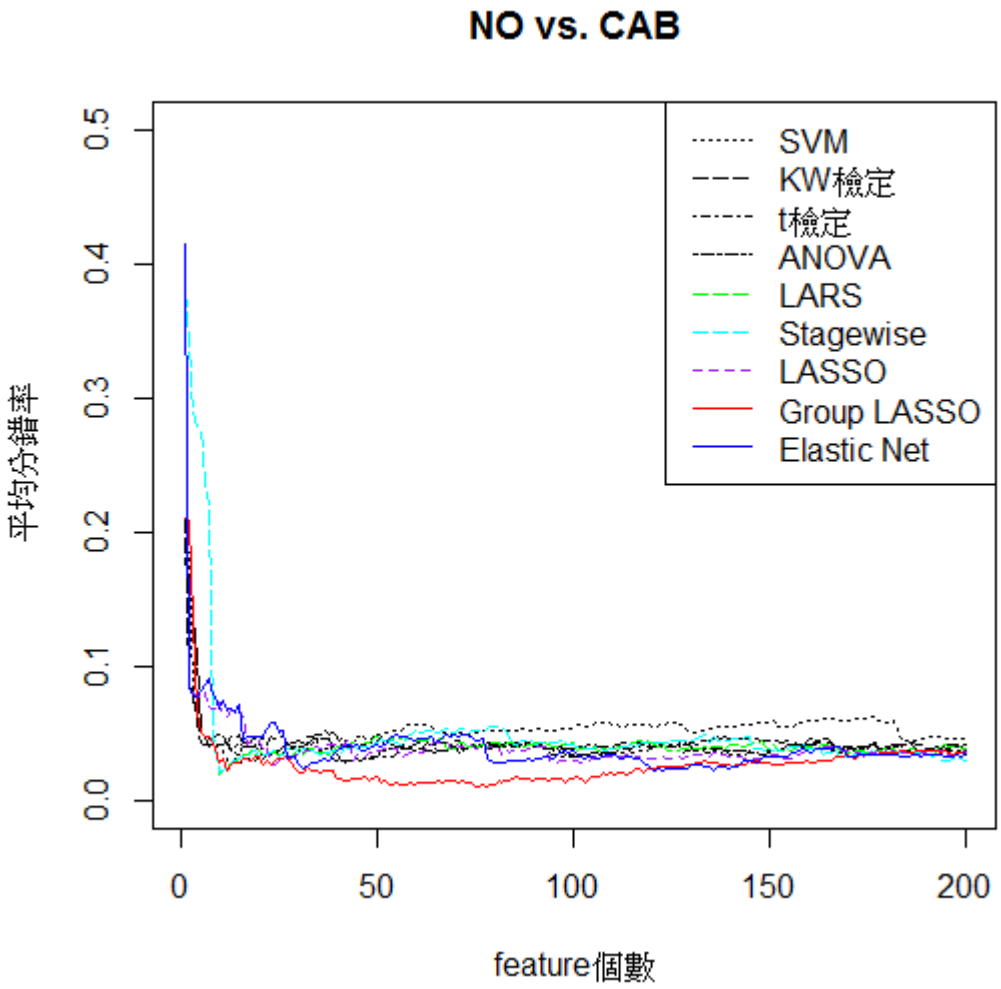

圖 5.2 各特徵選取方法下判別正常與癌症早期之分錯率趨勢圖

圖 5.3 為判別正常和癌症晚期的分錯率趨勢圖，發現在組合數大約介於 1 至 60 個之間時，不論何種特徵選取方法其分錯率有很明顯的波動，其中利用 LARS 和 Stagewise 之特徵選取方法在組合數由 10 增加為 11 個時的分錯率差異極大。另外，藉由 KW 檢定的「統計量排序」方法在組合數介於 39 至 70 個間的分錯率表現最不理想。

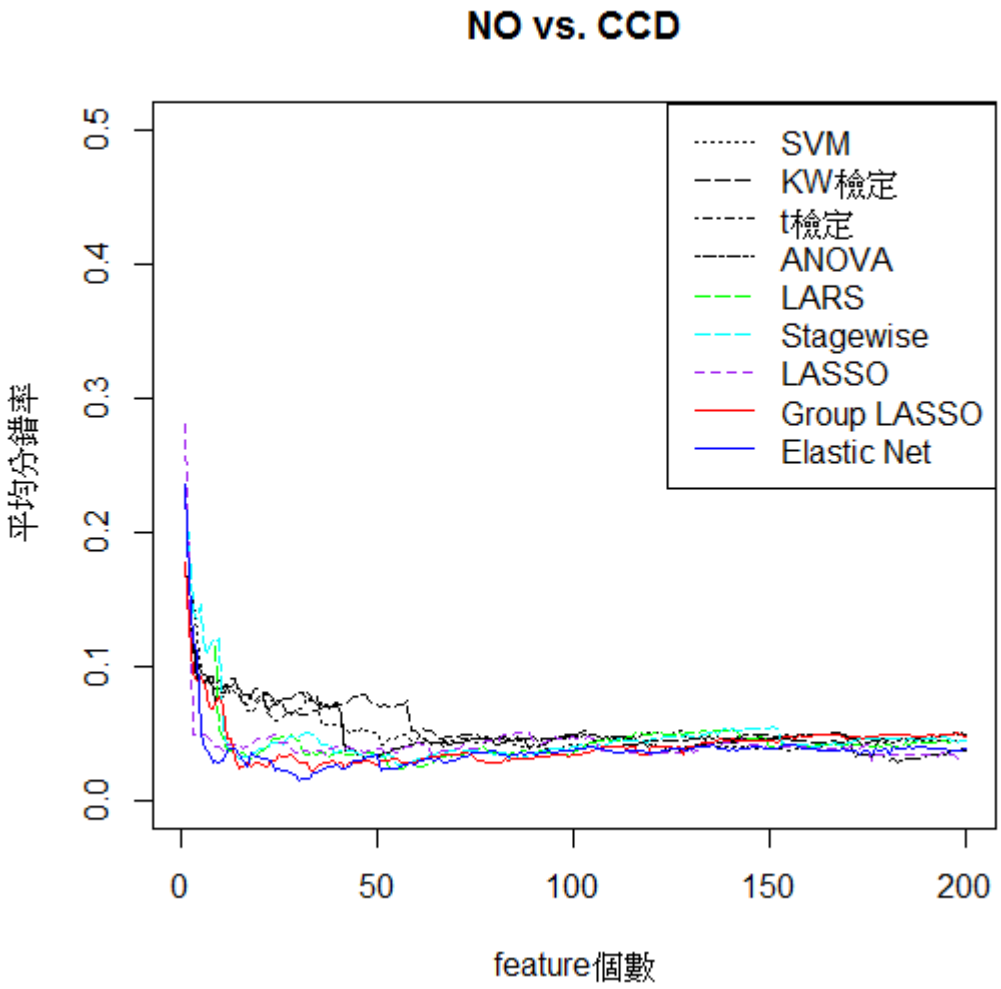

圖 5.3 各特徵選取方法下判別正常與癌症晚期之分錯率趨勢圖

圖 5.4 為判別良性腫瘤和癌症早期的分錯率趨勢圖，可發現當組合數為 1 個時，Elastic Net 之分錯率將近百分之五十，但組合數增為 2 個時其分錯率馬上降為百分之十三，此外 LARS、 Stagewise 和 LASSO 在組合數為 1 至 12 個間其分錯率快速下降，而 LASSO 在組合數為 33 至 60 個之間和組合數超過 108 個時其分錯率較其他方法高。

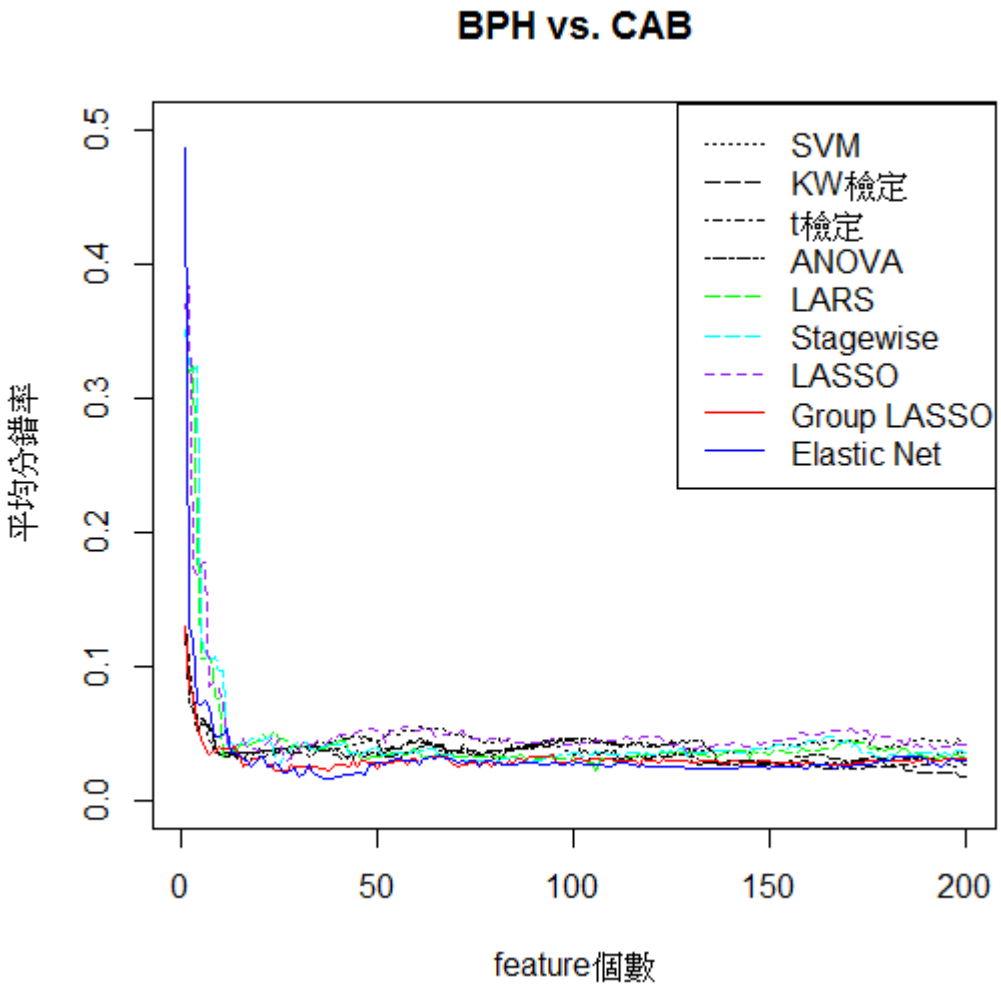

圖 5.4 各特徵選取方法下判別良性腫瘤與癌症早期之分錯率趨勢圖

圖 5.5 為判別良性腫瘤和癌症晚期的分錯率趨勢圖，當組合數為 1 個時，LARS、Stagewise、LASSO 以及 Elastic Net 的分錯率最高，不過當組合數到達 13 個時其分錯率已與其他方法相近，然而上述方法在組合數大約到達 140 個以後，其分錯率有明顯上升的跡象。

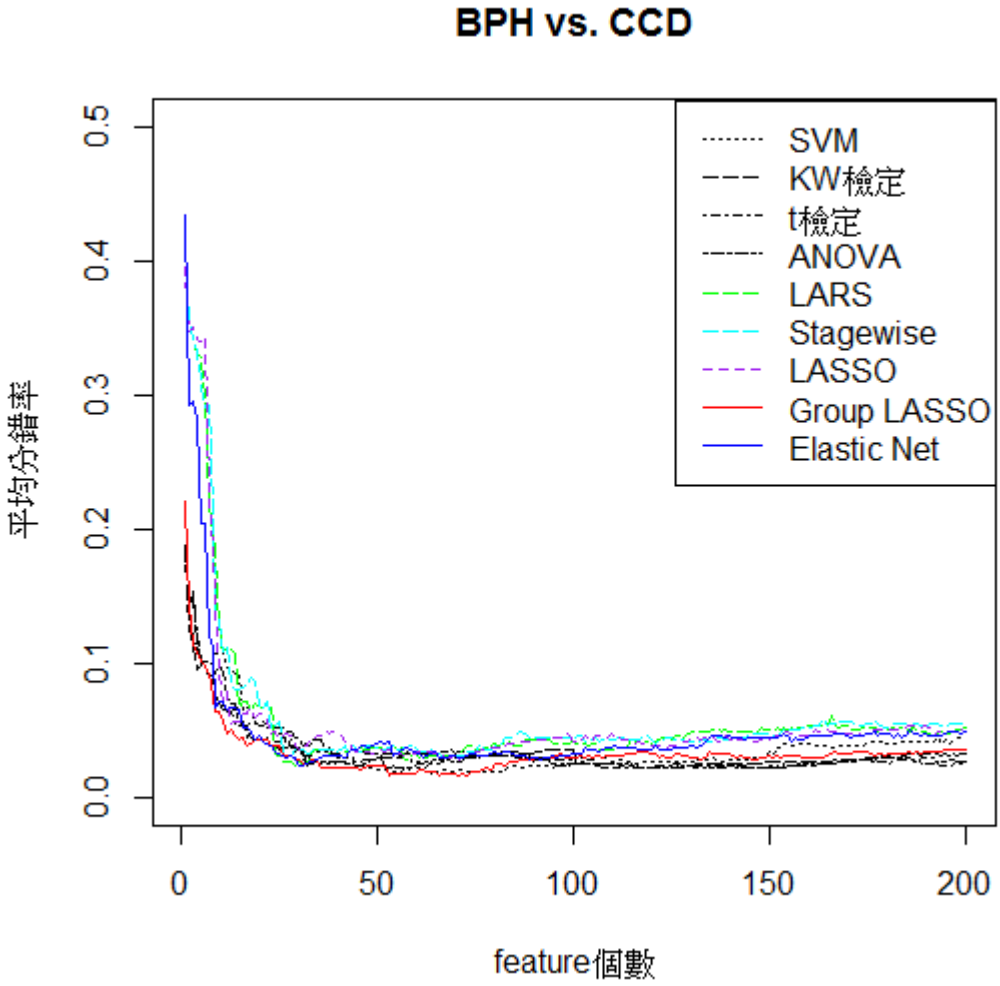

圖 5.5 各特徵選取方法下判別良性腫瘤與癌症晚期之分錯率趨勢圖

圖 5.6 判別癌症早期和癌症晚期的分錯率趨勢圖。普遍來看，五種「迴歸模型選取變數排序之方法」對於此分類的分錯率結果似乎都較「統計量排序」以及「分錯率排序」還要理想，而且可以很顯然的可以發現當組合數到達 13 個以後，Group LASSO 的分錯率的表現明顯的比其他特徵選取方法來得好。

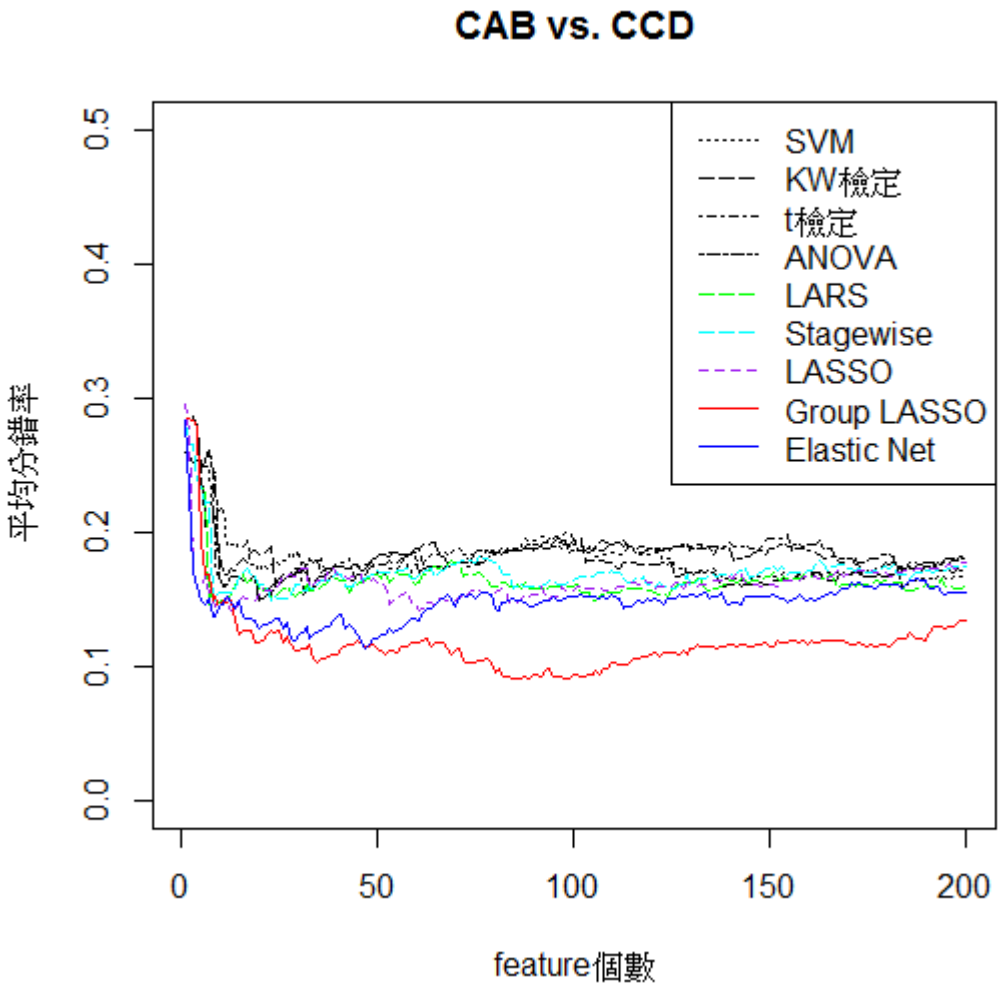

圖 5.6 各特徵選取方法下判別癌症早期與癌症晚期之分錯率趨勢圖

再來我們探討各特徵變數選取方式在六種兩兩分類上的結果。由於黃仁澤(2005)在其文章中提及利用 p 值排序的特徵選取方法，於是在本研究的附錄一中也附上在兩兩分類中利用 p 值排序的分錯率結果。而本研究的目的最終是要探討利用各迴歸模型選取變數的方法是否能比其他排序變數的方法在分錯率的表現上還要好。

表 5.3 中所呈現的是九種特徵選取方法在六個兩兩分類之最小分錯率以及組合數，其中我們最小分錯率的計算單位是百分比，而括號內的數值是為分錯率的標準差。那麼我們由表中可知 Group LASSO 特徵變數選取方法在分類「NO vs. BPH」、「NO vs. CAB」、「BPH vs. CCD」以及「CAB vs. CCD」時效果最好，分別可得到最小的分錯率為 1.82%、1.10%、1.65%和 9.09%，而其組合數分別為 14、76、53 和 47 個特徵變數。而 Elastic Net 是在「NO vs. CCD」以及「BPH vs. CAB」的分類中表現最佳，分別可得最小的分錯率為 1.45%以及 1.68%，而組合數分別為 30 和 37 個特徵變數。

若對於各兩兩分類的判別結果僅比較五種迴歸模型選取變數的方法時，LASSO 的表現在各兩兩分類中之分錯率幾乎都不盡理想。而 Group LASSO 在「NO vs. BPH」、「NO vs. CAB」、「BPH vs. CCD」以及「CAB vs. CCD」這四種兩兩分類之判別結果表現最好，而 Elastic Net 則在「NO vs. CCD」以及「BPH vs. CAB」中之分錯率表現最佳，那 Elastic Net 之所以能夠在上述兩個分類中勝過 Group LASSO 我認為是因為由圖 5.3 和圖 5.4 中我們可以發現此兩個圖形中之紅線(表示 Group LASSO)和藍線(表示 Elastic Net)的分錯率趨勢一路糾纏，有時紅線勝過藍線，有時藍線勝過紅線。不像其他分類，會在到達某一組合數後，表示 Group LASSO 的紅線就會開始明顯的維持在藍線下方。不過我們也可發現這兩種方法最大的共同點就是因為具備群集選擇的功能，故將其運用至變數間具有高度相關的資料中時其變數選取的結果可能較能夠讓人信任。

表 5.3

各特徵選取方法於兩兩分類上之最小分錯率與組合數

方法	NO vs. BPH		NO vs. CAB		NO vs. CCD	
	最小分錯率	組合數	最小分錯率	組合數	最小分錯率	組合數
SVM	2.68(0.59)	11	2.78(1.66)	15	3.79(1.87)	155
KW 檢定	2.65(0.73)	17	3.75(1.28)	167	2.92(1.95)	183
t 檢定	2.29(0.91)	36	2.91(1.93)	42	3.24(1.80)	49
ANOVA	2.36(0.89)	38	2.99(1.91)	41	3.24(1.82)	49
LARS	2.71(1.98)	18	1.99(4.71)	10	2.35(2.44)	54
Stagewise	2.78(1.98)	22	2.16(4.71)	10	2.68(2.46)	57
LASSO	3.19(0.68)	15	2.71(2.85)	24	2.95(1.97)	176
Group LASSO	1.82(0.72)	14	1.10(2.19)	76	2.26(1.67)	34
Elastic Net	2.51(1.67)	29	1.95(2.94)	111	1.45(1.95)	30
方法	BPH vs. CAB		BPH vs. CCD		CAB vs. CCD	
	最小分錯率	組合數	最小分錯率	組合數	最小分錯率	組合數
SVM	3.02(0.97)	17	1.87(2.38)	64	15.43(1.76)	31
KW 檢定	1.85(1.17)	199	2.22(2.08)	121	15.98(1.58)	138
t 檢定	2.55(1.06)	186	2.19(2.27)	148	14.88(1.98)	20
ANOVA	2.64(0.96)	137	2.20(2.29)	153	14.88(1.98)	20
LARS	2.11(3.90)	106	2.37(5.60)	29	14.57(1.7)	7
Stagewise	2.62(4.30)	25	2.99(5.60)	27	14.88(1.67)	9
LASSO	3.06(3.80)	27	2.88(5.60)	71	14.17(1.48)	61
Group LASSO	2.24(0.95)	38	1.65(2.09)	53	9.09(2.76)	98
Elastic Net	1.68(3.46)	37	2.37(4.53)	30	11.34(1.62)	47

註：最小分錯率單位為%、括號內為經由一百組測試資料所得之分錯率標準差。

另外，由於我們之前在第四章第三節中提過 Stagewise 和 LASSO 可以利用修改 LARS 來得其迴歸係數值。所以我們便藉由對於實際資料的分析來驗証它們各自的優缺點。

首先表 5.4 是呈現對於一組訓練資料，各迴歸模型的配適過程所需要花費的時間長度。很明顯的可以發現 LARS 的計算時間最快、Stagewise 的時間最長。雖然 LARS 和 Stagewise 的共同點是一旦在某步驟將某一變數選入模型後就不可能在之後的步驟中又將此變數從模型中去除，但它們不同的地方就在於 Stagewise 每一步驟中之迴歸係數的變動伏度都極小，以致於此方法的計算時間要更久。

表 5.4

對於一組訓練資料 LARS、Stagewise 以及 LASSO 配適迴歸模型過程之時間

分類	LARS	Stagewise	LASSO
NOvs.BPH	0.30	3.97	0.83
NOvs.CAB	0.31	4.28	0.82
NOvs.CCD	0.30	3.51	0.92
BPHvs.CAB	0.96	7.42	1.18
BPHvs.CCD	0.38	4.50	1.14
CABvs.CCD	0.28	3.96	0.84

註：單位時間為秒。

而表 5.5 是呈現在各兩兩分類下 LARS、Stagewise 以及 LASSO 對於每組訓練資料配適迴歸模型之過程中所須要的平均步驟數。可發現 LASSO 的步驟數目會介於 LARS 和 Stagewise 間，而且 LARS 的平均步驟數在高維度資料中會與訓練資料的樣本數相同，此外 Stagewise 的步驟數比其他兩者還要多，所以這也是為什麼此法在計算時間上較花時間的另一項原因。

表 5.5

LARS、Stagewise、LASSO 於各兩兩分類中之每組訓練資料配適迴歸模型過程的平均步驟數

分類	LARS	Stagewise	LASSO
NOvs.BPH	212	1689	555.25
NOvs.CAB	222	1769	574.92
NOvs.CCD	220	1753	587.12
BPHvs.CAB	214	1705	603.97
BPHvs.CCD	212	1689	611.09
CABvs.CCD	222	1769	631.3

表 5.6 是我們利用 NO vs. BPH 分類來分別配適 LARS、Stagewise 以及 LASSO 之第一組訓練資料的迴歸模型，而表中為部份變數在配適迴歸模型過程中之迴歸估計係數的變化情形，其中每一行是為特徵變數，而每一列是表第幾步驟，因此每個細格為某個特徵變數在第幾步驟的迴歸係數估計值。由 LARS 和 Stagewise 中 X_2 和 X_7 之估計值的變化相比可明顯發現 Stagewise 的迴歸係數估計值在每個步驟下有可能不會變動或是只有變動一點點值，然而 LASSO 是有可能將某變數先在某步驟加入模型，然後再刪除，如 X_6，就是先被 LASSO 選入然後到第 272 個步驟時又將其剔除，直到第 291 個步驟又將其選入模型中。

表 5.6

各變數在 LARS、Stagewise 以及 LASSO 中迴歸係數的變化

LARS							
步驟數	X_1	X_2	X_5	X_6	X_7	X_8	X_9
33	0	0	0	0	0	0	0
34	0	-5.43E-05	0	0	0	0	0
35	0	-0.00015	0	0	0	0	0
36	0	-0.00018	0	0	0	0	0
37	0	-0.00028	0	0	0	0	0
38	0	-0.00032	0	0	0	0	0

39	0	-0.00035	0	0	0	0	0
40	0	-0.00095	0	0	0	0	0
41	0	-0.00128	0	0	0	0	0
42	0	-0.00148	0	0	0	0	0
43	0	-0.0022	0	0	-0.00512	0	0
44	0	-0.00222	0	0	-0.00523	0	0
45	0	-0.00224	0	0	-0.00537	0	0
46	0	-0.00236	0	0	-0.00597	0	0
47	0	-0.00249	0	0	-0.00676	0	0
48	0	-0.00278	0	0	-0.00772	0	0
49	0	-0.00279	0	0	-0.00774	0	0
50	0	-0.0029	0	0	-0.00809	0	0
Stagewise							
步驟數	X_1	X_2	X_5	X_6	X_7	X_8	X_9
70	0	0	0	0	0	0	0
71	0	-0.00059	0	0	0	0	0
72	0	-0.00059	0	0	0	0	0
73	0	-0.00091	0	0	0	0	0
74	0	-0.00091	0	0	0	0	0
75	0	-0.00092	0	0	0	0	0
76	0	-0.0011	0	0	0	0	0
77	0	-0.00162	0	0	-0.00226	0	0
78	0	-0.00184	0	0	-0.00337	0	0
79	0	-0.00192	0	0	-0.00383	0	0
80	0	-0.002	0	0	-0.00422	0	0
81	0	-0.00203	0	0	-0.0044	0	0
82	0	-0.00203	0	0	-0.0044	0	0
83	0	-0.00215	0	0	-0.00491	0	0
84	0	-0.0022	0	0	-0.00505	0	0
85	0	-0.00232	0	0	-0.00562	0	0
86	0	-0.00236	0	0	-0.00576	0	5.90E-05
LASSO							
步驟數	X_1	X_2	X_5	X_6	X_7	X_8	X_9
270	0	-0.00153	-0.04349	-0.00029	-0.01804	0	0.005
271	0	-0.00149	-0.04374	-0.00015	-0.01795	0	0.005193
272	0	-0.00145	-0.04399	0	-0.01786	0	0.005397

273	0	-0.00145	-0.04405	0	-0.01784	0	0.005438
274	0	-0.00145	-0.04406	0	-0.01784	0	0.005444
275	0	-0.00142	-0.04422	0	-0.01777	0	0.005618
276	0	-0.00141	-0.04432	0	-0.01773	0	0.005687
277	0	-0.00141	-0.04433	0	-0.01773	0	0.005695
278	0	-0.00133	-0.04486	0	-0.0174	0	0.006262
279	0	-0.00132	-0.04489	0	-0.01738	0	0.006291
280	0	-0.00132	-0.04491	0	-0.01735	1.53E-05	0.006346
281	0	-0.00132	-0.04494	0	-0.01729	3.99E-05	0.00643
282	0	-0.00132	-0.04504	0	-0.01711	0.000127	0.006736
283	0	-0.00131	-0.04509	0	-0.017	0.000182	0.006908
284	0	-0.0013	-0.0452	0	-0.0168	0.000288	0.007245
285	0	-0.0013	-0.04524	0	-0.01666	0.000356	0.007378
286	0	-0.00129	-0.04527	0	-0.01655	0.00041	0.00749
287	0	-0.00129	-0.04528	0	-0.01653	0.00042	0.007511
288	0	-0.00129	-0.04527	0	-0.01646	0.000456	0.007583
289	0	-0.0013	-0.04524	0	-0.01636	0.000498	0.007649
290	0	-0.00131	-0.04521	0	-0.01625	0.000545	0.007729
291	0	-0.00143	-0.04511	-0.00069	-0.01557	0.000809	0.008165

另外，我們也藉由觀察單一變數迴歸係數估計值的變化，以便更了解 LARS、Stagewise 以及 LASSO 的迴歸係數變化趨勢。因此圖 5.7 中為呈現在 NO vs. BPH 分類中第一組訓練資料裡 X_6 之迴歸係數估計值的變化。可以發現當 LARS 之係數有向零變化的趨勢時，Stagewsie 會將其值儘可能的維持在負號的部分，並使其單調的發展；而 LASSO 也得將其值維持在負號的部分，因為必須符合凸函數最佳化的限制，且與 Stagewise 的穩定度相較下其值的變化會上下起伏不定，不過最終都不可能變為正號。而且，LASSO 和 Stagewise 兩種選取模型的方法在一開始時，其迴歸係數的變化可能極為相似，但經過越多步驟之後 Stagewise 的迴歸係數的變化會較 LASSO 平滑許多，也因為 LASSO 的迴歸係數的變化會呈現大幅度的波動而使得 LASSO 較 Stagewise 易產生過度配適的現象而使得其方法的預測結果較差。

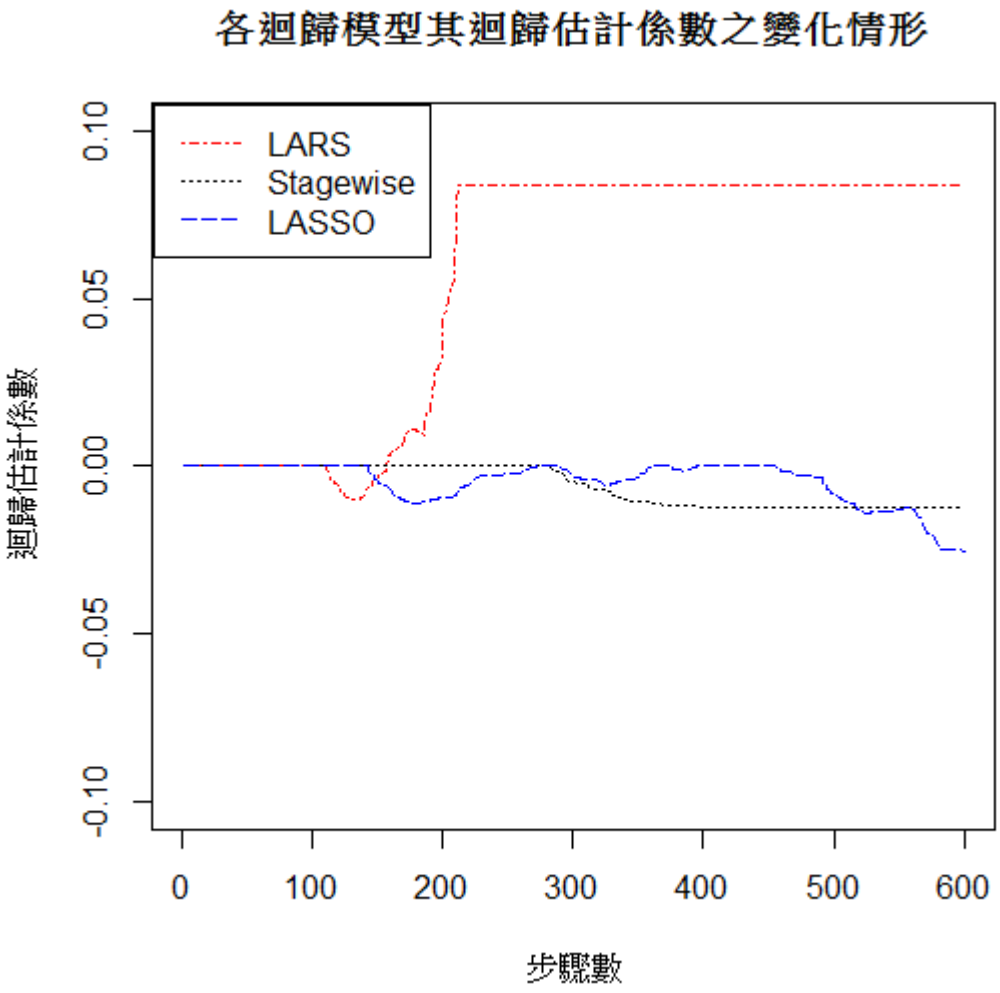

圖 5.7 LARS、Stagewise 以及 LASSO 在 NO vs. BPH 中之 X_6 的迴歸係數估計值之變化

由於六個兩兩分類中，Group LASSO 在「NO vs. BPH」、「NO vs. CAB」、「BPH vs. CCD」以及「CAB vs. CCD」中的分錯率表現最優，雖然在「NO vs. CCD」以及「BPH vs. CAB」中 Elastic Net 的分錯率較好，不過 Group LASSO 的分錯率也不至於太差，於是我們就將 Group LASSO 特徵選取方法之結果與 Adam 等人(2002)挑選的 AUC>=0.62 的那 124 個特徵變數放入決策樹中來建立分類模型所得之分錯率來進行比較。由表 5. 7 中可以發現不論為哪一種兩兩分類 Group LASSO 的方法所得的分錯率都較理想，尤其在分類「BPH vs. CABCCD」時，分錯率更低到 0.87%。而 AUC 決策樹的方法雖在「NO vs. BPH」的分錯率表現最佳但其分錯率最低仍然還有 4%，在「CAB vs. CCD」分類中

之分錯率更高達 17%，我們認為其原因是因不論對於哪種兩兩分類的判別都是用共同的 9 個特徵變數來分類而使得分錯率都較高，但 Group LASSO 則會根據不同的分類而導致模型的選取也會改變，因此其結果皆比 AUC 決策樹來得好。

表 5.7

各兩兩分類中 Elastic Net 以及 AUC 決策樹的分錯率結果

方法	NO vs. BPH		BPH vs. CAB		CAB vs. CCD	
	最小分錯率	組合數	最小分錯率	組合數	最小分錯率	組合數
Group LASSO	1.82	14	2.24	38	9.09	98
AUC 決策樹	4	9	14	9	17	9
方法	NO vs. CABCCD		BPH vs. CABCCD		NOBPH vs. CABCCD	
	最小分錯率	組合數	最小分錯率	組合數	最小分錯率	組合數
Group LASSO	2.08	23	0.87	51	2.18	44
AUC 決策樹	11	9	13	9	10	9

註：最小分錯率單位為%。

另外，我們也將 Group LASSO 的結果與陳詩佳(2007)提出先用判定係數萃取特徵變數後，再利用 Meta-Learning 的概念將 SVM 串聯起來的方法來判別其分類結果相比，表 5.8 即為兩種方法之分錯率結果，可以發現 Group LASSO 在「NO vs. CCD」、「BPH vs. CCD」以及「CAB vs. CCD」分類中之分錯率與判定係數 SVM 串聯法的分錯率相較下，Group LASSO 可以使其分錯率至少降低約 1%左右，其中「CAB vs. CCD」更能使分錯率降低約 3%，而對於「NO vs. BPH」、「NO vs. CAB」以及「BPH vs. CAB」分類的分錯率，Group LASSO 也能夠使其分錯率有些微的下降(約 0.5%)。因此 Group LASSO 特徵選取方法確實能夠得到較小的分錯率，雖然為了能夠降低分錯率可能須要較多的變數組合數，不過由上百個特徵變數中最後只選取到不超過 100 個特徵變數即可達成分錯率最小化的目的。

表 5.8

Elastic Net 特徵選取方法與判定係數萃取 SVM 串聯法之分錯率比較

方法	NO vs. BPH		NO vs. CAB		NO vs. CCD	
	最小分錯率	組合數	最小分錯率	組合數	最小分錯率	組合數
Group LASSO	1.82(0.72)	14	1.10(2.19)	76	2.26(1.67)	34
判定係數 SVM 串聯法	2.34(2.36)	7	1.77(1.62)	20	3.29(2.25)	31
方法	BPH vs. CAB		BPH vs. CCD		CAB vs. CCD	
	最小分錯率	組合數	最小分錯率	組合數	最小分錯率	組合數
Group LASSO	2.24(0.95)	38	1.65(2.09)	53	9.09(2.76)	98
判定係數 SVM 串聯法	2.72(2.01)	14	2.48(2.78)	31	12.14(3.91)	16

註：最小分錯率單位為%、括號內為經由一百組測試資料所得之分錯率標準差。

第三節 探討四分類之分錯率結果

在四分類的分析中由於 t 檢定只能用在檢定兩母體平均是否有差異，故我們無法將其應用至四分類的資料中，故在此部份探討四分類時，我們就對 SVM、KW 檢定、ANOVA、LARS、Stagewise、LASSO、Group LASSO 以及 Elastic Net 這八種特徵選取方法之判別結果作討論。另外，因為 Group LASSO 以及 Elastic Net 在面對應變數 y 有四個類別時是無法作分析的，故我們的作法是先利用這兩個迴歸方法在「正常 vs.良性腫瘤」、「正常 vs.癌症早期」、「正常 vs.癌症晚期」、「良性腫瘤 vs.癌症早期」「良性腫瘤 vs.癌症晚期」以及「癌症早期 vs.癌症晚期」中對於每個特徵變數所產生的等級平均值再作一次平均，並藉由再一次的平均來當作四分類中排序特徵變數的依據。而圖 5.8 為八種特徵選取方法下所產生的前兩百名特徵變數之分錯率趨勢圖，觀察出，迴歸方法中除了 Group LASSO 外，由 LARS、Stagewise、LASSO 以及 Elastic Net 對於前兩百名特徵變數的整體分錯率似乎都比 SVM、KW 檢定以及 ANOVA 方法的分錯率還要高，然而 Group LASSO 的整體分錯率又有更優於 SVM、KW 檢定以及 ANOVA 方法的跡象。

再由表 5.13 中得知，更能確定 LARS、Stagewise、LASSO 以及 Elastic Net 的最小分錯率 SVM、KW 檢定以及 ANOVA 方法還要差，此外，Group LASSO 在判別四分類時的最小分錯率為 10.72%較其他方法佳，其次是 SVM 方法 11.78%，再看其組合數的話 Group LASSO 需用至 107 個變數達成最小分錯率，但 SVM 只需用到 59 個，其實由圖 5.9 中觀察 Group LASSO 的分錯率趨勢會發現大約在組合數為 51 個時即可得到約 11.63%的分錯率了。因為我們的研究仍著重於追求最小分錯率，故在四分類的判別中我們仍認為 Group LASSO 為一個較佳的特徵選取方法。

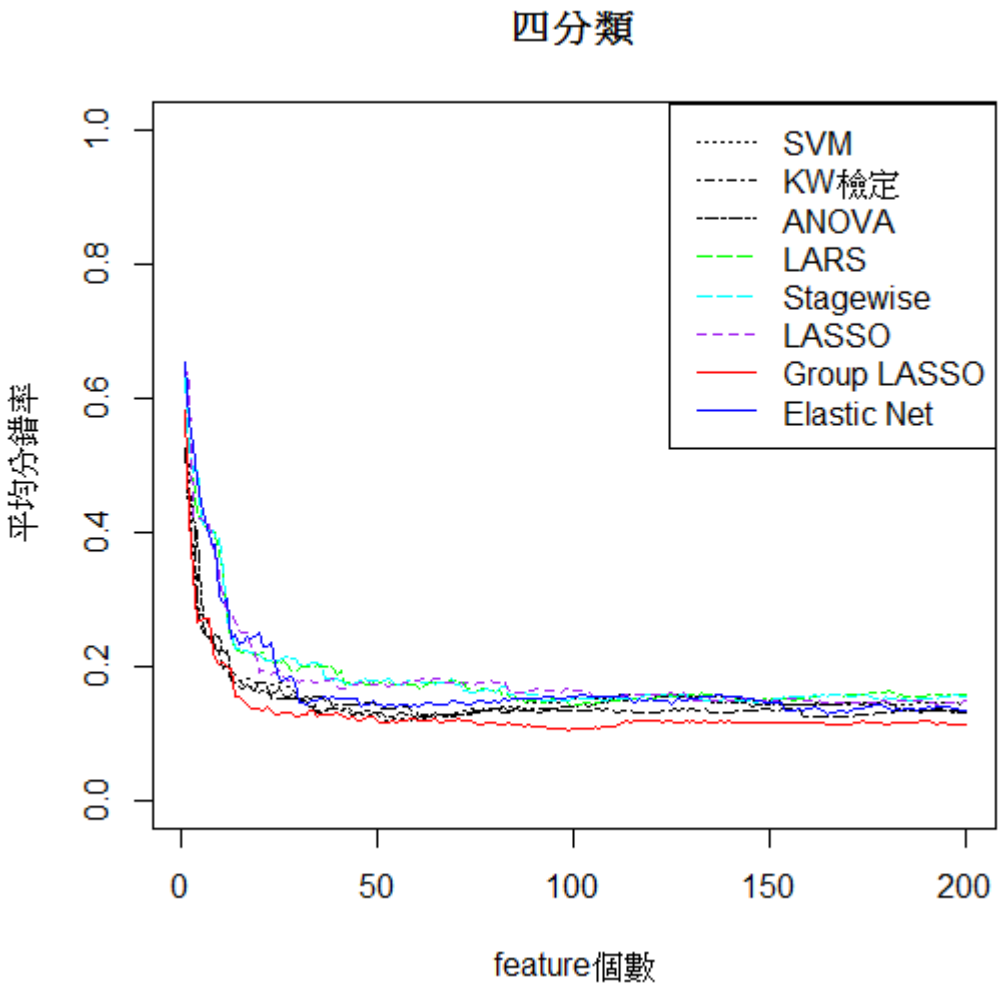

圖 5.9 各特徵選取方法下四分類之分錯率趨勢圖

表 5.9

各特徵選取方法於四分類上之最小分錯率與組合數

方法	四分類	
	最小分錯率	組合數
SVM	11.86(4.63)	59
KW 檢定	12.45(4.72)	166
ANOVA	12.44(4.83)	60
LARS	14.23(6.76)	100
Stagewise	14.78(6.95)	147
LASSO	14.36(6.92)	143
Group LASSO	10.59(4.93)	99
Elastic Net	12.94(7.43)	165

註：最小分錯率單位為%、括號內為經由一百組測試資料所得之分錯率標準差。

第六章 分析結果討論與建議

本研究是利用經過事前處理的攝護腺癌之蛋白質質譜資料來作分析，而此種資料通常是受測者的資料少於特徵變數個數的情形，故為一個高維度資料。所以若能設法從這類資料中來篩選到少數重要的特徵變數個數且不失去其判斷正確性的話，即為一個良好的特徵變數選取方法。因此本研究嘗試比較九種選取特徵變數的方法在六個兩兩分類和四分類上的分錯率表現，而九種選取特徵變數方法分別為 SVM 的「分錯率排序」、還有以 t 檢定、ANOVA F 檢定以及 KW 檢定的「統計量排序」、迴歸方法 LARS、Stagewise、LASSO、Group LASSO 以及 Elastic Net 的「選入迴歸模型之順序排序」，然後分別對各方法的前兩百名特徵變數依序入代入 SVM 中得出各種變數組合數下的分錯率，藉此產生各方法的分錯率趨勢圖以及各方法之最小分錯率即其對應之組合數。

由分析結果我們發現不論於哪一種分類，在運算速度的時間上 LARS 皆比 Stagewise 和 LASSO 還要快速，原因是因 LARS 一旦在某步驟將某個變數選入活動集合後就不會再將此變數從活動集合中移除，但 LASSO 即有可能會先將某變數選入，然後又移除它一段期間後又再次選入的情形發生，因此 LASSO 這樣的選取模型過程一定會比 LARS 更花時間。此外，LASSO 不論是在兩兩分類還是四分類中之判別結果的表現都是五種迴歸方法中最不盡理想的。我們認為可能是因資料中遇有兩變數相關程度很高的情況發生時 LASSO 就會”任意”將其中一個變數選入模型中，因此並不能確保先被選入模型的這個變數是兩個變數中與應變數較有關係而影響其預測準確度。而 Zou 和 Hastie(2003)年提出的 Elastic Net，除了包含 LASSO 的限制式外又加入迴歸係數平方絕對值加總的第二條限制式，因而使此法具有群集選取(grouped selection)的能力來選取變數。因此這種方法在高維度資料時更能看出其預測的結果確實較 LARS、Stagewise 和 LASSO 理想。另外，Yuan 和 Lin(2007)也察覺到 LASSO 在高維度資料中變數選取的問題因而發展出 Group LASSO。最後由本研究的分析中發現，同樣具有群集選取能力的兩種迴歸方法—Elastic Net 和 Group LASSO，在六種兩兩分類中，Group LASSO 於「正常和良性腫瘤」、「正常和癌症早期」」、「良性腫瘤和癌症晚期」以及「癌症早期和晚期」分類

的分錯率較 Elastic Net 好，而 Elastic Net 則是在「正常和癌症晚期」以及「良性腫瘤和癌症早期」優於 Group LASSO，於是我們再去觀察此兩種方法在這兩個兩兩分類的分錯率趨勢圖(圖 5.3 和圖 5.4)，可發現兩種方法的趨勢很常有交錯的現象不像其餘四個兩兩分類的趨勢圖(圖 5.1、圖 5.2、圖 5.5 和圖 5.6)兩種方法的走向很明顯的分開。而最後我們會認為 Group LASSO 還是較佳的原因就是考慮對於六種兩兩分類的平均表現以及其方法穩定性的緣故。因此之後我們又將 Group LASSO 的結果與 Adam 等人(2002)以及陳詩佳(2007)的方法比較，而 Group LASSO 也確實能夠得到較小的分錯率。

而對於四分類的判別結果 Group LASSO 比起其他八種方法也得到最小的分錯率(10.59%)，其次是 SVM(11.86%)，不過不難注意到 SVM 的所須變數組合數大約只需 59 個，而 Group LASSO 則需 99 個，但其實觀察其分錯率趨勢圖(圖 5.7)的話，Group LASSO 在組合數約為 51 個時其分錯率(11.63%)就優於 SVM 了，因此在我們追求極小分錯率的目標下，我們仍認為 Group LASSO 的結果很不錯。

然而本研究仍然有許多可以改進的地方，例如設定 Elastic Net 的參數時(亦即"min.lambda"和"relax.lambda")可能較為馬虎，而使得結果未必是最好的也不一定，故未來或許可以再找尋其他更有效率的演算法來計算 Elastic Net 之迴歸係數的路徑。此外，未來或許可以考慮將具有群集選取能力的迴歸模型或是能夠解決高維度資料變數間不獨立的迴歸模型選取方法一併作探討。

參考文獻

一. 中文部分

陳詩佳 (2007)，「使用 Meta-Learning 在蛋白質質譜資料特徵選取之探討」，國立政治大學統計系研究所碩士論文。

黃仁澤 (2005)，「對於高維度資料進行特徵選取–應用於分類蛋白質質譜儀資料」，國立政治大學統計系研究所碩士論文。

蒲永孝和黃昌淵，「認識男人的殺手–前列腺癌」，正中書局，1997 年。

潘荔錞、蔡志彥和簡志青，「蛋白質體學在臨床醫學之應用」，化工資訊與商情月刊第 3 期，2003 年 9 月號。

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 6 月 25 日。

簡邦平，「攝護腺健康新知」，原水文化，2006 年。

二. 英文部分

Adam, B. L., Qu, Y., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J.,Schellhammer, P. F., Yasui, Y., Feng, Z. and Wright, G. L. Jr. (2002), "Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men", *Cancer Research* 62(13) 3609-3614.

Degroeve, S., Baets, B. D.,Peer, Y. V. and Rouze, P. (2002), "Feature Subset Selection for Splice Site Prediction", *Bioinformatics* 18(2) 75-83.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani R. (2003), "Least Angle Regression", *Annals of Statistics* 32(2) 407-499.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment", *Journal of the American Statistical Association* 96(456) 1151-1160.

Fox, R. J. and Dimmic, M. W. (2006), ”A Two-Sample Bayesian t-test for Microarray Data”, *BMC Bioinformatics* 7:126.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), “A Note on the Group LASSO and a Sparse Group LASSO”.

Guyon, I., Westion, J. and Barnhill, S. (2002), “Gene Selection for Cancer Classification Using Support Vector Machines”, *Barnhill Bioinformatics* 46 389-422.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), ” The Elements of Statistical Learning. Springer”.

Hastie, T., Taylor, J., Tibshirani, R. and Walther, G. (2007), “Forward Stagewise Regression and the Monotone Lasso”, *Electronic Journal of Statistics* 1(1) 1-29.

Issaq, H. L., Veenstra, T. D., Conrads, T. P. and Felschow, D. (2002), “The SELDI-TOF MS Approach to Proteomics: Protein Profiling and Biomarker Identification”, *Biochemical* and *Biophysical Research Communications* 587-592.

Jiang, H., Deng, Y., Chen, H. S., Tao, L., Sha, Q., Chen, J., Tsai, C. J. and Zhang, S. (2004), ”Joint Analysis of Two Microarray Gene-Expression Data Sets to Select Lung Adenocarcinoma Marker Genes”, *BMC Bioinformatics* 5:81.

Leng, C., Lin, Y. and Wahba, G. (2006), “A Note on the Lasso and Related Procedures in Model Selection”, *Statistica Sinica* 16 1273-1284.

Ma, S. and Huang, J. (2005), ”Regularized ROC Method for Disease Classification and Biomarker Selection with Microarray Data”, *Bioinformatics* 21(24) 4356-4362.

Meier, L., Geer, S. V. D. and Buhlmann, P. (2008), “The Group LASSO for Logistic Regression”, *Journal of the Royal Statistical Society* 70(1) 53-71.

Park, M. Y. and Hastie, T. (2006), “L1 Regularization Path Algorithm for Generalized Linear Models”, *Journal of the Royal Statistical Society* 659-677.

Somorjai, R. L., Dolenko, B. and Baumgartner, R. (2003), ”Class Prediction and Discovery Using Gene Microarray and Proteomics Mass Spectroscopy Data: curses, caveats,

cautions", *Bioinformatics* 19(12) 1484-1491.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society* 58(1) 267-288.

West, M. (2003), "Bayesian Factor Regression Models in the Large p, Small n Paradigm", *Bayesian Statistics*.

Weston, J., Elisseeff, A. and Scholkopf, B. (2003), "Use of the Zero-Norm with Linear Models and Kernel Methods", *BIOwulf Technologies* 3 1439-1461.

Yuan, M. and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables", *Journal of the Royal Statistical Society* 68 49-67.

Zou, H. and Hastie, T. (2004), "Regularization and Variable Selection via the Elastic Net", *Journal of the Roual Statistical Society* 67 301-320.

附錄一

方法	NO vs. BPH		NO vs. CAB		NO vs. CCD	
	最小分錯率	組合數	最小分錯率	組合數	最小分錯率	組合數
KW 檢定	2.67(0.77)	10	3.40(1.33)	86	3.09(2.03)	196
t 檢定	2.32(0.91)	35	2.90(1.77)	122	2.91(1.73)	46
ANOVA	2.32(0.86)	35	2.91(1.77)	122	2.92(1.72)	47
方法	BPH vs. CAB		BPH vs. CCD		CAB vs. CCD	
	最小分錯率	組合數	最小分錯率	組合數	最小分錯率	組合數
KW 檢定	1.63(1.2)	192	1.97(2.03)	126	15.27(1.64)	40
t 檢定	2.20(1.12)	188	2.14(2.17)	133	15.48(2.01)	25
ANOVA	2.53(1.03)	136	2.12(2.17)	131	15.48(2.06)	25