行政院國家科學委員會專題研究計畫　成果報告

病例分類與資料縮減研究-應用蛋白質資料庫檢測癌症

(2/2)

計畫類別：整合型計畫
計畫編號：NSC94-2118-M-004-001-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：國立政治大學統計學系

計畫主持人：余清祥
共同主持人：黃貞瑛

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 12 月 15 日

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※ ※
※ 病例分類與資料縮減研究–應用蛋白質資料庫檢測癌症 ※
※ ※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫　■整合型計畫
計畫編號：NSC 94-2118-M-004-001
執行期間： 93 年 8 月 1 日至 94 年 7 月 31 日

計畫主持人：余清祥
共同主持人：
計畫參與人員：

成果報告類型(依經費核定清單規定繳交)：■精簡報告

執行單位：國立政治大學統計系

中 華 民 國 94 年 10 月 31 日

行政院國家科學委員會專題研究計畫成果報告

病例分類與資料縮減研究–應用蛋白質資料庫檢測癌症

Disease Classification and Data Reduction: Application to Cancer Detection Based on Proteomic Data

計畫編號：NSC 94-2118-M-004-001
執行期限：93 年 8 月 1 日至 94 年 7 月 31 日
主持人：余清祥　　執行單位：國立政治大學統計系

一、中文摘要

在資料庫內容龐大紛雜的現代社會中，時效性往往是最重要的考量因素，以期在最短的時間內獲取近似、可接受的解答，為後續發展提供即時的建議。例如：醫師根據癌症病患的檢體報告，儘快判斷病患是否需要立即實施手術、化學治療，或甚至不需要任何治療、但須持續追蹤觀察。因為資料量的縮減通常代表較低的分析時間與成本，縮減資料自然成為講求時效及近似解答的最佳選擇之一，其中常見的方法包括直方圖(Histogram)、歧異值分解(Singular Value Decomposition)、索引樹(Index Tree)、抽樣、小波(Wavelet)等等。

本計畫將使用攝護腺病人的蛋白質體資料庫(Proteomic data)，其中病例個數約 300 人、變數個數卻接近 5 萬個，以正確的病例分類為目標，比較幾種常見資料縮減方法的優劣。本計畫將預計分為三年進行：第一年使用人工篩選（錯誤較少、變數較少）過的蛋白質質譜儀數據，考慮以 Support Vector Machine (SVM)、類神經網路、Classification and Regression Tree (CART)、羅吉士迴歸四種常見的分類方法，尋求在二元、分類標準下的最佳分類方法；第二年使用變數個數約 5 萬個的原始資料，以二元分類為目標，配合之前較佳的分類方法，尋求可篩選出最多訊息的資料縮減方法；第三年則嘗試合併每位病人兩份檢體結果，以多元分 類為目標，獲得正確的病例診斷。

關鍵詞：資料縮減、分類、病例診斷、模擬

Abstract

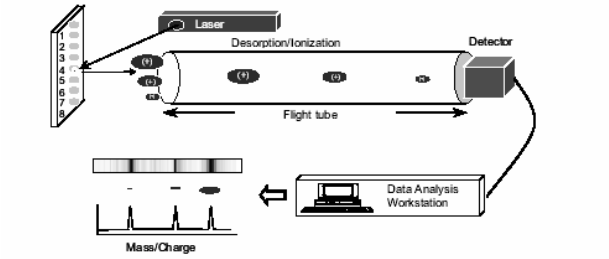
It is often needed to get quick approximate answers from large databases (i.e., data reduction), since obtaining answers quickly is important and it is acceptable to sacrifice the accuracy of the answer for speed. The reduction process is important in the exploratory data analysis, particularly when interactive response times are critical. For example, doctors need to decide from the medical exam if cancer patients need surgeries, chemical therapies, or thorough physical exam. Popular data reduction methods include histogram, singular value decomposition (SVD), index tree, sampling, and wavelet.

We will use data from prostate cancer patients (Proteomic data), which include records of about 300 patients and almost 50,000 variables. Our goal is to include the data reduction methods to minimize the classification error. The project will be divided into three years. The focus of the first year is to explore the performance of frequently used classification methods, such as support vector machine (SVM), neural network, classification and regression tree, and logistic regression. We shall use the pre-processed data with only 779 variables and possible errors corrected manually, and the goal of the first year is binary classification. Data reduction methods will be considered in the second year and the raw data (about 48,000 variables and errors not corrected) will be used as well. The focus will be on the diagnosis of patients and we shall consider methods of combining samples from the same patient.

Keywords: Data reduction, Classification, Diagnosis, Simulation

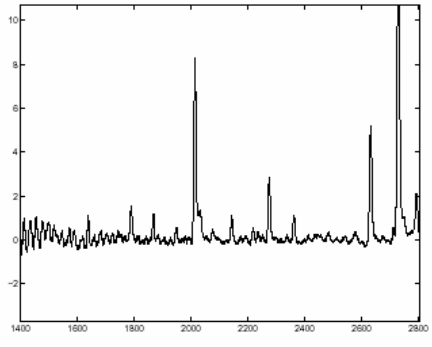
二、緣由與目的

本計畫考量的病歷診斷使用蛋白質資料庫，乃藉由質譜技術(Mass Spectrometry)之表面加強雷射脫附游離－飛行時間(Surface Enhanced Laser Desorption / Ionization–Time of Flight；SELDI-TOF)取得蛋白質體資料（詳見圖一），近來因質譜技術之高速提昇，預期有大量類似之資料庫產生，分析此類資料庫的需求將更為迫切。本研究使用的資料之產生是將病人之血清置於 Ciphergen Biosystems 公司之(SELDI-TOF)質譜儀中所獲得，目標在於分類攝護腺病人是否罹患癌症。



圖一、SELDI-TOF 圖

每一病人所產生之蛋白質數據圖表(Protein profiles)為一質量頻譜(Spectrum)，每一質譜約有 48,000 個維度，即其 x 軸為質量/電荷，其範圍為 1 至 200,000 daltons，而 y 軸為離子含量（詳見圖二），分析時一般將頻率較高者視為較為特殊，攜帶較多的訊息。每一病人有兩組實驗數據（同一血清，分別作兩次質譜分析），因為兩組資料的結果差異性不小，過去曾有幾個研究（例如：Adam et al., 2002；Yasui et al., 2003；Qu et al. 2003）嘗試以此資料進行病例分類，分類錯誤率都不小（最佳的分析結果也有超過 10%以上的錯誤率）。



圖二、質譜頻率圖

三、文獻探討及模型介紹

由於本研究目的在於增加病例診斷的準確率，故選擇有效的分類方法亦是重要的課題之一，下文將對支持向量機(Support Vector Machine, SVM)、類神經網路(Artificial Neural Network, ANN)和分類迴歸樹(Classification and Regression Tree, CART)來進行說明和比較。

1. 支持向量機

支持向量機為西元 1995 年由 Vapnik 及其研究夥伴所提出，其能展現有效的分類和迴歸估計。支持向量機的主要目標就是找到一超平面，使得兩類的分類最正確，同時使兩類資料距離分類面最遠，而其重要優點是可處理線性不可分。

2. 類神經網路

類神經網路是目前被廣泛應用的方法，其優點是可建構非線性模式，模型準確度高，亦不像迴歸分析有自由度之限制，且彌補其他模式須設立許多假設條件的困擾。類神經網路是利用電腦來模仿生物神經網路的處理系統，為一計算系統，使用大量簡單的類神經元(artificial neuron)，又稱為處理單元(processing unit)或節點(node)，來模仿生物神經網路的能力。

3. 分類迴歸樹

分類迴歸樹對於進行分類和預測結果是一項不錯的選擇，其目標是產生易了解又具解釋能力的結果。分類迴歸樹為由 Breiman、Friedman、Olshen 和 Stone (1984)所提出的一種樹型建構(Tree-building)技術，其結合反應變數發展一連串問題，讓解釋變數來回答對或錯，每個問題即詢問解釋變數是否滿足給定的條件，而回應的答案會經由樹狀圖之分枝帶領使用者直至觀測值已被分類完成，在每個節點上，分析過程會找出最佳問題來幫助使用者建立最佳決策，而至每個數枝之末端，每個觀測值之分類決策即被建立。

當資料變數的數目過多，其中若是彼此間存在高相關性，則會使形成的模式對於應變數的估算中，反映的訊息有所重疊；即使變數間相關性不高，變數多也會增加計算上的複雜及難度。故本研究朝向維度縮減方法

進行，而維度縮減的優點，一是可藉由降低資料維度來大幅減少計算量，二是可將資料投影到較低維度的子空間(subspace)，可以幫助分析者容易想像或形象化所分析的資料。

本研究採取兩種資料縮減方法：主成份分析和主成份分析網路。主成份分析和主成份分析網路以特徵抽取(feature extraction)進行線性和非線性維度縮減方法，前者為統計多變量方法，而後者為類神經網路的變形，皆是利用資料轉換來降低維度。

四、實證分析

1. 分類方法比較

針對訓練資料之平均分錯率(training error rate)，支持向量機表現佳，對於兩筆資料其分錯率皆為 0，而類神經網路針對人工處理資料亦皆為 0。針對測試資料之平均分錯率(testing error rate)，其人工處理資料兩兩分類中，三者表現差異不大，而分錯率最高的組別為癌初對癌末(CAB/CCD)，多重分類以分類迴歸樹表現略差；針對原始資料，以分類迴歸樹表現略差，其兩兩分類中，分錯率最高的組別為良腫對癌末(BPH/CCD)。

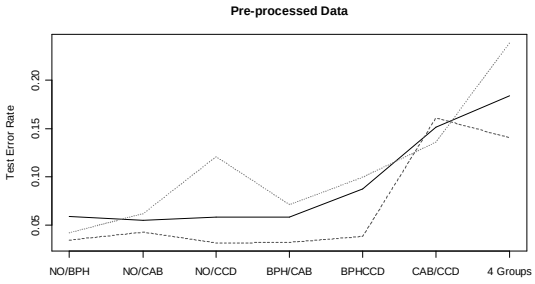


圖 1：不同分類方法錯誤率比較

因主成份分析網路模擬時間過長，所以此僅模擬 50 次。人工處理資料之主成份分析網路和主成份分析之分錯率差異較小，原始資料之主成份分析網路和主成份分析之分錯率差異較大，除人工處理資料正常對良腫之組別和原始資料正常對癌末之組別，兩者平均線有交叉傾向外，整體看來以主成份分析，隨著解釋變數個數增加，分錯率下降速率較快，且其分錯率較主成份分析網路低，故表現較佳。

2. 重疊法比較

本研究為探討重疊法之效果，針對人工處理資料擷取主成份 25 和 50 個，隱藏層節點個數 25、50 和 75 個來配對以進行重疊法；針對原始資料擷取主成份 25 和 50 個，隱藏層節點個數 50、100 和 150 個進行重疊法，因主成份分析之分錯率較小，故採取較少的主成份個數以進行分析。

為比較其效果，本由就將其和僅進行主成份分析一同比較，其分析圖形如下圖：

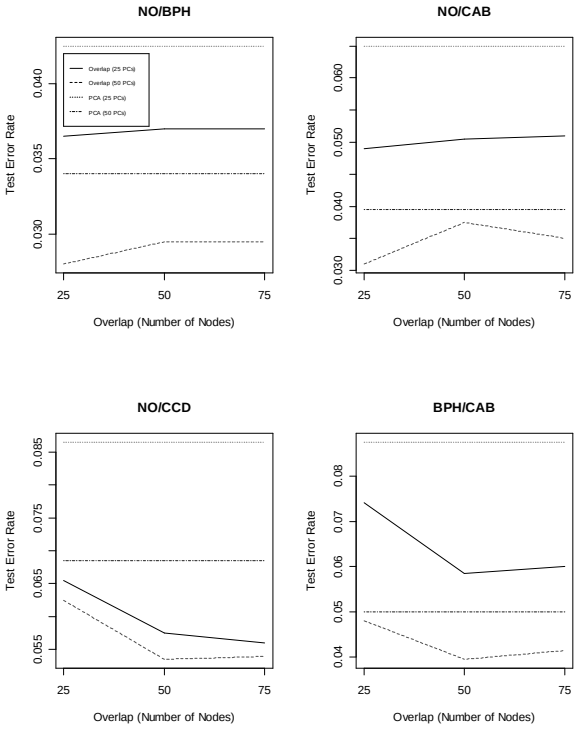


圖 2：重疊法和主成份分析之平均錯誤率

由圖形可知，重疊法對於人工處理資料皆有明顯改善，其改善後分錯率全部小於僅使用主成份之分錯率，且皆達到本研究目標，改善效果非常顯著，且正常對良腫、正常對癌初、良腫對癌初和良腫對癌末之組合，其重疊法之最低分錯率低於對主成份分析取特徵值大於 1 主成份個數之分錯率；而正常對良腫、正常對癌初和癌初對癌末低於未行維度縮減方法之分錯率，可見人工處理資料雖已進行消除雜訊動作，但其並未完全消除。

重疊法對原始資料改善效果較不顯著，但使用重疊法後，主成份和節點個數組合對於大部分病例類別分類，至少有一組小於僅使用主成份分析之分錯率，且正常對癌末類別之分錯率有低於取特徵值大於 1 者之主成份分析分錯率。

五、計劃結果自評

本論文研究的攝護腺癌症蛋白質資料庫，是經由表面強化雷射解吸電離飛行質譜技術的血清蛋白質強度資料，藉此資料判斷受測者是否罹患癌症(即疾病診斷，或分類問題)。因為蛋白質資料的變數較多，例如原始資料包含 48000 個區間(或變數)、人工處理資料也有 779 個區間，遠多於病例個數，多數傳統的分類方法無法直接應用，即使可直接套用在計算上難度也較高。因此本文主旨在不犧牲分類正確性的原則下，尋求有效的維度縮減方法，以去除不必要的雜訊。研究流程為先找出表現較佳的分類方法，再探討有效的維度縮減方法，

根據分類方法模擬結果可知，支持向量機、類神經網路和分類迴歸樹三者在人工處理資料與兩兩分類的表現類似，但在原始資料與兩兩分類上以分類迴歸樹表現較差，在多重分類(無論是原始或人工處理資料)亦是分類迴歸樹表現較差。推測其原因可能是分類迴歸樹原理採用二分樹法，對於較複雜資料會造成分枝過多而難以管理，故無法對本研究資料進行有效的分類。支持向量機原理起源雖然也是二分法，分類正確性卻勝過分類迴歸樹，整體的分類效果和類神經網路差不多；以計算時間而言，支持向量機需時較類神經網路短，或許這是近年支持向量機受歡迎的原因之一，然而支持向量機多重分類原理尚在發展中，故成效和類神經網路相比，亦無孰優孰劣。本研究建議分類方法可選用支持向量機和類神經網路。

在維度縮減的探討方面，本研究僅考慮主成份分析、主成份分析網路兩種方法，整體而言，維度縮減後的分類結果大致與使用全部資料接近。其中主成份分析對於原始資料其效果和未行維度縮減方法差不多，但其和類神經網路分類相結合，對原始資料卻有效去除雜訊和降低分錯率。主成份分析網路表現不如主成份分析理想，經維度縮減後整體分錯率亦有提升，可能原因是模擬時間過長造成的模擬次數太少，或節點數仍不夠多。

重疊法之應用在人工處理資料表現佳，而對原始資料重疊法效果不大，其原因可能是因為資料本身的複雜性，不像人工處理資料已經處理過雜訊。也有可能因主成份分析網路計算時間過久，在此只列出 50 次模擬的結果，由於模擬次數太少，所得結果的變異程度過大，或是應採用質量 2000-4000 dalton 以外的範圍。另一可能原因是原始資料之主成份分析和主成份分析網路之分錯率相差太大，因為之前對人工處理資料的分析中，發現當主成份分析主成份分析網路的分錯率差異較大時，重疊法的分錯率也偏高，將兩者差異拉近時，重疊法的效果才彰顯。

六、參考文獻

[1]Adam, B. et al. (2002), Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men. *Cancer Research* 62, 3609-3614.

[2] Ball, G. et al. (2002), An Integrated Approach Utilizing Artificial Neural Networks and SELDI Mass Spectromery for the Classification of Human Tumours and Rapid Idenfication of Potential Biomarkers. *Bioinformatics* 18(3), 395-404.

[3] Barbara, D. et al. (1997), The New Jersey Data Reduction Report. Bulletin of the Technical Committee on Data Engineering, vol. 20(4), December 1997, IEEE Computer Society.

[4] Carreira-Perpinan, M. A. (1997), A Review of Dimension Reduction Techniques. Technical Report CS-96-09, Department of Computer Science, University of Scheffield.

[5] Cazares, L. H. et al. (2002), Normal, Benign, Preneoplastic, and Malignant Prostate Cells Have Distinct Protein Expression Profiles Resolved by Surface Enhanced Laser Desorption/Ionization Mass Spectrometry, *Clinical Cancer Research* 8, 2541-2552.

[6] Petricoin III, E. F. et al. (2002), Use of Proteomic Patterns in Serum to Identify Ovarian Cancer. *The Lancet* 359, 572-577.

[7] Hastie, T. et al. (2001), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.