

東海大學資訊工程研究所

碩士論文

Department of Computer Science

Tunghai University

指導教授：石志雄博士

Advisor : Chihhsiong Shih, Ph. D.

以基因演算法、偏最小平方與 Apriori 演算法為基礎之階層式疾病轉化因子探討

-以中風轉化因子探討為例

Discussion on hierarchical disease transformation factors based on gene algorithm, partial least squares and Apriori algorithm--A case study of stroke transformation factor

研究生：張佑瑋

Gradute Student : You-Wei Chang

中華民國一〇八年一月

Jan, 2019

東海大學碩士學位論文考試審定書

東海大學資訊工程學系 研究所

研究生 張佑瑋 所提之論文

以基因演算法、偏最小平方與 Apriori 演算

法為基礎之階層式疾病轉化因子探討-以中

風轉化因子探討為例

經本委員會審查，符合碩士學位論文標準。

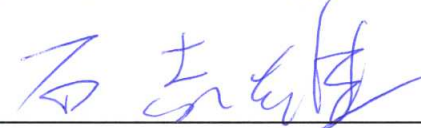
學位考試委員會

召集人

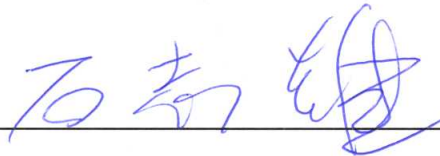


簽章

委員



指導教授



簽章

中華民國 108 年 1 月 8 日

摘要

中風是全台十大死因之一，也是造成人們失能的主要原因，且只有百分之十五的人可以完全康復，因此本研究欲利用全民健保資料庫來整理與分析，希望能找尋一些與中風相關的疾病模型，和相關之疾病關係圖。

本研究使用了 GA-PLS 的方法，提高準確度，並增加了關聯規則和貝氏網路，用常在商業中探討商品之間的關聯規則，替換成疾病，探討疾病之間的關係，之後再建構有向無循環圖，套到貝氏網路中計算該圖準確度為多少，增加說服力。

針對的目標為較少樣本的缺血性中風轉出血性中風，或者是較多樣本的中風的轉化因素，探討在不同大小的的資料中，是否都能有個很好的結果。

資料是使用 2000 年到 2013 年健保資料庫中病人看病結果的資料，根據 ICD-9-CM 編碼篩選出有得過中風和沒得過中風的病患，有得過中風病患病人的資料只擷取當他們得到中風前的看診紀錄，用疾病當作變數，找尋病患患有那些疾病時較容易得到中風。

最後結果為當使用缺血轉出血的資料時，GA-PLS 的預測率可達 0.855，其中會影響的疾病有頭暈、便秘、慢性腎衰竭、高血壓、糖尿病、高脂血症、焦慮、肌肉痛、前列腺肥大等，使用中風的資料時，GA-PLS 的預測率可達 0.828，在貝氏網路中平均 ROC 曲面下面積為 0.8274 其中會影響的疾病有腸胃疾病、其他上呼吸道疾病、皮膚疾病、口腔唾液腺及頷骨之疾病、高血壓和心臟病等。將結果和目前醫學所公布的相關疾病相比，可以發現找出來的這些疾病有符合醫學結果，且其中還有發現一些目前不在這其中的疾病，但他們又呈現高度相關，如缺血轉出血的相關影響疾病有頭暈、便秘、慢性腎衰竭等，中風的相關影響疾病有口腔唾液腺及頷骨之疾病、食道胃及十二指腸之疾病等，後續可以針對這些疾病作更深入的研究，探討這些疾病是因為甚麼原因導致他們之間是有相關的。

關鍵字：基因演算法、偏最小平方、關聯規則、貝氏網路、中風

Abstract

Stroke is one of the top ten causes of death in Taiwan, and it is also the main cause of people's disability, and only 15% of people can fully recover. Therefore, this study wants to use the National Health Insurance database to organize and analyze, hoping to find some disease model associated with stroke, and a related disease map.

This study used the GA-PLS method to improve accuracy, and increased association rules and Bayesian networks. We used the rules of association between commodities, which often use in in business, and replaced them with various diseases, explored the relationship between every diseases. Then we create the directed non-cyclic graph, and calculate the accuracy of the graph in the Bayesian network to increase the persuasiveness.

The target is less sample of ischemic stroke to hemorrhagic stroke, or a more sample of stroke conversion factors, to explore whether there are good results in different sizes of data.

The data is based on the results of patient visits in the health care database from 2000 to 2013. We pick out the patient who had stroke before and who had not, according to the ICD-9-CM code. The data of patient who had stroke before, we only pick their visit record before they had stroke, using the disease as a variable, it is easier to find out which disease they have that will make them easier to get stroke.

The final result is that when using ischemic hemorrhage data, the predictive rate of GA-PLS can reach 0.855, among which the diseases that may be affected are dizziness, constipation, chronic renal failure, hypertension, diabetes, hyperlipidemia, anxiety, muscle pain, prostatic hypertrophy, etc. When using the stroke data, the prediction rate of GA-PLS can reach 0.828, and the area under the ROC curved surface in the Bayesian network is 0.8274. The diseases that will be affected are gastrointestinal diseases, other upper respiratory diseases, and skin diseases, diseases of the oral salivary glands and tibia, hypertension and heart disease. Comparing the results to the related diseases published by the current medical science, it can be found that the diseases found are in line with medical results, and some of the diseases that are not currently found in list, but they are highly correlated to ischemic hemorrhage. Such as dizziness, constipation, chronic renal failure, etc. The related diseases affecting stroke include diseases of oral salivary glands and tibia, diseases of the esophagus and the duodenum, etc., and further research work can be conducted to explore what cause these diseases to be related.

目錄

中文摘要.....	I
英文摘要.....	II
第一章 研究動機	1
第二章 文獻探討	3
2.1 相關背景介紹	3
2.2 相關研究	5
2.3 中風	8
2.4 基因演算法	9
2.5 偏最小平方	10
2.6 粒子群演算法(PSO)	11
2.7 Seq2seq	11
2.8 Apriori-IFM	12
2.9 貝氏網路	12
2.10 卡方檢定	13
第三章 研究方法	14
3.1 資料整理	15
3.2 選擇相關變量	16
3.2.1 GA-PLS	16
3.2.2 最小偏差法(Partial Least Squares)	21
3.2.3 Seq2seq	23
3.3 PSO 找出疾病的權重	24
3.4 找出疾病之間的關係	27
3.4.1 Apriori 演算法	28
3.4.2 Apriori-IFM	31

3.5 貝氏網路	33
3.6 ROC 曲線	38
3.7 卡方檢定	39
第四章 實作	40
4.1 資料	40
4.2 實驗結果	41
4.2.1 缺血性轉出血性	41
4.2.2 中風	46
4.2.2.1 Seq2seq	47
4.2.2.2 GA-PLS	50
4.2.2.3 GA 後的 PSO	55
4.2.2.4 Aprioi-IFM	60
4.2.2.5 貝氏網路	65
4.2.2.6 關聯規則後 PSO	67
4.3 結果討論	79
4.4 實作結果	81
第五章 結論	84
參考文獻	86

圖目錄

圖 1.1 論文架構.....	2
圖 2.1 邏輯函數的分布圖.....	4
圖 3.1 研究架構.....	14
圖 3.2 相關軟體.....	15
圖 3.3 基因挑選.....	17
圖 3.4 GA-PLS 流程圖.....	17
圖 3.5 seq2seq 架構圖.....	23
圖 3.6 改良版的 PSO 流程圖.....	25
圖 3.7 Apriori 流程圖.....	28
圖 3.8 Apriori 演算法過程.....	30
圖 3.9 Apriori – IFM 流程圖.....	31
圖 3.10 疾病關係圖.....	33
圖 3.11 節點的聯合機率分配.....	36
圖 3.12 條件機率表格.....	36
圖 4.1 實驗流程圖.....	41
圖 4.2 缺血轉出血分群資料前(GA-PLS).....	42
圖 4.3 缺血轉出血分群資料後(GA-PLS 改良前).....	43
圖 4.4 缺血轉出血分群資料後(GA-PLS 改良後).....	43
圖 4.5 PSO 權重箱型圖(缺血轉出血).....	44
圖 4.6 PSO 長條圖(缺血轉出血).....	45
圖 4.7 seq2seq 資料樣本結構.....	48
圖 4.8 seq2seq 測試後輸出結果(左)有中風(右)沒有中風.....	49
圖 4.9 GA-PLS 分群資料(全年齡 PLS 5 維).....	51
圖 4.10 GA-PLS 分群資料(65 歲以下).....	52

圖 4.11 GA-PLS 分群資料(65 歲以上)	52
圖 4.12 GA-PLS 分群資料(65 歲以下男性)	52
圖 4.13 GA-PLS 分群資料(65 歲以下女性)	53
圖 4.14 GA-PLS 分群資料(65 歲以下男性)	53
圖 4.15 GA-PLS 分群資料(65 歲以下女性)	53
圖 4.16 GA-PLS 分群資料(累計次數)	54
圖 4.17 GA-PLS 分群資料(PLS 2 維)	54
圖 4.18 GA-PLS 分群資料(PLS 10 維)	54
圖 4.19 GA 後 PSO(全年齡)	55
圖 4.20 GA 後 PSO(65 歲以下)	55
圖 4.21 GA 後 PSO(65 歲以上)	56
圖 4.22 未刪迴圈前(全年齡)	60
圖 4.23 關聯圖(全年齡)	61
圖 4.24 關聯圖(65 歲以下)	61
圖 4.25 關聯圖(65 歲以上)	62
圖 4.26 關聯圖(65 歲以下男性)	62
圖 4.27 關聯圖(65 歲以下女性)	63
圖 4.28 關聯圖(65 歲以上男性)	63
圖 4.29 關聯圖(65 歲以上女性)	64
圖 4.30 關聯圖(累計次數)	64
圖 4.31 ROC 曲線	67
圖 4.32 關聯規則 PSO 權重	69
圖 4.33 UI 程式介面(全年齡未選取)	82
圖 4.34 UI 程式介面(全年齡 analyze1)	83
圖 4.35 UI 程式介面(全年齡 analyze2)	83

表目錄

表 2.1 相關研究比較表	6
表 2.2 相關研究方法比較	7
表 3.1 樣本的示意表	16
表 3.2 GA+PLS+KNN Algorithm	19
表 3.3 SIMPLS-T Algorithm	22
表 3.4 Modified swarm (PSO) algorithm	25
表 3.5 病人資料(關聯規則)	29
表 3.6 病人資料(貝氏)	35
表 3.7 聯合機率分配	35
表 3.8 貝氏網路的聯合機率	37
表 3.9 卡方檢定	39
表 4.1 缺血轉出血 GA-PLS 的正確度	44
表 4.2 缺血轉出血的相關因子	45
表 4.3 卡方檢定(缺血轉出血)	46
表 4.4 seq2seq 的準確度	50
表 4.5 GA-PLS 的準確度	51
表 4.6 GA-PSO 中風的相關因子(全年齡)	57
表 4.7 GA-PSO 中風的相關因子(65 歲以下)	58
表 4.8 GA-PSO 中風的相關因子(65 歲以下)	59
表 4.9 貝氏網路準確度	66
表 4.10 關聯規則卡方檢定(全年齡)	70
表 4.11 關聯規則卡方檢定(65 歲以下)	71
表 4.12 關聯規則卡方檢定(65 歲以上)	72
表 4.13 關聯規則卡方檢定(65 歲以下男性)	73

表 4.14 關聯規則卡方檢定(65 歲以下女性).....	74
表 4.15 關聯規則卡方檢定(65 歲以上男性).....	75
表 4.16 關聯規則卡方檢定(65 歲以上女性).....	76
表 4.17 關聯規則卡方檢定(累計次數).....	77
表 4.18 PSO 和卡方的相關係數.....	77
表 4.19 PSO 和卡方前三種疾病比對.....	78
表 4.20 中風可能之疾病.....	80
表 5.1 中風因素比較表.....	85

第一章 研究動機

尋找那些疾病會影響那些疾病的發生，絕大部分都是依據過去的研究或者是經驗，然後再針對有可能相關的疾病做研究，且所使用的方法大部分都是回歸分析、羅吉斯回歸等，雖然近年來已有一些是用機器學習的方法，但所使用的特徵也都是根據過去的研究或者是經驗再做資料收集，而本研究是想用歷年健保資料庫的資料，做機器學習的架構，以中風為例子，找尋那些疾病會影響中風的發生，不透過過去的研究或者是經驗，而是讓這些巨量資料告訴我們，那些疾病是會影響中風的，且或許還可以找出一些與目前影響中風有關疾病以外的疾病，但它又是從健保資料裡面所找到的，如果模型預測度夠高的話，那該疾病很有可就是新影響中風的因素。

而為甚麼要以中風 (stroke) 為主呢，是因為它是我國的十大死因之一。依據衛生福利部 105 年國人十大死因統計顯示[3]，腦血管疾病為國人十大死因的第 4 位，在台灣 35 歲以上的成年人，每年約 3 萬人會發生第一次中風，其中 25% 到 45% 會在 5 年內二度中風，估計全台中風人數達 15 至 20 萬人，近年來中風有年輕化趨勢，顯示中風已不再只是老年人的疾病[10]。

所以希望能運用此方法，找尋一些與中風相關的疾病模型，和相關之疾病關係圖，找尋中風的轉化因素。

而本文依據前一篇所寫的方法[6]，拿掉了 PSO 加入 Apriori-IFM 和貝氏網路，改良了 GA-PLS 提高準確度，下圖為論文的架構：

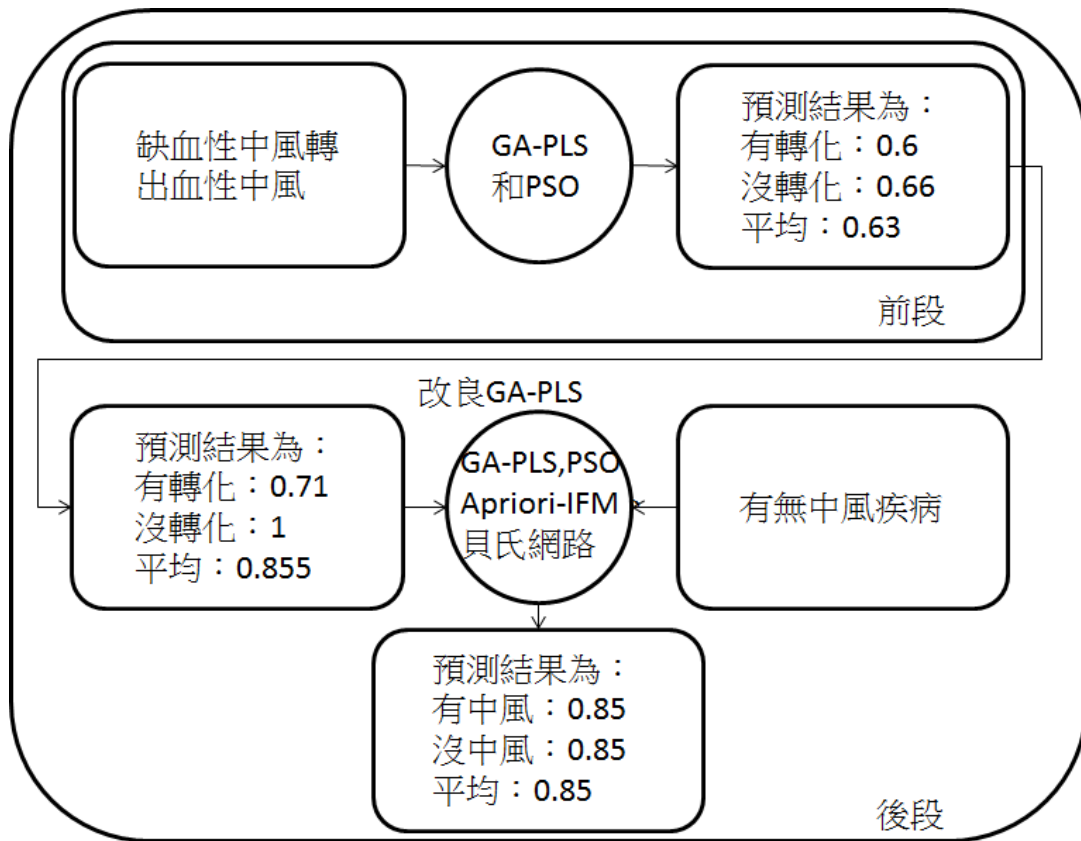


圖 1.1 論文架構

第二章 文獻探討

2.1 相關背景介紹

近年來隨著科技的進步，新的統計理論不斷被提出，在建立疾病預測模型上，方法及種類繁多，下面就簡單介紹幾個常用在預測疾病上的模型預測方法：

一、迴歸分析

迴歸分析 (Regression Analysis) 是一種統計學上分析數據的方法，可以探討連續隨機變數對連續變量的影響，目的在於了解兩個或多個變數間是否相關、相關方向與強度，以便觀察特定變數變動時，來預測研究者感興趣的變數。

迴歸分析是建立應變數 Y 與自變數 X 之間關係的模型 [2]，但在實際研究中，通常自變數會不只一個，這是為了模型能預測得更準確，此種同時由多個自變數來預測應變樹的方法稱為多元迴歸分析，而多元迴歸方程式的數學模型如下：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

其中 Y 為應變數， $X_1, X_2, X_3, \dots, X_n$ 為自變數， ε 為誤差值。

二、邏輯回歸

邏輯回歸 (logit regression) 適用於變數為二元類別的資料，其基本假設與線性回歸類似，在邏輯回歸分析中，自變數可為連續型變數或者是類別變數 [25]。由於類別變數是屬於離散型資料，所以必須把資料轉換成介於 0 到 1 之間的連續型資料型態，才可以做回歸。和迴歸分析最大的差異在於反映變數的型態不同，所以在運用上也須符合傳統的迴歸分析的假設，避免共線性問題，以及符合常態分配等相關統計假設。

邏輯回歸模型在統計上的運用已經極為普遍，尤其在醫學方面使用得更為廣泛。

邏輯分布公式如下：

$$P(Y = 1|X = x) = \frac{e^{x\beta}}{1 + e^{x\beta}}$$

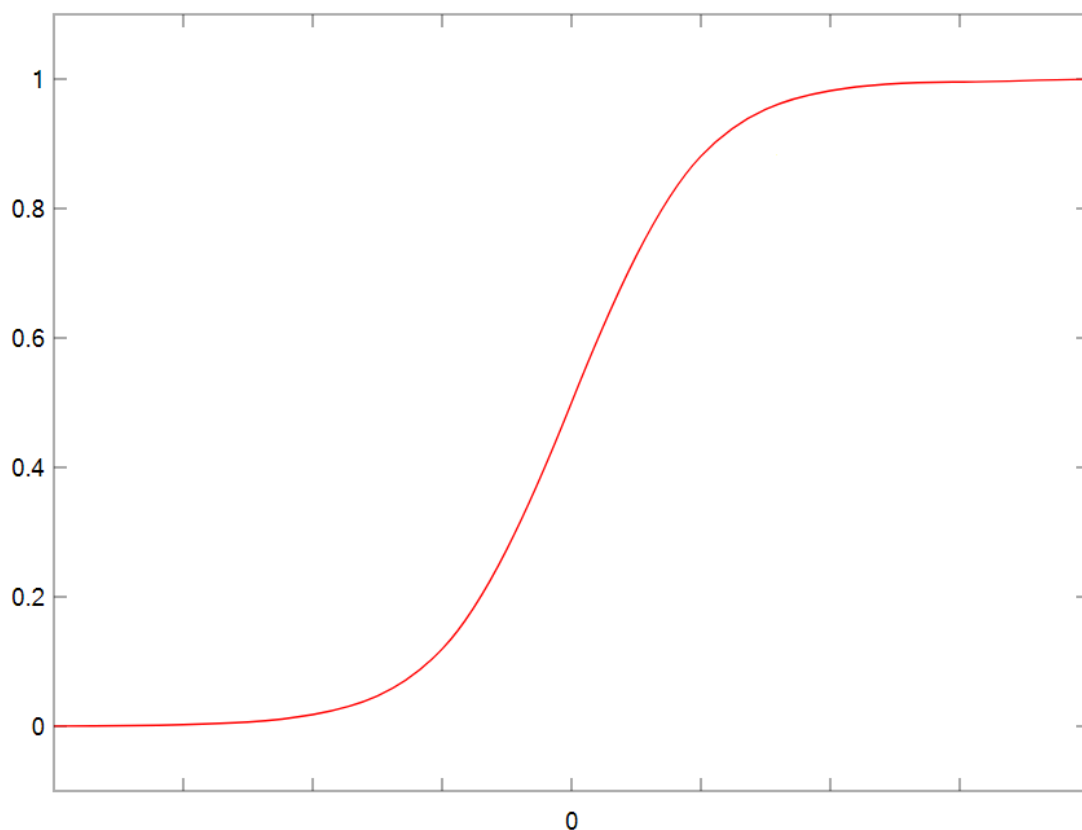


圖 2.1 邏輯函數的分布圖

三、時間序列

時間序列 (Time series) 是一種依照時間順序排列的數據集，根據資料所反映出來的發展過程、方向和趨勢，進行推演，藉以預測下一段時間可能達到的水平。通常時間序列的時間間隔為一恆定值 (如 1 秒，10 分鐘，12 小時，7 天，1 年)，因此可以做為離散型資料進行分析處理[14]。

時間序列分析，運用過去的歷史資料，通過統計分析，推測未來的發展趨勢 [30]。事物的過去會延續到未來這個假設為前提，包含兩種含義：一是不會發生跳躍性變化；二是過去發生的現象可能是未來的發展趨勢。

四、機器學習

機器學習主要是設計和分析一些資料讓電腦可以自動學習的演算法。我們只要給定問題的範圍加上要訓練的資料 (Training Data)，然後從資料中選擇特徵再建構資料模型，最後把這個模型作為學習成果，就可以拿來作預測。

機器學習簡單來說就是像人們教育下一代一樣，如果我們給小孩子們看過很多蘋果的圖片或實物，並且和他說這是蘋果，那麼他下一次就會認得哪一些是蘋果了，而預測則是利用過去的資料來預測未來可能的資料。

本研究就是基於機器學習方法，收集有關中風疾病之相關疾病，之後用關聯規則找尋這些疾病的關聯，最後使用貝氏網路的方法，使其呈現視覺化的圖形，驗證這個網路的預測率。

2.2 相關研究

作者	方法	針對的目標	變數	預測準度
Li, H.W., et al. (2011) [37]	時間依賴性 Cox 回歸模型	因中風再入院的相關因子	年齡、性別和合併症	有高度相關
Jae-woo Lee, et al. (2017) [33]	Cox 的比例風險回歸模型	10 年中風預測模型	年齡，BMI，膽固醇，高血壓，糖尿病，吸煙狀況和強度，體力活動，飲酒，既往史和家族史	平均 AUC 值 0.825
Tsai-Chung Li, et al. (2018) [46]	Cox 的比例風險回歸模型	預測缺血性卒中發病率	年齡，性別，吸煙習慣，2 型糖尿病病程，血壓，HbA1c 水平，總膽固醇與高密度脂蛋白比值，肌酸酐，空腹血糖變異，動脈栓塞和血栓形成，糖尿病視網膜病變，低血糖症，抗糖尿病藥物使用	ROC 曲線下之面積為 0.72

			和心血管疾病用藥	
慕哈曼 (2013)[9]	階層式迴 歸分析、 後傳播網 路、支持 向量機以 及決策樹	一多階數資 料探勘方法 建立腦中風 風險預測模 型	腦部影像資料	ROC 曲線下 之面積為 0.8067
尤哲威 (2015)[1]	類神經網 路、二元 邏輯斯迴 歸	利用類神經 網路建構中 風患者再復 發中風之預 測模型	利用卡方檢定與 T 檢定其 因素與中風復發之間關聯 分析之檢定結果	ROC 曲線下 之面積為 0.721
張佑璋	基因演算 法、關聯 規則和貝 氏網路	影響中風發 生的相關因 子	所有疾病變數	GA 平均可達 0.828、平均 ROC 曲面下 面積為 0.8274

表 2.1 相關研究比較表

在許多相關研究中，已經有很多研究針對生理資訊，但很少對所有疾病為變量做研究，最多也只有對已知的相關併發症做研究，上面列舉了一些相關研究的比較，表 2.1。

Li, H. W. [37] 等人使用 2004-2007 年國家健康保險局索賠數據庫進行了回顧性分析，分別建立了時間依賴性 Cox 回歸模型，以確定缺血性中風病患出院後 1 個月，6 個月和 1 年內再入院的預測因素。最後結果發現如果更多患者長時間使

用抗血小板藥物，可能會避免再次發生中風的機率。

作者	階層性	相關背景 知識有無	快速處理 問題	因素之間 的關聯性
Li, H. W., et al. (2011)[37]			V	
Jae-woo Lee, et al. (2017)[33]			V	
Tsai-ChungLi, et al. (2018)[46]			V	
慕哈曼(2013)[9]	V		V	
尤哲威(2015)[1]		V		
張佑瑋	V	V		V

表 2.2 相關研究方法比較

Jae-woo Lee[33]等人使用 2002 - 2003 年國家衛生檢查人員的資料，建立一個 10 年中風預測模型，用 Cox 的比例風險回歸模型，最後得到使用外部驗證數據集的中風風險預測模型的 AUC 值在男性為 0.83，在女性為 0.82。

Tsai-ChungLi[46]等人使用 2001-2004 年期間 2 型糖尿病患者，年齡在 30-84 歲之間，建立 2 型糖尿病患者缺血性中風風險預測模型，使用 Cox 的比例風險回歸模型來確定推導缺血性中風發病的危險因素，最後的結果為在 3 年、5 年和 8 年缺血性中風發病風險的曲線下面積分別為 0.72, 0.71 和 0.68。

慕哈曼[9]使用的資料為 2004 至 2011 年的腦部影像資料，來對於中風進行預測，先藉由比較樣本方式對於資料進行“再平衡”處理，再藉由階層式迴歸分析對於平衡資料的重要特色加以篩選，最後使用後傳播網路、支持向量機以及決策樹來處理被選擇的特徵以對於中風進行預測，結果為 ROC 曲線下之面積為 0.8067。

尤哲威[1]藉由全民健康保險研究資料庫找出中風患者且再次復發之研究對象，利用類神經網路建構中風患者再復發中風之預測模型，而兒最後的結果為

Sensitivity 可達 0.699、Specificity 可達 0.562、ROC 曲線下之面積可達 0.721。

本研究和上述列舉出來的研究相比，大部分都是使用回歸分析的作法，這個做法的優點是可直接快速處理問題，且這個方法非常有名，但需要嚴格的假說，且有對相關問題的背景知識，且須處理異常值。而用我的方法，研究者可不需要有相關的背景知識，但需要龐大且巨大資料量，且不須給他一些已知的規則，它會自己導出他自己得規則，依據所給的巨量資料，有可能會出現和現有已知的規則一樣，但也有可能會出現新的規則。

和前面的研究相比可以發現預測準確度結果都是本研究較佳且其中也有一些是使用健保資料庫的資料。較為不一樣的地方是本研究用所有疾病當成變量，主要是想透過巨量資料，找出其新的相關因子，不須依靠過去的研究或者是經驗，只針對特定疾病做分析，希望可以找到新的變量。

2.3 中風

中風又稱作腦血管意外（英語：cerebrovascular accident，簡稱 CVA）、腦血管病變（英語：cerebrovascular incident，CVI），是腦部血管有局部性的阻塞或出血，使得依靠這條血管供給血液及營養的腦組織受損，造成腦細胞傷害、壞死、局部神經性機能障礙。

中風又分成兩類缺血性腦中風和出血性腦中風[16]。

缺血性腦中風（又稱腦梗塞）是由於腦部供血不足，導致腦組織功能障礙及壞死[34]。可能的原因有下面兩種，腦血栓：形成的原因有兩種—(1)血液凝固異常使血液黏稠度變大而形成血栓。(2)腦血管（動脈）發生粥狀硬化，形成斑塊，使動脈的管腔變狹窄，而產生血栓，血液流通受阻，因而造成腦部缺氧性壞死。腦栓塞 腦部以外的地方（最常見的為心臟）來的栓子(如：血塊、硬化斑塊、脂肪、氣泡等)阻塞腦血管，而導致腦部缺血性壞死[48]。

出血性腦中風（俗稱腦溢血）是顱骨內任何地方的血液積累，腦內出血：最主要因素是高血壓（佔 70-80%），其他由於外傷、血液方面的疾病或腦瘤亦會引起。蜘蛛膜下腔出血：常見的原因是動脈瘤破裂所引起。

出血性轉化（Hemorrhagic transformation, HT）可以是腦梗死患者的自然轉歸過程，也可以出現於中風治療之後[38]。目前的研究報導中，對於出血性轉化的發生率說法不一，然而嚴重的出血性轉化致殘、致死率高，使得臨床醫生不得不防。其危險成因有抗凝指徵、人口資料（性別，種族和年齡）、醫學因素（血壓，血糖控制，脂質分佈和腎功能）。

2.4 基因演算法

基因演算法是一種借鑒生物進化而演化出來的搜尋方法，也就是達爾文的進化論（適者生存，不適者淘汰）。這種學說認為，生物要生存下去，就必須進行生存鬥爭。生存鬥爭包括種內鬥爭、種間鬥爭以及生物跟無機環境之間的鬥爭三個方面。在爭鬥過程中，適應良好的物種將會留下，並且有更多機會遺傳給後代；適應不良的物種就容易被淘汰，且產生後代的機會也比較少。達爾文把這種在生存鬥爭中適者生存，不適者淘汰的過程叫做自然選擇。它表明，遺傳和變異是決定生物進化的內在因素。自然界中的多種生物之所以能夠適應環境而得以生存進化，是和遺傳和變異生命現象分不開的。正是生物的這種遺傳特性，使生物界的物種能夠保持相對的穩定；而生物的變異特性，使生物個體產生新的性狀，以致於形成新的物種，推動了生物的進化和發展。

基因演算法是模擬達爾文進化論的生物進化過程的計算模型。它的構想是來自於生物遺傳學和適者生存的自然法則，是具有（生存+檢測）迭代過程的搜尋演算法。它是由美國的 J. Holland 教授 1975 年首先提出[31]，是一種高效、並行、全局搜索的方法，能自動獲取和累積有關搜索空間的知識，自適應地調整搜索方向，不需要確定的規則。遺傳演算法的這些性質，已被人們廣泛地應用於組

合優化、機器學習、信號處理、自適應控制和人工生命等領域。它是現代有關智能計算中的關鍵技術之一。

2.5 偏最小平方

偏最小平方 (Partial least squares) 來源於瑞典統計學家 Herman Wold，然後由他的兒子 Svante Wold 發展[47]。

在迴歸分析中，利用自變數群去解釋應變數的改變，由於通常在分析時，會發現所在意的自變數眾多，且有些資料中的自變數會有高度的共線性 (multi-collinearity) 或是自變數個數多於觀察值個數，此時迴歸分析的結果可能會產生錯誤，這些錯誤可透過對資料預先處理或是其他統計方法來改善，偏最小平方迴歸分析即是一種[23]。此方法是認為自變數群可以由數個潛在變數 (latent variable) 或稱成份 (component) 所解釋，使用結構方程式 (structure equation modeling) 的技巧確認潛在變數的存在，並同時進行迴歸分析尋找解釋能力最佳的潛在變數[13]。

也就是說，相較於其他的迴歸分析方法，更適合解決多重共線性的問題，也就上面所提的，由於變量之間存在高度相關而使迴歸分析估計不準的現象。

PLS 與 PCA 的原理類似，但皆以資料的維度縮減為目的，相對於 PCA 只在原始解釋變數 X 內，尋找最大變異的線性組合，但當有一些有用的變數的相關性小時，選取主成分時就容易被忽略，使得預測可靠性降低，PLS 則在進行轉換原始變數 X 成為潛在變數時，另外引入依變數 Y 的訊息步驟，同時從 X 、 Y 中尋找影響較大的變數建立預測模型。

因此在本篇中 PLS 是來藉由縮減資料維度，從較少的解釋變數中，來獲得最多資訊預測反應變數。

2.6 粒子群演算法(PSO)

粒子群演算法(Particle Swarm Optimization, PSO)是由 Eberhart 和 Kennedy 於 1995 年所提出[20][43]，是一種具有群體智慧的方法，也是屬於演化式計算的一個新分支。其優點在於快速收斂、較少的參數設定、適用於動態環境的能力等優點。此搜尋技術主要是模擬鳥群覓食的社會行為；一群鳥隨機的分佈在一個區域中，在這個區域裡只有一塊食物。所有的鳥都不知道食物在哪裡。

2.7 Seq2seq

Seq2Seq 即 Sequence to Sequence，是一種時序對映射的過程，實現了深度學習模型在序列問題中的應用，其中比較突出的是機器翻譯和機器人問答。Seq2Seq 被提出於 2014 年，最早由兩篇論文引出，分別是 Google Brain 團隊的《Sequence to Sequence Learning with Neural Networks》[45]和 Yoshua Bengio 團隊的《Learning Phrase Representation using RNN Encoder-Decoder for Statistical Machine Translation》[19]。

Seq2Seq 主要由兩個 RNN 模型組成，一個用於 encode 輸入序列（將詞序列轉變為一個固定大小的向量），一個是將 encode 的結果 decode（將獲得的語義向量轉為一個 token 序列，即詞序列），故稱為 RNN Encoder-Decoder。它的運作原理其實與人類的思維很相似，當我們看到一段話時，會先將這句話理解吸收，再根據我們理解的內容說出回覆，Sequence to Sequence 就是在模擬這個過程。

總而言之，Sequence to Sequence 最漂亮的地方便是串接了兩個 RNN，第一個 RNN 負責將長度為 M 的序列給壓成 1 個向量，第二個 RNN 則根據這 1 個向量產生出 N 個輸出，這 $M \rightarrow 1$ 與 $1 \rightarrow N$ 相輔相成下就構建出了 M to N 的模型，能夠處理任何不定長的輸入與輸出序列。

2.8 Apriori—IFM

Apriori 演算法是一種經典的關聯規則數據挖掘演算法，它是由 Rakesh Agrawal 和 Ramakrishnan Skrikant 所提出的[11]。通過對數據的關聯性做分析和挖掘，找出這些訊息的支持度和信賴度，在決策制定過程中具有重要的參考價值。

Apriori 算法可用於商業中，用於分析消費市場價格，它能夠很快的提供決策人各種產品之間的價格關係和它們之間的影响。百貨公司、超市和一些老字號的零售店也在進行數據挖掘，以便猜測近年來顧客的消費習慣。

Apriori 算法可用於各種行業上，只要決策者想要知道，某一個變數和另一個變數或多個變數之間的關係，就可以使用這個方法[36]。

在本研究中，是基於 Apriori 演算法去做修改，增設了 IFM(Importance Factor Method)[8]，讓使用者不須去設定最小支持度和最小信賴度，而改為設定所需的規則數量，因為在設定最小支持度和最小信賴度是很重要的，如果設的太高會找不到規則，如過設的太低規則數會過多，需要多次的調整才會達到理想的結果。

使用此方法我們主要是用於找尋疾病之間的關係，一般來說，假設你知道這個病人有糖尿病，那你可能就會認為說，該病患也有很大的機率患有高血壓，所以基於這個概念，我們用 Apriori—IFM 找出疾病之間的關係，使它具備有重要的參考價值。

2.9 貝氏網路

貝氏網路(Bayesian networks, BN)也叫做信賴網路(belief networks)、機率網路(probabilistic networks)、因果網路(causal networks)或者為知識地圖(knowledgemap)，是一套可以用來精進預測的方法，在資料不是很多、又想盡量

發揮預測能力時特別有用[40]。

主要以有向的無迴路圖 (directed acycle graph, DAG)為基礎，應用其變數之間的因果關係與其相互影響的機率，DAG 包含兩個部分，分別是節點(node)及連結(link)[29]。在貝氏網路中，節點表示變項；連結表示變項之間的相互關係。連結的有無即代表其節點之間的關係是否為條件相依或條件獨立的情形，其影響程度則是以條件機率來表示[13]。

貝氏網路大多應用於醫療診斷、人工智慧、啟發式搜尋與垃圾郵件攔截等等領域，BN 擁有數學上的嚴謹度和直覺式理解，並且能在一組隨機變數上，有效的表示與計算聯合機率分配 (joint probability distribution, JPD)[39]。

且根據多項研究顯示，關聯規則所找出的規則就是一個有方向性的圖，經過一些處理過後，就可作為有向的無迴路圖，可接續做貝氏網路，這是一個可行的辦法，當不知道這些因子的關係時可用[17][21][24][28]。

2.10 卡方檢定

卡方獨立性檢定適用於分析兩組類別變數的關聯性。同一樣本中，兩個變項的關聯性檢定，也就是探討兩個類別變項(例如：性別和結婚狀態)之間，是否為相互獨立，或者是有相依的關係存在，是否達到顯著。

在這裡使用的是皮爾森卡方檢定，最早由卡爾·皮爾森在 1900 年發表[35]，用於類別變數的檢定。皮爾森卡方檢定可用於兩種情境的變項比較：適配度檢定，和獨立性檢定。

「適配度檢定」驗證一組觀察值的次數分配是否異於理論上的分配。

「獨立性檢定」驗證從兩個變數抽出的配對觀察值組是否互相獨立。

第三章 研究方法

本研究使用上一篇論文的 GA-PLS 並加以改良，加入關聯規則和貝氏網路，針對中風病人，找尋其相關會影響中風的疾病，尋找他們之間的關係，再測試此關係的可信度，其架構如下圖 3.1。

所使用工具為 SAS EG 和 MATLAB 如圖 3.2，使用 SAS EG 做資料整理，MATLAB 撰寫相關程式 GA-PLS、Apriori-IFM、貝氏網路和畫出疾病關聯圖。

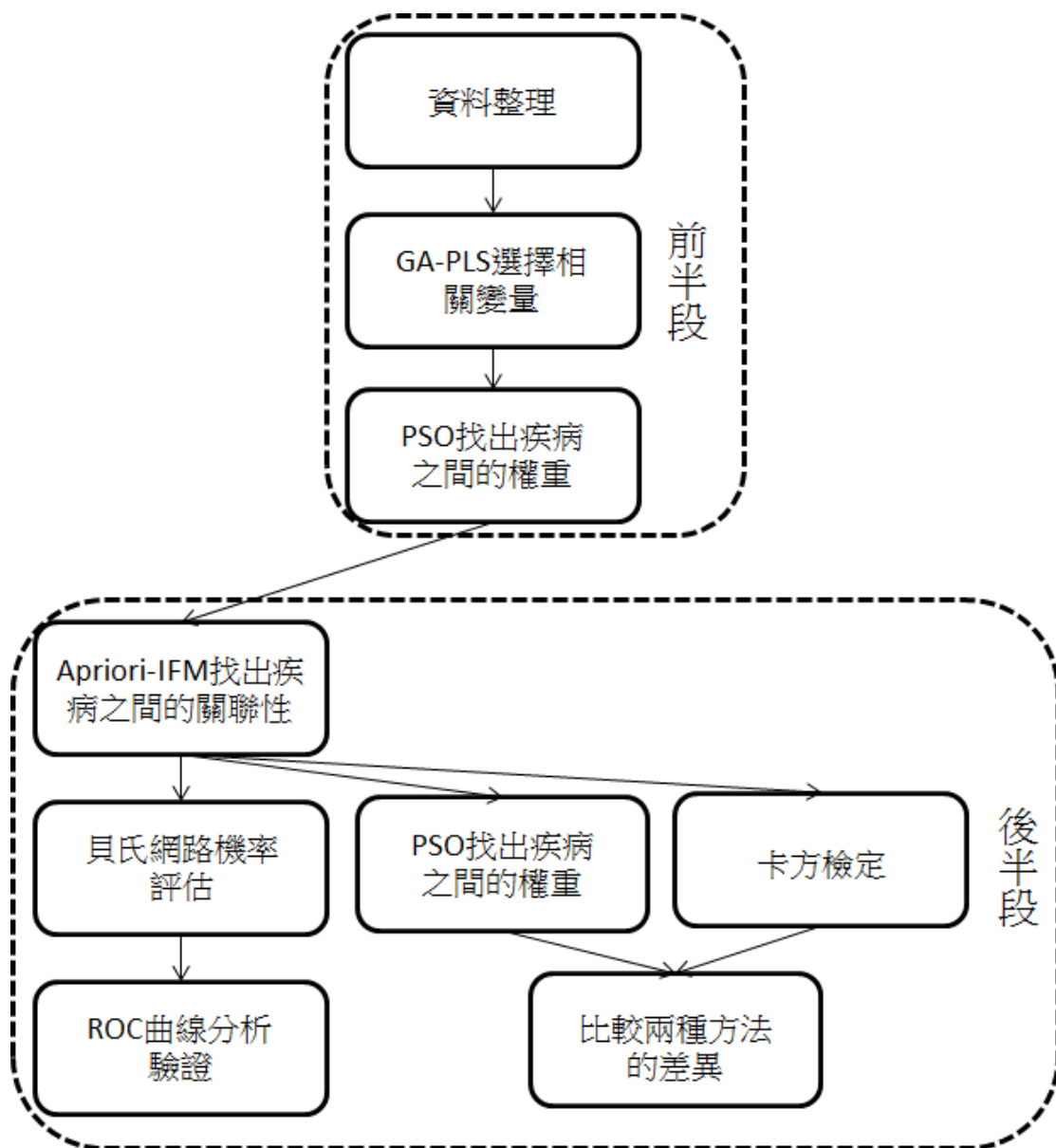


圖 3.1 研究架構



圖 3.2 相關軟體

3.1 資料整理

先用 SAS EG 把從健保資料庫中的中風和非中風患者的病歷給取出，作法是使用查詢產生器將疾病編碼縮減為 3 個字元，主要是讓相近的疾病可被分類在一起，使用查詢產生器將病人的出生日期轉換成年齡，且去除 18~110 以外的病人，之後對病人名稱的編碼做排序，可以讓之後 matlab 做整理時少花很多時間，再將所用不到的資訊剷除，減少所佔的記憶體，就可以匯出成 csv 檔，用 matlab 讀取 csv 檔，將資料轉換成 0, 1 的資料格式，最後分成有中風和沒有中風，在這裡中風使用的是 icd 9 cm code：430、431、432、433、434、435、436、437、438，使其呈現表 3-1：

表 3.1 樣本的示意表

	ACODE_ICD9_1	ACODE_ICD9_2	ACODE_ICD9_3
病人 1	疾病編號	疾病編號	疾病編號
病人 1	疾病編號	疾病編號	疾病編號
⋮	⋮	⋮	⋮
病人 2	疾病編號	疾病編號	疾病編號
⋮	⋮	⋮	⋮

轉換成下面的表格

	疾病 1	⋯	疾病 M
患者 1	0 or 1	⋯	0 or 1
⋮	⋮		⋮
患者 N	0 or 1	⋯	0 or 1

3.2 選擇相關變量

3.2.1 GA-PLS

GA-PLS 流程圖如圖 3.4 所示。基因演算法一開始會隨機生成數個長度等同原始訓練資料長度的二進制字串，稱為第零代(generation)或是初代編碼，其中的每一組二元字串被稱為一個染色體(chromosome)，每組編碼皆代表一個可行的解(可能包含最佳解)；對於一個染色體，所有被設置為 1 的資料會被抽出並包含到一個子集裡，反之則不包括如下圖 3.3。

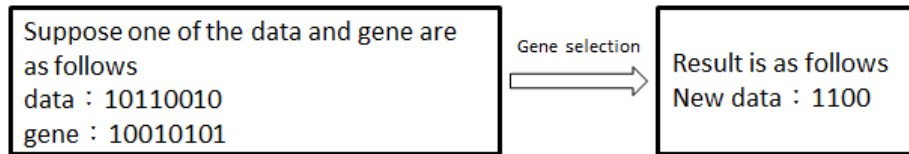


圖 3.3 基因挑選

之後以 PLS 降維成 5 維，挑選全部的維度做為分群標準，在經過基因演算法挑選樣本後，可以有效減少疾病數量與計算的資料數量，以提高程式的執行效率。

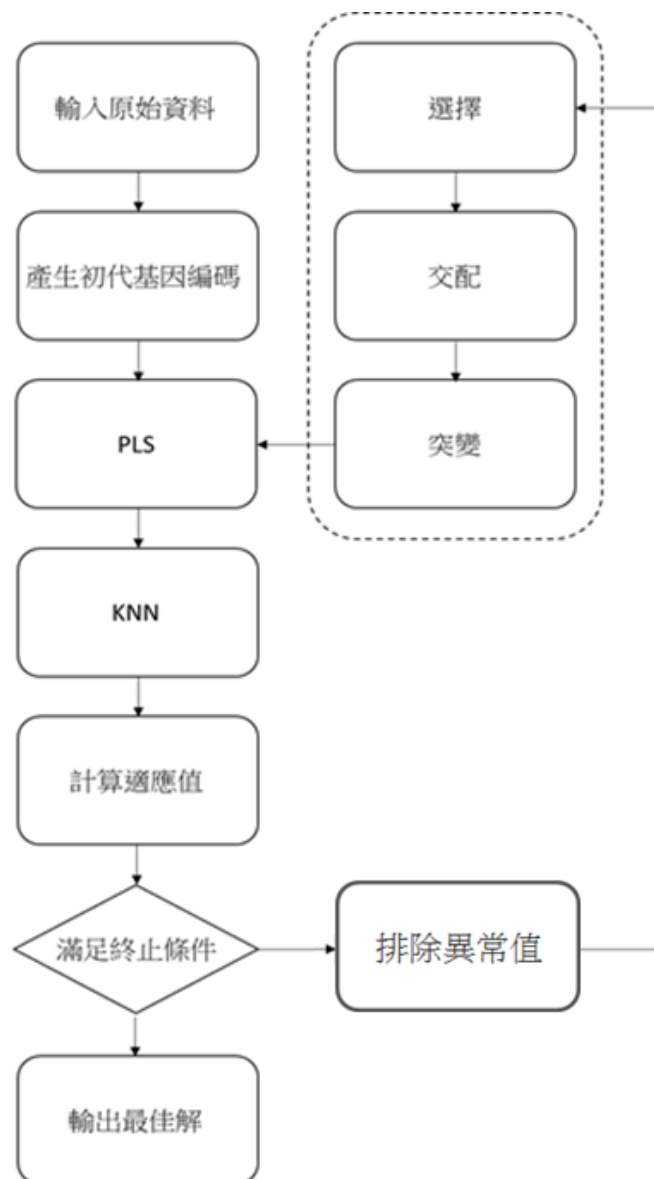


圖 3.4 GA-PLS 流程圖

其中基因演算法的適應函數定義如下：

$$Fd = (P1 + P2)/2 \quad (1)$$

每次每個染色體都會做一次適應度函數，P1 是被分到第一類的機率且是對的機率，也就是有中風的患者，被分對的機率。P2 是被分到第二類的機率且是對的機率，也就是沒有中風的患者，被分對的機率。而 P1 和 P2 是用 K-nearest neighbor 演算法(KNN)來計算距離該樣本最近的同類別的樣本個數有幾個，再除以總樣本個數，公式(2)。

$$Pi = \frac{d_i}{N_i} \quad i = 1,2 \quad (2)$$

d_i 是距離該樣本最近的同類別的樣本個數有幾個，N 是總樣本個數， N_i 是各個類別的樣本數。去計算加了權重的值是否大於零 $s_{i,j,k}$ ， $s_{i,j,k}$ 是計算該樣本附近同類別的樣本分數， $r_{i,j,k}$ 是附近同類別樣本的鄰居是第幾個，去加一個權重，越靠近則加的越多，如果不是同類別時，為 $-s_{i,j,k}$ ，公式(3)。

$$d_i = \begin{cases} d_i + 1 & \text{sum}(s_{i,j}) > 0 \\ d_i & \text{sum}(s_{i,j}) \leq 0 \end{cases} \quad i = 1,2 \quad j = 1 \sim N_i$$

$$s_{i,j,k} = 1 + e^{\frac{-r_{i,j,k}}{N_i}} \quad i = 1,2 \quad j = 1 \sim N_i \quad k = 1 \sim N_i \quad (3)$$

在這裡設一個變數為異常值，紀錄當使用最佳基因解時，判斷病人是否被分配錯誤，如果是則+1，如果不是則為 0。

之後去計算適應度函數，適應函數在不同的應用會對應到不同的計算方式，而基因的改變主要透過三種機制，挑選機制、交配機制、突變機制。

挑選機制主要選擇適應函數較高的染色體，已排除不好的染色體，其實就是探討如何從群體（樣本空間）挑選出個體（樣本）的取樣方式，本文的實驗中使用 Holland 提出的輪盤式 (roulette wheel selection) 選擇，由計算完的適應值做排序，淘汰較差的 50%，餘下前 50%者，適應值越高的，優先被挑選的機率越高。

交配機制是由挑選機制所選出來的親代染色體做合併產生出子代，讓子代含

有親代的部分特性，希望可以有更高的適應函數，但也可能會產生出只遺傳到缺點的子代，不過有挑選的機制，較差的子代會被淘汰，而優良的子代則會因為被挑選的機率較大，而可以繼續繁衍出下一代的子代。

突變機制會導致染色體發生變化，不過如果突變發生的太過頻繁，則會導致基因演算法變成隨機演算法，但如果又太少，則會陷入區域最佳解，因此基因演算法將突變視為次要的遺傳運算子，以較低的機率來反轉子代的某一個位元(0 變 1 或 1 變 0)。

當異常值大於等於迭代次數的一半時，代表這筆資料已經被連續分類錯誤，所以我判斷這個病人資料為異常，把這筆資料剔除，減少雜訊，之後繼續執行 GA-PLS。

表 3.2 GA+PLS+KNN Algorithm

// Initialise generation 0:

$$M_{ij} = \begin{bmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} \end{bmatrix}$$

Input: D Training data set, $\alpha_{ij} = \{0,1\}$, p= number of patient, m=total

disease number, 0 stands for without the disease, while 1 stands for with such disease. Binary

strings of chromosomes, $G_{ij}(i=1..n, j=1..m)$, n := the number of individual chromosomes in the

population;

k := 0;

χ := the fraction of the population to be replaced by crossover in each iteration, 0.5 in this

application;

μ := the mutation rate;

P_k := a population of n randomly-generated individuals;

Output: One string in G_{ij} with the best fitness function

$K_c(c)$: sample number in class c

M_c : Misclassification count

// Evaluate P_k :

do { // Create generation $k + 1$:

 // 1. Copy:

 Select $(1 - \gamma) \times n$ members of the sorted P_k and insert into P_{k+1} ;

 // 2. Crossover:

 Select $\gamma \times n$ members of the sorted P_k based on $F_k(d)$;

 pair them up;

 produce offspring;

 insert the offspring into P_{k+1} ;

 // 3. Mutate:

 Select $\mu \times n$ members of P_{k+1} ;

 invert a randomly-selected bit in each;

 // Evaluate P_{k+1} :

 Compute $\text{fitness}_{k+1}(i)$ for each chromosome $i \in P_{k+1}$ via eq.(6)~(12);

For ($i = 1$ to G_{col} size)

 Dim % Dimensions

$X_{scores} \leftarrow \text{Pls}(D_{train}(\text{find } G(i)=1), \text{Dim})$

 Count $\leftarrow 0$

 For ($j = 1$ to $D_{train\ col}$ size)

 IDX $\leftarrow \text{Knnsearch}(X_{scores}, Kc(D_{train}(j).class))$

```

    For (k = 1 to IDXrow size)

        If (IDX(j,k) ∈ Dtrain(j).class)

            
$$N_s(k) \leftarrow \left( 1 + e^{\frac{-k}{Kc(D_{train}(j).class)}} \right)$$


        Else

            
$$N_s(k) \leftarrow - \left( 1 + e^{\frac{-k}{Kc(D_{train}(j).class)}} \right)$$


        End

    End

    If (sum(Ns) > 0)

        Count++

        Mc(i,j) ← 0

    Else

        Mc(i,j)++

    End

    End

    Fd(i) ← Count/ Dtrain col size

End

// Increment:

k := k + 1;

} while fitness of fittest individual in Pk is not high enough;

return the fittest individual from Pk;

```

3.2.2 最小偏差法(Partial Least Squares)

PLS 是一種多元統計分析的方法，由 Herman Wold (1975) 提出，近幾年來在理論、方法與應用上都迅速的發展。PLS 在模型的建立上，融合了多元線性迴

歸、主成分分析及典型相關分析等重要的統計技術，實現了多種數據分析的應用。

但傳統的 PLS 必須經過反覆的疊代運算以及資料的壓縮來求得正確的權重向量與分數向量，是比較耗時的[7]。所以本文是用 De Jong, S. (1993) 提出了一種方法[22]，直接由原始資料取得分數向量，不需經過反覆的疊代運算，此種方法即為 SIMPLS。將權重向量 r 、 q 單位化即 $r'r=1$ 、 $q'q=1$ ，在結束第一輪運算後 r_1 為 $S=X'Y$ 的第一個左奇異向量(left singular vector)，則 q_1 為 $S=X'Y$ 第一個右奇異向量(right singular vector)，不同於我們取決權重向量為最大變異的分數向量。所以在 SIMPLS 中，令權重向量 r 為 S 的第一個左奇異向量。由 $Y=TC'+F*$ 和 $C'=T'Y$ 可得到 Y 的線性模型估計式：

$$\hat{Y} = TT'Y = XRT'Y \quad (4)$$

其估計的迴歸係數矩陣 B_{PLS} ：

$$B_{PLS} = RT'Y = RT^{-1}Y \quad \text{且 } T'T = 1$$

演算法如下表示：

表 3.3 SIMPLS-T Algorithm

(1)	$S = X_0'Y_0$
(2)	For $i=1, \dots, A$ $i=1$: compute SVD of S $i>1$: compute SVD of $S = S - P(P'P)^{-1}P'S$
(3)	$r =$ first left singular vector
(4)	$t=X_0*r$
(5)	$p=X_0't/(t't)$
(6)	Store r , t and p into \mathbf{R} , \mathbf{T} , \mathbf{P} respectively

(7) Compute regression coefficients

$$B_{PLS} = RT'Y$$

3.2.3 Seq2seq

Seq2seq 這個類神經網路是 Encoder Decoder network，是由稱為 Encoder 和 Decoder 的兩個分開的 RNN 組成的模型。Encoder 每次讀取一個輸入序列，並在每個步驟輸出一個向量。Encoder 的最終輸出保持為 context vector。Decoder 使用這個上下文向量來一次一步地產生輸出序列。

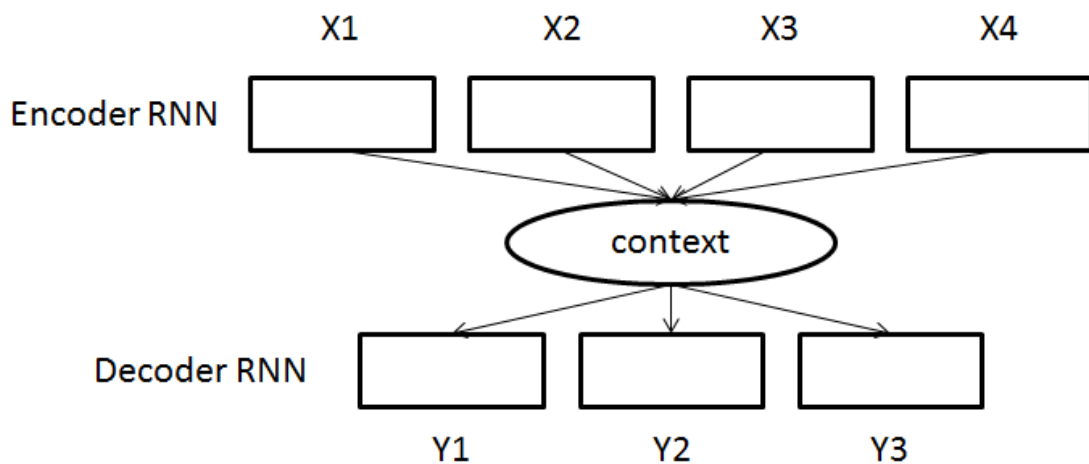


圖 3.5 seq2seq 架構圖

總結來說就是串接了兩個 RNN，第一個 RNN 負責將長度為 M 的序列給壓成 1 個向量，第二個 RNN 則根據這 1 個向量產生出 N 個輸出，這 M → 1 與 1 → N 相輔相成下就建構出了 M to N 的模型，能夠處理任何不定長的輸入與輸出序列。

所以這次我們設定輸入一位病人的過往病例，輸出一個答案(1 or 0)，判斷是否有可能會得到中風。

這次的模型參數 encoder and decoder 各 1 layer，畢竟我只要一個輸出即可，Hidden size in encoder and decoder 總共有 512 layers，Batch size 有 64 個，Learning rate 有 0.0001 的成功率，Train the model with it iterations 是 5000 次，總共迭代 5000 次。

資料格式設定為基數欄是一位病人的過往病例，偶數欄是這位病人是否有得過中風，也就是說 Encoder 一位病人的過往病例，產出 context vector，最後在 Decoder，之後的答案要等於為偶數欄的值。

這次所使用的是各 5000 筆隨機挑選出有中風和沒中風的資料，共 10000 筆資料下去做訓練。

3.3 PSO 找出疾病的權重

做完 GA-PLS 或 Apriori-IFM 之後，依據所得出來的疾病，做 PSO 疾病的權重分配，主要是想知到到底哪一些疾病影響最大，流程如下圖 3.6。資料的變量是依據最好的基因或關聯規則所挑出的疾病而做改變，再選出兩類各 5000 筆的資料，依據 GA-PLS 中的適應函數來做分類結果的判斷，選出當權重為多少時其適應函數最大。

改良 PSO 的主要原因是因為 PSO 有易陷入局部最佳解的情況[41][32]，所以對 PSO 稍作修改，在收斂至個體與局部最佳解產生黏著時，保留局部最佳解，增加是否有粒子與群體最佳解碰撞的判斷式，主要就是當有粒子非常靠近群體最佳解時，會把那個粒子隨機丟到其他地方，使得有效的避開陷入局部極值的狀況，表 3.4 為改進 PSO 演算法。[4][7]

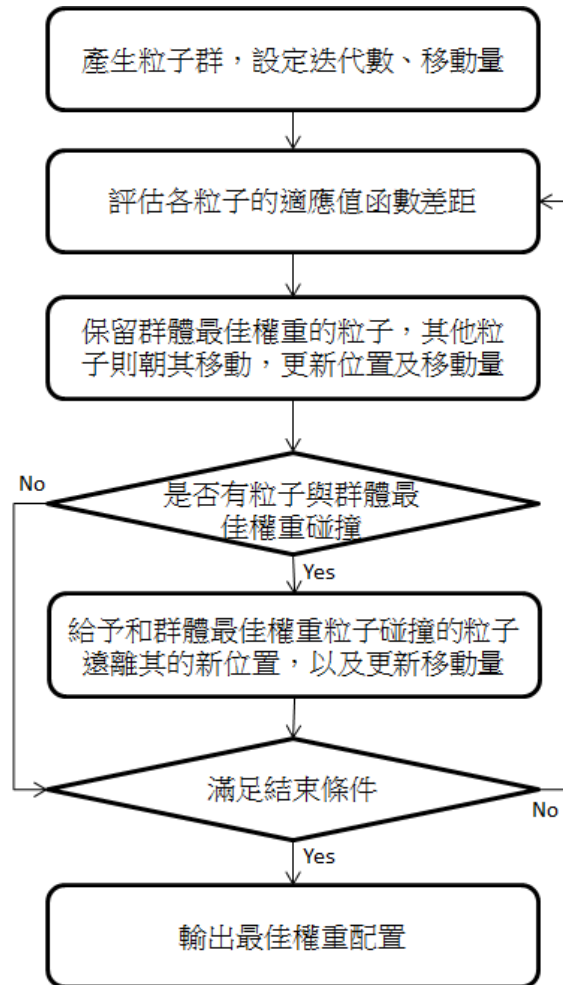


圖 3.6 改良版的 PSO 流程圖

表 3.4 Modified swarm (PSO) algorithm

Input: $C_{ik} = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ik} \end{bmatrix}$ the binary matrix of all selected diseases for all patients after PLS/GA process, k being the reduced number of 750 from the original total of 2770 diseases.

Output: $[w_1, w_2, \dots, w_k]$ for selected diseases with max. F(d) cost with

$$X_0 = C_{ik} = \begin{bmatrix} w_1 * a_{11} & \cdots & w_k * a_{1k} \\ \vdots & \ddots & \vdots \\ w_1 * a_{i1} & \cdots & w_k * a_{ik} \end{bmatrix}$$

, and $Y_0 = [1..j]$ in the SIMPLS.

```

Population ← 0

Pg_best ← 0

c1=c2←2.0

rand←-1~1

P={P1, P2···P10}; %particle 1~particle 10

X={X1,X2···X10}; %Ten weight distributions of k diseases for particle 1~ 10

Pwt={Xi1, Xi2···Xik}; %The weight distributions of k diseases for particle i

V={V1,V2···V10}; %particle velocity for each particle 1~10

Pvelocity={Vi1, Vi2···Vik}; %The changing velocity for weights, xij, in particle i

PopulationSize← 10

For (i = 1 to PopulationSize)

    Pvel ← Random Velocity()

    Pwt(i,k) ← Random Position(PopulationSize)

    Pp_best ← Pwt(i,k)

    % Cost(p)←is Fitness function from GA-PLS

    If (Cost(Pp_best) ≅ Cost(Pg_best))

        Pg_best ← Pp_best

    End

End

End

While (~StopCondition())

    For (P ∈ population)

        If (Pwt(i,k) = Pg_best)

            Pvelocity ← 0

        Else

```

```


$$P_{velocity} \leftarrow v(t+1) = w \times P_{velocity}(t) + c_1 \times rand \times (P_{best} - P_{wt}(t)) + c_2 \times$$


$$rand(P_{g\_best} - P_{wt}(t))$$


$$P_{wt}(l,k) \leftarrow P_{wt}(t+1) = P_{wt}(t) + P_{velocity}(t+1)$$

End
If (Cost(Pwt) ≥ Cost(Pp_best))
    Pp_best ← Pwt(i,k)
    If (Cost(Pp_best) ≥ Cost(Pg_best))
        Pg_best ← Pp_best
    End
End
End
End
Return(Pg_best)

```

3.4 找出疾病之間的關係

由前面的方法找出了和中風有關的疾病，但本研究的目的是要找出疾病和中風之間的關係，所以在這邊本研究使用關聯規則，找出上面的疾病做為可能有關之屬性，而在關聯規則中一般都會設置門檻值，如最小支持度及最小信賴度，但由於設定門檻值太過主觀，且會影響到輸出的結果，所以在這邊我們提出了一個不用設門檻值的方法 Apriori-IFM，只需要設所需的規則數量即可，以下為 Apriori 演算法和 Apriori-IFM 流程圖：

3.4.1 Apriori 演算法

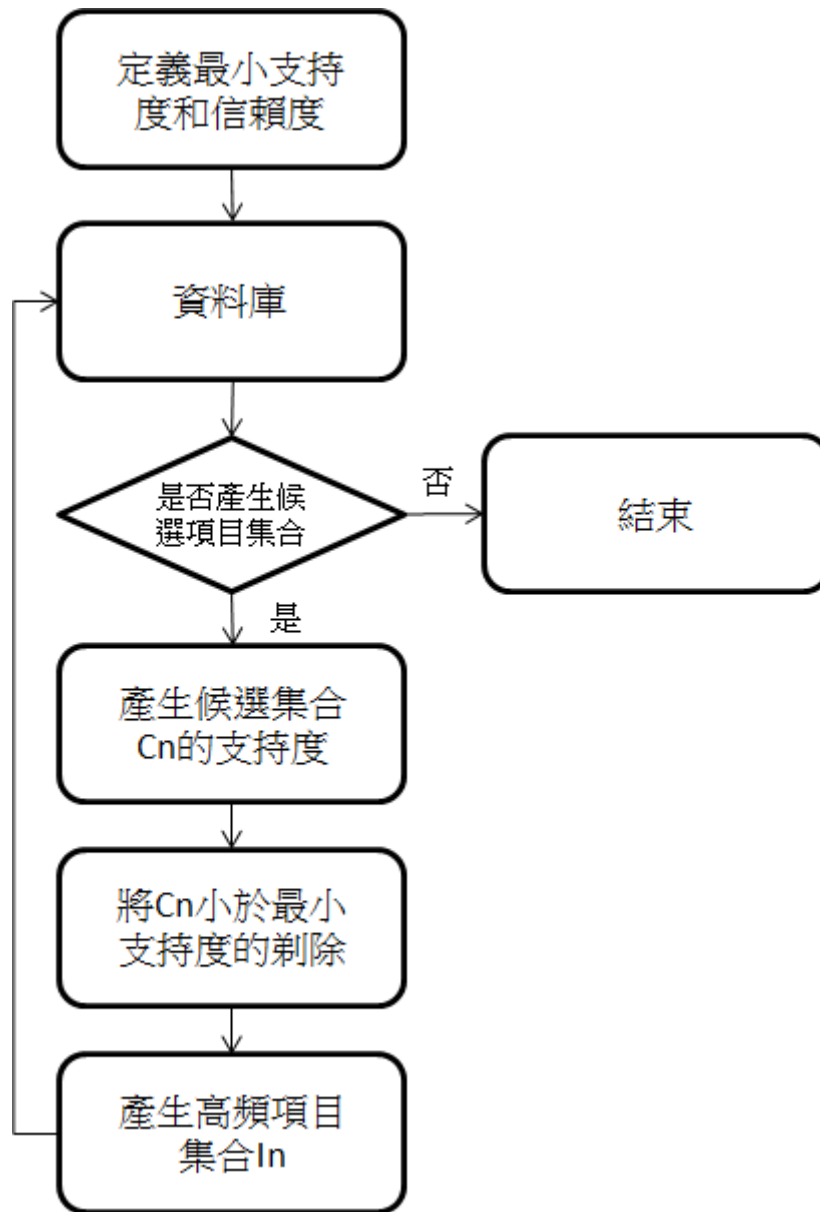


圖 3.7 Apriori 流程圖

一開始先定義最小支持度和信賴度，這是挑選候選集合的重要指標，當候選集合的支持度大於最小支持度時，則把這些候選集合當成高頻項目集合，之後再由資料庫讀取候選集合的支持度，找出高頻項目集合，並利用這些高頻項目集合的結合，產生出新的候選集合，之後再一直重複上述動作，直到生產不出新的候

選集合為止。

支持度的公式為：

$$\text{Support}(A, B) = P(A \cap B) \quad (5)$$

信賴度的公式為：

$$\text{Confidence}(A \rightarrow B) = P(B | A) = P(A \cap B) / P(A) \quad (6)$$

例子假設我有 4 筆病人的資料，疾病十種，如下表 3.5：

病人	疾病基因編碼
病人 A	1111000000
病人 B	0000111110
病人 C	0000010111
病人 D	0110000010

表 3.5 病人資料(關聯規則)

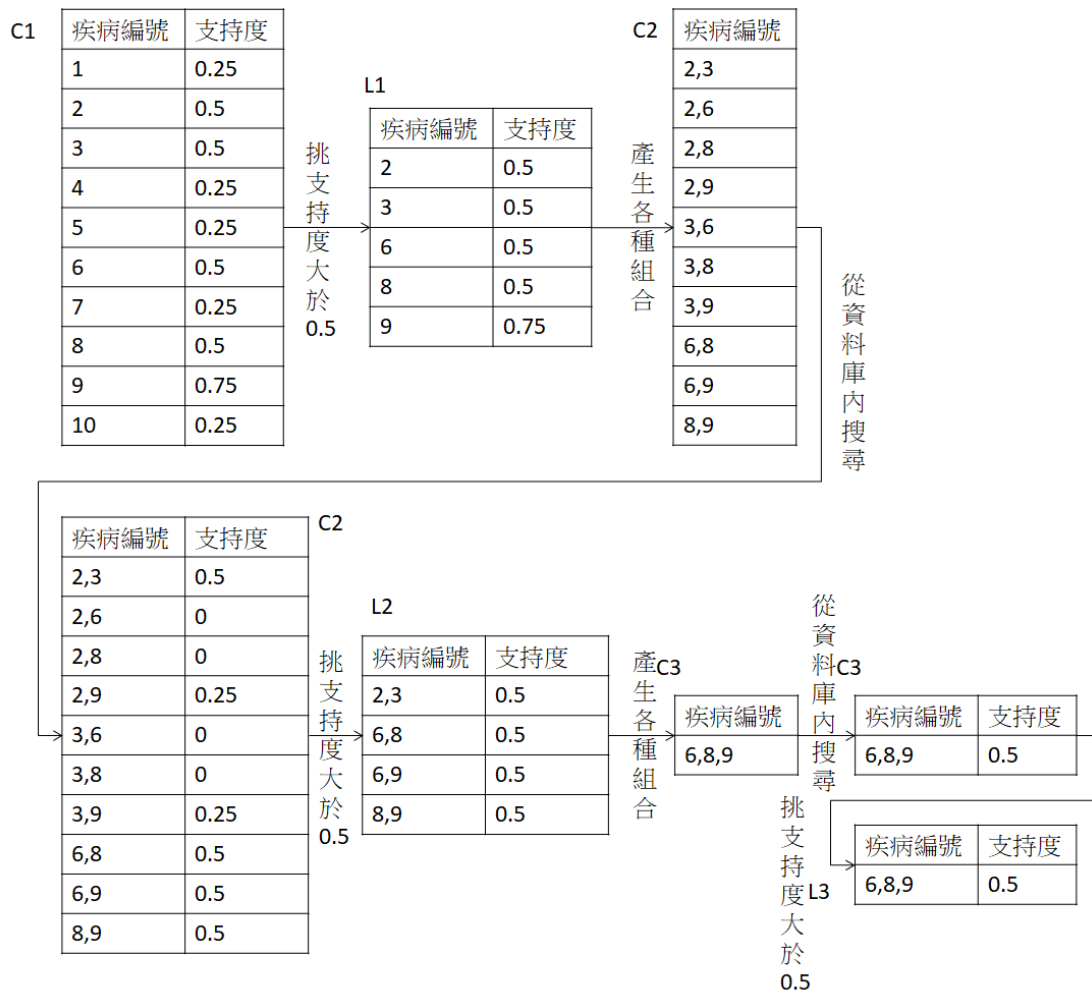


圖 3.8 Apriori 演算法過程

1. 首先計算每一項的支持度如 C1，接著挑出支持度為 0.5（研究者自訂）以上的疾病如 L1。
 2. 再以 L1 為集合，經過排列組合產生 C2 及支持度；再挑出支持度 2 以上的疾病組合如 L2。
 3. 再以 L2 為集合，經過排列組合產生 C3 及支持度；再挑出支持度 2 以上的疾病組合如 L3。
 4. 以此類推，直到無任何項目集合產生為止。
- 最後輸出 I 的非空子集合有 {2}、{3}、{6}、{8}、{9}、{2, 3}、{6, 8}、{6, 9}、{8, 9}、{6, 8, 9}，之後去對 I 集合做排列組合，再根據所定義的信賴度(假設為

0.5)來輸出關聯規則，如： $\{2\} \Rightarrow \{2, 3\}$ 信賴度為 100%，這個就輸出； $\{3\} \Rightarrow \{6, 8, 9\}$ 信賴度為 0%，這個就不輸出。

3.4.2 Apriori-IFM

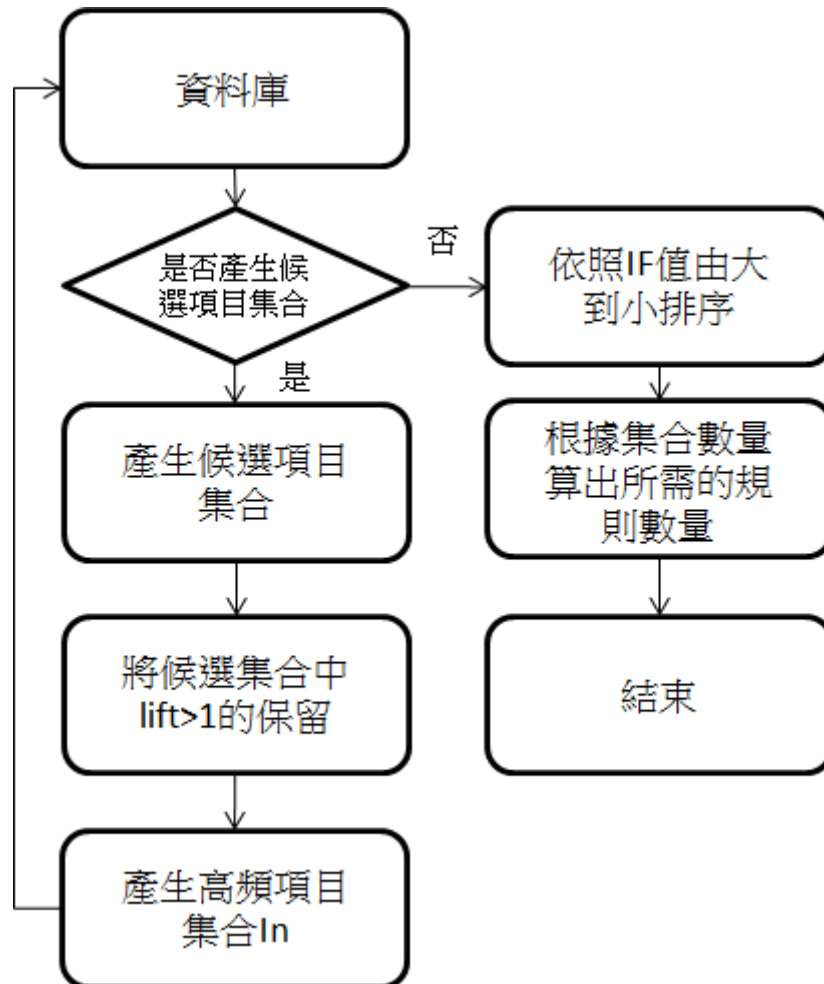


圖 3.9 Apriori-IFM 流程圖

一開始根據資料產生候選集合，去計算每個候選集合 lift(增益值)，保留 lift 大於 1 的候選集合，利用找出的集合結合出新的集合，之後再一直重複上述動作，直到生產不出新的候選集合為止，和 Apriori 演算法的差別在於不使用支持度改成增益值，然後依據 IF(importance factor)由大到小做排序，輸出所需的規則數量。

規則數量是依據關聯規則中所找到所有集合數所去做的計算，依統計中 Dillman 於 2000 提出的公式，去算得最小需多少樣本數，而在這邊變成需多少規則數，計算方式如下：

$$N_s = \frac{N_p * p * (1 - p)}{(N_p - 1) * \left(\frac{B}{C}\right)^2 + p * (1 - p)} \quad (7)$$

N_s 是須完成的樣本數； N_p 是所有集合數； p 是母群體異質性程度，通常設 $p=0.5$ （此為最保守策略，因為變異數最大的情況會發生在母體兩半互為不同類別）； B ：可容忍的抽樣誤差，通常設 0.03（抽樣誤差正負 3%）； C ：可接受的信賴區間（信心水準），通常設 1.96（該數值為可接受的信賴區間 95% 所對應的 Z 分數）。

Lift(增益值)的公式：

$$\text{Lift}(A \rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A) * P(B)} \quad (8)$$

IF(重要性係數)的公式：

$$\text{IF}(A \rightarrow B) = \text{Support}(A, B) * \text{Confidence}(A \rightarrow B) \quad (9)$$

其中 IF 值是輸出規則的重要依據，所以是計算支持度和信賴度相乘，依照值的大小由大到小排序，一次輸出直到所需的最小規則數為止。lift 值是代表此項目集對於完整資料集有趣程度的分數。先取得兩個同時發生之項目的機率，然後除以兩個單獨出現之項目的機率，就可以計算出增益值，大於 1 為正相關，等於 1 為無相關，小於 1 為負相關。因此，我們可以得知有些關聯規則是無用的，如高信賴低支持。

產生完關聯規則後，去判斷和刪減，當關聯規則產生出迴圈時，先比較支持度的大小，刪除較小支持度的規則，但如果規則有兩筆以上，則比較信賴度大小，刪除信賴度最小的規則，使他不產生迴圈，因為本研究假定疾病關聯是一個有向無循環的關係圖。

3.5 貝氏網路

依據上面的關聯規則整理出來的結果，我們可以得到一個有向無循環圖的模型，由於在進行貝氏網路前，必須先建立一個有效的網路，而在本研究中，我們把從關聯規則所得出的模型當成一個有效網路。

在貝氏網路中，當條件獨立成立時，聯合機率分布[42]即為所有條件機率的乘積為：

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1})$$

$$= \prod_{i=1}^n P(X_i | \text{pa}(X_i)) \quad \text{pa}(X_i) \text{ 是 } X_i \text{ 的父節點} \quad (10)$$

貝氏網路可以使預測的工作更加容易，因為我們可以給定一些已知的變數和欲求得的結果。一開始不知道目標事件 $\tilde{\theta}$ 的真實狀態，但知道 $\tilde{\theta}$ 服從機率分布 $P(\tilde{\theta})$ ，稱為事前機率。當得到新的樣本資訊或證據E後，可以根據貝氏定理更新事後機率 $P(\tilde{\theta} | E)$ 。

$$P(\tilde{\theta} = \theta_j | E) = P(\theta_j | E) = \frac{P(\theta_j \cap E)}{P(E)} = \frac{P(E|\theta_j) * P(\theta_j)}{\sum_{k=1}^m P(E|\theta_k) * P(\theta_k)} \quad (11)$$

$$P(\theta_j \cap E) = P(\theta_j | E) * P(E) = P(E|\theta_j) * P(\theta_j) \quad (12)$$

$$P(E) = P(E|\theta_1) * P(\theta_1) + P(E|\theta_2) * P(\theta_2) + \dots + P(E|\theta_m) * P(\theta_m)$$

$$= \sum_{j=1}^m P(E|\theta_j) * P(\theta_j) \quad (13)$$

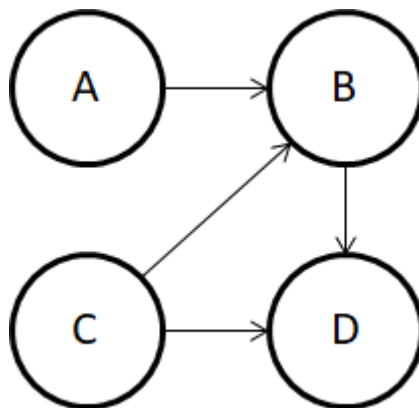


圖 3.10 疾病關係圖

其做法為，首先依據資料整理出整個網路的經驗聯合機率分配，再根據整個網路的經驗聯合機率分配，求出各個節點的聯合機率分配，再利用節點的邊際機率和聯合機率去計算節點的事後機率，最後依照公式(10)即可得出貝氏網路的聯合機率。

假設我們有 4 種疾病，他們的關係如上圖 3.10：

先去計算各節點的機率 $P(A)$ 、 $P(B)$ 、 $P(C)$ 、 $P(D)$ ，再去計算各節點的條件機率，A 疾病的條件機率為 $P(A)$ ，B 疾病的條件機率為 $P(B|A, C)$ ，C 疾病的條件機率為 $P(C)$ ，D 疾病的條件機率為 $P(D|B, C)$ ，其中 D 疾病的條件機率根據公式

(11)可寫成 $P(D|B, C) = \frac{P(B, C, D)}{P(B, C)} = \frac{P(B|C, D) * P(C|D)}{P(B, C)}$ ，因為是貝氏網路所以條件獨立

$P(A, B) = P(A) * P(B)$ ，根據公式(10)計算聯合機率分布：

$$\begin{aligned} P(A, B, C, D) &= P(A) * P(B|A, C) * P(C) * P(D|B, C) \\ &= P(A) * \frac{P(A, B, C)}{P(A, C)} * P(C) * \frac{P(A, B, C, D)}{P(A, B, C)} \\ &= P(A, B, C, D) \end{aligned}$$

所以最後就可以求得貝氏網路的聯合機率分布。

例子：假設我們有 10 筆病人資料，4 種疾病，疾病關聯如上圖 3.10，資料如下表 3.6：

表 3.6 病人資料(貝氏)

	A	B	C	D
病人 1	1	0	1	0
病人 2	0	0	1	0
病人 3	1	1	0	1
病人 4	1	0	0	0
病人 5	1	1	0	1
病人 6	0	0	1	0
病人 7	0	0	0	1
病人 8	1	0	1	0
病人 9	0	1	0	0
病人 10	0	0	1	0

依據病人資料整理出整個網路的經驗聯合機率分配如下表 3.7：

表 3.7 聯合機率分配

情況	A	B	C	D	機率
1	0	0	0	1	0.1
2	0	0	1	0	0.3
3	0	1	0	0	0.1
4	1	0	0	0	0.1
5	1	0	1	0	0.2
6	1	1	0	1	0.2

接著用表 3.7 和圖 3.10 建構出節點的聯合機率分配如下圖 3.11：

A	P(A)
0	0.5
1	0.5

C	P(C)
0	0.5
1	0.5

A	B	C	P(A,B,C)
0	0	0	0.1
0	0	1	0.3
0	1	0	0.1
0	1	1	0
1	0	0	0.1
1	0	1	0.2
1	1	0	0.2
1	1	1	0

B	C	D	P(B,C,D)
0	0	0	0.1
0	0	1	0.1
0	1	0	0.5
0	1	1	0
1	0	0	0.1
1	0	1	0.2
1	1	0	0
1	1	1	0

圖 3.11 節點的聯合機率分配

運用貝氏公式、節點的邊際機率和聯合機率去計算節點的條件機率如下圖

3.12 :

P(A)	
A=0	A=1
0.5	0.5

P(C)	
C=0	C=1
0.5	0.5

		P(B A,C)	
		B=0	B=1
A=0	C=0	0.5	0.5
A=0	C=1	1	0
A=1	C=0	0.33	0.67
A=1	C=1	1	0

		P(D B,C)	
		D=0	D=1
B=0	C=0	0.5	0.5
B=0	C=1	1	0
B=1	C=0	0.33	0.67
B=1	C=1	0.5	0.5

圖 3.12 條件機率表格

最後依據聯合機率分布公式計算出貝氏網路的聯合機率如下表 3.8 :

表 3.8 貝氏網路的聯合機率

編號	A	B	C	D	機率
1	0	0	0	0	0.0625
2	0	0	0	1	0.0625
3	0	0	1	0	0.25
4	0	0	1	1	0
5	0	1	0	0	0.04125
6	0	1	0	1	0.08375
7	0	1	1	0	0
8	0	1	1	1	0
9	1	0	0	0	0.04125
10	1	0	0	1	0.04125
11	1	0	1	0	0.25
12	1	0	1	1	0
13	1	1	0	0	0.055275
14	1	1	0	1	0.112225
15	1	1	1	0	0
16	1	1	1	1	0

依據貝氏網路最後得出的聯合機率，當我們知道如果這位病人有得過哪些疾病時，我們便可預測病人未來可能會得到那些疾病，也就是我們可以去驗證，當拿已知資料時，可以去驗證這個模型預測的準確率。訓練資料集包含 n 筆資料， $i=1, 2, \dots, n$ ，訓練資料有 m 個類別 $\tilde{\theta} = \{\theta_1, \theta_2, \dots, \theta_m\}$ ，定義第 i 筆資料中 k 個屬性的觀察值為 $E_i = \{E_{i1}, E_{i2}, \dots, E_{ik}\}$ ，公式為：

$$P(\theta_j | E_1, E_2, E_3 \dots E_k) = \frac{P(E_1, E_2, E_3 \dots E_k | \theta_j) * P(\theta_j)}{P(E_1, E_2, E_3 \dots E_k)} \quad (14)$$

$$P(E_1, E_2, E_3 \dots E_k | \theta_j) = \prod_{l=1}^k P(E_l | \theta_j) \quad (15)$$

$$P(E_1, E_2, E_3 \dots E_k) = P(E^*) = \sum_{j=1}^m P(E^* | \theta_j) * P(\theta_j) \quad (16)$$

$$P(\theta_j | E_1, E_2, E_3 \dots E_k) = \frac{\prod_{l=1}^k P(E_l | \theta_j) * P(\theta_j)}{\sum_{j=1}^m \prod_{l=1}^k P(E_l | \theta_j) * P(\theta_j)} \quad (17)$$

$\tilde{\theta} = H$ 為有興趣的目標，在這邊就是中風變量； E_i 為有關證據，也就是病人的過去所得過的相關疾病。

3.6 ROC 曲線

再依據公式(17)所求得出的機率，判斷是否分類正確，計算出真陽性 (TP)：診斷為有，實際上也有中風病患的數量；偽陽性 (FP)：診斷為有，實際卻沒有中風病患的數量；真陰性 (TN)：診斷為沒有，實際上也沒有中風病患的數量；偽陰性 (FN)：診斷為沒有，實際卻有中風病患的數量。

TPR：在所有實際為陽性的樣本中，被正確地判斷為陽性之比率。

$$TPR = \frac{TP}{TP + FN} \quad (18)$$

FPR：在所有實際為陰性的樣本中，被錯誤地判斷為陽性之比率。

$$FPR = \frac{FP}{FP + TN} \quad (19)$$

然後去計算 ROC 曲線下面積：

$$ACC = (TP + TN) / (P + N) \quad (20)$$

$P=TP+FN$ ， $N=FP+TN$ ，AUC 是假若隨機抽取一個陽性樣本和一個陰性樣本，分類器正確判斷陽性樣本的值高於陰性樣本之機率，所以當 AUC 的值越大時，代表這個分類器正確率越高。把這些資料統計起來，就可以知道這一個模型它的正確性有多少是否可信。

3.7 卡方檢定

對 PSO 找出有影響性較高的疾病作卡方檢定，來比較兩種不同的方法，是否有其差異或相似。

變項是病人得到 PSO 找出有影響性較高的疾病的有無，和有無中風的判斷，卡方值的計算如下：

$$\sum x_{i,j}^2 = \frac{(O - E)^2}{E} \quad (21)$$

O 為觀察值的次數，E 為期望次數，期望次數的計算是以每行與每列之交乘值除以總數(Total)便得到期望值，例如： $[(A+B)*(A+C)]/Total$ 為 A 之期望值，如下表 3.9。

表 3.9 卡方檢定

		Disease(B)		
		yes	no	total
Disease(A)	yes	A	B	A+B
	no	C	D	C+D
	total	A+C	B+D	A+B+C+D

其假說檢定如下：

H0: χ^2 為 0

H1: χ^2 不為 0

自由度(df)=(A-1)* (B-1)，A 與 B 為行列分組數目。

第四章 實作

4.1 資料

本研究使用榮總健保資料庫的資料，2000 年到 2013 年的資料，把資料分成有得過中風的病人，和沒得過中風的病人，共兩組，且由於疾病資料的種類過多，例如：中風，在疾病中就有數十種不同的稱呼，所以在一開始必須先把性質相近的疾病分成同一類，這樣才不會導致最後的結果非常奇怪，最後疾病變量共有 2770 個。

在這裡需要注意的是，取得過中風的病人資料時，只能記錄這位病人在得到中風之前的疾病，不能紀錄他得了中風之後又得了甚麼疾病，因為本研究是要探討那些疾病是會影響中風的發生，和這些疾病和中風之間的關係去作探討。

以及需要掃除年齡小於 18 歲和大於 110 歲的病人，和性別不詳的病人，這樣定的原因是，通常藥品臨床試驗選擇受試者時，一般都會將年齡下限定為十八歲以下，其主要原因是美國規定滿十八歲為成人，法律上已經可以自行決定參與試驗，並簽具受試者同意書，法律效力不成問題，所以本研究也依照這個慣例，至於上限定為一百一十歲，則主要是為了剔除一些較為極端的個案，因為目前人類壽命上限為一百一十五歲，所以本研究就取了一個整數值一百一十歲，來確保資料不會出現極端的點。由之後需要做貝氏網路，所以把年齡這個變數分成兩組，分別為青壯年(18~65 歲)和老年(65 歲以後)。

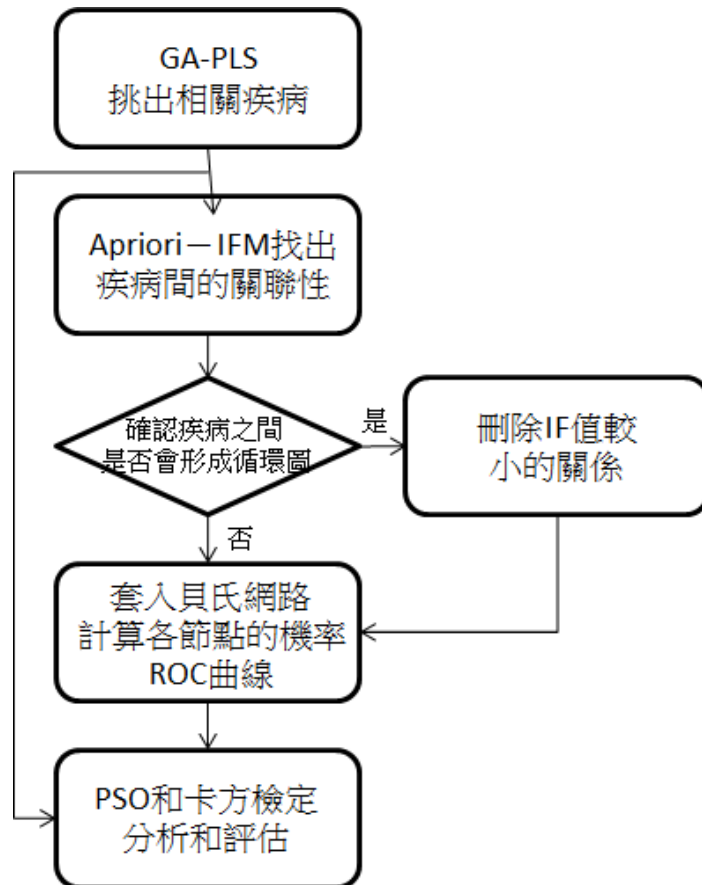


圖 4.1 實驗流程圖

以及上一次所使用的缺血性中風和出血性中風的資料，來比較對於改良過後的 GA-PLS 方法是否有較好的結果。

4.2 實驗結果

4.2.1 缺血性轉出血性

在這邊主要是根據前一篇會議論文，探討改進過後的 GA-PLS 和改進前的 GA-PLS 比較兩者之間的結果，所以在這邊所使用的資料處理方法是和上一篇一樣，和現在的資料前處理方法不太一樣。

使用預先整理好的兩組樣本，分別是只有缺血性中風的患者，和缺血性中風轉出血性中風的患者，在兩組樣本中個隨機取出 1000 筆當作訓練資料，使用 GA-PLS 的方法，迭代 500 次。

如下圖 4.2 和圖 4.4，n 是只有缺血性中風的患者資料，h 是缺血性中風轉出血性中風的患者資料，左邊是在每一個迭代過程中所計算出的適應函數，值介於 0 到 1 之間，越高代表分類的越好，在 250 代時突然飆高是因為我們把一些離群值給移除了，是根據前文所提的異常值來做判斷，使它不影響分類的結果。圖 4.3 是為改良前的。

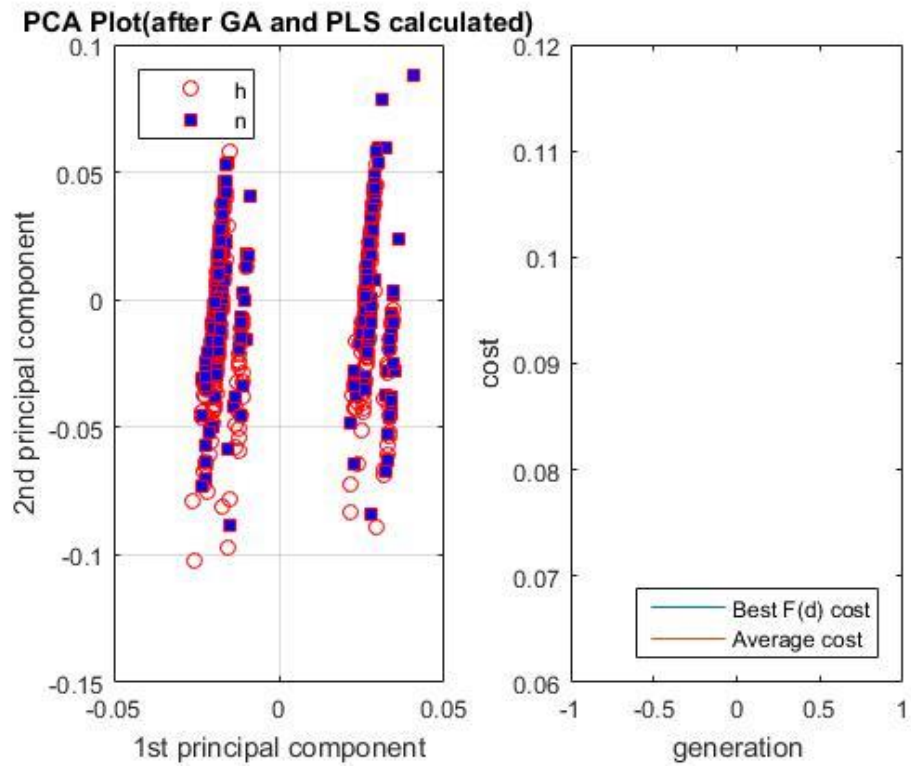


圖 4.2 缺血轉出血分群資料前(GA-PLS)

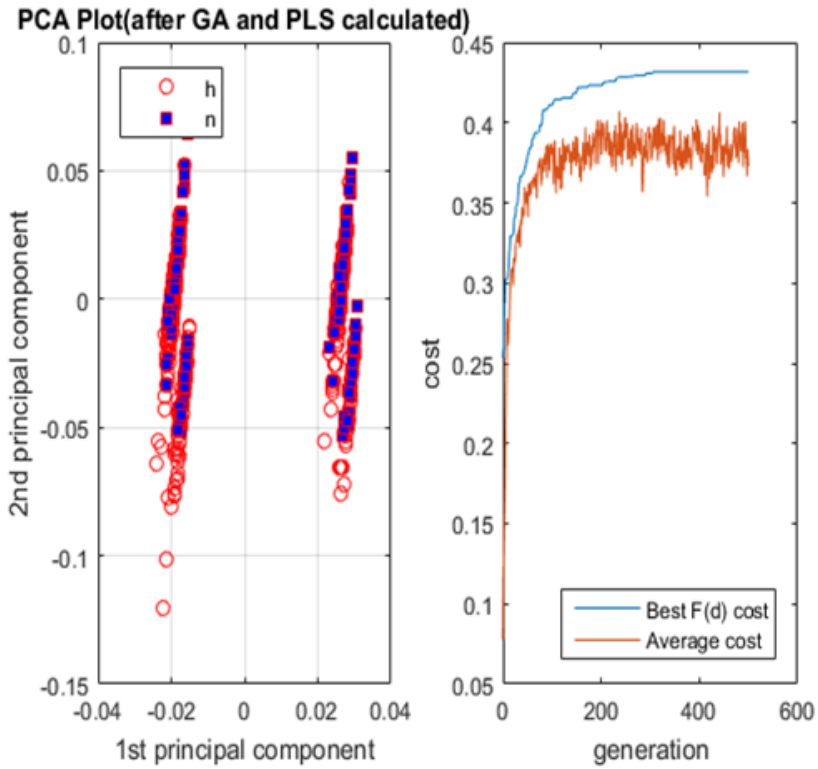


圖 4.3 缺血轉出血分群資料後(GA-PLS 改良前)

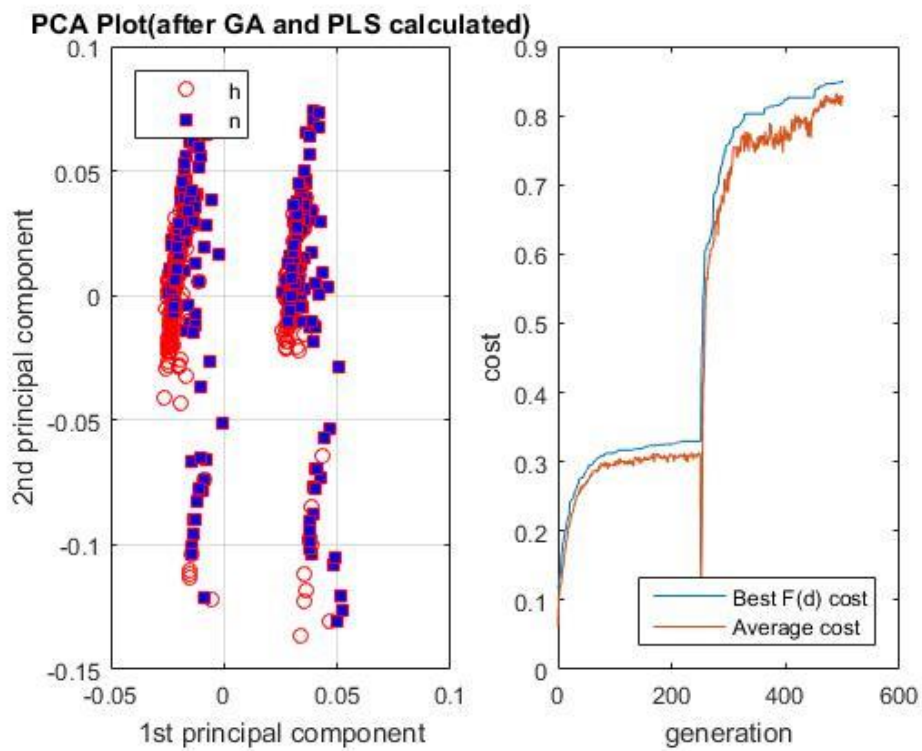


圖 4.4 缺血轉出血分群資料後(GA-PLS 改良後)

接著把測試資料各 100 放入模型中，由於本來就是已知的資料，所以是去計算資料在該模型中被分對的準確度，結果為下表 4.1：

正確率	只有缺血性	缺血性轉出血性	平均
改良前	0.6	0.66	0.63
改良後	0.71	1	0.855

表 4.1 缺血轉出血 GA-PLS 的正確度

和之前的結果相比，發現使用改良過後的 GA-PLS 結果好很多，原本準確率只有 0.63，現在有 0.855，所以代表著改良過後的 GA-PLS 是比原本的好。

接著使用 PSO 來找出影響中風較高了因子，資料沿用 GA-PLS。

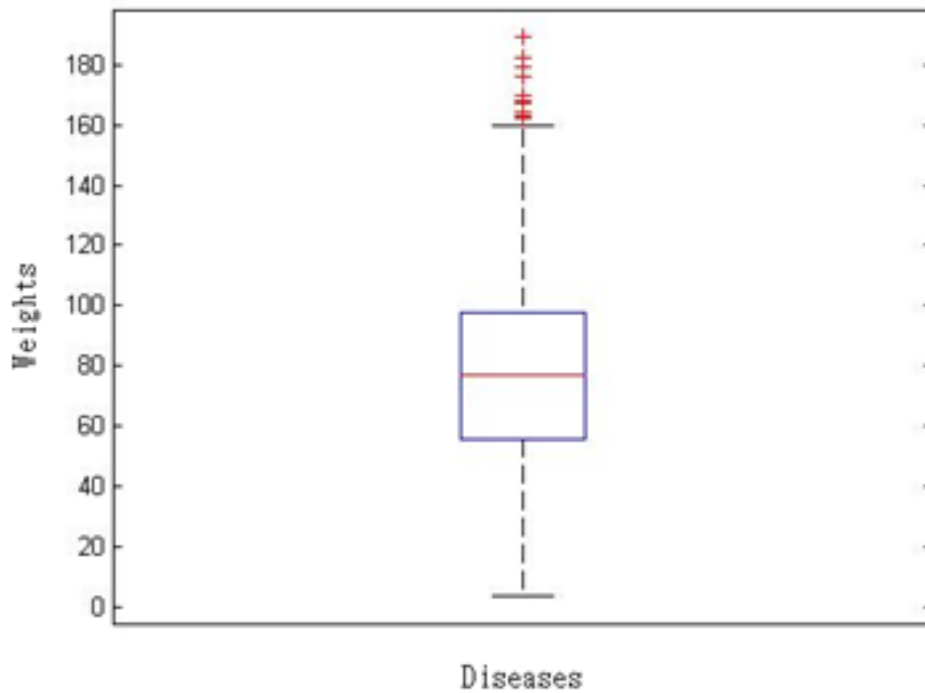


圖 4.5 PSO 權重箱型圖(缺血轉出血)

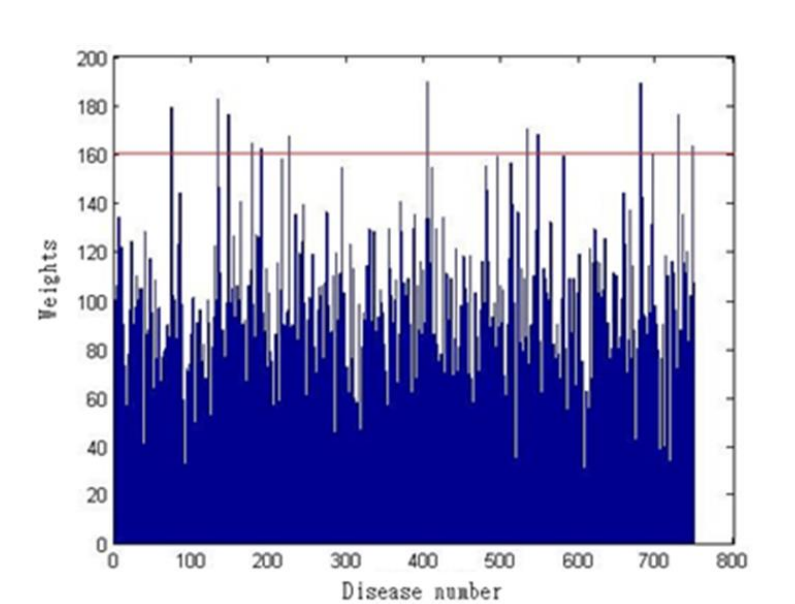


圖 4.6 PSO 長條圖(缺血轉出血)

下表 4.2 是用 PSO 挑權重，透過離群點來挑出影響較高的疾病：

7804	Dizziness and giddiness
5640	Constipation
585	Chronic renal failure
401	*Essential hypertension
25090	Diabetes with unspecified complication, Type II [non-insulin dependent type][NIDDM type][adult-onset type] or unspecified type, not stated as uncontrolled
78009	Other alteration of consciousness
2722	Mixed hyperlipidemia
30000	Anxiety state, unspecified
7291	Myalgia and myositis, unspecified
6000	Hypertrophy (benign) of prostate
40210	Benign hypertensive heart disease without congestive heart

表 4.2 缺血轉出血的相關因子

從表 4.2 中可以發現到頭暈、便秘、慢性腎衰竭、高血壓、糖尿病、高脂血症、焦慮、肌肉痛、前列腺肥大等，是會影響缺血性中風轉出血性中風的。

接者使用卡方檢定來驗證看該疾病是否和缺血轉出血有顯著相關，如下表 4.3：

表 4.3 卡方檢定(缺血轉出血)

疾病編號	卡方檢定	p-value
7804	0.875	0.349
5640	0.522	0.469
585	1.046	0.306
401	4.646	0.031
25090	0.218	0.640
78009	0.522	0.469
2722	1.915	0.166
30000	0.522	0.469
7291	0.522	0.469
6000	1.915	0.166
40210	1.915	0.166

可以發現對缺血轉出血較有顯著影響的是疾病編號 401 高血壓疾病，而在目前的醫療報告中高血壓和高血糖是促發的危險因子，和上面結果比對發現是有符合的，代表著方法找出的結果，可信度還是挺高的。

4.2.2 中風

我們將中風資料依年齡和性別做不同組的實驗，分為全年齡、65 歲以下、65 歲以上、65 歲以下男性、65 歲以下女性、65 歲以上男性、65 歲以上女性、

累計疾病次數(全年齡)，不再是以 0, 1 資料來表達，而是病人得到該疾病次數來表達，並對於 PLS 降維取多少維度來比較看看是取多少較好，這裡分成 2 維、5 維(目前最後使用的維度)和 10 維，看在對不同資料組成中是否有哪些疾病會對中風造成影響。

我們從兩組中風資料(有中風合肥有中風)中隨機挑選各 1000 筆資料當成訓練資料，再從兩組資料中隨機挑出，不含訓練資料的各 100 筆資料當成測試資料。

4.2.2.1 Seq2seq

本研究用另外一種方法 Seq2seq，來比較看哪一種比較好，只針對全年齡的資料。這個方法的資料型態和 GS-PLS 方法不太一樣，由於 Seq2seq 一般是用來做自動翻譯語句和回答機器人，所以是先輸入一串句子或對話，接者再輸入一串上一個句子的翻譯或者是接續上一段對話的下一個對話。在這裡本研究把上句變為一位病人所得過的所有疾病，當然中風是不會放在裡面的，下句變為那一位病人是否由得過中風的判斷。也就是說我要輸入一段病人的病歷，讓它告訴我這位病人是否有中風，如下圖 4.7：

```

1 1 425 521 524 556 604 904 958 994 1021 1025 2686
2 no
3 1 278 315 317 329 347 398 433 435 439 441 442 456 469 470 484 486 521 525 526 538 556 596 600 602
604 605 634 642 652 654 664 667 668 671 691 692 693 704 705 806 812 814 815 817 822 873 877 879
881 882 939 973 1025 1026 1027 1037 2634 2698 2704
4 no
5 1 422 424 434 514 520 521 522 523 524 525 526 533 535 538 548 589 590 591 604 627 667 671 688 693
700 704 769 773 791 792 806 812 873 879 880 882 944 982 1006 1027 2652 2654 2700 2704
6 no
7 1 93 130 251 315 317 347 35393 403 417 424 426 456 457 469 484 505 520 525 526 538 550 556 589
590 591 595 600 601 602 604 605 627 634 644 676 765 769 775 779 781 792 794 802 805 809 812 814
817 873 875 877 879 880 881 882 883 889 918 919 942 946 980 991 992 1025 1026 1027 1100 2634 2698
8 no
9 1 93 180 278 347 424 431 523 525 526 550 588 589 590 591 593 596 602 605 627 634 645 646 688 693
694 769 774 775 810 811 814 815 817 879 944 1027 2693 2698 2704
10 no
11 1 589 591 605 828 873
12 no
13 1 17 131 251 315 424 425 427 431 456 520 524 525 538 550 588 589 591 593 596 604 605 624 634 642
664 775 801 804 806 809 812 815 816 873 877 879 946 982 1026 1034 2686
14 no

```

圖 4.7 seq2seq 資料樣本結構

在這裡空格和空隔間為一個疾病的編號，no 和 have 則是告訴它這個病人是屬於哪一類。

Seq2seq 的訓練資料是有無中風病人各 5000 筆資料，測試資料是各 1000 筆，這次的模型參數 encoder and decoder 各 1 layer，畢竟我只要一個輸入（病人資料）和一個輸出（判斷是否為中風）即可，Hidden size in encoder and decoder 總共有 512 layers，Batch size 有 64 個，Learning rate 有 0.0001 的成功率，Train the model with it iterations 是 5000 次，總共迭代 5000 次。

測試時也是使用此概念，先輸入一段未在訓練時所放入的資料，接者依據我們原本就知道的答案，比對一下它給出的答案和原本的答案是否一致，看它答對的比例是多少，來判斷這個方法是否可靠。下圖 4.8 測試後所輸出的部分結果：

1 , have <EOS> 1 , no <EOS>
2 , have <EOS> 2 , no <EOS>
3 , no <EOS> 3 , no <EOS>
4 , have <EOS> 4 , no <EOS>
5 , no <EOS> 5 , no <EOS>
6 , no <EOS> 6 , no <EOS>
7 , have <EOS> 7 , no <EOS>
8 , have <EOS> 8 , no <EOS>
9 , have <EOS> 9 , no <EOS>
10 , no <EOS> 10 , no <EOS>
11 , no <EOS> 11 , have <EOS>
12 , have <EOS> 12 , no <EOS>
13 , have <EOS> 13 , have <EOS>
14 , no <EOS> 14 , no <EOS>
15 , no <EOS> 15 , no <EOS>
16 , no <EOS> 16 , no <EOS>
17 , have <EOS> 17 , no <EOS>
18 , have <EOS> 18 , no <EOS>
19 , have <EOS> 19 , have <EOS>
20 , have <EOS> 20 , have <EOS>
21 , have <EOS> 21 , no <EOS>
22 , have <EOS> 22 , have <EOS>
23 , have <EOS> 23 , no <EOS>
24 , have <EOS> 24 , no <EOS>
25 , have <EOS> 25 , have <EOS>

圖 4.8 seq2seq 測試後輸出結果(左)有中風(右)沒有中風

圖 4.8 左邊是有中風的部分測試資料結果，右邊是沒有中風的部分測試資料結果，最後經過統計整理過後，如下表 4.4：

	有中風	沒有中風	平均
正確率	0.846	0.777	0.811

表 4.4 seq2seq 的準確度

4.2.2.2 GA-PLS

GA-PLS 的方法迭代 500 次，如下圖 4.9 到圖 4.18，n 是沒有中風的病患資料，h 是有中風的病患資料，並對不同組都各做一次。

圖中只顯示了經過 PLS 降為處理過後影響最大的兩個變數，和每一代的適應函數，這可以使我們看出這個方法分類情形，由圖 4.9 以看到這兩類的病人一開始分類的適應函數只有 0.625，但和做完 GA-PLS 可以比較明顯的發現，適應函數至少大於 0.85，這可證明說這個方法是有用的。其中在 250 代時突然飆高的原因和在做缺血性轉出血性時是一樣的，剔除了一些離群值，使得它不影響分類結果。

接者我們再把測試用的那各 100 筆資料，放入該模型中，去判斷該病人是屬於哪一類，是有中風還是沒有中風的，但因為我們原本就知道這些測試資料的類別，所以可以根據這些類別來確認這個方法的正確率，最後的結果如下表 4.5：可以看到不管是哪一組資料基本上都有 7、8 成的準確度，有十分強的分辨效果。

再接者比對一下，降維過程中 PLS 是要取幾維是較為準確的，我們這裡測試了三種情形，分別是 2 維、5 維和 10 維，在表 4.5 中可看到 2 維和 5 維的結果是差不多的，但在 10 維的結果中是較差強人意的，且我們知道取越多維，對於原本的原始資料解釋是更多的，故後續研究皆是用 PLS 取 5 維的方式做實驗。

正確率	有中風	沒有中風	平均
全年齡(PLS 5 維)	0.85	0.85	0.85
65 歲以下	0.79	0.86	0.825
65 歲以上	0.70	0.85	0.775
65 歲以下男性	0.90	0.86	0.88
65 歲以下女性	0.71	0.87	0.79
65 歲以上男性	0.78	0.86	0.82
65 歲以上女性	0.84	0.88	0.86
累計次數(全年齡)	0.78	0.87	0.825
PLS 2 維	0.88	0.83	0.855
PLS 10 維	0.64	0.88	0.76

表 4.5 GA-PLS 的準確度

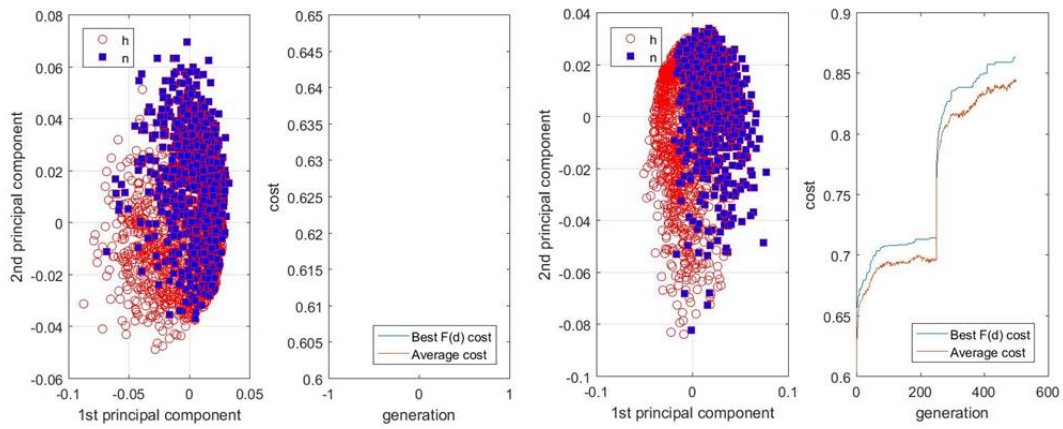


圖 4.9 GA-PLS 分群資料(全年齡 PLS 5 維)

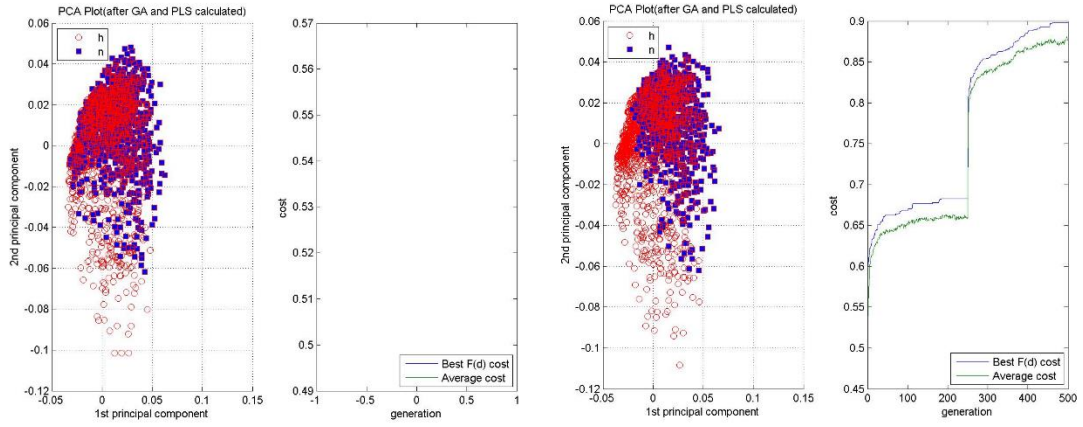


圖 4.10 GA-PLS 分群資料(65 歲以下)

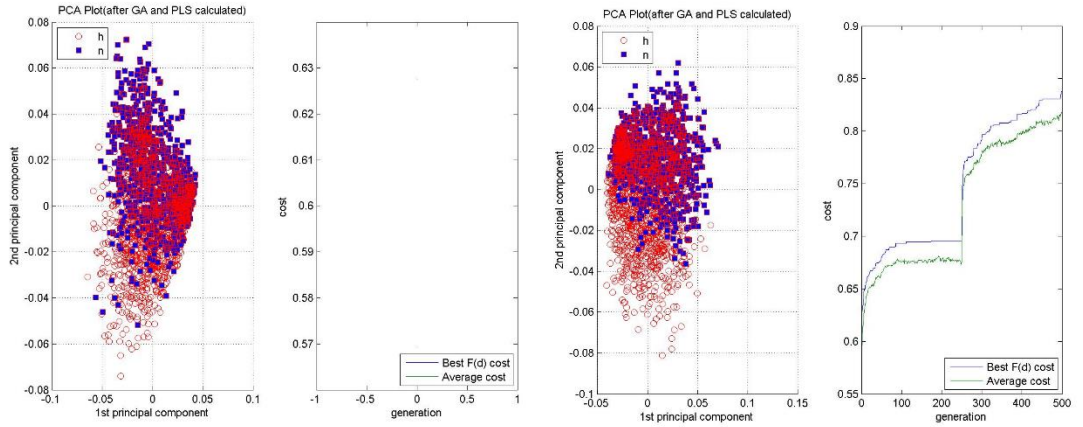


圖 4.11 GA-PLS 分群資料(65 歲以上)

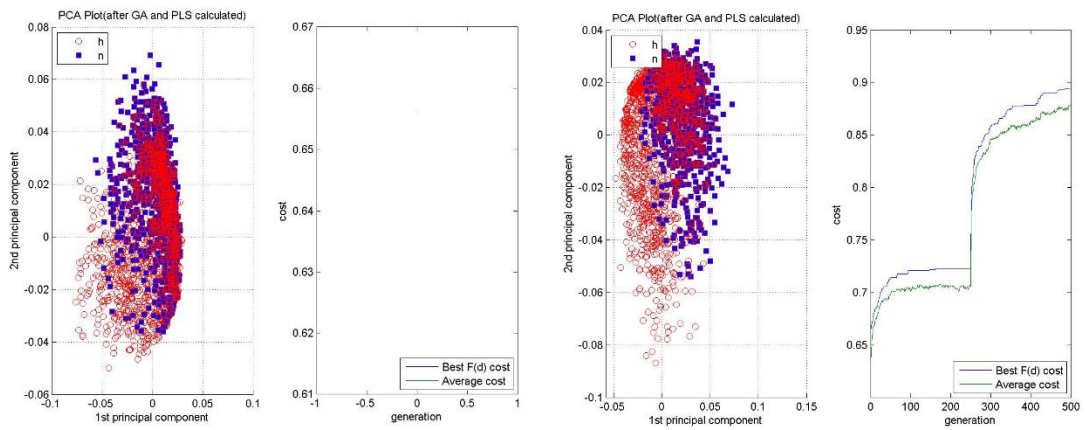


圖 4.12 GA-PLS 分群資料(65 歲以下男性)

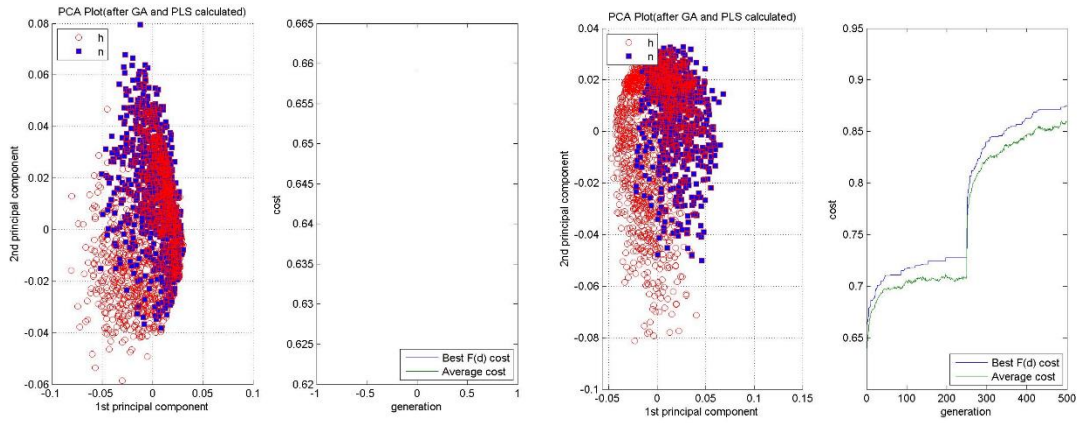


圖 4.13 GA-PLS 分群資料(65 歲以下女性)

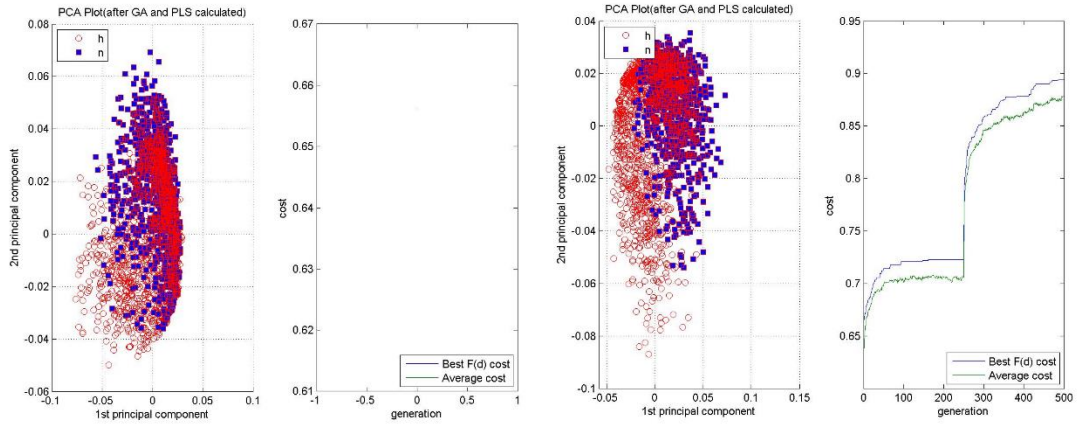


圖 4.14 GA-PLS 分群資料(65 歲以下男性)

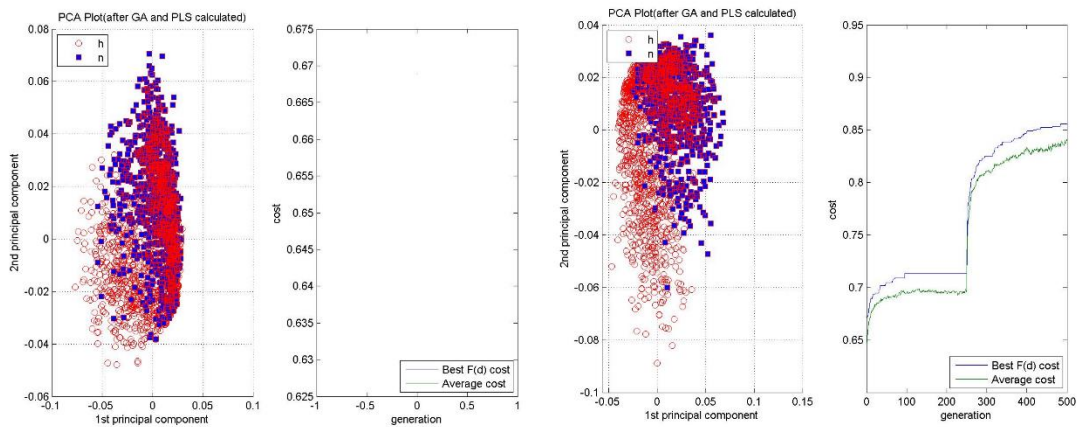


圖 4.15 GA-PLS 分群資料(65 歲以下女性)

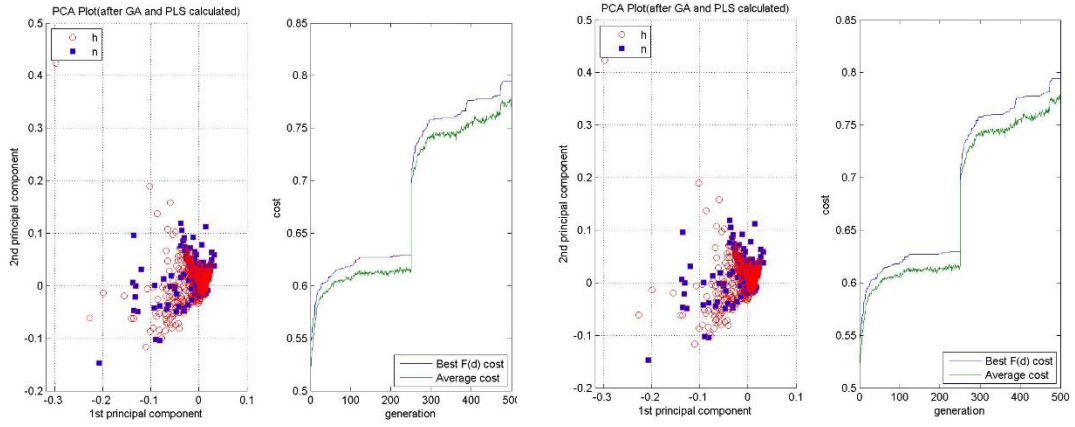


圖 4.16 GA-PLS 分群資料(累計次數)

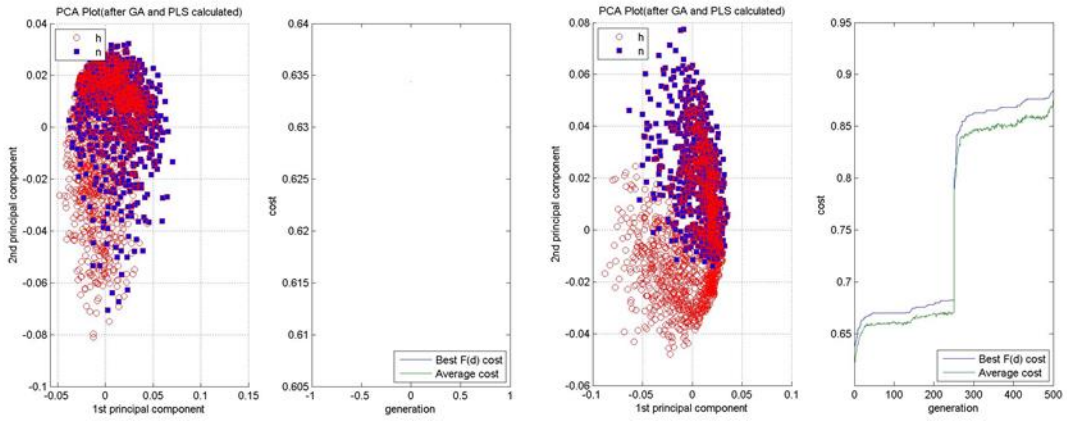


圖 4.17 GA-PLS 分群資料(PLS 2 維)

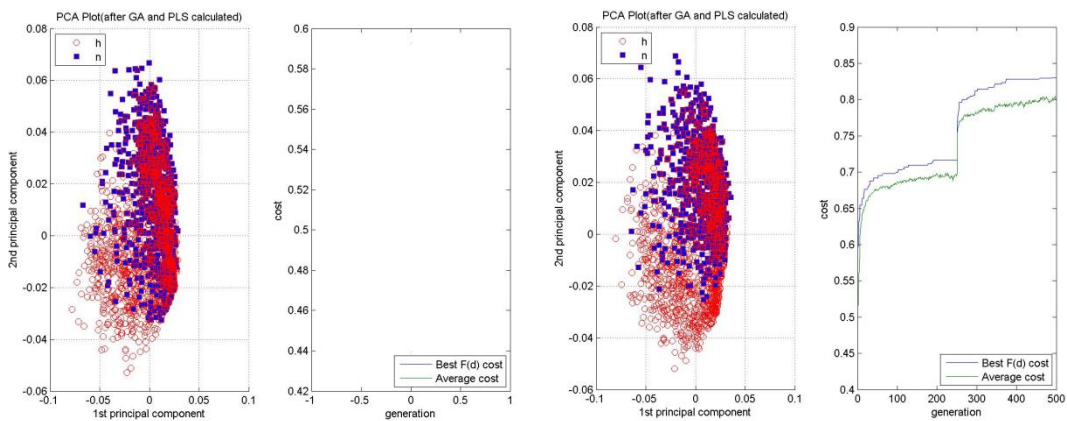


圖 4.18 GA-PLS 分群資料(PLS 10 維)

4.2.2.3 GA 後的 PSO

之後使用 PSO 找疾病權重，資料格式完全沿用之前的 GA-PLS，所輸入的資料格式為 0,1 資料，並設 30 個資料點，總共跑 200 代，找出權重後使用統計的四分法，找出離群點，因為這些疾病皆有可能是影響中風疾病之一。而在這邊只有對全年齡、65 歲以下和 65 歲以上的組別去做，因為後來發現對這三組來說，皆沒有找到目前有列出的中風相關因子，結果如下圖 4.19~21 和表 4.6~4.8：

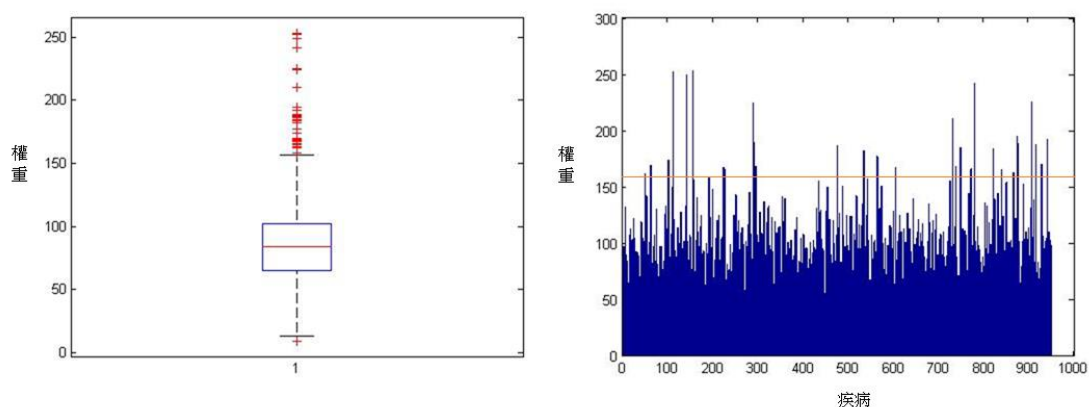


圖 4.19 GA 後 PSO(全年齡)

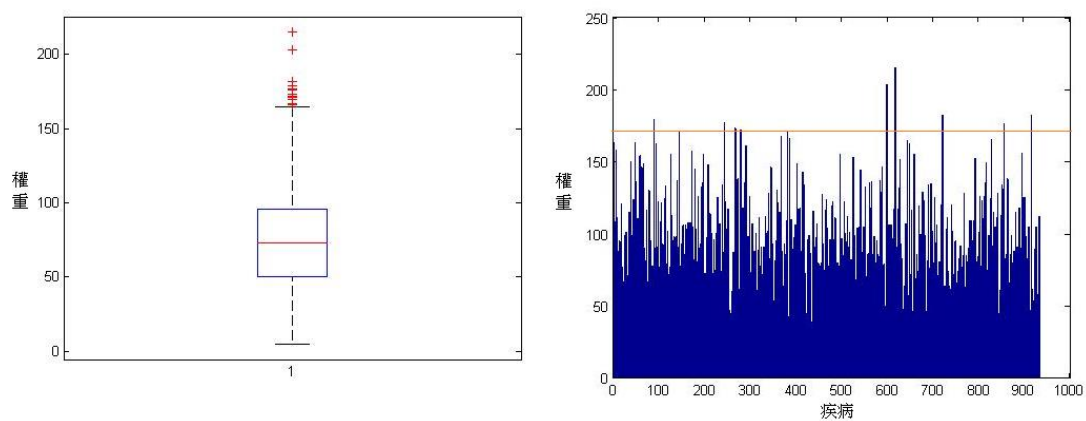


圖 4.20 GA 後 PSO(65 歲以下)

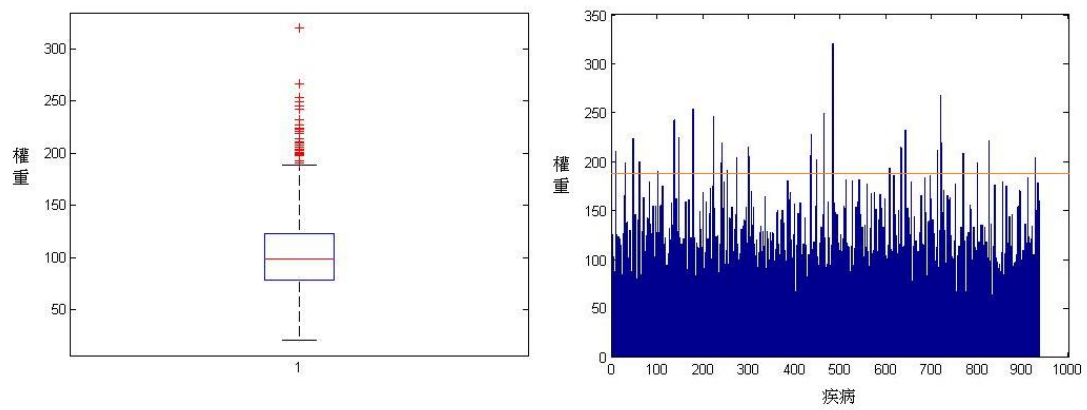


圖 4.21 GA 後 PSO(65 歲以上)

如表 4.6，找到的大部分是都是腫瘤、精神疾病、潰瘍和開放傷口等疾病，與目前中風相關因子較無明顯相關。接著對這些疾病做卡方檢定，來看是否和中風有顯著關係。由下表可發現有接近一半的疾病 p-value 都小於 0.05，可得出這些疾病可能和中風是有顯著關係的。

表 4.6 GA-PSO 中風的相關因子(全年齡)

疾病 編號	疾病名稱	卡方 檢定	p-value
103	*Pinta	41.14	1.415e-10
114	*Coccidioidomycosis	3.67	0.055
143	*Malignant neoplasm of gum	1.95	0.162
157	*Malignant neoplasm of pancreas	96.92	7.212e-23
193	Malignant neoplasm of thyroid gland	17.18	3.399e-05
225	*Benign neoplasm of brain and other parts of nervous system	18.38	1.805e-05
228	*Hemangioma and lymphangioma, any site	3.148	0.076
290	*Senile and presenile organic psychotic conditions	0.57	0.447
293	*Transient organic psychotic conditions	12.08	0.0005
296	*Affective psychoses	21.29	3.937e-06
476	*Chronic laryngitis and laryngotracheitis	22.44	2.159e-06
533	*Peptic ulcer, site unspecified	19.02	1.286e-05
534	*Gastrojejunal ulcer	6.88	0.0086
564	*Functional digestive disorders, not elsewhere classified	7.02	0.008
604	*Orchitis and epididymitis	3.34	0.067
730	*Osteomyelitis, periostitis and other infections involving bone	2.31	0.128
737	*Curvature of spine	0.19	0.655
748	*Congenital anomalies of respiratory system	2.33	0.126
771	*Infections specific to the perinatal period	0.011	0.913
779	*Other and ill-defined conditions originating in the perinatal period	0.85	0.354
820	*Fracture of neck of femur	1.71	0.190
838	*Dislocation of foot	1.16	0.279
865	*Injury to spleen	1.16	0.279
872	*Open wound of ear	1.16	0.279
875	*Open wound of chest(wall)	1.16	0.279
905	*Late effects of musculoskeletal and connective tissue injuries	0.85	0.354
927	*Crushing injury of upper limb	1.16	0.279
939	*Foreign body in genitourinary tract	41.14	1.415e-10

表 4.7 GA-PSO 中風的相關因子(65 歲以下)

疾病 編號	疾病名稱	卡方 檢定	p-value
722	*Intervertebral disc disorders	50.82	1.007e-12
496	Chronic airways obstruction, not elsewhere classified	8.041	0.004
519	*Other diseases of respiratory system	7.65	0.005
816	*Fracture of one or more phalanges of hand	7.83	0.005
355	*Mononeuritis of lower limb	2.59	0.107
447	*Other disorders of arteries and arterioles	0.09	0.758
451	*Phlebitis and thrombophlebitis	1.71	0.190
255	*Disorders of adrenal glands	0.77	0.380
040	*Other bacterial diseases	0.52	0.470
127	*Other intestinal helminthiasis	3.53	0.059
598	*Urethral stricture	0.52	0.470
999	*Complications of medical care, not elsewhere classified	0.84	0.356

如表 4.7，找到的大部分是和椎間盤疾病、呼吸道、骨折、神經炎、血管疾病、腎上腺疾病、細菌疾病、腸道蠕蟲病、尿道狹窄等疾病，與目前中風相關因子較無明顯相關。接著對這些疾病做卡方檢定，來看是否和中風有顯著關係。由下表可發現只有三種疾病 p-value 小於 0.05，可發現在很多種疾病時做 PSO 找權重是時好時壞的。

表 4.8 GA-PSO 中風的相關因子(65 歲以下)

疾病 編號	疾病名稱	卡方 檢定	p-value
715	*Osteoarthritis and allied disorders	5.69	0.017
528	*Diseases of the oral soft tissues, excluding lesions specific for gingiva and tongue	7.97	0.0047
571	*Chronic liver disease and cirrhosis	0.002	0.958
733	*Other disorders of bone and cartilage	3.02	0.082
496	Chronic airways obstruction, not elsewhere classified	83.95	5.068e-20
382	*Suppurative and unspecified otitis media	0.072	0.788
706	*Diseases of sebaceous glands	3.54	0.059
879	*Open wound of other and unspecified sites, except limbs	0.088	0.765
791	*Nonspecific findings on examination of urine	0.0003	0.986
813	*Fracture of radius and ulna	3.47	0.062
821	*Fracture of other and unspecified parts of femur	13.32	0.0002
577	*Diseases of pancreas	4.014	0.045
584	*Acute renal failure	6.83	0.0089
529	*Diseases and other conditions of the tongue	0.43	0.511
456	*Varicose veins of other sites	0.055	0.812
731	*Osteitis deformans and osteopathies associated with other disorders classified elsewhere	0.26	0.609
V71	*Observation and evaluation for suspected conditions not found	0.32	0.567
286	*Coagulation defects	0.075	0.783
198	*Secondary malignant neoplasm of other specified sites	0.18	0.666
220	Benign neoplasm of ovary	0.88	0.345
383	*Mastoiditis and related conditions	0.0015	0.968
140	*Malignant neoplasm of lip	2.083	0.148
502	Pneumoconiosis due to other silica or silicates	0.0007	0.977
508	*Respiratory conditions due to other and unspecified external agents	0.0007	0.977
207	*Other specified leukemia	1.04	0.307
V22	*Normal pregnancy	0.96	0.326

如表 4.8，找到的大部分是和骨、肝、腎、口腔、呼吸、凝血缺陷等疾病，與目前中風相關因子較無明顯相關。接著對這些疾病做卡方檢定，來看是否和中風有顯著關係。由下表可發現只有六種疾病 p-value 小於 0.05，可發現在很多種疾病時做 PSO 找權重是時好時壞的，且找出的疾病與目前中風相關因子較無明顯相關。

4.2.2.4 Apriori-IFM

接著做 Apriori-IFM，找出疾病之間的關聯性，在這邊資料輸入只能有得過中風疾病的病人資料，且資料必須為 0,1 資料。找完關聯規則後必須去除一些會產生迴圈的邊，目的有兩個，其一是去完迴圈後可明顯看出疾病的走向，其二是供後續的貝氏網路使用。圖 4.22 和 4.23 可以看到明顯的區別。

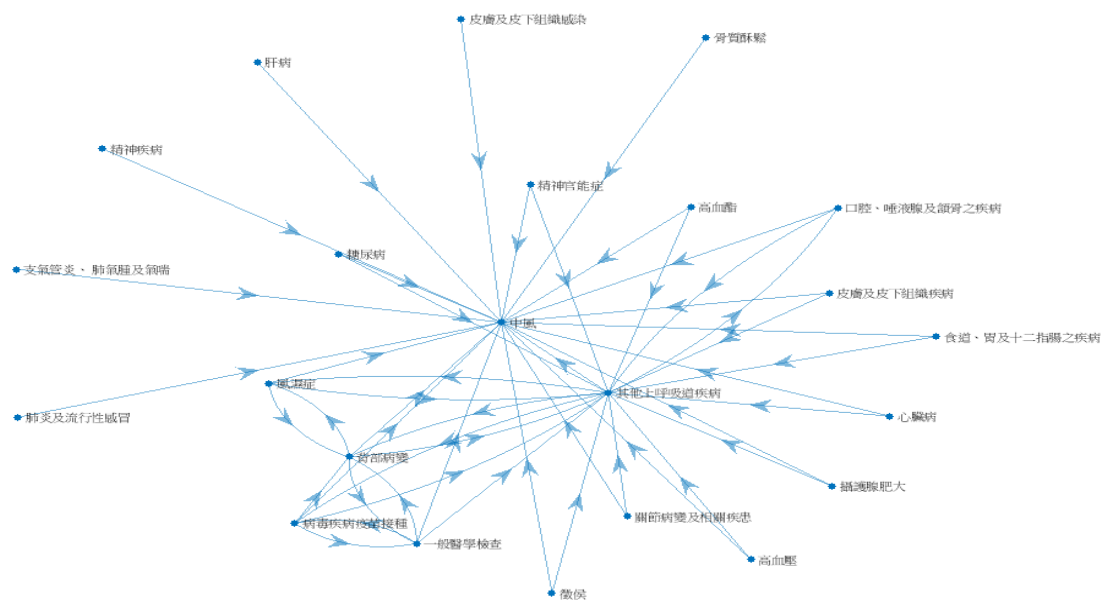


圖 4.22 未刪迴圈前(全年齡)

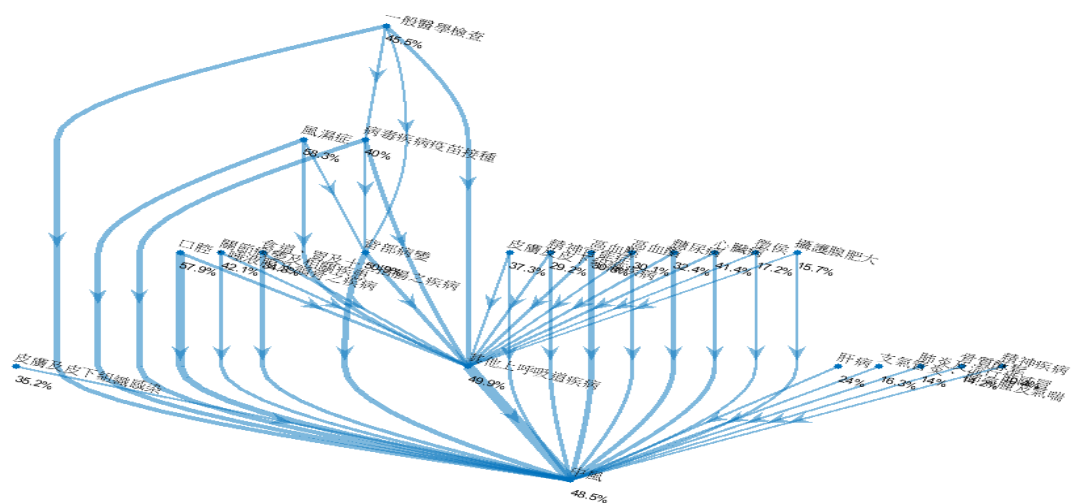


圖 4.23 關聯圖(全年齡)

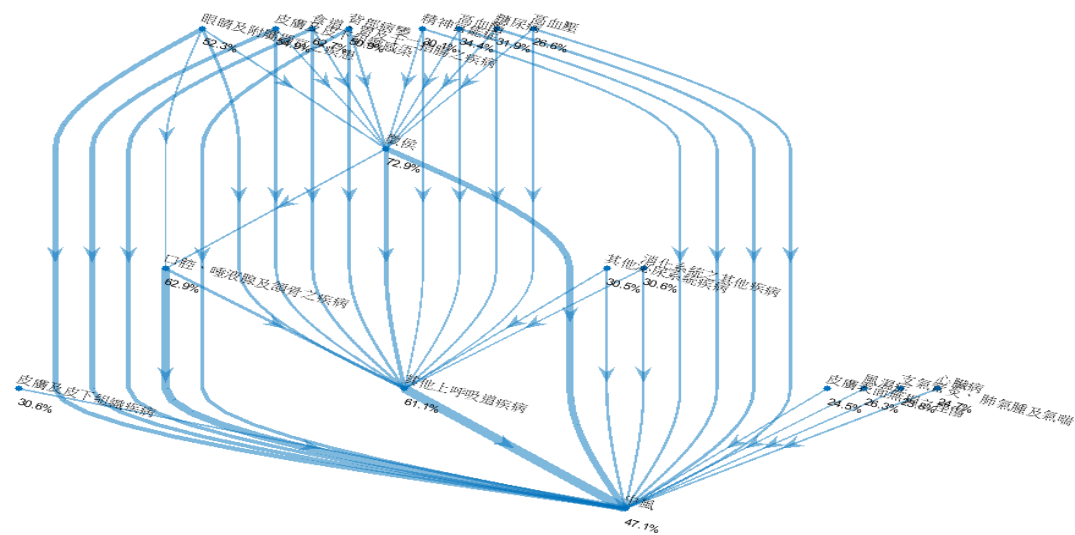


圖 4.24 關聯圖(65歲以下)

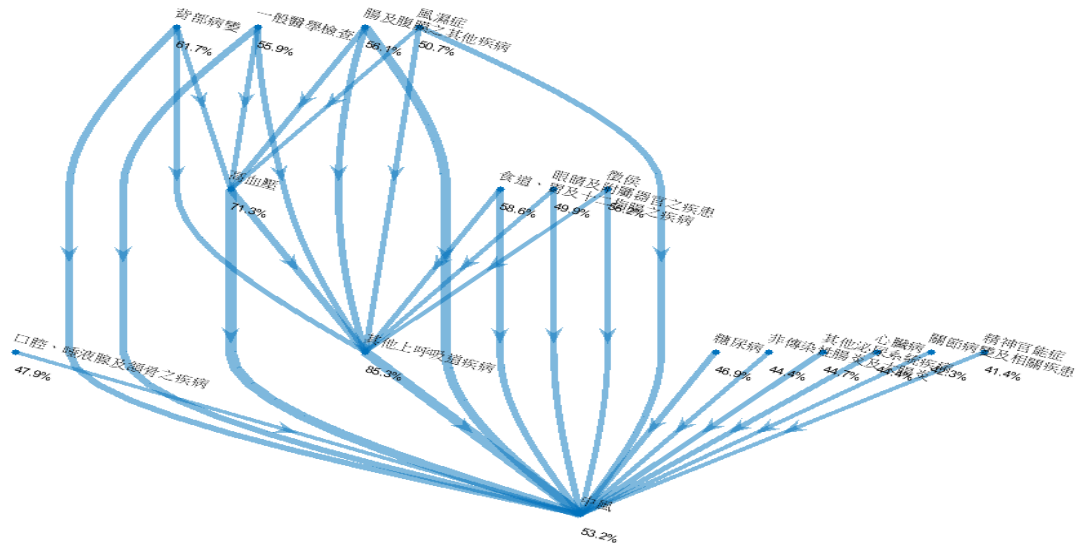


圖 4.25 關聯圖(65歲以上)

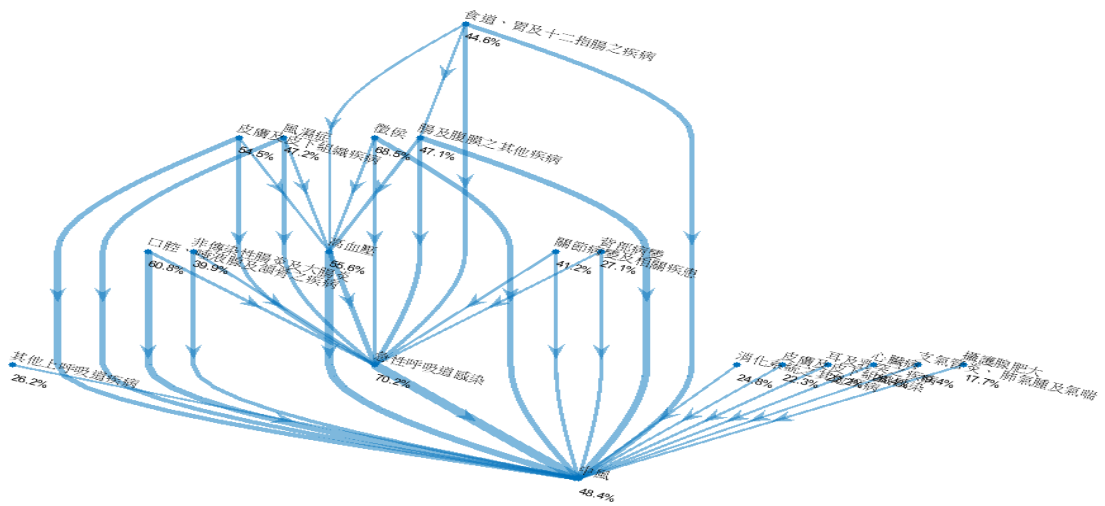


圖 4.26 關聯圖(65歲以下男性)

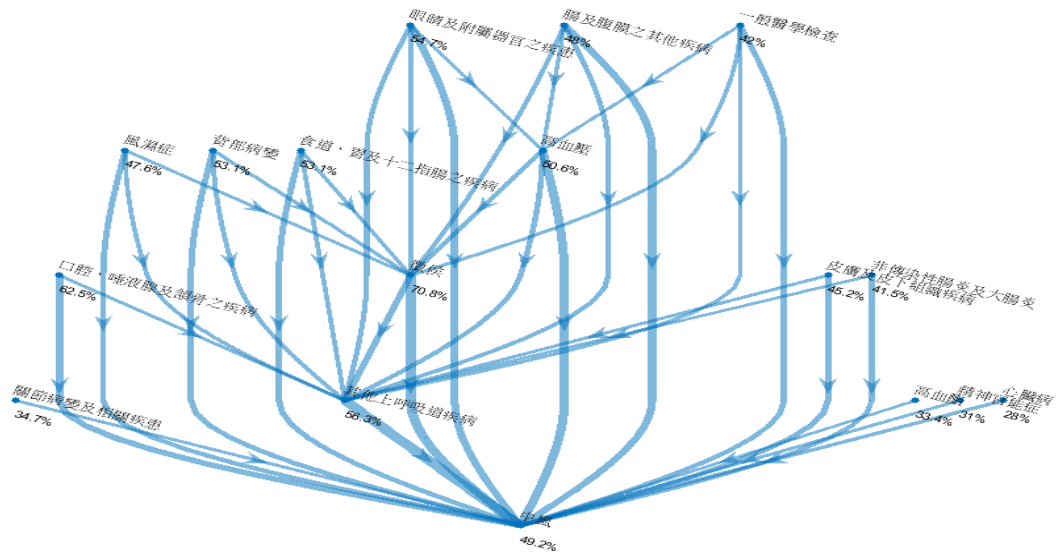


圖 4.27 關聯圖(65歲以下女性)

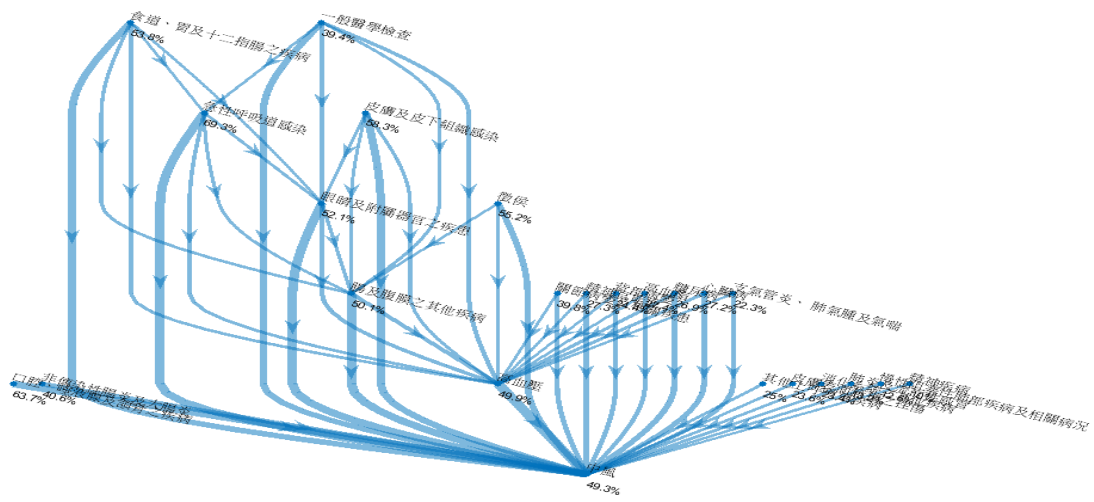


圖 4.28 關聯圖(65歲以上男性)

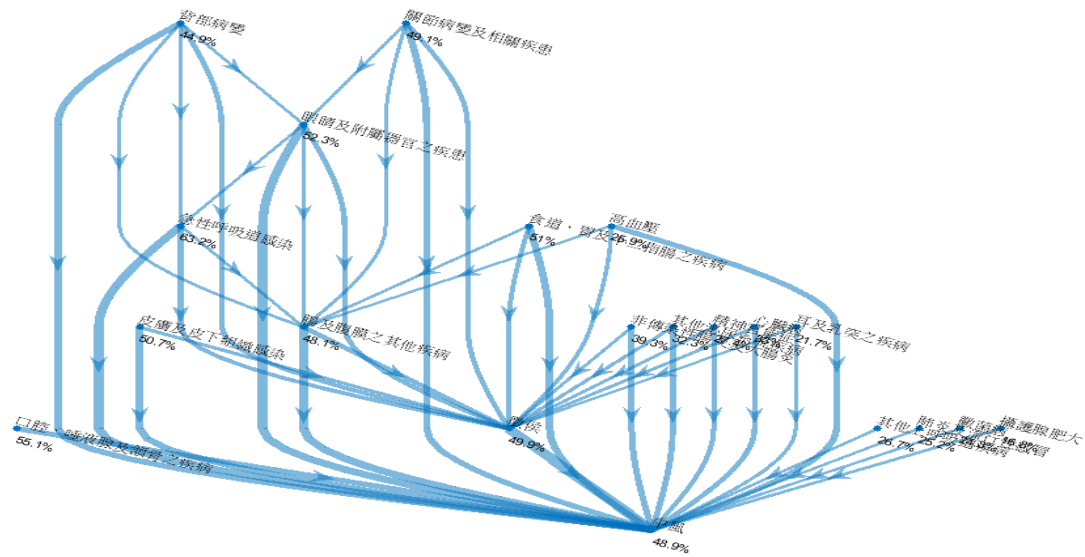


圖 4.29 關聯圖(65 歲以上女性)

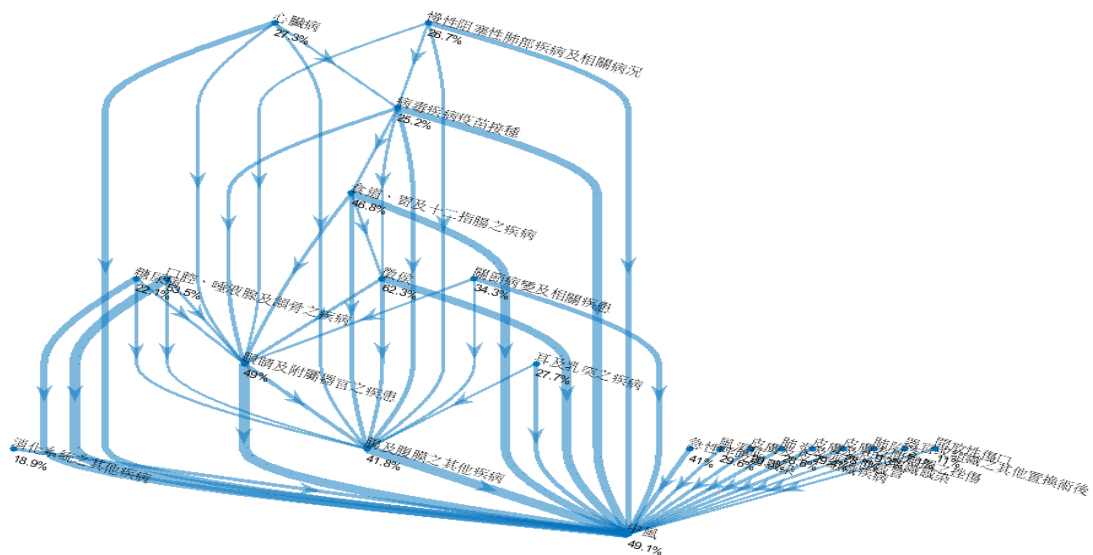


圖 4.30 關聯圖(累計次數)

由上圖 4.23 可發現對全年齡的病人來說大多數會影響中風的疾病都會先經過其他上呼吸道疾病，然後再導致中風。由上圖 4.24 可發現對 65 歲以下的病人來說大多數會影響中風的疾病都會先經過其他上呼吸道疾病和徵候，然後再導致中風。由上圖 4.25 可發現對 65 歲以上的病人來說大多數會影響中風的疾病都會先經過其他上呼吸道疾病和高血壓，然後再導致中風。由上圖 4.26 可發現對 65

歲以下男性的病人來說大多數會影響中風的疾病都會先經過急性呼吸道感染和高血壓，然後再導致中風。由上圖 4.27 可發現對 65 歲以下女性的病人來說大多數會影響中風的疾病都會先經過其他上呼吸道疾病和徵候，然後再導致中風。由上圖 4.28 可發現對 65 歲以上男性的病人來說大多數會影響中風的疾病都會先經過高血壓、腸及腹膜之其他疾病、眼睛及附屬器官之疾病和急性呼吸道感染，然後再導致中風。由上圖 4.29 可發現對 65 歲以上女性的病人來說大多數會影響中風的疾病都會先經過徵候、腸及腹膜之其他疾病、其他上呼吸道疾病和眼睛及附屬器官之疾病，然後再導致中風。由上圖 4.30 可發現對 65 歲以上女性的病人來說大多數會影響中風的疾病都會先經過腸及腹膜之其他疾病、眼睛及附屬器官之疾病、徵候和食道胃及十二指腸之疾病，然後再導致中風。

4.2.2.5 貝氏網路

由於已經從上面的關聯規則中，得到貝氏網路的結構，就是有向無循環圖 (DAG)，接下來整理樣本資料，資料先分成訓練資料和測試資料，訓練資料包含有無中風病人各 30000 筆資料，測試資料包含有無中風病人各 3000 筆資料。

之後把訓練資料中出現過的所有疾病資料列出，再根據它出現的頻率計算出機率，也就是計算整個網路的經驗聯合機率分配，再根據關聯規則中所得到的網路，建構節點的聯合機率分配，之後再依據貝氏公式：

$$P(\theta_j|E) = \frac{P(E|\theta_j) * P(\theta_j)}{\sum_{j=1}^m P(E|\theta_j) * P(\theta_j)} \quad (11)$$

求出各節點的條件機率，最後再依照貝氏網路的數學定義：

$$P(X) = \prod_{i \in I} P(X_i | X_{pa(i)}) \quad (10)$$

我們就可以得到，貝氏網路的條件機率表格 (Conditional probability Table; CPT)。

做完貝氏網路之後，我們把測試用的資料放入貝氏網路中，只把該病人有得

過的疾病放入，然後預測他們得到目標疾病的可能性，可以寫成：

$$P(\theta_j|E_1, E_2, E_3 \dots E_k) = \frac{\prod_{l=1}^k P(E_l|\theta_j) * P(\theta_j)}{\sum_{j=1}^m \prod_{l=1}^k P(E_l|\theta_j) * P(\theta_j)} \quad (17)$$

求出機率後，去看原本測試資料是屬於哪一類，如果是屬於原本就是該疾病的患者，當 $P \geq 0.5$ 時那麼 TP 值加 1，當 $P < 0.5$ 時那麼 FN 值加 1，如果不是屬於原本就是該疾病的患者，當 $P \geq 0.5$ 時那麼 TN 值加 1，當 $P < 0.5$ 時那麼 FP 值加 1，然後算出 TPR、FPR 和 AUC，結果如下表 4.9：

表 4.9 貝氏網路準確度

	全年齡	65 歲以下	65 歲以上	65 歲以下男性	65 歲以下女性	65 歲以上男性	65 歲以上女性	累計次數
TPR(敏感度)	0.78	0.857	0.836	0.5946	0.999	0.811	0.754	0.711
FPR(錯誤命中率)	0.164	0.103	0.354	0.173	0.001	0.163	0.29	0.231
AUC(準確度)	0.8271	0.921	0.7355	0.7164	0.999	0.8642	0.7588	0.7971

在圖 4.31 中可以看到每一組的 ROC 曲線，基本上準確度都在 70% 以上，平均的準確度有 82.74%，特別在 65 歲以下及 65 歲以下女性，準確度達 9 成多，是非常準確的結果。

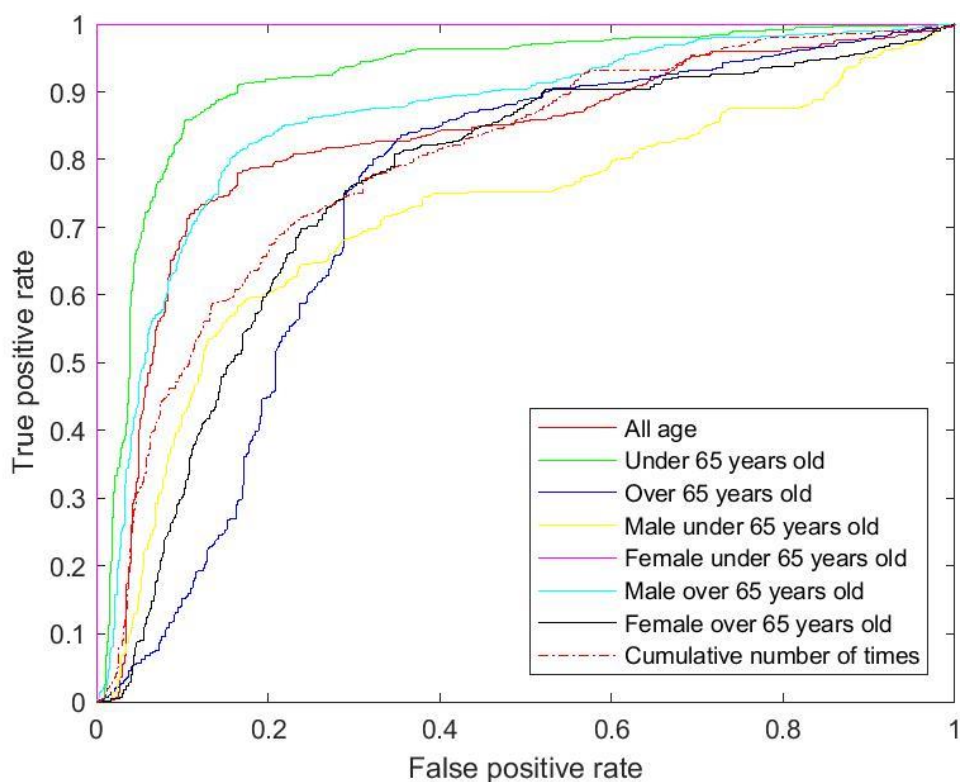


圖 4.31 ROC 曲線

4.2.2.6 關聯規則後 PSO

依據關聯規則所找出的疾病，對它們做 PSO 找出權重大小，來判斷哪些疾病影響程度較高，基本上到這個步驟找出的疾病個數大概落在 15~25 之間，在這裡 PSO 的設定都和前面一樣 30 個資料點，做 200 代。接著對這些疾病做卡方檢定，比較 PSO 權重和卡方檢定是否有差異。

在卡方檢定中，從表 4.10~4.17 中可以看出這幾個疾病對中風都是非常顯著的。除了在全年齡中精神官能症和高血脂是不顯著外其他都是非常顯著；在 65 歲以下中每一個都是非常顯著；在 65 歲以上中其他上呼吸道疾病和食道、胃及十二指腸之疾病是不顯著外其他都是非常顯著；在 65 歲以下男性中腸及腹膜之其他疾病和皮膚及皮下組織感染是不顯著外其他都是非常顯著；在 65 歲以下女性中腸及腹膜之其他疾病、高血脂和精神官能症是不顯著外其他都是非常顯著；

在 65 歲以上男性中腸及腹膜之其他疾病、一般醫學檢查和精神官能症是不顯著外其他都是非常顯著；在 65 歲以上女性中精神官能症是不顯著外其他都是非常顯著；在累計次數中開放性傷口是不顯著外其他都是非常顯著。

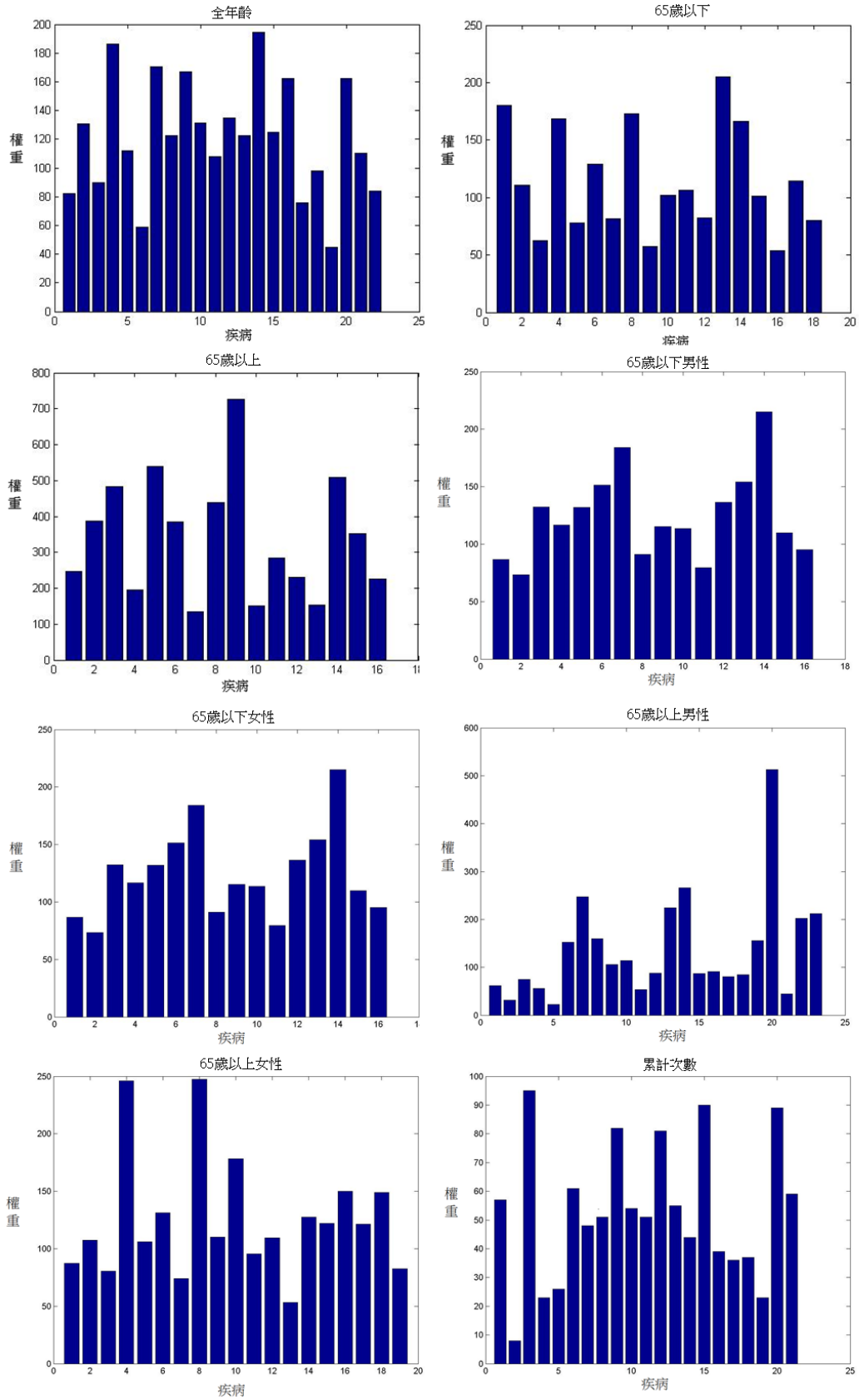


圖 4.32 關聯規則 PSO 權重

表 4.10 關聯規則卡方檢定(全年齡)

疾病名稱	PSO 權重	卡方檢定	p-value
皮膚及皮下組織感染	194.37	1080.02	7.243e-237
背部病變	186.38	2190.209	0
食道、胃及十二指腸之疾病	170.22	63.107	1.957e-15
病毒疾病疫苗接種	167.15	16045.44	0
肺炎及流行性感冒	162.29	9595.12	0
肝病	162.29	563.47	1.473e-124
高血脂	134.89	2.38	0.122
精神官能症	131.07	1.91	0.166
口腔、唾液腺及顎骨之疾病	130.43	33568.901	0
心臟病	124.63	13459.89	0
糖尿病	122.67	8346.009	0
皮膚及皮下組織疾病	122.57	5621.69	0
一般醫學檢查	111.87	50.42	1.238e-12
骨質酥鬆	110.44	3888.64	0
高血壓	108.12	19818.07	0
攝護腺肥大	97.67	10009.96	0
風濕症	89.91	6503.85	0
精神疾病	84.009	43435.38	0
其他上呼吸道疾病	82.1	12766.84	0
徵侯 ¹	75.56	2019.89	0
關節病變及相關疾患	58.48	617.58	2.501e-136
支氣管炎、肺氣腫及氣喘	44.49	771.82	7.208e-170

註 1：代表的是「來自病人的主觀感受」，如頭暈、失眠、幻覺、肢體暫時性麻痺、浮腫等。來源為 ICD-9-CM code 780-789。

表 4.11 關聯規則卡方檢定(65 歲以下)

疾病名稱	PSO 權重	卡方檢定	p-value
消化系統之其他疾病	204.76	789.23	1.180e-173
其他上呼吸道疾病	179.93	13656.59	0
皮膚及皮下組織疾病	172.64	5295.02	0
眼睛及附屬器官之疾患	168.31	8086.03	0
風濕症	166.03	2031.27	0
食道、胃及十二指腸之疾病	128.79	8788.93	0
支氣管炎、肺氣腫及氣喘	114.43	6.18	0.012
口腔、唾液腺及顎骨之疾病	110.81	20409.55	0
精神官能症	106.7	738.07	1.573e-162
皮膚表面無損之挫傷	101.79	3118.104	0
糖尿病	101.29	907.77	2.006e-199
高血酯	82.56	317.61	4.793e-71
背部病變	81.66	5633.67	0
心臟病	79.92	1006.709	6.248e-221
皮膚及皮下組織感染	77.75	10010.24	0
徵候 ¹	62.46	12170.12	0
其他泌尿系統疾病	57.31	1368.43	1.519e-299
高血壓	53.44	1966.88	0

註 1：代表的是「來自病人的主觀感受」，如頭暈、失眠、幻覺、肢體暫時性麻痺、浮腫等。來源為 ICD-9-CM code 780-789。

表 4.12 關聯規則卡方檢定(65 歲以上)

疾病名稱	PSO 權重	卡方檢定	p-value
徵候 ¹	725.5	161.83	4.504e-37
腸及腹膜之其他疾病	539.4	1445.48	2.751e-316
心臟病	507.83	1691.36	0
背部病變	482.73	238.36	8.95e-54
眼睛及附屬器官之疾患	437.74	132.708	1.047e-30
高血壓	387.13	2414.32	0
風濕症	385.53	463.302	9.188e-103
關節病變及相關疾患	352.45	369.29	2.663e-82
糖尿病	283.94	717.07	5.79e-158
其他上呼吸道疾病	247.17	1.8	0.179
非傳染性腸炎及大腸炎	230.94	250.66	1.86e-56
精神官能症	224.59	98.07	4.0374e-23
一般醫學檢查	195.56	576.57	2.089e-127
其他泌尿系統疾病	152.52	760.14	2.495e-167
口腔、唾液腺及頷骨之疾病	150.1	2934.19	0
食道、胃及十二指腸之疾病	133.87	1.54	0.213

註 1：代表的是「來自病人的主觀感受」，如頭暈、失眠、幻覺、肢體暫時性麻痺、浮腫等。來源為 ICD-9-CM code 780-789。

表 4.13 關聯規則卡方檢定(65 歲以下男性)

疾病名稱	PSO 權重	卡方檢定	p-value
其他上呼吸道疾病	126.46	1815.69	0
非傳染性腸炎及大腸炎	119.69	3090.76	0
食道、胃及十二指腸之疾病	118.69	1612.43	0
消化系統之其他疾病	115.43	234.85	5.201e-53
關節病變及相關疾患	105.64	247.0007	1.17e-55
耳及乳突之疾病	85.99	193.38	5.807e-44
背部病變	84.42	111.41	4.806e-26
心臟病	83.93	3407.56	0
攝護腺肥大	83.61	3216.909	0
風濕症	82.72	1571.709	0
皮膚及皮下組織感染	79.7	2203.48	0
腸及腹膜之其他疾病	69.75	0.0016	0.967
口腔、唾液腺及頷骨之疾病	68.08	11608.38	0
皮膚及皮下組織疾病	63.54	1.25	0.262
高血壓	58.68	6951.31	0
急性呼吸道感染	54.82	5256.18	0
支氣管炎、肺氣腫及氣喘	41.24	291.18	2.75e-65
徵候 ¹	40.46	2946.01	0

註 1：代表的是「來自病人的主觀感受」，如頭暈、失眠、幻覺、肢體暫時性麻痺、浮腫等。來源為 ICD-9-CM code 780-789。

表 4.14 關聯規則卡方檢定(65 歲以下女性)

疾病名稱	PSO 權重	卡方檢定	p-value
高血脂	215.18	1.58	0.208
背部病變	184.13	625.85	3.98e-138
關節病變及相關疾患	154.19	332.85	2.296e-74
風濕症	151.39	1274.19	4.58e-279
一般醫學檢查	136.32	11.304	0.0007
徵候 ¹	132.3	964.82	7.946e-212
皮膚及皮下組織疾病	132.09	1469.14	1.98e-321
眼睛及附屬器官之疾患	116.53	1513.44	0
非傳染性腸炎及大腸炎	115.31	2265.85	0
食道、胃及十二指腸之疾病	113.46	267.63	3.715e-60
精神官能症	109.71	1.26	0.26
心臟病	95.15	3528.46	0
腸及腹膜之其他疾病	91.19	0.21	0.643
其他上呼吸道疾病	86.6	4350.57	0
高血壓	79.54	5850.59	0
口腔、唾液腺及頷骨之疾病	73.43	8864.04	0

註 1：代表的是「來自病人的主觀感受」，如頭暈、失眠、幻覺、肢體暫時性麻痺、浮腫等。來源為 ICD-9-CM code 780-789。

表 4.15 關聯規則卡方檢定(65 歲以上男性)

疾病名稱	PSO 權重	卡方檢定	p-value
支氣管炎、肺氣腫及氣喘	512.72	1654.88	0
皮膚表面無損之挫傷	266.23	1071.07	6.404e-235
食道、胃及十二指腸之疾病	247.66	181.8	1.953e-41
精神官能症	224.23	0.51	0.47
精神疾病	212.36	5987.901	0
慢性阻塞性肺部疾病及相關病況	202.59	3357.99	0
徵候 ¹	160.04	1025.05	6.415e-225
心臟病	155.45	4461.04	0
非傳染性腸炎及大腸炎	152.12	2183.78	0
一般醫學檢查	114.34	3.04	0.081
高血壓	105.95	3891.65	0
高血脂症	91.18	13.61	0.0002
其他上呼吸道疾病	88.1	1473.89	1.83e-322
背部病變	87.41	102.78	3.731e-24
糖尿病	84.35	1695.61	0
消化系統之其他疾病	80.82	156.59	6.262e-36
眼睛及附屬器官之疾患	74.9	1288.31	3.90e-282
口腔、唾液腺及頷骨之疾病	62.01	7030.33	0
皮膚及皮下組織感染	55.77	1856.69	0
關節病變及相關疾患	53.74	164.55	1.14e-37
肺炎及流行性感冒	44.44	2221.85	0
急性呼吸道感染	31.59	3759.3	0
腸及腹膜之其他疾病	22.17	3.44	0.063

註 1：代表的是「來自病人的主觀感受」，如頭暈、失眠、幻覺、肢體暫時性麻痺、浮腫等。來源為 ICD-9-CM code 780-789。

表 4.16 關聯規則卡方檢定(65 歲以上女性)

疾病名稱	PSO 權重	卡方檢定	p-value
腸及腹膜之其他疾病	247.27	5.31	0.021
眼睛及附屬器官之疾患	246.2	1107.37	8.249e-243
關節病變及相關疾患	178.21	14.7	0.0001
黴菌病	149.96	269.36	1.563e-60
耳及乳突之疾病	148.89	167.81	2.219e-38
背部病變	131.42	563.52	1.441e-124
高血壓	127.42	3954.44	0
肺炎及流行性感冒	121.94	678.89	1.161e-149
心臟病	121.43	2719.52	0
非傳染性腸炎及大腸炎	110.17	1783.68	0
其他上呼吸道疾病	109.65	1200.05	5.933e-263
口腔、唾液腺及顎骨之疾病	107.62	5843.52	0
皮膚及皮下組織感染	106.01	1380.67	3.321e-302
其他泌尿系統疾病	95.6	11.38	0.0007
徵候 ¹	87.3	778.02	3.229e-171
攝護腺肥大	82.65	1811.35	0
急性呼吸道感染	80.67	2484.52	0
食道、胃及十二指腸之疾病	73.99	504.45	1.018e-111
精神官能症	53.24	2.042	0.152

註 1：代表的是「來自病人的主觀感受」，如頭暈、失眠、幻覺、肢體暫時性麻痺、浮腫等。來源為 ICD-9-CM code 780-789。

表 4.17 關聯規則卡方檢定(累計次數)

疾病名稱	PSO 權重	卡方檢定	p-value
腸及腹膜之其他疾病	95	39.63	3.065e-10
肺炎及流行性感冒	90	105.12	1.148e-24
器官或組織之其他置換術後	89	6810.87	0
關節病變及相關疾患	82	452.48	2.072e-100
風濕症，背部除外	81	3293.15	0
食道、胃及十二指腸之疾病	61	448.37	1.626e-99
開放性傷口	59	0.68	0.406
糖尿病	57	6379.52	0
耳及乳突之疾病	55	487.63	4.656e-108
急性呼吸道感染	54	19103.81	0
慢性阻塞性肺部疾病及相關病況	51	4385.62	0
消化系統之其他疾病	51	425.66	1.429e-94
徵候 ¹	48	4967.4	0
其他皮膚及皮下組織疾病	44	3387.55	0
病毒疾病疫苗接種	39	14333.64	0
皮膚表面無損之挫傷	37	4086.56	0
皮膚及皮下組織感染	36	857.5	1.697e-188
心臟病	26	12434.34	0
口腔、唾液腺及頷骨之疾病	23	26884.8	0
肺陷落	23	756.09	1.896e-166
眼睛及附屬器官之疾患	8	4251.07	0

註 1：代表的是「來自病人的主觀感受」，如頭暈、失眠、幻覺、肢體暫時性麻痺、浮腫等。來源為 ICD-9-CM code 780-789。

表 4.18 PSO 和卡方的相關係數

	全年齡	65 歲以下	65 歲以上	65 歲以下男性	65 歲以下女性	65 歲以上男性	65 歲以上女性	累計次數
相關係數	0.0681	0.0588	0.0972	0.0733	0.2412	0.0859	0.0821	0.05
p-value	0.7632	0.8167	0.7203	0.7726	0.4258	0.7895	0.7384	0.8296

從上表 4.18 可看到 PSO 和卡方檢定的相關係數很低，基本上可以看成兩者毫無相關，而從結果來看，挑選 PSO 前三個權重較高的疾病和卡方檢定的前三個疾病來做比對如下表 4.19：

表 4.19 PSO 和卡方前三種疾病比對

	全年齡	65 歲以下	65 歲以上	65 歲以下男性
PSO	皮膚及皮下組織感染、背部病變、食道、胃及十二指腸之疾病	其他上呼吸道疾病、皮膚疾病、風濕症	徵候、腸及腹膜之其他疾病、心臟病	其他上呼吸道疾病、非傳染性腸炎及大腸炎、食道、胃及十二指腸之疾病
卡方檢定	精神疾病、口腔唾液腺及頷骨之疾病、高血壓	口腔唾液腺及頷骨之疾病、其他上呼吸道疾病、徵候	口腔唾液腺及頷骨之疾病、高血壓、心臟病	口腔唾液腺及頷骨之疾病、高血壓、急性呼吸道感染
	65 歲以下女性	65 歲以上男性	65 歲以上女性	累計次數
PSO	高血脂、背部病變、關節病變及相關疾患	支氣管炎肺氣腫及氣喘、皮膚表面無損之挫傷、食道胃及十二指腸之疾病	腸及腹膜之其他疾病、眼睛及附屬器官之疾患、關節病變及相關疾患	腸及腹膜之其他疾病、肺炎及流行性感冒、器官或組織之其他置換術後
卡方檢定	口腔唾液腺及頷骨之疾病、高血壓、其他上呼吸道疾病	口腔唾液腺及頷骨之疾病、精神病、心臟病	口腔唾液腺及頷骨之疾病、高血壓、心臟病	口腔唾液腺及頷骨之疾病、急性呼吸道感染、病毒疾病疫苗接種

由上表 4.19 可發現 PSO 找到的較會影響中風的疾病偏向腸胃疾病、其他上呼吸道疾病和皮膚疾病等，而卡方檢定是口腔唾液腺及頷骨之疾病、高血壓、其他上呼吸道疾病和心臟病等，且比 PSO 較為集中。

4.3 結果討論

在 GA-PLS 改良過後的比較中可以發現，在資料都相同的情況下，在表 4.1 可以明顯的發現改良過後的方法明顯好於沒改良過後的方法。且找出的轉化因素也比上次的結果無疑更符合，代表著如果有以下這些疾病頭暈、便秘、慢性腎衰竭、高血壓、糖尿病、高脂血症、焦慮、肌肉痛、前列腺肥大等，是高危險族群，需多加注意。

在實驗中，我們發現使用 GA-PLS 和 seq2seq，GA-PLS 的結果較 seq2seq 好上一點，但 seq2seq 所用的資料量是 GA-PLS 的 5 倍之多，且訓練次數是 GA-PLS 的 10 倍，所以由此可知 GA-PLS 方法在本研究中是比較適合的。

在表 4.5 中可以看到不同組的資料在 GA-PLS 中的準確度，而其中全年齡和累計次數的資料是一樣的，只差在一種是 0,1 的資料，一種是累計次數的資料，可發現全年齡的準確度較累計次數的高，猜測其可能的原因為一些較為常得到的疾病所影響，如大腸炎、開放性傷口等，這些會導致多次來看診的疾病會影響到累計次數的資料，並且這些疾病又不會導致中風，最後使準確率比使用 0,1 資料的低，從表 4.20 中可在累計次數找到流行性感冒，更可證實我的猜想。

在 GA 後的 PSO 中可發現大部分找出的疾病和最後關聯規則找出的疾病有很大的不同，且在卡方檢定中前半部分的疾病 p-value 都未小於 0.05，但在後半部分中幾乎所有疾病 p-value 都小於 0.05，猜測可能是因為前半段疾病變數過多使 PSO 效果較為不佳所導致的。

在之後的關聯規則和貝氏網路中，我們再次縮減了疾病變數，並使用 ROC 曲線判斷準確度，平均有達 82.74%如表 4.9，且還可以觀察到疾病的方向，找到疾病中的節點，發現不管是哪一組基本上都會經過其他上呼吸道疾病或者是腸及腹膜之其他疾病等，而在後續 PSO 和卡方檢定中腸及腹膜之其他疾病和其他上呼吸道疾病在 PSO 中權重都較高，在卡方檢定中較高的是口腔唾液腺及頷骨之疾病、高血壓、其他上呼吸道疾病和心臟病等，兩種方法所得出結果只在部分上有些許

的相同，但以目前會影響中風的疾病來比較的話，會發現卡方檢定的結果會較為貼近許多。

表 4.20 為所有組合在關聯規則後所挑選的疾病，來比較在不同年齡和性別中，得到中風疾病的相關疾病是否有不相同之處或相同之處。由表中可看到全部都有的有口腔唾液腺及頷骨之疾病、食道胃及十二指腸之疾病和心臟病，可看出這三種疾病對中風是非常有相關的，而根據相關研究顯示，目前有關中風的危險因子有包含高血壓、糖尿病、心臟病和高血脂等疾病[5]，所以這邊多了一個假說口腔唾液腺及頷骨之疾病、食道胃及十二指腸之疾病可能是中風的危險因子，所以當有以上這些疾病的病人須多加注意。

表 4.20 中風可能之疾病

	全年齡	65歲以下	65歲以上	65歲以下男性	65歲以下女性	65歲以上男性	65歲以上女性	累計次數
其他上呼吸道疾病	√	√	√	√	√	√	√	
口腔、唾液腺及頷骨之疾病	√	√	√	√	√	√	√	√
風濕症	√	√	√	√	√			√
背部病變	√	√	√	√	√	√	√	
一般醫學檢查	√		√		√	√		
關節病變及相關疾患	√		√	√	√	√	√	√
食道、胃及十二指腸之疾病	√	√	√	√	√	√	√	√
皮膚及皮下組織疾病	√	√		√	√			√
病毒疾病疫苗接種	√							√
精神官能症	√	√	√		√	√	√	
高血壓	√	√	√	√	√	√	√	
高血脂	√	√			√	√		
糖尿病	√	√	√			√		√
皮膚及皮下組織感染	√	√		√		√	√	√
心臟病	√	√	√	√	√	√	√	√
肝病	√							

徵候	V	V	V	V	V	V	V	V
攝護腺肥大	V			V			V	
支氣管炎、肺氣腫及氣喘	V	V		V		V		
肺炎及流行性感冒	V					V	V	V
骨質酥鬆	V							
精神疾病	V					V		
眼睛及附屬器官之疾患		V	V		V	V	V	V
其他泌尿系統疾病		V	V				V	
皮膚表面無損之挫傷		V				V		V
消化系統之其他疾病		V		V		V		V
腸及腹膜之其他疾病			V	V	V	V	V	V
非傳染性腸炎及大腸炎			V	V	V	V	V	
急性呼吸道感染				V		V	V	V
耳及乳突之疾病				V			V	V
慢性阻塞性肺部疾病及相關病況						V		V
黴菌病							V	
肺陷落								V
器官或組織之其他置換術後								V
開放性傷口								V

4.4 實作結果

根據結果，本研究編寫了一個 UI 介面，方便使用者使用，當使用者填入自己的資料後，可以非常直觀的了解，他未來可能得到那些疾病，以及該疾病的機率是多少，且在圖 4.33 下排我們顯示了機率較高的前四種疾病，便可較為明顯的了解自己將來得到哪些疾病的機會較高。

操作方法是先選擇性別和年齡，如果都沒選擇的話就是全年齡的資料如下圖 4.34，選好之後按下下方的確定按鈕，會更新疾病關聯圖、可選的疾病和最下面機率較高的前四種疾病，這個時候顯示的是當你未得到左邊疾病列表中任意一種疾病時，所得到這些疾病的機率是多少，此數字是根據此前所做的貝氏網路所得的結果，假如是有左列疾病列表中任意一種疾病時，請點選所得過的疾病並按下

analyze1 按鈕，這時會更新疾病關聯圖中疾病底下的機率，並在最下面顯示前四種機率最高的疾病，且旁邊的關聯圖會顯示你所點選疾病所產生的關聯，用黃色來標示，若按 analyze2 按鈕，則只會僅顯示所選之疾病的關聯圖，如圖 4.35。

此次實驗基本上都是在 matlab 實做出來的，除了在資料的前處理上是使用 SAS EG 去挑選及搜尋資料。基本上只要在資料前處理上，挑選不同的疾病，是可以直接套用這個方法，而產生出新的疾病分析。

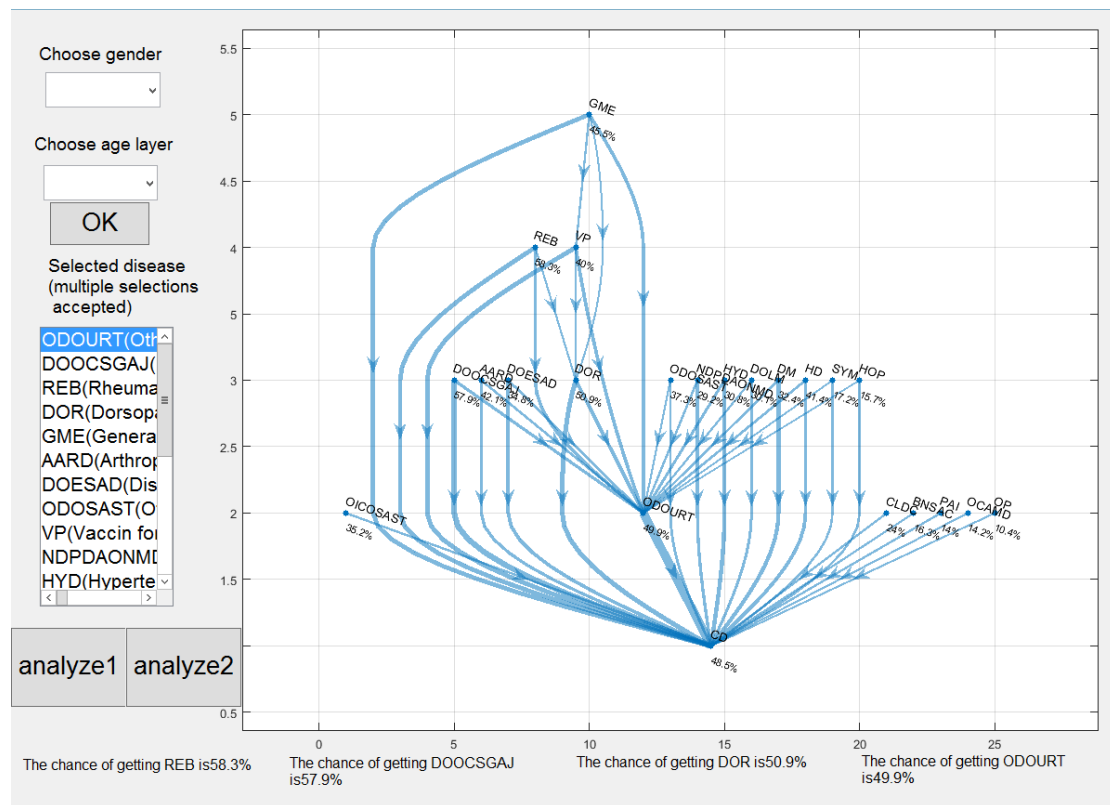


圖 4.33 UI 程式介面(全年齡未選取)

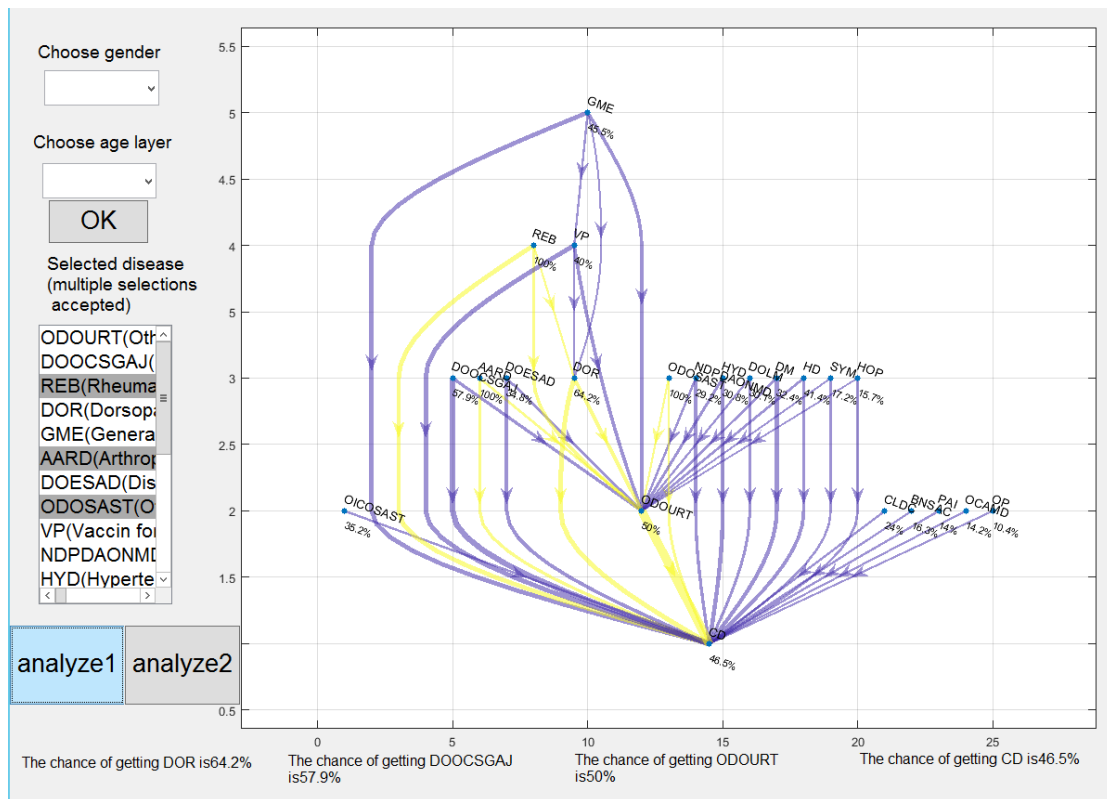


圖 4.34 UI 程式介面(全年齡 analyze1)

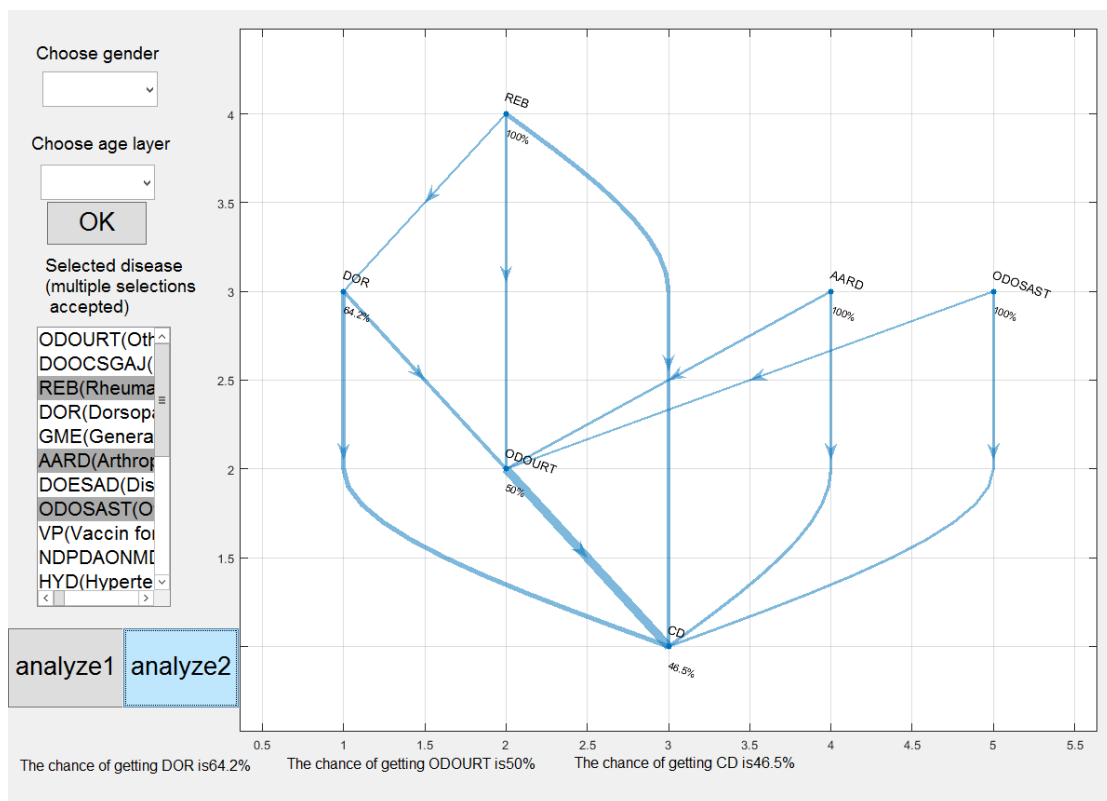


圖 4.35 UI 程式介面(全年齡 analyze2)

第五章 結論

近年來隨著大數據和機器學習的流行，預測模型的方法也不再限制只有統計方法，越來越多人投入到這一塊當中，但在選擇特徵的時候，始終都會去詢問專業人士後，再去做預測模型。而本研究在特徵選擇方面，完全交給機器去做選擇，只要求可以把中風和非中風的病人分開，在這一方面幾乎很少有關類似的論文出現，而最後的結果也可以達到 0.85 的準確度，比其他相關論文的準確度來的高，也證實是可以使用疾病當特徵變項去使用。

雖然我們提出了一個準確度達 0.85 的預測模型，但由於其變數過多，會讓人不了解到底哪些疾病的相關程度較高，且也不好提供相關建議供人們參考，所以在之後我們做了與中風相關疾病之關係圖及貝氏網路，減少變數的數量，以供人們做參考及建議，結果為 ROC 曲面下平均面積為 0.8274，表示此模型分類的準確度，最後找出的相關疾病有口腔唾液腺及頷骨之疾病、食道胃及十二指腸之疾病和心臟病等相關疾病可能和中風有關，經過和相關論文比對的結果 [5][12][18][26][27][44]，得出口腔唾液腺及頷骨之疾病、食道胃及十二指腸之疾病等可能是新的的中風危險因子，如表 5.1 絕大部分與中風有關的疾病因素，都是高血壓、心臟病、糖尿病等。

且此框架亦為 GA-PLS->Apriori-IFM->貝氏網路->PSO、卡方檢定，可套入不同的疾病上面，在缺血性中風轉出血性中風的問題上，其預測結果平均可達 0.855，基本上只要資料的前處理有做好，那麼套入到此架構因該都會有一個不錯的結果。找到的影響因素為頭暈、便秘、慢性腎衰竭、高血壓、糖尿病、高脂血症、焦慮、肌肉痛、前列腺肥大等，和相關研究比對過後發現除了高血壓、糖尿病和高脂血症的疾病以外，其他的有可能是新影響出血性轉化的因素。

後續的發展方向可以針對找出新的中風危險因子做更深入的研究，又或者可以持續改進模型，提高準確率，這些都是可以研究或探討的，或者更進一步來說可結合現在最流行的深度學習 Deep Learning，來去做更進一步的研究，亦可去

預測病人未來 1 年、5 年或 10 年，中風的發病率，又或者可依據該病人所得之疾病，假若按照醫師所規定治療或減緩，來去預測他得到中風的機率，這些都是後續可發展的方向。

中風因素	當前 認知 [5]	本研究得 出(8 個取 7 個以上 有的)	陳適安 (心房顫 動患 者)[44]	Gabriel Yiin[24]	Aigner A. et al[12]	Chang T. et al[18]	Hopewell JC. et al[27]
心臟病	V	V	V	V			V
高血壓	V	V	V	V	V	V	V
糖尿病	V		V			V	V
高血酯	V						
慢性腎病	V						
口腔唾液腺及 頷骨之疾病		V					
食道胃及十二 指腸之疾病		V					
其他上呼吸道 疾病		V					
徵侯		V					
背部病變		V					
關節病變及相 關疾患		V					
短暫性腦缺血						V	

表 5.1 中風因素比較表

參考文獻

- [1] 尤哲威(2015)，“以健保資料庫對中風患者再復發之預測”，臺北醫學大學碩士論文。
- [2] 王濟川、郭志剛(2005)。“Logistic 迴歸模型－方法及應用”，台北市：五南圖書。
- [3] 行政院衛生署 (2016)。死因統計表。檢自 <https://www.mohw.gov.tw/cp-16-33598-1.html>。
- [4] 李維平, 李元傑, 謝明勳, "以群中心策略改良人工蜂群演算法", 資訊管理學報, Vol. 21, no.1, pp. 25-44, 2014.
- [5] 邱弘毅, 腦中風之現況與流行病學特徵, 台灣腦中風學會會訊, 第 15 卷第 3 期, 2-4。
- [6] 張佑璋(2017 年 7 月), “以最小偏差基因及粒子群演算法分析缺血性中風轉出血性中風成因探討”, 2017 13th 台灣軟體工程研討會, 逢甲大學。
- [7] 梁昭隆(2016), "使用最小偏差法及蜂群法之無線室內定位精度改進探討與比較", 全國碩博士論文.
- [8] 黃仁鵬、熊浩志、紀美吟、郭煌政(2004), "不用設定最小門檻值的關聯規則探勘方法－IFM", 第一屆創新與管理學術研討會, pp. 193。
- [9] 慕哈曼(2013), “發展一多階數資料探勘方法建立腦中風風險預測模型”, 國立臺灣科技大學碩士論文。
- [10] 鄭建興、賴達明 (2009), “腦中風 每天奪走 30 命”, 元氣周報, 檢自 <https://health.udn.com/health/story/5977/343894>。
- [11] Agrawal R. and Srikant R..(1994), “Fast Algorithms for Mining Association Rulse in Large Database.”, Proceedings of the 20th International Conference on Very Large Data Bases , pp:487-499.
- [12] Aigner A, Grittner U, Rolfs A, Norrving B, Siegerink B, Busch MA.

Contribution of established stroke risk factors to the burden of stroke in young adults. *Stroke*. 2017 doi: 10.1161/STROKEAHA.117.016599.

- [13] Bastien P., Vinzi V.E., Tenenhaus M. (2005), "PLS generalised linear regression." , *Comput. Statist. Data Anal* , pp. 17-46.
- [14] Ben-Gal, I. (2007) "Bayesian networks, In:Ruggeri F., Faltin F. and Kenett R. (Eds.)" , *Encyclopedia of Statistics in Quality and Reliability*, John Wiley and Sons.
- [15] Bhaskaran, A. Gasparrini, S. Hajat, L. Smeeth, B. Armstrong(2013), "Time series regression studies in environmental epidemiology." , *Int. J. Epidemiol* , pp. 1187-1195.
- [16] Bilic, I., G. Dzamonja, I. Lusic, M. Matijaca, and K. Caljkusic(2009). "Risk factors and outcome differences between ischemic and hemorrhagic stroke." , *Acta Clinica Croatica* 48 (4): 399-403.
- [17] Bowes, J., Neufeld, E., Greer, J. E. and Cooke, J. (2000)"A Comparison of Association Rule Discovery and Bayesian Network Causal Inference Algorithms to Discover Relationships in Discrete Data." *Proceedings of the Thirteenth Canadian Artificial Intelligence Conference (AI' 2000)*, Montreal, Canada.
- [18] Chang T, Gajasinghe S, Arambepola C. Prevalence of stroke and its risk factors in urban Sri Lanka: population-based study. *Stroke*. 2015;46:2965 – 2968.
- [19] Cho, K. et al. (2014), "Learning phrase representations using RNN encoder-decoder for statistical machine translation." , *In Proc. Conference on Empirical Methods in Natural Language*

Processing, pp:1724 - 1734.

- [20] Chuang CS, Ho SC, Lin CL, Lin MC, Kao CH. (2016) " Risk of Cerebrovascular Events in Pneumoconiosis Patients: A Population-based Study, 1996-2011." *Medicine (Baltimore)*. Mar; 95(9): e2944.
- [21] David Tian, Ann Gledson, Andreas Antoniadis, Aristo Aristodimou, Ntalaperas Dimitrios, Ratnesh Sahay, Jianxin Pan, Stavros Stivaros, Goran Nenadic, Xiao-jun Zeng et al., (2013) "A bayesian association rule mining algorithm." *Systems Man and Cybernetics (SMC) 2013 IEEE International Conference on*, pp. 3258-3264.
- [22] De Jong, S. (1993). SIMPLS : an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251 - 263.
- [23] Dong-Yan Huang, Zhengchen Zhang, and Shuzhi Sam Ge(2014), "Speaker state classification based on fusion of asymmetric simple partial least squares (simpls) and support vector machines." *Computer Speech & Language*, vol. 28, no. 2, pp. 392 - 419.
- [24] Ersel, D., Günay, S., (2012) "Bayesian networks and association analysis in knowledge discovery process." *Istatistikciler Dergisi* 5, pp 51-64.
- [25] G. Tripepi; C. Zoccali; K. J. Jager; F. W. Dekker(2008), "Linear and logistic regression analysis." ,*Kidney international*, Vol: 73, Issue: 7, pp: 806-10.
- [26] Gabriel Yiin(2002). "Time Trends in Atrial Fibrillation-Associated Stroke and Premorbid Anticoagulation: Population-Based Study and

- Systematic Review". Stroke 50(1). DOI: 10.1161/STROKEAHA.118.022249
- [27] Hopewell JC, Clarke R. Emerging risk factors for stroke: what have we learned from Mendelian randomization studies? Stroke. 2016;47(6):1673 – 8.
- [28] Huang, Z., Li, J., Su, H., Watts, G. S., & Chen, H. (2007)"Large-scale regulatory network analysis from microarray data:modified Bayesian network learning and association rule mining." Decision Support Systems, 43(4), 1207–1225.
- [29] Huang, Y. (1999). "Learning bayesian networks guided by decomposable markov networks." , Unpublished MSc, The University of Regina (Canda). 3–54.
- [30] Imai, Chisato; Armstrong, Ben; Chalabi, Zaid; Mangtani, Punam; Hashizume, Masahiro(2015), "Time series regression model for infectious disease and weather." , Environmental research, Vol: 142, pp: 319–27.
- [31] J. H. Holland(1975), Adaptation in Natural and Artificial Systems, Univ. of Michigan Press, Ann Arbor.
- [32] J. Kennedy, and R. C. Eberhart, (1995) "Particle swarm optimization" , Proc. IEEE International Conference on Neural Networks (Perth, Australia), IEEE Service Center, Piscataway, NJ, pp. IV: 1942–1948.
- [33] Jae-woo Lee, Hyun-sun Lim, Dong-wook Kimb Soon-ae Shin, Jinkwon Kim, Bora Yoo, Kyung-hee Cho (2017), "The development and implementation of stroke risk prediction model in National Health Insurance Service's personal health record" , Computer methods and programs in biomedicine, Vol. 153, pp. 253–257.

- [34] K. Toyoda, Y. Okada, S. Kobayashi(2007), “Early recurrence of ischemic stroke in Japanese patients: the Japan standard stroke registry study.” ,*Cerebrovasc* , pp. 289-295.
- [35] Karl Pearson. X.(2009)” On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*: 157 - 175.
- [36] Khan, M. S., Muyeba, M., and Coenen, F.(2008), “A Weighted Utility Framework for Mining Association Rules,” in *Proceedings of European Modeling Symposium (IEEE)*, pp.87-92.
- [37] Li, H. W., Yang, M. C., Chung, K. P.(2011) “Predictors for readmission of acute ischemic stroke in Taiwan.” ,*J Formos Med Assoc*, Vol.110, pp.627 - 633.
- [38] Marsh EB, Llinas RH, Schneider AL, Hillis AE, Lawrence E, Dziedzic P, et al(2016). Predicting hemorrhagic transformation of acute ischemic stroke: prospective validation of the HeRS Score. *Medicine (Baltimore)*. 2016;95:e2430.
- [39] Pearl, J. (1988) “Probabilistic Reasoning in Intelligent Systems” , Morgan Kaufmann, San Francisco.
- [40] Pearl, J.(1997), “Graphical Models for Probabilistic and Causal Reasoning.” ,*The Computer Science and Engineering Handbook*, pp:697-714.
- [41] R. C. Eberhart, and J. Kennedy,(1995) “new optimizer using particle swarm theory” , *Proc. Sixth International Symposium on Micro Machine*

and Human Science, Nagoya, Japan, pp. 39–43.

- [42] R. C. Sato, G. T. K. Sato(2015), “Probabilistic graphic models applied to identification of diseases.” , Einstein (São Paulo), 13 (2) (2015), pp. 330–333.
- [43] SHI, Guangyi, et al. (2016)” The human body characteristic parameters extraction and disease tendency prediction based on multi-sensing fusion algorithms.” In: Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2016 IEEE International Conference on. IEEE, p. 126–130.
- [44] Shih-Ann Chen, MD *;Tze-Fan Chao, MD; Chern-En Chiang, MD; Tzeng-Ji Chen, MD; Gregory Y. H. Lip, MD * .(2019) “Reassessment of Risk for Stroke During Follow-up of Patients With Atrial Fibrillation.” Ann Intern Med. DOI: 10.7326/M18-1177.
- [45] Sutskever, I. Vinyals, O. & Le. Q. V. (2014), “Sequence to sequence learning with neural networks.” , In Proc. Advances in Neural Information Processing Systems 27, pp:3104 – 3112.
- [46] Tsai-Chung Li, Hsiang-Chi Wang, Chia-Ing Li, Chiu-Shong Liu, Wen-Yuan Lin, Chih-Hsueh Lin, Sing-Yu Yang, Cheng-Chieh Lin (2018), “Establishment and validation of a prediction model for ischemic stroke risks in patients with type 2 diabetes” , Diabetes research and clinical practice, Vol.138, pp. 220–228.
- [47] Wold, Svante, Michael Sjöström, and Lennart Eriksson (2001), “PLS-Regression: A Basic Tool of Chemometrics.” , Chemometrics and Intelligent Laboratory Systems, 58 (2), 109–130.
- [48] Y. Y. Lee, K. L. Lin, H. S. Wang, M. L. Chou, P. C. Hung, M. Y. Hsieh, et

al. (2008), "Risk factors and outcomes of childhood ischemic stroke in Taiwan." ,Brain, pp. 14-19.