國立政治大學統計學系研究所

碩士學位論文

基於支持向量機計算的相互熵之特徵選取

A Feature Selection Study based on SVM and Mutual Entropy

指導教授：周珮婷 博士

研究生：游上葦 撰

中華民國 108 年 7 月

DOI:10.6814/NCCU201900996

摘要

特徵選取為機器學習領域中一重要部分，適當的選取特徵(變數)，除了減少機器運算時間、人力與金錢外，也可以避免模型過度配適或是欠擬和的情況發生。雖然經過多年發展已有很多特徵選取的方法，但同一種模型，不一定適用所有資料情況，因此提出新方法希望在特徵選取上會有更多選擇。

本文提出一新方法衡量變數之間的關係，使用資訊理論中熵的概念，結合分類器支持向量機，獲取變數間關係，並將變數分群，給予每群適當的相對應分數，以此篩選變數。本文採用半監督式學習，計算敏感度、特異度與準確度之平均及所使用變數數量，並以高斯混合模型搭配 EM 演算法利用 KS 檢定之檢定統計量定義資料重要變數，評估方法能否選取重要變數，本文一共使用 2 筆模擬資料與 5 筆真實資料，並將結果與各大方法比較，結果顯示在各資料集中皆有穩定表現，即使在變數少的情況下也能有不錯表現。

關鍵字：機器學習、特徵選取、維度縮減，支持向量機、支援向量機、熵、相互熵。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

Abstract

Feature selection technique plays a significant role in machine learning. Selecting features (variables) adequately can not only reduce the expenditure, operating time in machine and the cost of labor but also prevent under fitting or overfitting. Although lots of feature selection methods have been developed for decades, it is impossible to apply a unique method to all types of data sets. In this study, we propose a new method to calculate the correlation between variables based on the Shannon entropy from information theory and SVM classifier. Variables are grouped into several clusters and selected by the new correlation measurement. Besides, we define the importance of variable by the test statistic of KS test using Gaussian mixed model and E-M algorithm for the propose of result assessment. The performance of proposed method on two simulated data and five real data are demonstrated and compared with other feature selection methods. The predicted results are stable through the proposed method with a reduced dataset.

Keyword：Machine learning, Feature selection, Dimension reduction, Support Vector Machine, SVM, Entropy, Shannon Entropy, Mutual Entropy

DOI:10.6814/NCCU201900996

目　次

第一章 緒論.....1
第二章 文獻探討.....3
第三章 研究目的.....5
一、特徵選取(feature selection).....5
第四章 研究方法.....6
第一節 ENTROPY 特徵選取.....6
一、Shannon Entropy.....6
二、Mutual Entropy.....7
三、Support Vector Machine(SVM).....10
四、Feature selection via Mutual Entropy based on classified model.....11
五、分群規則.....13
六、方法步驟.....15
第二節 其他特徵選取方法.....17
一、Random forest(RF).....17
二、F-score.....18
三、Least Absolute Shrinkage and Selection Operation(LASSO).....19
四、Fast Correlation-Based Filter(FCBF).....20
第五章 資料介紹.....22
一、模擬資料一.....22
二、模擬資料二.....24
三、Wisconsin Diagnostic Breast Cancer.....26
四、Connectionist Bench(Sonar, Mines vs. Rocks).....27
五、Credit Card Fraud Detection.....28
六、Oxford Parkinson's Disease.....29
七、APS Failure at Scania Trucks.....30
第六章 研究結果.....33
第一節 資料結果.....35
一、模擬資料一及模擬資料二.....35
二、Wisconsin Diagnostic Breast Cancer.....38
三、Connectionist Bench(Sonar, Mines vs. Rocks).....41
四、Credit Card Fraud Detection.....44

DOI:10.6814/NCCU201900996

五、Oxford Parkinson’s Disease.....47
六、APS Failure at Scania Trucks.....50
第二節 結論.....53
第七章 未來展望.....54
第八章 參考資料.....55

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

表　次

表 4-1 本文所用資料之分群數 14
表 5-1 各資料集樣本與變數概況 32
表 6-1 模擬資料 1 各方法於平均 50 次 bootstrap 平均選取之變數數量 36
表 6-2 模擬資料 2 各方法於平均 50 次 bootstrap 平均選取之變數數量 36
表 6-3 Wisconsin Diagnostic Breast Cancer 各項指標.....38
表 6-4 Connectionist Bench(Sonar, Mines vs. Rocks) 各項指標 41
表 6-5 Credit Card Fraud Detection 各項指標 44
表 6-6 Oxford Parkinson's Disease 各項指標.....47
表 6-7 APS Failure at Scania Trucks 各項指標.....50

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

圖　次

圖 4-1 二元分類 7
圖 4-2 基於變數 B 計算變數 A entropy 示意圖 8
圖 4-3 基於變數 A 計算變數 B entropy 示意圖 9
圖 4-4 SVM 示意圖 10
圖 4-5 本研究方法建模及取得估計值之示意圖 11
圖 4-6 Mutual Entropy 矩陣示意圖 12
圖 5-1 模擬資料 1 視覺化圖形 23
圖 5-2 模擬資料 2 視覺化圖形 25
圖 5-3 乳癌資料集視覺化圖形 26
圖 5-4 聲納資料集視覺化圖形 27
圖 5-5 信用卡詐騙資料集視覺化圖形 28
圖 5-6 帕金森氏症資料集視覺化圖形 29
圖 5-7 空氣壓力系統資料集視覺化圖形 31
圖 6-1 定義重要性變數方法之評估能力 33
圖 6-2 ME 於模擬資料 1 及模擬資料 2 之表現 36
圖 6-3 RF 於模擬資料 1 及模擬資料 2 之表現 37
圖 6-4 F-SCORE 於模擬資料 1 及模擬資料 2 之表現 37
圖 6-5 FCBF 於模擬資料 1 及模擬資料 2 之表現 37
圖 6-6 LASSO 於模擬資料 1 及模擬資料 2 之表現 37
圖 6-7 Wisconsin Diagnostic Breast Cancer　ME 重要變數選取情況 39
圖 6-8 Wisconsin Diagnostic Breast Cancer　RF 重要變數選取情況 39
圖 6-9 Wisconsin Diagnostic Breast Cancer　F-SCORE 重要變數選取情況 39
圖 6-10 Wisconsin Diagnostic Breast Cancer　FCBF 重要變數選取情況 39
圖 6-11 Wisconsin Diagnostic Breast Cancer　LASSO 重要變數選取情況 40
圖 6-12 Connectionist Bench(Sonar, Mines vs. Rocks)　ME 重要變數選取情況 42
圖 6-13 Connectionist Bench(Sonar, Mines vs. Rocks)　RF 重要變數選取情況 42
圖 6-14 Connectionist Bench(Sonar, Mines vs. Rocks)　F-SCORE 重要變數選取情況 42
圖 6-15 Connectionist Bench(Sonar, Mines vs. Rocks)　FCBF 重要變數選取情況 42
圖 6-16 Connectionist Bench(Sonar, Mines vs. Rocks)　LASSO 重要變數

DOI:10.6814/NCCU201900996

選取情況.....43
圖 6-17 Credit Card Fraud Detection ME 重要變數選取情況.....45
圖 6-18 Credit Card Fraud Detection RF 重要變數選取情況.....45
圖 6-19 Credit Card Fraud Detection F-SCORE 重要變數選取情況.....45
圖 6-20 Credit Card Fraud Detection FCBF 重要變數選取情況.....45
圖 6-21 Credit Card Fraud Detection LASSO 重要變數選取情況.....46
圖 6-22 Oxford Parkinson's Disease ME 重要變數選取情況.....48
圖 6-23 Oxford Parkinson's Disease RF 重要變數選取情況.....48
圖 6-24 Oxford Parkinson's Disease F-SCORE 重要變數選取情況.....48
圖 6-25 Oxford Parkinson's Disease FCBF 重要變數選取情況.....48
圖 6-26 Oxford Parkinson's Disease LASSO 重要變數選取情況.....49
圖 6-27 APS Failure at Scania Trucks ME 重要變數選取情況.....51
圖 6-28 APS Failure at Scania Trucks RF 重要變數選取情況.....51
圖 6-29 APS Failure at Scania Trucks F-SCORE 重要變數選取情況.....51
圖 6-30 APS Failure at Scania Trucks FCBF 重要變數選取情況.....51
圖 6-31 APS Failure at Scania Trucks LASSO 重要變數選取情況.....52

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

第一章 緒論

降維(dimensionality reduction)被大量應用在機器學習上、其概念在於，藉由刪除較無解釋力之變數，減少建模時運算時間，也減少金錢、人力等。因變數過多，可能導致模型對於訓練資料過度配適(overfitting)，而特徵選取能避免此情況，或是因某些變數含有過多雜訊，反而使得以這類變數建立的模型降低在預測上的準確度，因此特徵選取也防止欠擬合(under fitting)的情況。維度降低，可方便預測結果之解釋，或是資料的視覺化呈現，特徵選取方法雖刪除某些變數但以最大程度保留資料原始樣貌，因其有著諸多優點，因此選定特徵選取為研究方向。

統計學降維之特徵選取(feature selection)最常見的為迴歸中之逐步迴歸(stepwise)法，其基本概念在於以某一標準(如模型的R^2或是加入新變數後模型或係數檢定之檢定統計量或 P-value 做為標準等)決定變數是否納入模型，主成分分析(PCA)也為一歷史悠久之降維方式，將原始變數以線性組合為新的變數，稱為特徵萃取(feature extraction)，解釋力較強的變數將給予較大的係數，現今也已有多種降維方式，如隨機森林、LASSO 及基因演算法(genetic algorithm)等。

本研究提出一特徵選取方法，其基本概念在於，將變數分為數個子集合，並給定每一個子集合一門檻值，根據此值選取適當子集合用以建模預測，因此如何將解釋力強的變數集做正確選取以及如何設定特徵選取停止值為一重要課題。

本文藉由資訊理論中的 Shannon Entropy 計算而成的 Mutual Entropy 並結合分類器支持向量機(Support Vector Machine, SVM)，計算每個變數間之相關係數，得出 Mutual Entropy 矩陣，並以此為基礎使用階層式分群法(hierarchical

DOI:10.6814/NCCU201900996

cluster)分群並搭配 silhouette coefficient 值，取得最佳分群數，其次，經由 Mutual Entropy 矩陣計算並給定每一群之變數一個公平且足以代表該群變數資訊量的值，作為特徵選取的標準，本研究主要探討半監督式學習，最後與其他方法比較敏感度(sensitivity)、特異度(specificity)、準確度(accuracy)以及建模所使用維度等指標，期望成為一易於應用的降維方法，對於降維在機器學習與統計學領域上做出貢獻。

第三章將先敘述研究目的，第四章詳述本文所使用方法與比較之方法，第五章為資料介紹，共 2 筆模擬資料與 5 筆真實資料，第六章為研究結果，在與其他降維方法比較下，本研究結果並不會與之相差太多，第七章為未來展望，希望在 Mutual Entropy 矩陣分群規則以及更多類型的資料上能有更多的研究。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

第二章 文獻探討

特徵選取長久以來一直是機器學習研究裡重要的議題，這些年來有許多相關的方法及其衍生的研究。Shannon(1948)提出了資訊理論 Shannon Entropy 的概念，此概念至今已被廣泛應用於特徵選取方面。 Hsieh 等人(2018)以 DCG (Data Could Geometry)計算 Mutual Entropy，Mutual Entropy 考慮的一個概念是基於 X 和 Y 兩特徵，如果 X 能夠傳達關係中不可忽略的信息量到 Y 且 Y 亦同，則 X 和 Y 具有相互依賴的關係。並且提到如果資料內的差異被適當的計算並適當的表示，那麼應該可以得知資料內的完整資訊，兩變數擁有較低的 Mutual Entropy 應該給予類似的資訊值，且兩兩變數間建立的 Mutual Entropy 矩陣可以用於特徵分組。

在降維方面，至今已有多種方法被提出及應用，Tibshirani(1996)提出 LASSO，將原本傳統迴歸最小化觀察值與預測值距離的想法加入以模型係數絕對值組成的懲罰項，將某些接近零的係數縮減為 0，並保留優秀的變數。Breiman(2001)結合 Breiman(1996)以及 Ho(1995, 1998)提出的隨機森林，也同樣使用資訊理論概念，Genuer 等(2010)使用隨機森林之變數重要性篩選變數，大致分為兩步驟，第一，計算每個變數重要性並刪除重要性較低的變數，找到足夠數量變數用以預測或解釋，第二，分為兩目的，其一，若目的為解釋結果，經由 OOB(out of bag)概念選取 OOB 最低的變數組合建模，其二，從保留於解釋目的的變數中選取適當變數，逐步增減來觀察模型表現。Yu & Liu (2003) 提出根據資訊理論 Entropy 進行特徵選取的方法 FCBF，除了衡量自變數間的關係也衡量應變數與自變數的關係，並以此初步篩選變數，將表現較佳的變數蒐集起來，並再以自變數間的關係作為基準進一步篩選，Guyon & Elisseeff(2003)提出給變數一值後排列，例如計算相關係數等，但各有其缺點，如皮爾森相關係數只能表達變數間之線性關係，因此需選擇適當的方法計算，

DOI:10.6814/NCCU201900996

並依此選擇刪去或留下變數，與本文篩選變數概念大致相同，Guyon & Elisseeff(2003)指出特徵選取有很多淺在的益處，若是三維以下可以方便資料視覺化，便於瞭解資料，減少儲存空間，減少模型訓練時間，增進預測表現。

在變數選取研究中，更有許多演算法是與 SVM 結合，作為選取依據，如 Chen & Lin(2005)以 F-score 根據每個變數含有的資料特性，給定每個變數一個值，雖不考慮變數間關係，但根據 Chen & Lin 的結論，F-score 與 SVM 並用通常情況下為一簡單且有效的特徵選取方法，並在 Akay(2009)的乳癌研究中得到很好的準確度，Huang & Wang(2006)以基於基因演算法(Genetic Algorithm)篩選變數並有效提升準確度，並以 SVM 當作分類器，其有四步驟，將資料標準化，將特徵染色體從基因型轉為表型，獲取特徵子集合，每組染色體都會計算及配適值(fitness value)，Xu & Wang(2006)使用 PCA 進行降維後搭配 SVM 評估準確度，在此研究中發現 PCA 結合 SVM 為一個相當不錯的方法，但發現在在訓練集中若某類別的樣本數較少，則以 PCA 進行特徵萃取時會造成訊息遺失使得該類別在預測上表現較差，並且 PCA 搭配 SVM 時可以有效提升訓練及測試速度，因此不是每筆資料都適合作特徵萃取，也就是說降維方法須依資料特性而定。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

第三章 研究目的

一、特徵選取(feature selection)

維度災難或稱為維度詛咒(the curse of dimensionality)一直是統計學建模上與機器學習會面臨到的重要問題，包括在處理資料上面對高維度(大量變數)時，難以下手的困境、程式運算耗費大量時間、模型過度配適等。過度配適會造成對於測試資料與往後蒐集的新進樣本造成低估(underestimation)。若以高維度來分析資料會造成在蒐集資料時耗費更多金錢、時間對樣本蒐集大量特徵，因此維度縮減(dimension reduction)的研究與開發一直都是感興趣的研究方向，且希望維度縮減後能最大程度的保留該筆資料訊息，時至今日已經有數種選取特徵的方法，例如 RF、FCBF 等，選取方法大致分為線性與非線性的方法，本文希望藉由提出一新非線性特徵選取方法，往後在處理資料及建模時在處理維度(變數)上多一個有效率、有選取依據且實用的選擇。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

第四章 研究方法

第一節 ENTROPY 特徵選取

一、Shannon Entropy

Shannon Entropy 為計算一變數根據一標準(criterion)進行分類後，其分類結果是否單一的指標，為一介於 0 到 1 之間的值，數字愈大代表分類結果愈亂，亦即分類結果愈不單一，數字愈小代表分類結果愈單純，亦即分類結果愈單一，舉例而言某一節點只有兩類 0 與 1，且共 100 個樣本其中，0 有 40 個，1 有 60 個，則此結點之資訊亂度(entropy)為$\frac{-40}{100} * log_2\left(\frac{40}{100}\right) + \frac{-60}{100} * log_2\left(\frac{60}{100}\right)$。本文加入考慮建模結果的分類情況後，重新衡量變數與變數之間的關係，並建立變數間之 entropy 矩陣。

以二元分類資料為例，由圖 4-1 可知在$P(x) = 1,\ x = \frac{1}{2}$，變數於二元分類結果剛好為標記為 1 與標記為 0 為各半時，此時 Shannon entropy 具有最大值，代表在這樣的結果下，此分類結果的亂度最大，亦即分類結果愈不一致，而在$P(x) = 0,\ x = 0\ \ or\ \ 1$時，Shannon entropy 具有最小值，代表在這樣的結果下，此分類結果的亂度最小，亦即分類結果愈一致。

Shannon entropy 函數可表示為

$$H(X) = \sum_{i}^{c} -P(X = i) * log_c P(X = i) \qquad (4\text{-}1)$$

i 為第幾個類別，$\forall \mathrm{i} = 0, 1, \ldots, \mathrm{c} - 1$

c 為類別數量(the number of class)

DOI:10.6814/NCCU201900996

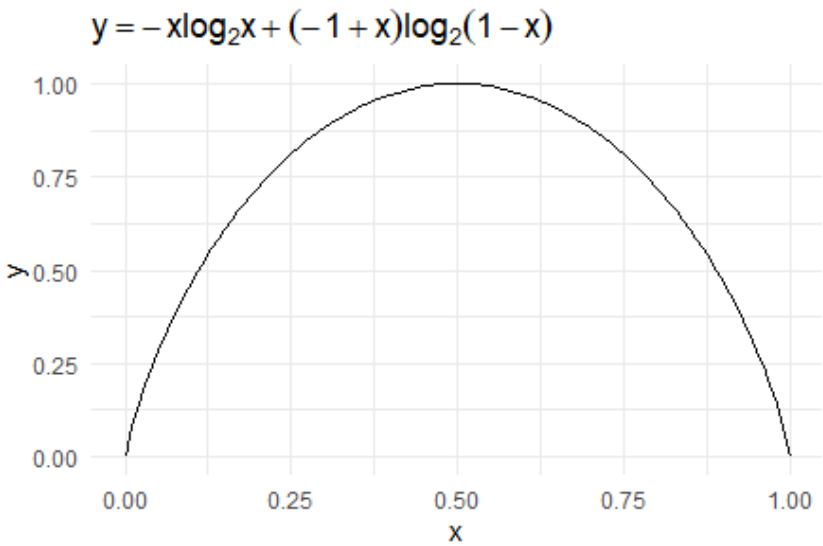


圖 4-1 二元分類 Shannon Entropy 函數示意圖

二、Mutual Entropy

Mutual Entropy 可視為一衡量兩變數之相關程度之方法，為一介於 0 至 1 之間的連續非負數值，Mutual Entropy 數字愈大，兩個變數相關性愈小，其理論基於 Shannon Entropy 並計算 Conditional Entropy

Conditional Entropy 函數可表示為

$$H(X_k|X_m = i) = \sum_i^c -P(X_k|X_m = i) * log_c P(X_k|X_m = i) \qquad (4\text{-}2)$$

$\forall$k, m=1, 2, ⋯, v

v 為自變數數量

DOI:10.6814/NCCU201900996

基於各分類樣本個數給定各分類權重$P(X_m = i)$、$P(X_k = i)$，並分別計算 Entropy 理論最大值$H(X_k)$、$H(X_m)$，則 Mutual Entropy

$$ME(X_k, X_m) = \frac{\frac{\sum_i P(X_m = i)H(X_k|X_m = i)}{H(X_k)} + \frac{\sum_i P(X_k = i)H(X_m|X_k = i)}{H(X_m)}}{2} \quad (4\text{-}3)$$

$$0 \le ME(X_k, X_m) \le 1$$

隨機生成一筆含有 20 樣本的模擬資料做為例子，此筆資料包含 10 個類別為 1 以及 10 個類別為 0 之樣本，計算第此筆抽取後之模擬資料變數 A 與變數 B 之 Mutual Entropy 值。

此計算結果可視為將變數 A 分為兩類別，如圖 4-2，其中綠色代表分類為 1，紅色代表分類為 0。

$$\sum_{i=0}^{1} P(X_B = i)H(X_A|X_B = i) =$$

$$\frac{8}{20}\left(\frac{-5}{8}log_2\frac{5}{8} + \frac{-3}{8}log_2\frac{3}{8}\right) + \frac{12}{20}\left(\frac{-6}{12}log_2\frac{6}{12} + \frac{-6}{12}log_2\frac{6}{12}\right)$$

$$= 0.3817736 + 0.6 = 0.9817736 \approx 0.9818$$

$$H(X_A) = \frac{1}{1}\left(\frac{-11}{20}log_2\frac{11}{20} + \frac{-9}{20}log_2\frac{9}{20}\right) = 0.9927745 \approx 0.9928$$

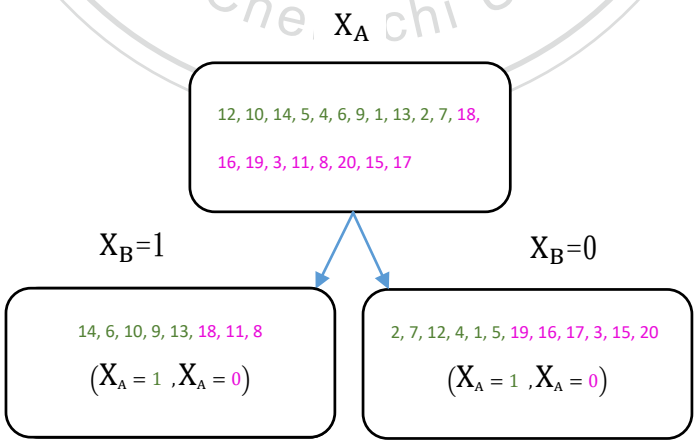


圖 4-2 基於變數 B 計算變數 A entropy 示意圖

DOI:10.6814/NCCU201900996

此計算結果可視為將變數 B 分為兩類別，如圖 4-3，其中綠色代表分類為 1，紅色代表分類為 0。

$$\sum_{i=0}^{1} P(X_A = i)H(X_B|X_A = i) =$$

$$\frac{11}{20}\left(\frac{-5}{11}log_2\frac{5}{11} + \frac{-6}{11}log_2\frac{6}{11}\right) + \frac{9}{20}\left(\frac{-6}{9}log_2\frac{6}{9} + \frac{-3}{9}log_2\frac{3}{9}\right)$$

$$= 0.5467166 + 0.4132331 = 0.9599497 \approx 0.9599$$

$$H(X_B) = \frac{1}{1}\left(\frac{-8}{20}log_2\frac{8}{20} + \frac{-12}{20}log_2\frac{12}{20}\right) = 0.9709506 \approx 0.9710$$

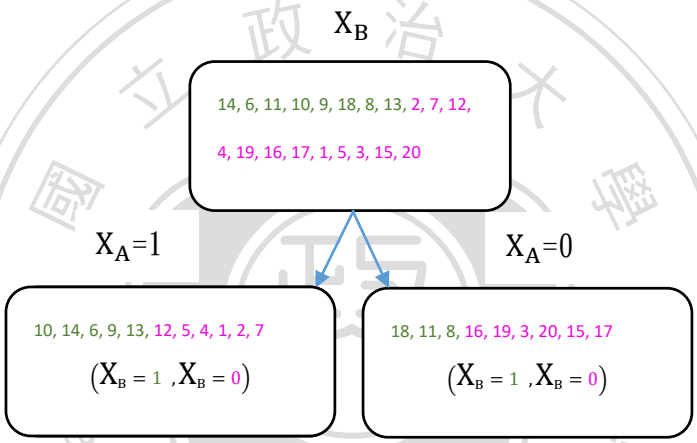


圖 4-3 基於變數 A 計算變數 B entropy 示意圖

因此變數 A 與變數 B 之 Mutual Entropy 為

$$ME(X_A, X_B) = \frac{\frac{\sum_i P(X_B = i)H(X_A|X_B = i)}{H(X_A)} + \frac{\sum_i P(X_A = i)H(X_B|X_A = i)}{H(X_B)}}{2} =$$

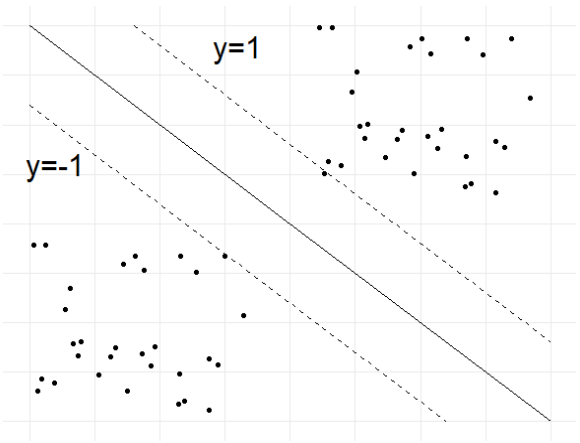
$$\frac{\frac{0.9818}{0.9928} + \frac{0.9599}{0.9710}}{2} = \frac{0.9890 + 0.9886}{2} = \frac{1.9776}{2} = 0.9888$$

DOI:10.6814/NCCU201900996

三、Support Vector Machine(SVM)

Support Vector Machine(SVM)最初由 Vapnik 等人於 1963 年提出，後經由 Boser(1992)、Cortes(1995)、Guyon(1993)、Scholkopf(1997)及 Vapnik(1998)等人研究補充，被大量應用在機器學習領域、分類與迴歸等。SVM 為一優秀分類器，其主要概念為找到一標準使得每個類別有最大程度的分離，因在某些維度下資料可能無法清楚分類，為使得每個類別有最大程度的分離，SVM 會經由 kernal function 將資料投影到更高維度，以平面或是超平面將資料分開，並且會先在分類間之分別最相近的資料點建立兩平行之線(Support Vector)、平面(Support Plane)或是超平面(Support Hyperplane)，分別記為 C1、C2，而此之中點所形成的線、平面或是超平面即最佳解，記為 C，此過程嘗試極大化 C 與 C1 即 C2 的距離如圖 4-4。

以分為兩類為例，$y = \begin{cases} 1, & if\ \ x_i\ \ in\ \ class1 \\ -1, & if\ \ x_j\ \ in\ \ class2 \end{cases}$，$C1: w`x + b = 1$，$C2: w`x + b = -1$，$C: w`x + b = 0$，目標即極大化 Margin1：$|C - C1|$以及$|C - C2|$。



$C1:\ w`x + b = 1$

$C2:\ w`x + b = -1$

$C: w`x + b = 0$

圖 4-4 SVM 示意圖

DOI:10.6814/NCCU201900996

四、Feature selection via Mutual Entropy based on classified model

假設有一筆m x n資料，分別取X_1至X_n搭配Y以 Support Vector machine(SVM) 分類模型建立模型，此時會有 n 個 SVM 模型，模型建立後再將X_1至X_n之樣本分別帶入相對應的X_1至X_n SVM 模型(圖 4-5)，藉由模型分類結果 $\hat{Y}$以 Mutual Entropy 計算變數間之相關值。建立變數間 Mutual Entropy 矩陣(為一 n by n 矩陣)(圖 4-6)，取矩陣之行和$\mathrm{colsum}_{Xi}, i = 1 \ldots n$，若該行之和為 0 代表此變數無解釋力，並給定該行之和為 n，以colsum分為數群，記為$C_k, 1 \le k \le n$。以各變數群最小colsum量當作該群指標$\mathrm{Volume}(C_k) = \min\{\mathrm{colsum}_{Xi}, X_i \epsilon C_k\}$，先行選取$\mathrm{Volume}(C_k)$最小值所對應群之變數，評估重要性變數篩選，以及準確度等指標後，再加入次大的變數群評估，以此類推。

因各變數之colsum計算來自 Mutual Entropy 矩陣，而此矩陣代表變數間的關係，又 Mutual Entropy 特性為數字愈小代表兩變數關係愈大，因此colsum愈小代表該變數與除了本身以外變數之關係愈大，亦即此變數帶有較多其他變數之訊息，於建模評估上會有較好效果，因此以最小值開始選取。

	X_1	X_2	...	X_{n-1}	X_n	Y
Sample 1						
Sample 2						
Sample 3						
...						
Sample m						
model	SVM model based on X_1	SVM model based on X_2	...	SVM model based on X_{n-1}	SVM model based on X_n	
output	$\hat{Y}_1$, based on above model with data X_1	$\hat{Y}_2$, based on above model with data X_2	...	$\hat{Y}_{n-1}$, based on above model with data X_{n-1}	$\hat{Y}_n$, based on above model with data X_n	

圖 4-5 本研究方法建模及取得估計值之示意圖

$\widehat{Y_K} := Y_K \mid model\ X_K$：*以$X_K$變數建立模型後再將$X_K$代入的結果*

DOI:10.6814/NCCU201900996

$$\begin{bmatrix} E_{1,1} & E_{1,2} & \cdots & E_{1,n-1} & E_{1,n} \\ E_{2,1} & E_{2,2} & \cdots & E_{2,n-1} & E_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ E_{n-1,1} & E_{n-1,2} & \cdots & E_{n-1,n-1} & E_{n-1,n} \\ E_{n,1} & E_{n,2} & \cdots & E_{n,n-1} & E_{n,n} \end{bmatrix}$$

圖 4-6 Mutual Entropy 矩陣示意圖

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

五、分群規則

1. Silhouette Coefficient：

測量該筆樣本在分為 n 群時，n=1, 2, 3…，其對應的適合程度

$$s(k) = \frac{a(k) - b(k)}{\max\{b(k), a(k)\}} \qquad (4\text{-}4)$$

$$-1 \leq s(k) \leq 1$$

b(k)：資料點 k（第 k 筆樣本)與群內其他點的平均距離

a(k)：資料點 k(第 k 筆樣本)與群外其他點的平均距離(最近群)

s(k)愈大表示該群之樣本與最近群之距離愈遠。

2. 本研究以階層式分群為分群方法，為方便比較且某些資料大多情況 Silhouette Coefficient 建議之分群結果為 2 至 3 群，導致每一群依然具有大量變數，如此便會失去特徵選取在觀察少量變數下對預測結果的能力，因此將資料限定群數，根據變數數量選定 k 群中 Silhouette Coefficient 最大的群數，當作該矩陣最適合之分群數。

其中 k 會依照以下規則，除變數數量小於等於 25，其餘以變數數量除以 5 對應之區間，選取分群數量，每 5 為一區間，所使用之資料分群數如 4-1。

變數數量	分群數量
小於等於 25	3 至 5
26-50	6-10
51-75	11-15
76-100	16-20
…	…

DOI:10.6814/NCCU201900996

本研究使用資料之分群數整理如表 4-1

表 4-1 本文所用資料之分群數

Dataset	Features	Clusters
Simulated data 1	1000	196-200
Simulated data 2	1000	196-200
Breast Cancer	30	6-10
Sonar	60	11-15
Fraud Detection	30	6-10
Parkinson's Disease	22	3-5
APS Failure	112	21-25

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

六、方法步驟

本文提出之基於 Mutual Entropy 及 SVM 之特徵選取方法之步驟為

1. 每個變數 X_i, i=1, …, n 分別以 SVM 模型建模。

2. 將X_i之變數資料帶入X_i建立之 SVM 模型，得到 $\widehat{Y}_i$。

4. 以此結果計算 Mutual Entropy 值$ME(X_k, X_m)$並建立 Mutual Entropy 矩陣如圖 4-6。

5. 以 Mutual Entropy 矩陣計算colsum，將colsum分群並取得最佳分群數。

6. 取每一群變數$\{X_i \mid X_i \in C_k\}$之最小colsum作為該群變數選取指標，

$Volume(C_k) = \min\{colsum_{Xi}, X_i \epsilon C_k\}$。

7. 由最小colsum對應之變數群先行選取建模。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

本研究結果以半監督式學習執行並依照下列步驟：

1. 以半監督式學習建立模型與預測，模擬資料 bootstrap 50 次，低維度資料 bootstrap 500 次，而高維度資料則不進行 bootstrap，並切割 70%為訓練資料與 30%為測試資料。

2. 每個變數 X_i, $i=1, \cdots, n$ 分別以 SVM 模型建模。

3. 將建立該建立 SVM 模型之變數資料帶入此模型得到 $\hat{Y}_i$。

4. 以此結果計算 Mutual Entropy 值$ME(X_k, X_m)$並建立 Mutual Entropy 矩陣。

5. 將數個 Mutual Entropy 矩陣分別依照研究方法所述之方法給予colsum，colsum以階層式分群做為分群方式，並以 Silhouette Coefficient 取得最佳分群數 c_i，將之相加取平均得到平均群數$\bar{c}$，並取四捨五入，每個矩陣皆分為$\bar{c}$群。

6. 取每一群變數$\{X_i \mid X_i \in C_k\}$之最小colsum作為該群變數選取指標，$\mathrm{Volume}(C_k) = \min\{\mathrm{colsum}_{Xi}, X_i \epsilon C_k\}$。

7. 取最小colsum對應之變數建模，並帶入測試資料計算，評估重要變數選取結果、測試集之平均敏感度及其標準差、平均特異度及其標準差與平均準確度及其標準差以及所使用之平均變數數量，再加入次大評估比較並重複此步驟。

DOI:10.6814/NCCU201900996

第二節 其他特徵選取方法

一、Random forest(RF)

基於 Leo Breiman (1996)提出之 bagging 算法(bootstrap aggregating)與 Tin Kam Ho(1995)隨機子集法(random subspace method)所產生的一種分類演算法，在本文中，隨機森林(random forest)將被用來進行特徵選取(feature selection)。

對原始資料之變數(variable)與樣本(sample)進行取後放回的隨機抽樣，舉例而言有一筆 m 個樣本與 n 個變數的資料集 D，經由 bootstrap 步驟會隨機產生 k, k≥2 筆樣本大小為 m 的新資料集，記為$D_i, \forall i = 1 \ldots k$，此 k 筆資料分別以隨機子集合方式以決策樹(decision tree)建模，並且在各個節點只考慮隨機抽取的 p個變數，且$p < n$，當有新資料進來時分別代入 k 個決策樹並進行投票或平均，得到最後的預測結果。

本文 k 設定 500，p設定為$\sqrt{n}$，以 Gini 計算節點資訊亂度，並計算節點到節點的 Gini 減少量，第 i 個變數之減少量記為$I\ (V_i), \forall i = 1 \ldots n$ ，以此減少量當作衡量變數重要性的指標，並依$I\ (V_i)$來選取變數。

DOI:10.6814/NCCU201900996

二、F-score

F-score 為一簡單評估變數對於二元分類能力之指標，一變數有著愈大之 F-score 代表此變數有著愈佳的分類能力，愈小表示此變數分類能力較弱，F-score 因只考慮單一維度(變數本身)之結果，雖簡單執行、便於運算，但缺乏了更高維度下(變數間關係)的評估，設一筆資料有兩種分類，+代表 1、-代表 0，則變數X_k之 F-score 值可定義為

$$\frac{\left(\overline{X_k}^{[+]}-\overline{X_k}\right)^2+\left(\overline{X_k}^{[-]}-\overline{X_k}\right)^2}{\frac{1}{n^+-1}\sum_{i=1}^{n^+}\left(X_{k,i}^{[+]}-\overline{X_k}^{[+]}\right)^2+\frac{1}{n^--1}\sum_{i=1}^{n^-}\left(X_{k,i}^{[-]}-\overline{X_k}^{[-]}\right)^2} \quad (4\text{-}5)$$

$\overline{X_k}^{[+]}$：該變數屬於分類 1 之樣本平均數

$\overline{X_k}^{[-]}$：該變數屬於分類 0 之樣本平均數

$\overline{X_k}$：該變數之樣本平均數

n^+：該變數屬於分類 1 之樣本數

n^-：該變數屬於分類 0 之樣本數

$X_{k,i}^{[+]}$：該變數屬於分類 1 之第 i 個樣本數

$X_{k,i}^{[-]}$：該變數屬於分類 0 之第 i 個樣本數

本研究分別計算每個變數之 F-score 值，並將 F-score 值分群，分群規則與 Mutual Entropy 特徵選取相同，分別計算每群 F-score 值總和，從最大群開始選取建模並帶入測試資料評估準確度等準則，再加入次大評估，以此類推。

DOI:10.6814/NCCU201900996

三、Least Absolute Shrinkage and Selection Operation(LASSO)

LASSO 算法最早由 Tibshirani(1996)提出，其概念為修改線性迴歸之 MSE，不同於迴歸$\min_{\beta_0,\beta}\sum_{i=1}^{n}(y_i-\beta_0-x_i^T\beta)^2$，LASSO 為最小化$\min_{\beta_0,\beta}\sum_{i=1}^{n}(y_i-\beta_0-x_i^T\beta)^2+\lambda\sum_{i=1}^{p}|\beta_i|$，此即 L1 正規化，將係數大小加入 MSE 做評估，若模型中有較大之係數，對於在最小化 MSE 時將會產生影響，並可以避免以此估計出的係數建立之模型對於訓練資料過度配適(overfitting)，雖然脊迴歸也是修改 MSE 對於最小化 MSE 估計參數時產生影響，但其並不會將接近 0 的係數所減為 0，而 LASSO 會使得接近 0 的係數往 0 的方去修正，因此 LASSO 也可以用於特徵選取，而λ可經由 cross validation 方式去估計出較適合的λ值。

本研究先以 cross validation 取得較適合之λ值，並觀察在該λ值時那些變數對應之係數不為 0，這些不為 0 之係數即本次特徵選取欲選取之目標變數，將這些變數取出以訓練資料建模並將測試資料代入評估準確度等準則。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

四、Fast Correlation-Based Filter(FCBF)

FCBF 為 Lei Yu(2003)等提出之特徵選取方法，與本文同樣使用 Entropy 概念，基於評估變數間相關性並考慮類別應變數影響而建立的一方法，包含兩方向，第一，如何決定特徵與類別應變數是否相關，第二，如何決定某特徵當考慮其他相關特徵時此特徵是否為冗餘(redundant)的，也同樣計算$H(X_k)$以及$H(X_k|X_m)$並定義資訊獲得(information gain)為

$$IG(X_k|X_m) = H(X_k) - H(X_k|X_m) \tag{4-6}$$

為保持基準，即為使值介於 0 至 1 之間，定義 symmetrical uncertainty

$$SU(X_k, X_m) = 2\left[\frac{IG(X_k|X_m)}{H(X_k) + H(X_m)}\right] \tag{4-7}$$

基於上式，根據第一方向，為加速搜尋速度，先分別計算每個變數與類別 C 之$SU(X_i, C)$並給定門檻值δ，先將所有$SU(X_i, C) > \delta$所對應的變數取出來，接著進行第二方向，定義 Predominant Correlation 與 Predominant Feature

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

1. Predominant Correlation

變數X_i與類別 C 為 Predominant 若且為若$SU(X_i, C) > \delta$以及對於所有X_i沒有一個X_j使得$SU(X_i, X_j) > SU(X_i, C)$ $\forall i \neq j$，若是對於X_i有此類X_j存在，便稱X_j為X_i的 redundant peer，並使用符號Sp_i來定義X_i的 redundant peer 的兩集合，Sp_i分為兩類Sp_i^+以及Sp_i^-，$Sp_i^+ = \{X_j | X_j \epsilon Sp_i, SU(X_j, C) > SU(X_i, C)\}$以及$Sp_i^- = \{X_j | X_j \epsilon Sp_i, SU(X_j, C) \leq SU(X_i, C)\}$。

2. Predominant Feature

若一變數對於類別是 predominant 若且為若他對類別的相關性是 predominant 或是藉由移除他的 redundant peer 變成 predominant。

若兩變數發現彼此是冗餘的，先刪除與類別變數較不相關的那個，接著進行三個規則搜尋變數

1. 若$Sp_i^+ = \emptyset$，則X_i為 predominant，刪除Sp_i^-的所有變數。

2. 若$Sp_i^+ \neq \emptyset$，在決定X_i前會先行處理在Sp_i^+內的變數，若沒有一個變成 predominant 則執行第 1 點，否則只刪除X_i以及決定是否刪除Sp_i^-內的變數。

3. 擁有最大$SU(X_i, C)$值的變數，可以當作起始點來刪除其他變數。

DOI:10.6814/NCCU201900996

第五章 資料介紹

一、模擬資料一

資料包含 1000 個連續自變數與 1 個類別應變數，其資料抽取自以下兩分配，共抽取 200 個樣本，其中標記為 0 與標記為 1 之樣本各 100 個，標記為 0 之樣本抽自分配$X+Y\times Z_1$，標記為 1 之樣本前 950 個變數抽自$X+Y\times Z_1$，後 50 個變數抽自$X+Y\times Z_2$，並定義此 50 個變數為重要變數，資料視覺化圖形如圖 5-1，樣本大致交錯分散，標記為 1 的樣本相對於標記為 0 的樣本有較大離散。

variable		label
1000 $(X+Y\times Z_1)$		0
950 $(X+Y\times Z_1)$	50 $(X+Y\times Z_2)$	1

令 X_1, X_2, Y, Z_1, Z_2 為隨機變數，分別來自不同的分配如下，

$$X \sim \mathrm{Normal}(0.5, \mathrm{sd}=0.01)$$

$$Y=\begin{cases} 1, & \frac{1}{2} \\ -1, & \frac{1}{2} \end{cases}$$

$$Z_1 \sim \mathrm{Exp}(\beta=\frac{1}{0.05})$$

$$Z_2 \sim \frac{1}{2}\mathrm{Beta}(5,1)$$

DOI:10.6814/NCCU201900996

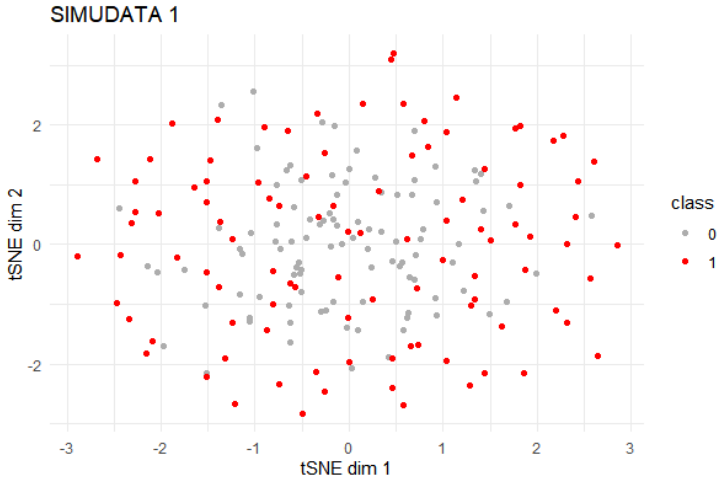


圖 5-1 模擬資料 1 視覺化圖形

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

二、模擬資料二

資料包含 1000 個連續自變數與 1 個類別應變數，其資料抽取自以下兩分配，共抽取 200 個樣本，其中標記為 0 與標記為 1 之樣本各 100 個，標記為 0 之樣本抽自分配$X_1 + Y \times Z$，標記為 1 之樣本前 950 個變數抽自$X_1 + Y \times Z$，後 50 個變數抽自$X_2 + Y \times Z$，並定義此 50 個變數為重要變數，資料視覺化圖形如圖 5-2，兩群樣本有非常明顯的群聚。

<table>
<tr><td colspan="2">variable</td><td>label</td></tr>
<tr><td colspan="2">1000
$(X_1 + Y \times Z)$</td><td>0</td></tr>
<tr><td>950
$(X_1 + Y \times Z)$</td><td>50
$(X_2 + Y \times Z)$</td><td>1</td></tr>
</table>

令 X_1, X_2, Y, Z_1, Z_2 為隨機變數，分別來自不同的分配如下，

$$X_1 \sim \frac{9}{10}\mathrm{Beta}(5,1)$$

$$X_2 \sim \frac{9}{10}\mathrm{Beta}(2,5)$$

$$Y = \begin{cases} 1, & \frac{1}{2} \\ -1, & \frac{1}{2} \end{cases}$$

$$Z \sim \mathrm{Normal}(0.1, \mathrm{sd} = 0.005)$$

DOI:10.6814/NCCU201900996

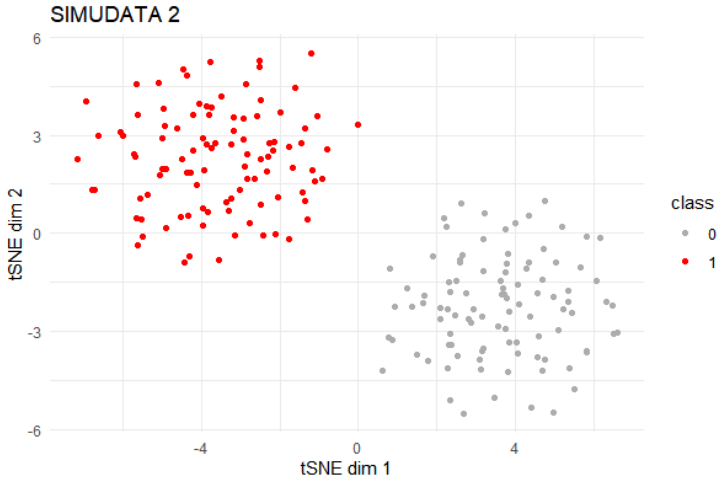


圖 5-2 模擬資料 2 視覺化圖形

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

三、Wisconsin Diagnostic Breast Cancer

威斯康辛乳癌診斷資料集，取自 UCI Machine Learning Repository，為 William H. Wolberg 等人於 1995 年所建立之資料，各變數之資料來自乳房腫塊之醫學影像計算而得，此筆資料包含 569 個樣本，30 個連續自變數(dependent variable)與 1 個類別應變數(independent variable)，其中連續自變數包含半徑、周長、凹點、緊密程度等，類別應變數為一二元分類之變數，將陽性樣本標記為 M(malignant)共 212 個，陰性樣本標記為 B(benign)共 357 個，此筆資料並無 NA 值，圖 5-3 為將資料標準化後視覺化的呈現，可能有少量陽性樣本被錯誤辨識，但整體敏感度與特異度表現應都不錯，且因 B 跟 M 兩類樣本有明顯群聚，且彼此間有明顯分別，故最終準確度表現應也會不錯。

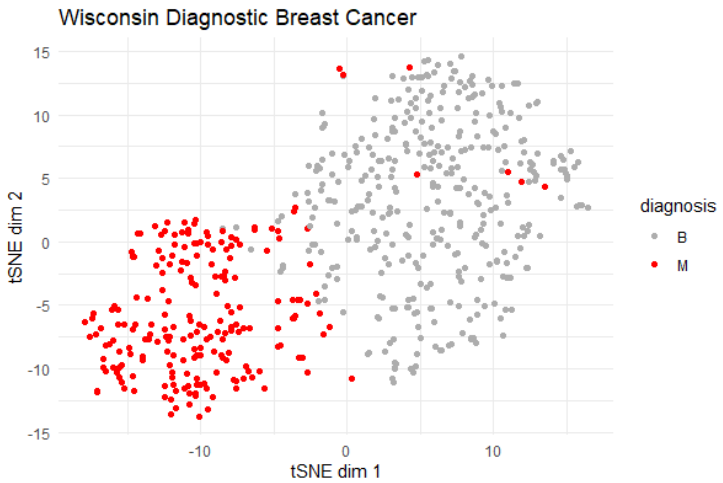


圖 5-3 乳癌資料集視覺化圖形

DOI:10.6814/NCCU201900996

四、Connectionist Bench(Sonar, Mines vs. Rocks)

聲納探測資料集，取自 UCI Machine Learning Repository，為 Gorman Sejnowski 等人於研究中使用，用以分別聲納回波是否來自金屬圓柱，此筆資料包含 208 個樣本，60 個連續自變數與 1 個類別應變數，每組樣本數字代表特定頻帶(particular frequency band)內的能量在一段時間內的積分，連續自變數之樣本數值都介於 0 到 1 之間，類別應變數為一二元分類之變數，將判斷結果為金屬圓柱標記為 M(Metal Cylinder)共 111 個、石頭標記為 R(Rock)共 97 個，此筆資料並無 NA 值，將資料標準化後視覺化的呈現如圖 5-4，因標記為 M 及標記為 R 的樣本大致交錯，可能會造成敏感度與特異度都不佳，或是其中之一極佳另一極差的情況，因此在使用少量變數建模的情況下預測準確度應不會太高。

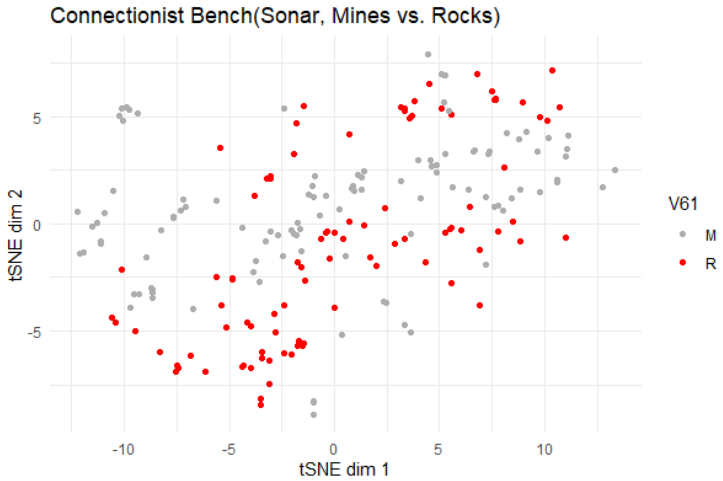


圖 5-4 聲納資料集視覺化圖形

DOI:10.6814/NCCU201900996

五、Credit Card Fraud Detection

信用卡詐騙資料集，取自 Kaggle，為 Worldline 及 ULB（Université Libre de Bruxelles)機器學習小組蒐集之資料，此筆資料包含 285299 個樣本，30 個連續自變數與 1 個類別應變數，此筆資料為保護當事人因此為經由 PCA 轉換後取 30 主成分而得的結果，類別應變數為一二元分類之變數，將詐欺結果標記為 1，正常結果標記為 0，其中標記為 1 的樣本有 492 筆，標記為 0 的樣本有 284807 筆，因此筆資料為極度不平衡資料且本研究過程對資料 bootstrap500 次，因此以 R 軟體固定 seed 對標記為 0 的樣本進行抽樣 1476 筆，形成一筆 30 個連續自變數與 1 個類別應變數及 1968 個樣本之資料，此筆資料並無 NA 值，如圖 5-5，雖有部分標記為 1 與標記為 0 的樣本交錯，可能造成敏感度較低，但整體上兩類樣本依然有明顯群聚，故最終準確度表現應會不錯。

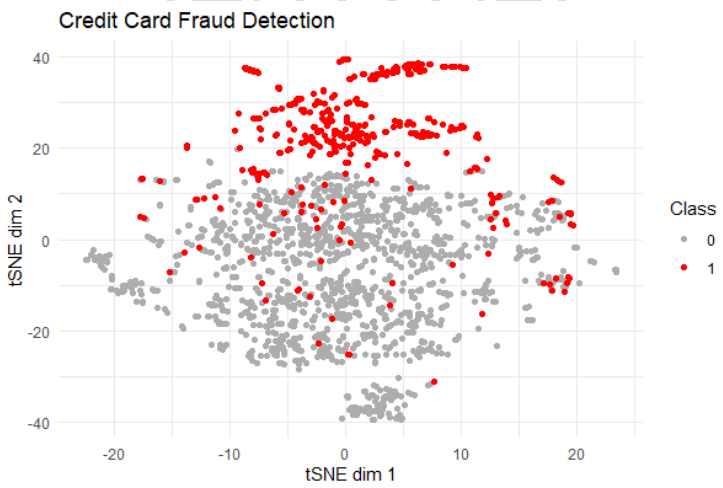


圖 5-5 信用卡詐騙資料集視覺化圖形

DOI:10.6814/NCCU201900996

六、Oxford Parkinson's Disease

帕金森氏症資料集為 Max Little 建立，藉由生物醫學語音測量(biomedical voice measurements)研究該名人員是否患有帕金森氏症，共測量 31 名人員，每名人員測量若干次，此筆資料共包含 195 個樣本，22 個連續自變數與 1 個類別應變數，每個變數代表一種特定語音測量(particular voice measure)，每組樣本為人員經由 22 種量測方式後的數值結果，類別應變數為一二元分類之變數，患病之結果標記為 1 共有 147 個，無患病標記為 0 共有 48 個，此筆資料並無 NA 值，在資料視覺化的圖 5-6 中標記為 0 的情形下樣本有明顯群聚，為 1 的情形下雖有部份也有明顯群聚，但有大量樣本與 status 為 0 的樣本重疊，此資料有可能敏感度不錯特異度差，因會將大部分樣本會傾向預測為標記為 1 的部分，標記為 0 的樣本可能無法正確預測，或是敏感度差特異度佳，因在左上角群之資料可能被視為 0，那麼變有大量標記為 1 之樣本被錯誤預測。

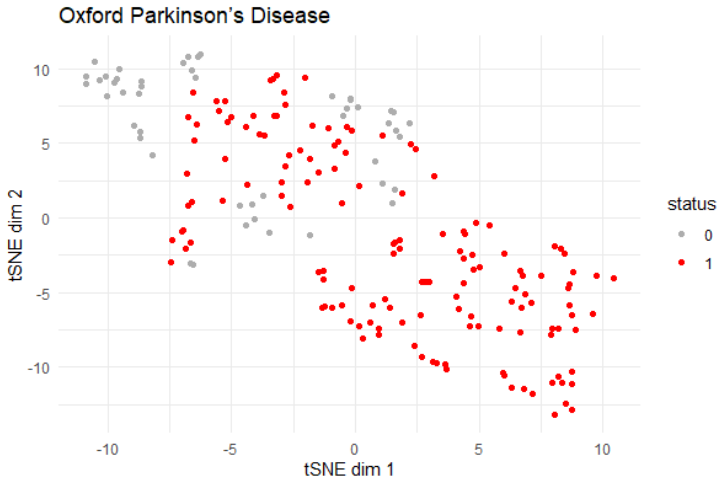


圖 5-6 帕金森氏症資料集視覺化圖形

DOI:10.6814/NCCU201900996

七、APS Failure at Scania Trucks

空氣壓力系統資料集為 Scania 汽車工業股份有限公司建立，取自 UCI Machine Learning Repository，蒐集 Scania 卡車日常數據，關於空氣壓力系統應(Air Pressure System, APS)用於卡車的功能，例如煞車和換檔，對於卡車故障情況，判斷是否是與空氣壓力系統有關，此筆資料原始共包含 76000 個樣本，170 個自變數與 1 個類別應變數，若故障與空氣壓力系統有關，則標記為 pos.（positive)，若與空氣壓力系統無關，則標記為 neg.（negative)，此筆資料含有 NA 值及某些變數包含大量 0，且本研究過程對資料 bootstrap500 次，因此以 R 軟體固定 seed 對標記為 neg. 的樣本進行抽樣 1359 筆，進行以下處理先行初步篩選變數及樣本，對於 NA 值不插補

1. 將 NA 值超過樣本數兩成之變數刪除，亦即刪除含有超過 15200 個 NA 值之變數。
2. 刪除含有 NA 值之樣本，此時資料已無 NA 值。
3. 對標記為 neg. 之樣本抽樣 1359 筆
4. 刪除含有現有樣本數八成以上 0 之變數

形成一筆 112 個自變數與 1 個類別應變數含有 1812 筆樣本之資料，其中類別自變數包含 453 個 pos(軟體標記為 1)與 1359 個 neg(軟體標記為 0)，此時資料並無 NA 值。

圖 5-7 為將資料標準化後視覺化的呈現，可看出有某些 pos. 與 neg. 被錯誤識別，但因標記為 pos 與標記為 neg 兩類樣本依然有有明顯群聚，應不致影響最終準確度太多。

DOI:10.6814/NCCU201900996

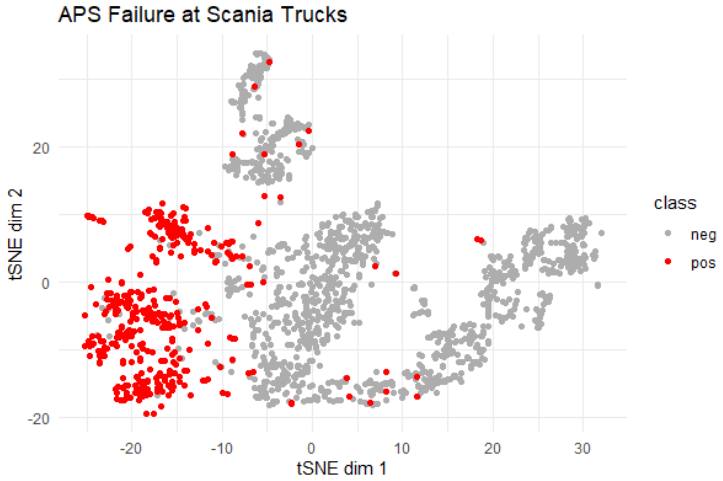


圖 5-7 空氣壓力系統資料集視覺化圖形

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

本研究使用之資料概況，如樣本數、變數數量等的整理如表 5-1。

表 5-1 各資料集樣本與變數概況

Dataset	Samples	Features	Classes	
Simulation data 1	200	1000	2	(1, 0)=(100, 100)
Simulation data 2	200	1000	2	(1, 0)=(100, 100)
Breast Cancer	569	30	2	(M, B)=(212, 357)
Sonar	208	60	2	(M, R)=(111, 97)
Fraud Detection	1968	30	2	(1, 0)=(492, 1476)
Parkinson's Disease	195	22	2	(1, 0)=(147, 48)
Aps failure	1812	112	2	(pos, neg)=(453, 1359)

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

第六章 研究結果

為評估特徵選取方法是否能正確選擇出重要性變數，且真實資料母體分配未知，因此欲定義真實資料重要性變數，本文以高斯混合模型(Gaussian mixed model)搭配 EM 演算法分別估計真實資料每個變數之分配，再將每個變數對應之標記為 1、M 及 pos.之樣本與估計出的分配帶入 KS 檢定，取其檢定統計量(test statistic)，檢定統計量愈大代表該變數重要性愈高。以本文使用之兩筆模擬資料測試此方法評估能力如圖 6-1，在後 50 個重要變數(橘色部分)，檢定統計量普遍較大，在前 950 個其他變數中(紫色部分)，檢定統計量普遍較小，此想法應能有效定義重要變數。

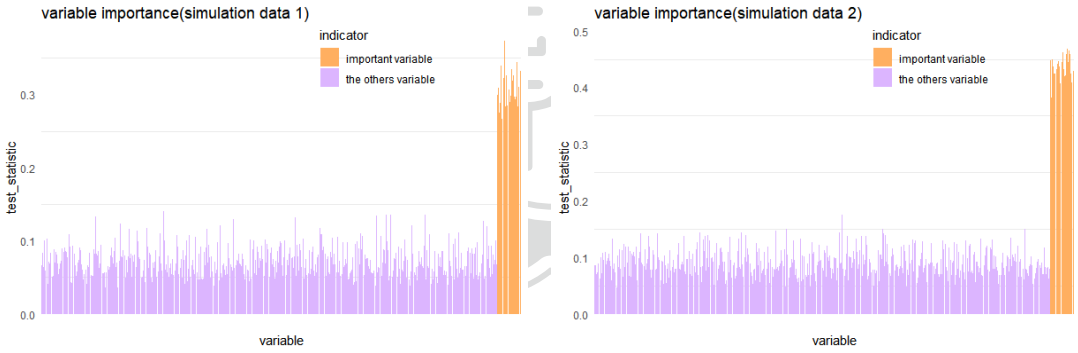


圖 6-1 定義重要性變數方法之評估能力

DOI:10.6814/NCCU201900996

圖 6-2 至圖 6-6 為此兩筆模擬資料於 50 次 bootstrap 中 50 個重要變數選取次數結果，結果顯示本研究方法能穩定選取重要變數。

第一節將五筆真實資料結果視覺化如圖 6-7 至圖 6-31 及表 6-3 至表 6-7，左圖為該筆資料各個變數之檢定統計量值由大排至小，以 bar chart 表示，左圖橘色部分顏色愈深，代表該變數在 bootstrap 多次選取中，被選取到愈多次，右圖代表該變數在 bootstrap 多次選取中，被選取實際次數之 bar chart，在每個方法使用差不多變數下進行結果評估，結果顯示本研究方法在選取重要性變數方面不但穩定且有不錯的效果，五筆真實資料選取之變數以 SVM 建模預測，並評估 500 次 bootstrap 之平均敏感度、平均特異度與平均準確度並提供一標準差區間，結果如表 6-3 至表 6-7，結果顯示使用本研究方法選取之變數，建模後預測也有不錯結果。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

第一節 資料結果

一、模擬資料一及模擬資料二

產生資料過程中已定義重要變數，兩筆模擬資料各 1000 個自變數，其中 50 個與產生過程中已定義為重要變數，因此不進行分配估計，圖 6-2 至圖 6-6 為模擬資料一及模擬資料二於 bootstrap 50 次中 50 個重要變數被選取次數，其於 1000 個變數中平均選取變數之數量如表 6-1 與 6-2，愈少變數條件下選擇到愈多次重要變數則結果愈佳，本文以兩種模擬資料評估方法在不同資料中選取重要變數之能力，於五種方法中，ME(圖 6-2)與 RF(圖 6-3)於兩筆模擬資料皆有穩定且不錯的表現，每個重要變數都被多次選取到，F-SCORE(圖 6-4)在模擬資料二(圖 6-4 右)有著極佳表現，但在模擬資料一(圖 6-4 左)之選取重要性變數表現不佳，應與此方法特性有關，F-SCORE 不考慮兩變數間關係，因此在樣本大致交錯的資料中有著較差的表現(圖 5-1)，於明顯群聚的資料中(圖 5-2)則有很好的表現，FCBF(圖 6-5)於兩筆模擬資料中也有不錯選取能力，LASSO(圖 6-6)選取能力則較差，於模擬資料二(表 6-2)中，LASSO 算法於 50 個重要變數平均只選取 4.14 變數，與產生資料中設定之 50 個重要變數有較大差距。

於兩筆模擬資料結果中，考量到變數間關係且使用了資訊理論相關之基礎的算法 ME、RF 與 FCBF 表現較為優秀，於本資料結果中 ME 與 RF 又稍為優於 FCBF。ME 在兩種不同資料，不論樣本間是否具有明顯群聚關係，皆有穩定表現，且有不錯的選取能力。

DOI:10.6814/NCCU201900996

表 6-1 模擬資料 1 各方法於平均 50 次 bootstrap 平均選取之變數數量

SIMUDATA1	Average num. of variable
ME	48.84
RF	49.12
F-SCORE	50.98
FCBF	51.66
LASSO	67.68

表 6-2 模擬資料 2 各方法於平均 50 次 bootstrap 平均選取之變數數量

SIMUDATA2	Average num. of variable
ME	55.6
RF	51.58
F-SCORE	50.9
FCBF	54.42
LASSO	4.14

國立政治大學 National Chengchi University

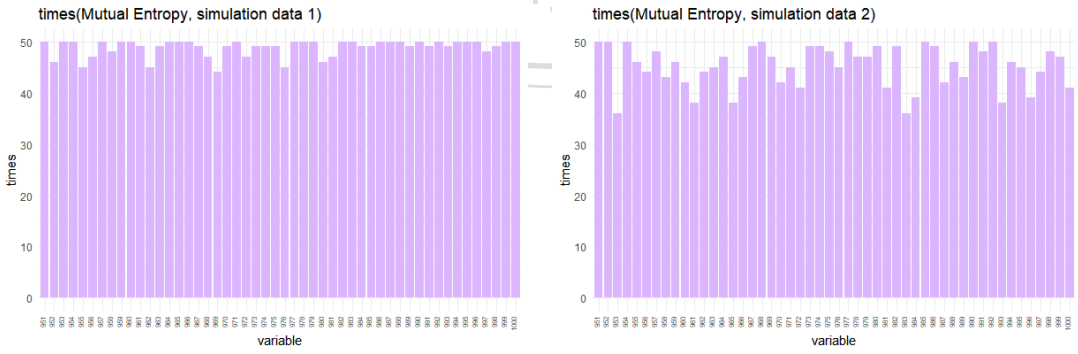


圖 6-2 ME 於模擬資料 1 及模擬資料 2 之表現

DOI:10.6814/NCCU201900996

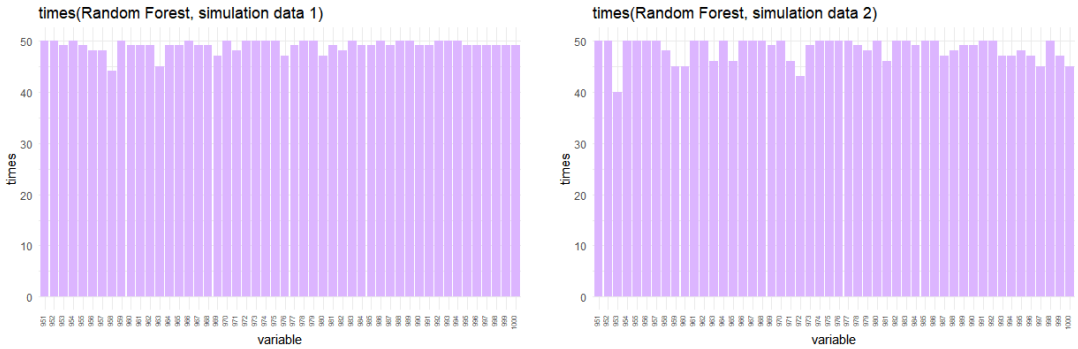


圖 6-3 RF 於模擬資料 1 及模擬資料 2 之表現

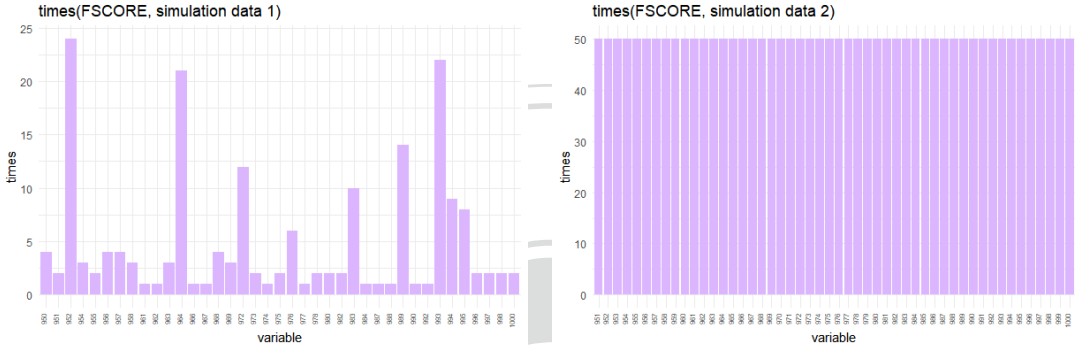


圖 6-4 F-SCORE 於模擬資料 1 及模擬資料 2 之表現

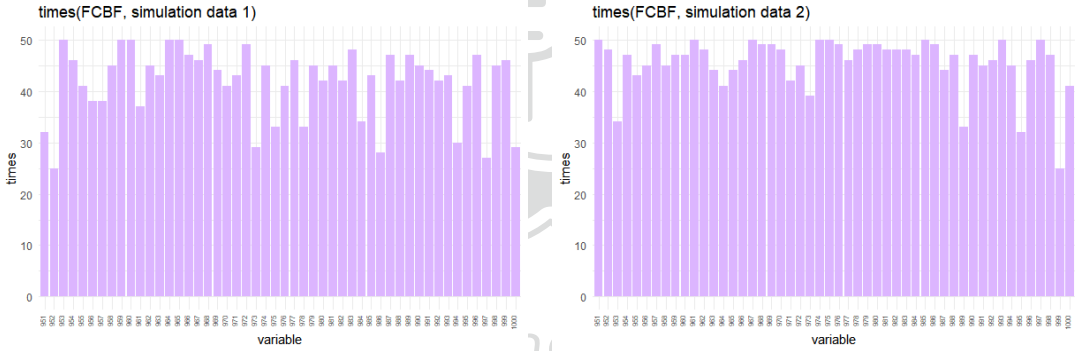


圖 6-5 FCBF 於模擬資料 1 及模擬資料 2 之表現

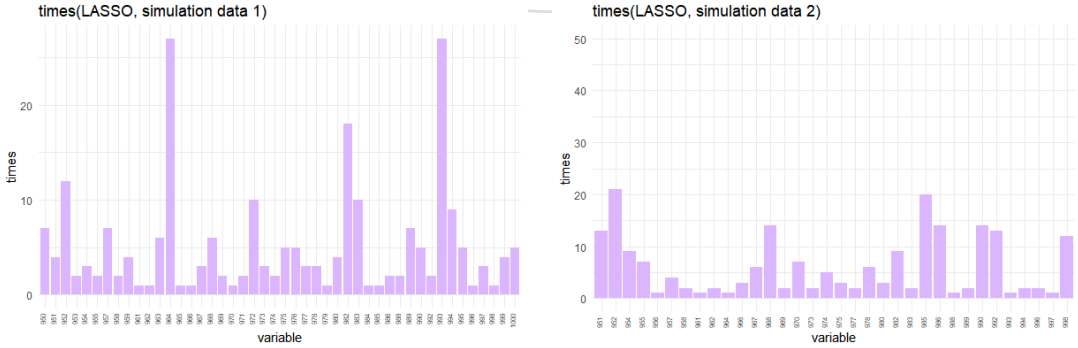


圖 6-6 LASSO 於模擬資料 1 及模擬資料 2 之表現

DOI:10.6814/NCCU201900996

二、Wisconsin Diagnostic Breast Cancer

在乳癌資料集中，在使用差不多變數情況下，ME(Mutual Entropy)、RF(Random Forest)、F-SCORE 與 FCBF，在重要性變數選取方面，選取到的變數皆為在圖 6-7 至圖 6-11 中左圖於有較大檢定統計量的變數，尤其以 ME 與 F-SCORE 在幾個前幾大變數中，於 500 次 bootstrap 中幾乎皆準確選擇(橘色較深部分)，ME 法平均選取多於 F-SCORE 兩個變數，也因此在前幾重要變數選取較為全面， RF 雖也有不錯的選擇能力，但可能因其演算法過程中的隨機性，並不是每次都能選擇到同樣的變數，而 FCBF 在選擇上於左圖雖也有靠左集中的趨勢，但在前幾個重要變數大約只有 400 次左右，LASSO 算法由於選擇較多變數，在各個變數的選擇呈現較為分散的情形，在此筆資料中，其重要性變數選擇上與其他四種方法比較稍差，於表 6-3 中，因此筆資料鑑別度高，各方法之敏感度與特異度及準確度皆不差，皆在一個標準差內，以 LASSO 為最高，但也用了比其他方法多兩倍左右的變數。

表 6-3 Wisconsin Diagnostic Breast Cancer 各項指標

ME	ME	RF	F-SCORE	FCBF	LASSO
Sensitivity	0.8841±0.0506	0.9077±0.0448	0.9075±0.0420	0.9418±0.0328	0.9523±0.0292
Specificity	0.9736±0.0219	0.9704±0.0220	0.9703±0.0216	0.9839±0.0141	0.989±0.012
Accuracy	0.9404±0.0216	0.9473±0.0183	0.9471±0.0176	0.9683±0.0146	0.9754±0.0123
Dimension	6.522	5.458	4.48	5.35	11.89

DOI:10.6814/NCCU201900996

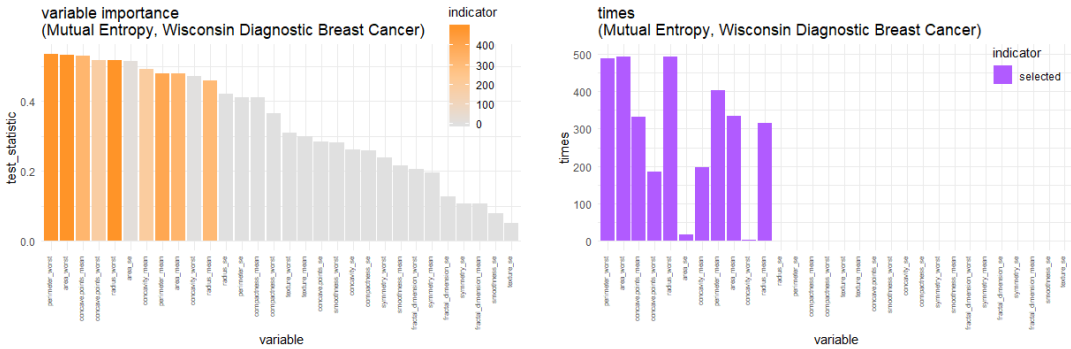


圖 6-7 Wisconsin Diagnostic Breast Cancer　ME 重要變數選取情況

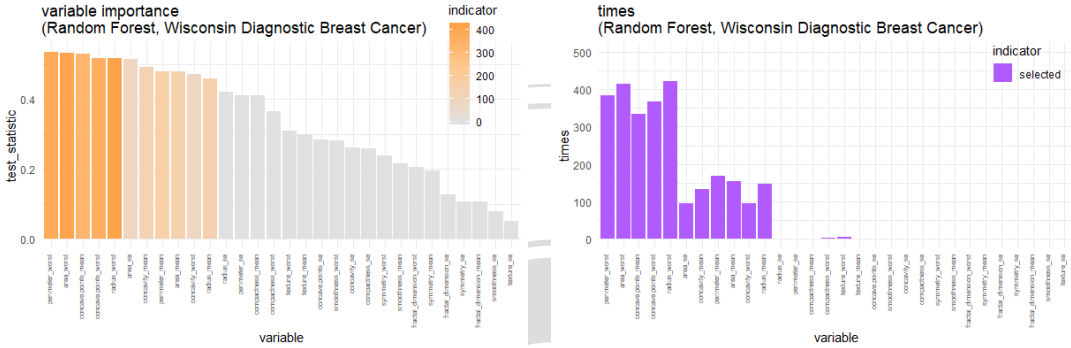


圖 6-8 Wisconsin Diagnostic Breast Cancer　RF 重要變數選取情況

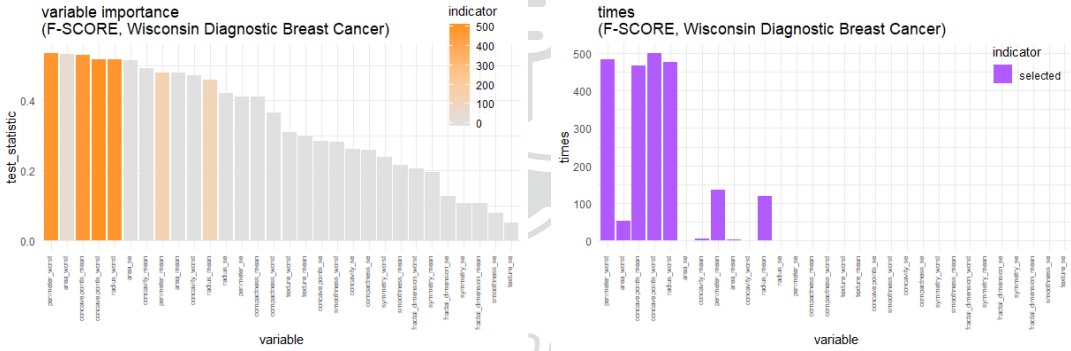


圖 6-9 Wisconsin Diagnostic Breast Cancer　F-SCORE 重要變數選取情況

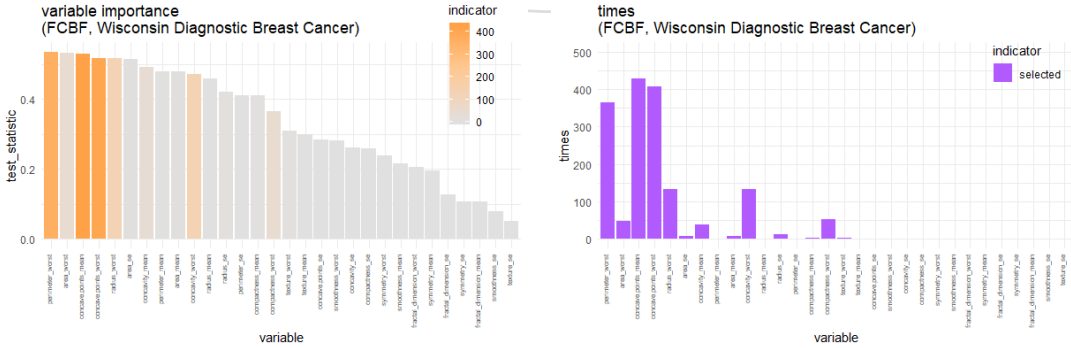


圖 6-10 Wisconsin Diagnostic Breast Cancer　FCBF 重要變數選取情況

DOI:10.6814/NCCU201900996

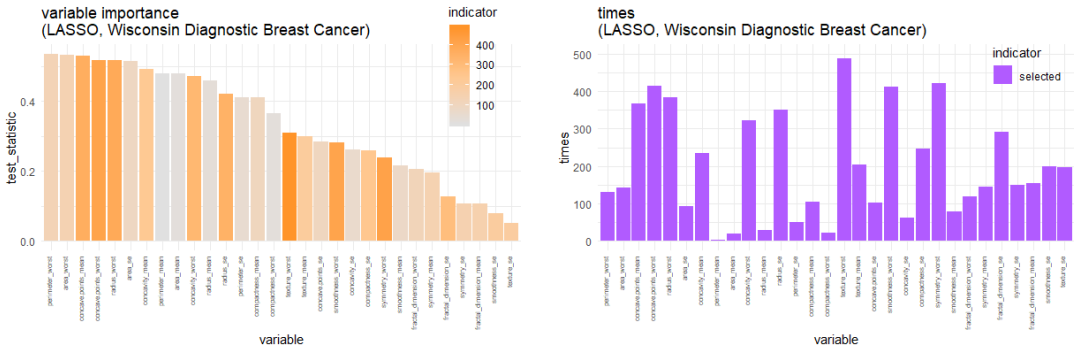


圖 6-11 Wisconsin Diagnostic Breast Cancer　LASSO 重要變數選取情況

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

三、Connectionist Bench(Sonar, Mines vs. Rocks)

在聲納資料集中，在使用差不多變數情況下，ME(Mutual Entropy)、F-SCORE 與 FCBF，在重要性變數選取方面，視覺化於圖 6-12、圖 6-14 與圖 6-15，在前幾大重要變數中，此三個方法皆能有效選取，於 500 次 bootstrap 中皆能選擇多次，F-SCORE 與 FCBF 選擇次數較多，但平均也比 ME 多出一個左右的變數，不過也可以看出這三個方法在前幾個重要變數選取方面皆大至能正確選取且選取全面，RF 在選取較為分散，應與其建立過程中的隨機性有關，LASSO 各變數皆選取多次，在本資料中較無法有效只選取重要變數，於表 6-4 中，準確度部分除了使用大量變數的 LASSO 外其他方法皆無超過 0.8，RF、FSCORE 與 FCBF 皆較 ME 高，不過也比 ME 多使用 1-2 個變數，各方法之一標準差區間也有所重疊。

表 6-4 Connectionist Bench(Sonar, Mines vs. Rocks) 各項指標

ME	ME	RF	F-SCORE	FCBF	LASSO
Sensitivity	0.8044±0.1416	0.8232±0.0950	0.8365±0.0943	0.8413±0.0877	0.9076±0.0641
Specificity	0.6510±0.1754	0.7324±0.1150	0.7047±0.1091	0.7357±0.1043	0.8510±0.0898
Accuracy	0.7368±0.0849	0.7832±0.0643	0.7772±0.0608	0.7938±0.0601	0.8828±0.0462
Dimension	4.878	6.714	5.426	5.506	30.146

DOI:10.6814/NCCU201900996

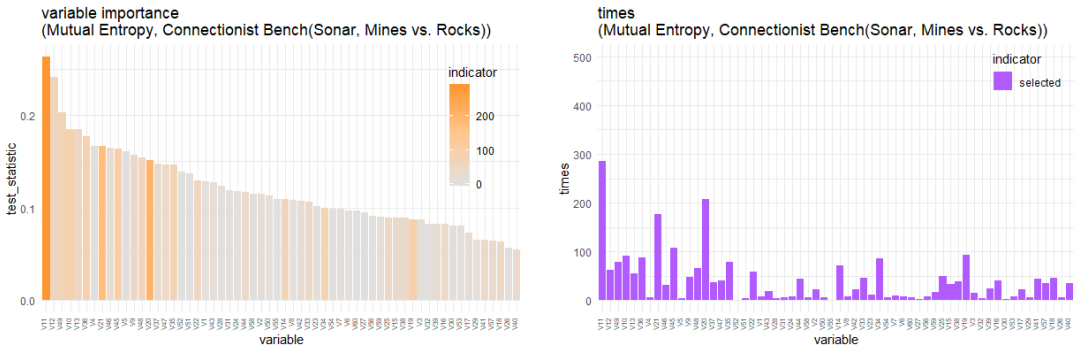


圖 6-12 Connectionist Bench(Sonar, Mines vs. Rocks) ME 重要變數選取情況

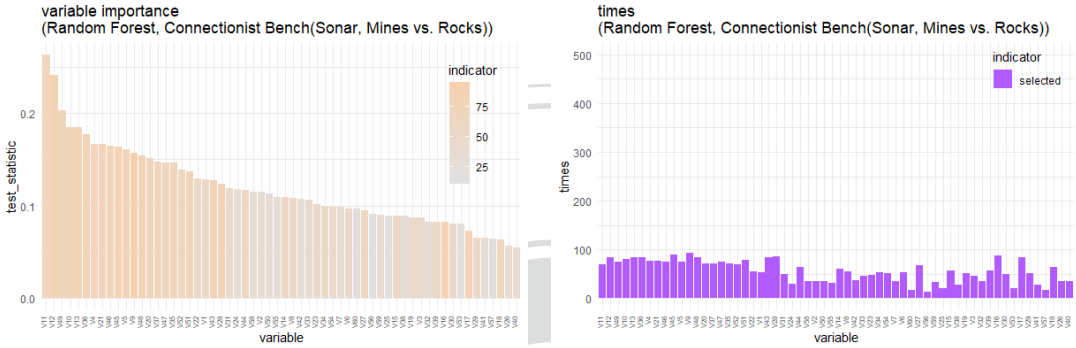


圖 6-13 Connectionist Bench(Sonar, Mines vs. Rocks) RF 重要變數選取情況

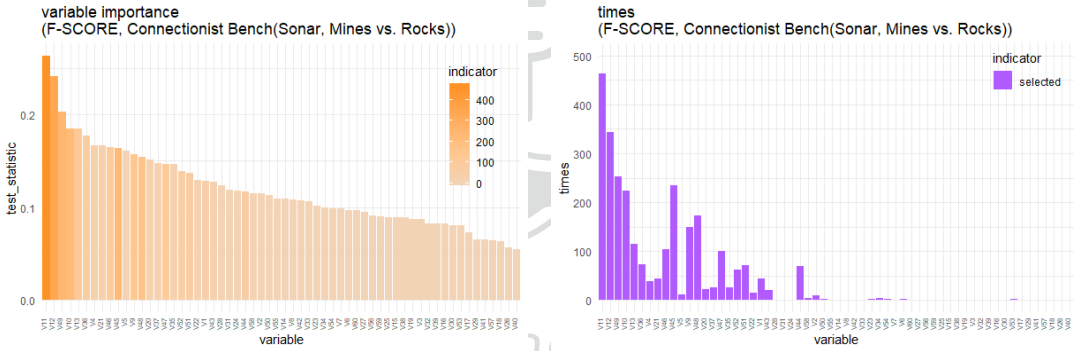


圖 6-14 Connectionist Bench(Sonar, Mines vs. Rocks) F-SCORE 重要變數選取情況

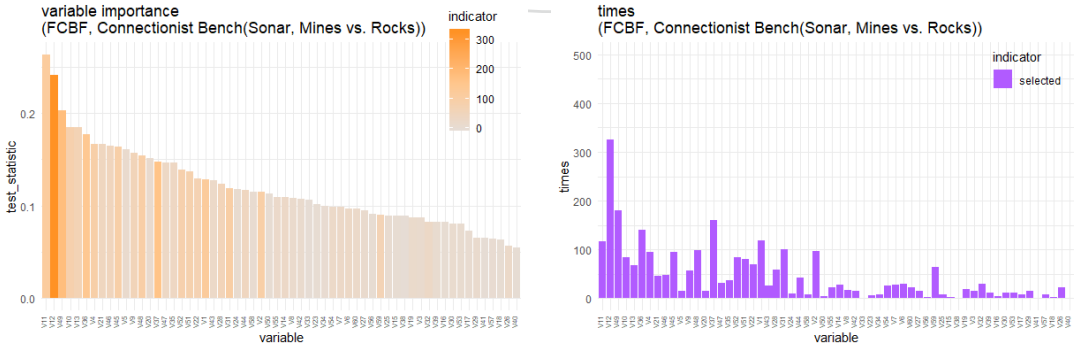


圖 6-15 Connectionist Bench(Sonar, Mines vs. Rocks) FCBF 重要變數選取情況

DOI:10.6814/NCCU201900996

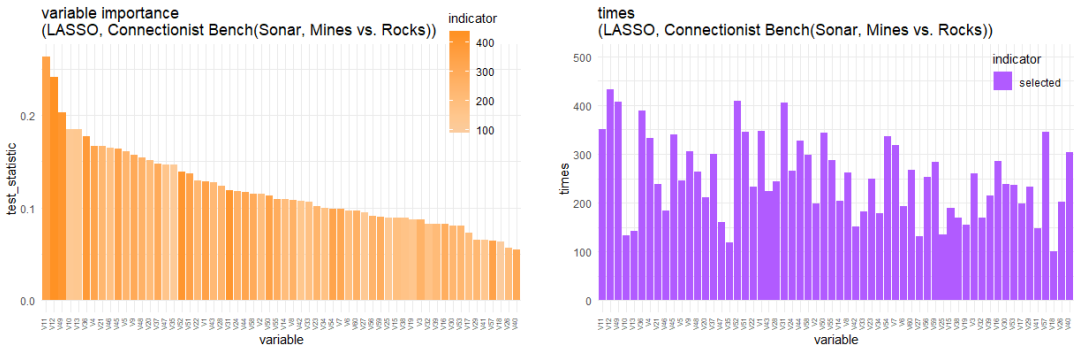


圖 6-16 Connectionist Bench(Sonar, Mines vs. Rocks) LASSO 重要變數選取情況

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

四、Credit Card Fraud Detection

在信用卡詐騙資料集中，在使用差不多變數情況下，ME(Mutual Entropy)、F-SCORE，在重要性變數選取方面，如圖 6-17 及圖 6-19，在前幾大重要變數中，皆能有效選取，於 500 次 bootstrap 中皆能選擇多次，ME 除了多次選取也更為全面，RF(圖 6-18)也有不錯選擇結果，FCBF(圖 6-20)在此筆資料中有些許情況在變數選擇上較為分散，LASSO(圖 6-21)於各個變數皆選取多次，在本資料中較無法有效判斷其選擇能力，於表 6-5 中，此筆資料視覺化圖形具明顯群聚，因此結果不錯，各方法敏感度、特異度與準確度皆有不錯結果，在本資料中各方法之預測能力無明顯區別。

表 6-5 Credit Card Fraud Detection 各項指標

ME	ME	RF	F-SCORE	FCBF	LASSO
Sensitivity	0.8397±0.0312	0.8508±0.0347	0.8689±0.0309	0.8209±0.0367	0.8817±0.0297
Specificity	0.9991±0.0022	0.9981±0.0028	0.9962±0.0035	0.9896±0.0064	0.9938±0.0046
Accuracy	0.9592±0.0081	0.9613±0.0084	0.9644±0.0077	0.9474±0.0101	0.9658±0.0075
Dimension	7.178	5.142	5.562	5.098	17.318

DOI:10.6814/NCCU201900996

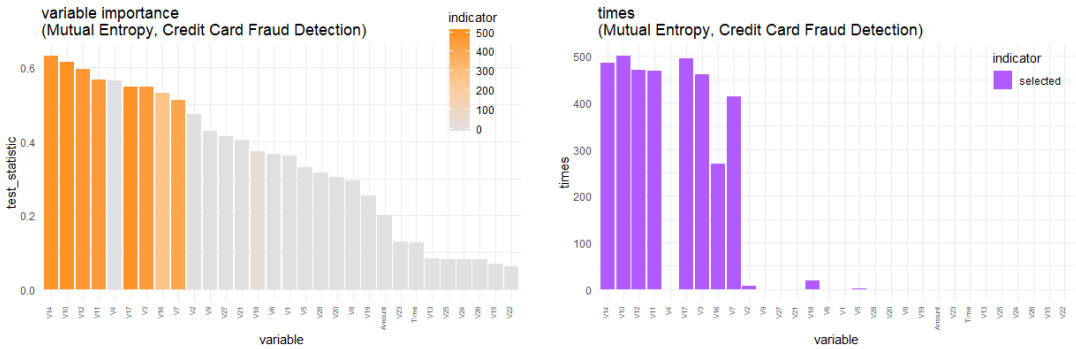


圖 6-17 Credit Card Fraud Detection　ME 重要變數選取情況

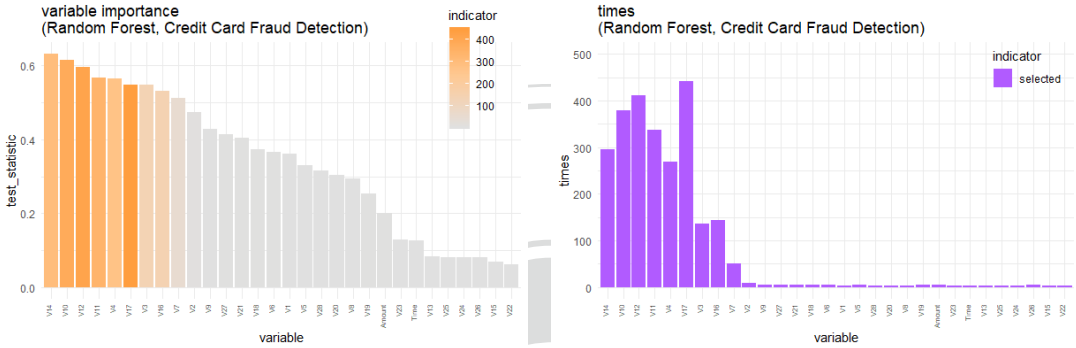


圖 6-18 Credit Card Fraud Detection　RF 重要變數選取情況

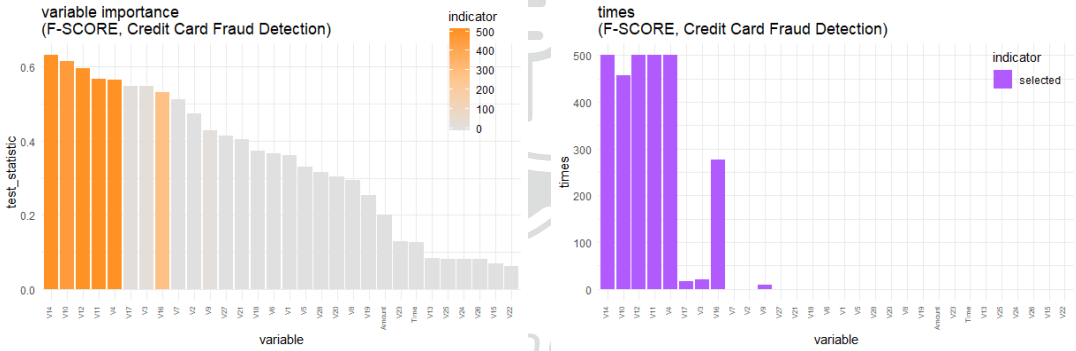


圖 6-19 Credit Card Fraud Detection　F-SCORE 重要變數選取情況

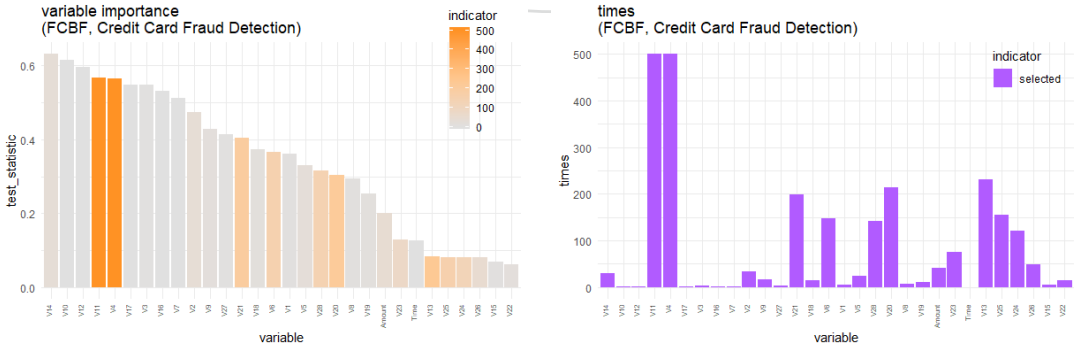


圖 6-20 Credit Card Fraud Detection　FCBF 重要變數選取情況

DOI:10.6814/NCCU201900996

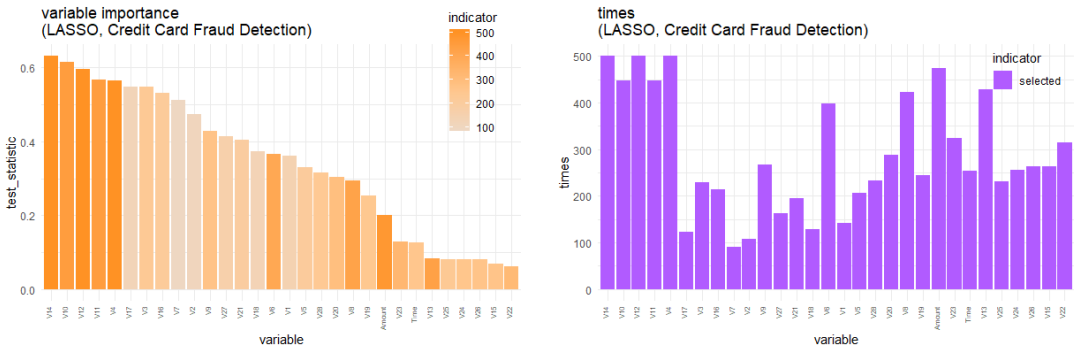


圖 6-21 Credit Card Fraud Detection　LASSO 重要變數選取情況

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

五、Oxford Parkinson's Disease

在帕金森氏症資料集中，ME(圖 6-22)與 F-SCORE(圖 6-24)在選擇重要變數上有著不錯的表現，其中 F-SCORE 的選擇較為集中，RF(圖 6-23)及 FCBF(圖 6-25)，在選擇上較為分散，LASSO(圖 6-26)則是選擇較多變數，較不容易看出其選擇能力，於表 6-6，在準確度部分，除了 LASSO 外各方法皆為 0.85 至 0.86 左右，LASSO 較高但也使用較多變數，敏感度各方法表現良好，特異度部分因資料關係(圖 5-6)，各方法皆不高，各方法之各項指標皆在一標準差區間內。

表 6-6 Oxford Parkinson's Disease 各項指標

ME	ME	RF	F-SCORE	FCBF	LASSO
Sensitivity	0.9602±0.0556	0.9702±0.0410	0.9409±0.0424	0.9498±0.0433	0.9813±0.0307
Specificity	0.5296±0.1842	0.5039±0.2071	0.5750±0.1619	0.6164±0.1614	0.5968±0.156
Accuracy	0.8540±0.0495	0.8545±0.0507	0.8510±0.0441	0.8680±0.0443	0.886±0.0405
Dimension	4.206	6.77	2.212	3.536	12.412

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

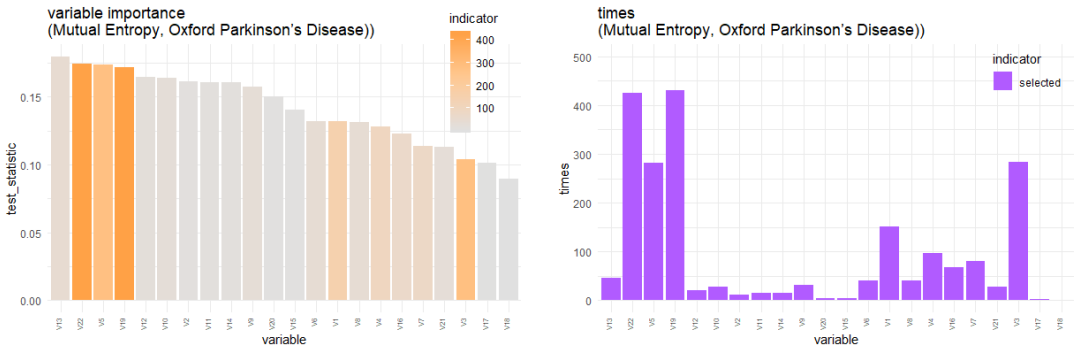


圖 6-22 Oxford Parkinson's Disease　ME 重要變數選取情況

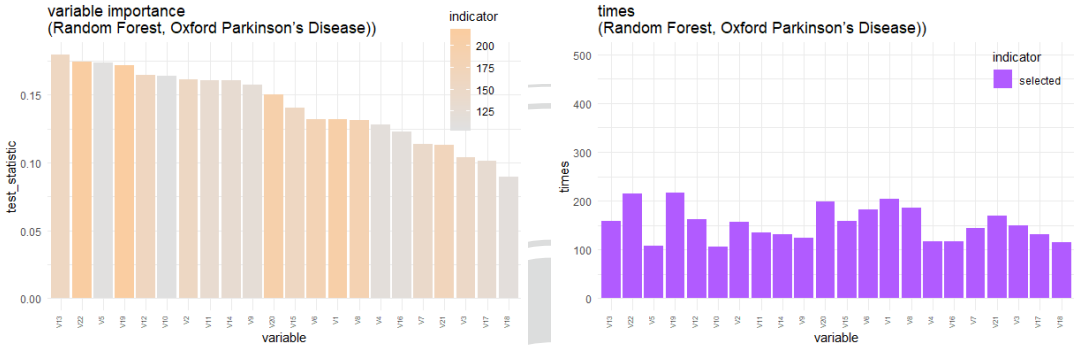


圖 6-23 Oxford Parkinson's Disease　RF 重要變數選取情況

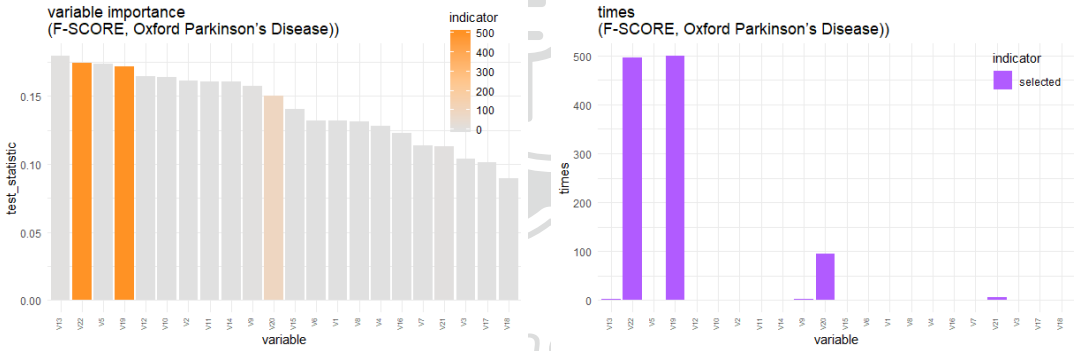


圖 6-24 Oxford Parkinson's Disease　F-SCORE 重要變數選取情況

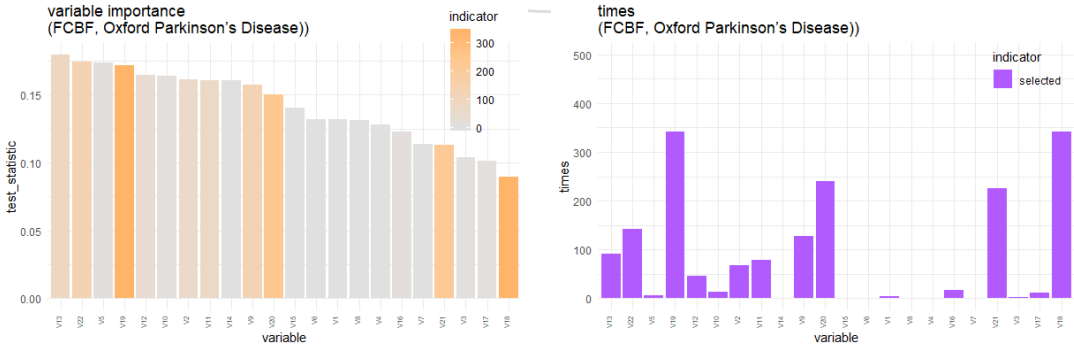


圖 6-25 Oxford Parkinson's Disease　FCBF 重要變數選取情況

DOI:10.6814/NCCU201900996

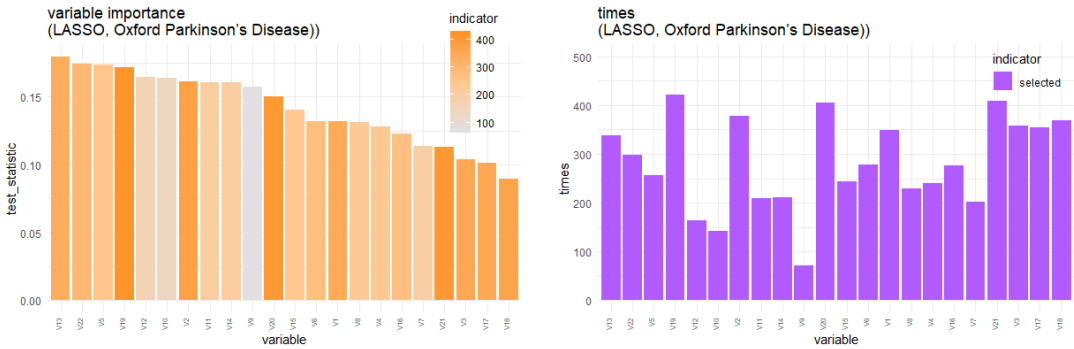


圖 6-26 Oxford Parkinson's Disease LASSO 重要變數選取情況

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

六、APS Failure at Scania Trucks

在空氣壓力系統資料集，依重要性視覺化資料前 100 變數如圖 6-27 至 6-31，ME(圖 6-27)、RF(圖 6-28)與 F-SCORE(圖 6-29)選取皆較為集中，F-SCORE 在 500 次 bootstrap 抽取中幾乎皆抽取到相同變數，FCBF(圖 6-30)在選擇上較為分散，RF 選取之變數不像 F-SCORE 集中於單一，而是在前幾重要變數中分別皆有選取，LASSO(圖 6-31)依然選取大量變數，選擇也較為分散，較無法判斷其選擇能力，從敏感度、特異度與準確度角度等指標來看(表 6-7)，除了 LASSO 外各方法在使用變數差不多下，結果相近，各自一標準差區間皆有重疊。

表 6-7 APS Failure at Scania Trucks 各項指標

ME	ME	RF	F-SCORE	FCBF	LASSO
Sensitivity	0.8428±0.0710	0.8752±0.0333	0.8781±0.0322	0.8773±0.031	0.9056±0.0279
Specificity	0.9737±0.0080	0.9769±0.0084	0.9737±0.0089	0.979±0.008	0.984±0.0069
Accuracy	0.9409±0.0189	0.9515±0.0093	0.9498±0.0094	0.9535±0.0095	0.9644±0.0087
Dimension	7.486	6.928	7.53	6.726	36.444

DOI:10.6814/NCCU201900996

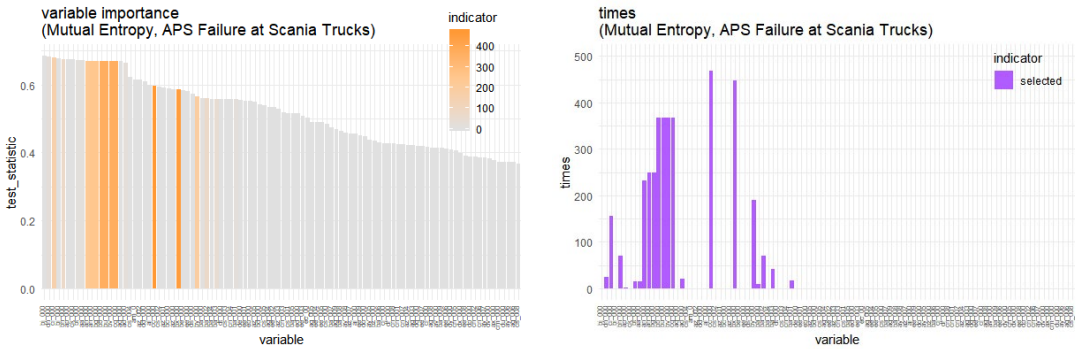


圖 6-27 APS Failure at Scania Trucks　ME 重要變數選取情況

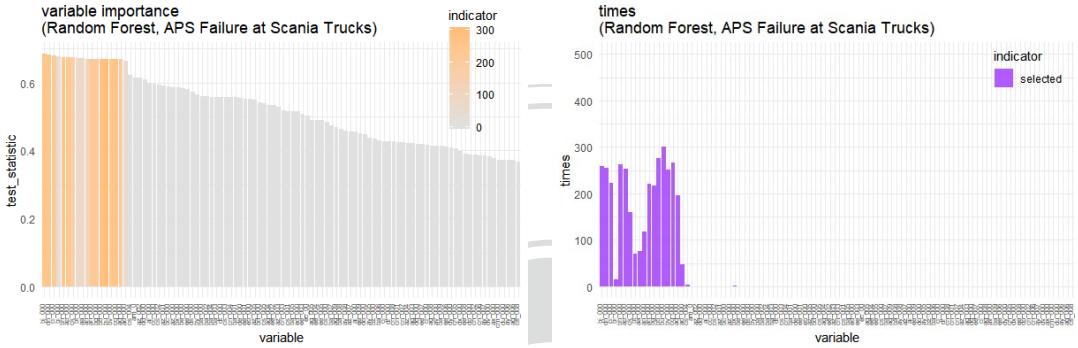


圖 6-28 APS Failure at Scania Trucks　RF 重要變數選取情況

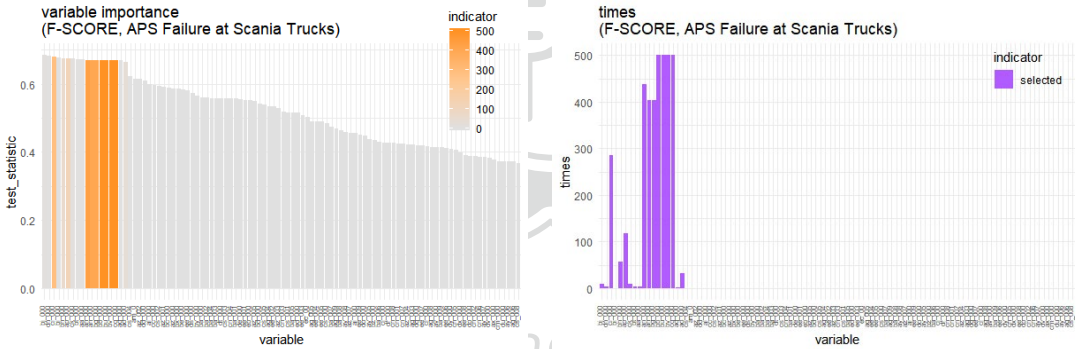


圖 6-29 APS Failure at Scania Trucks　F-SCORE 重要變數選取情況

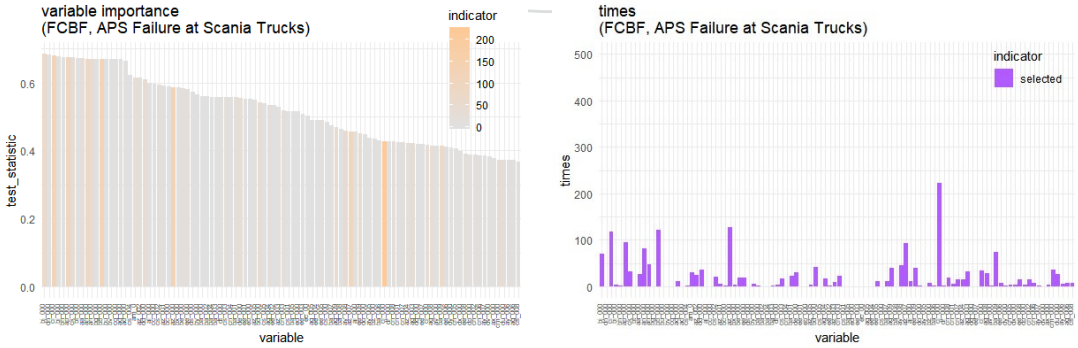


圖 6-30 APS Failure at Scania Trucks　FCBF 重要變數選取情況

DOI:10.6814/NCCU201900996

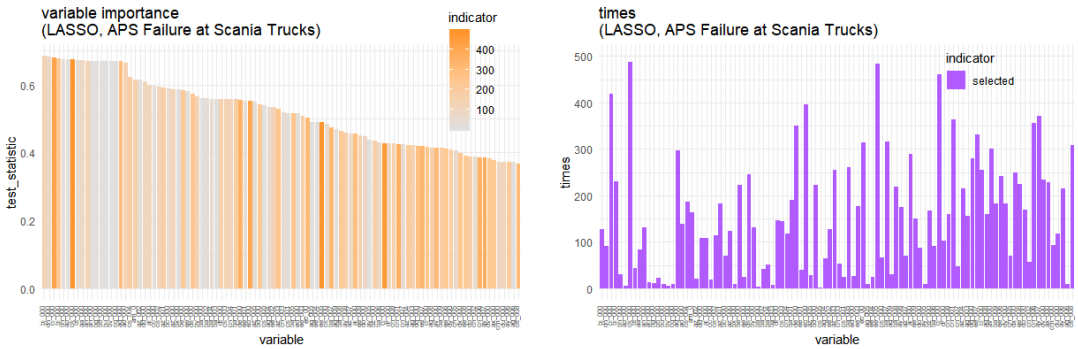


圖 6-31 APS Failure at Scania Trucks　LASSO 重要變數選取情況

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996

第二節 結論

根據以上結果，本文提出之基於 Mutual Entropy 及 SVM 之特徵選取方法，為一有效、穩定且能正確選取重要變數之方法，在需要增加變數時有根據，可依Volume(C_k)逐步加入變數，其他同樣使用資訊理論且考慮變數間關係的 RF 與 FCBF 皆為優秀方法，於多次 bootstrap 選取中，數個根據本文定義方式的重要變數皆能被多次選取，且有不錯的預測結果。於本文中使用的兩筆模擬資料與五筆真實資料中，雖 ME 並不在所有資料結果皆優於其他方法，但因沒有一個方法能適用於所有資料，因此 ME 不論資料型態為何皆能有穩定表現，為一良好發現，於兩筆型態不同的模擬資料中也能說明這點，ME 考慮變數關係，並利用此關係之特性做變數選取，也有著理論基礎。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

第七章 未來展望

在分群方面，因本文採用階層式分群並將變數數量除以 5 在其對應的區間內選取 silhouette coefficient 最大值做為分群規則，因分群標準會直接影響每一群內變數，若群數太少則大量變數集中同一群，無法評估少量變數的預測能力，因此分群多寡也為一研究方向，在分群標準上應可嘗試其他分群方法，如 k-means 等，以及嘗試別的計算分群方法之指標，如 Elbow method 等，並嘗試是否能從中找出更適合的分群標準。

在資料使用上，本為皆採用應變數為二元分類之資料，可以嘗試多元分類資料，以及連續應變數之資料，將去除連續應變數之資料分群，選取適當群數，並給定標記，計算選取而得變數預測上的 MSE 等指標，另外，在含有大量因名義尺度(nominal scale)變數所形成含有虛擬變數(dummy variable)的資料中也是一嘗試方向。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU201900996

第八章 參考資料

Akay, M.F., (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36 (2), 3240–3247.

Arunasakthi, K., KamatchiPriya, & L., Askerunisa, A., (2014). Fisher Score Dimensionality Reduction for SVM Classification. *2014 International Conference on Innovations in Engineering and Technology*, 3(3), 1900-1904.

Blum, A., Langley, P., Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2), 245-271.

Bennasar, M., Hicks, Y., & Setchi, R., (2015). Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications*, 42, 8520-8532.

Boser, B.E., Guyon, I.M., & Vapnil, V.N., (1992). A Training Algorithm for Optimal Margin Classiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.

Breiman, L., (1996). Bagging Predictors. *Machine Learning*, 24, 123–140.

Breiman, L., (2001). Random Forests. *Machine Learning*, 45(1), 5-32. doi:10.1023/a:1010933404324

Chen, Y.W., & Lin, C.J., (2005). Combining SVMs with Various Feature Selection Strategies. *Department of Computer Science, National Taiwan University*. https://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf

Cortes, C., & Vapnik, V., (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. doi:10.1007/bf00994018

Dash, M., & Liu, H., (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1-4), 131-156.

DOI:10.6814/NCCU201900996

Díaz-Uriarte, R., & Alvarez de Andrés, S., (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1), 3. doi:10.1186/1471-2105-7-3

Guyon, I., & Elisseeff, A., (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(March), 1157-1182.

Haldurai, H., Madhubala, T., & Rajalakshmi, R., (2016). A Study on Genetic Algorithm and its Applications. *International Journal of Computer Sciences and Engineering*, 4(10), 139-143.

Ho, T.K., (1995). Random decision forests. *Proceedings of the Third International Conference on Document Analysis and Recognition*, 1, 278.

Ho, T.K., (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832-844.

Holland, J.H., (1992). Genetic Algorithms. *Scientific American*, 267(1), 66-73.

Hsieh, F., Liu, S.Y., Hsieh, Y.C., McCowan, B., (2018). From patterned response dependency to structured covariate dependency: Entropy based categorical-pattern-matching. *PLoS ONE*, 13(6).

Hsu, C.W., Chang, C.C., & Lin, C.J., (2003). Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Department of Computer Science, National Taiwan University*. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Huang, C.L., & Wang, C.J., (2006). A GA-based feature selection and parameters optimization for support vector machines, *Expert Systems with Applications*, 31, 231-240.

DOI:10.6814/NCCU201900996

Jolliffe, I.T., & Cadima, J., (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences*, 374(2065), 20150202.

Kumbhar, P., & Mali, M., (2016). A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification. *International Journal of Science and Research*, 5(5), 1267-1275.

Popovic, M., (2018). Researchers in an Entropy Wonderland: A Review of the Entropy Concept. *Thermal Science*, 22(2), 1163-1178.

Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., & Jain, A.K., (2000). Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 4(2), 164-171.

Kohavi, R., John, G., (1997). Wrappers for feature selection. *Artificial Intelligence*, 97(1-2), 273-324.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C., (2010). Variable selection using Random Forests. *Pattern Recognition Letters*, 31(14), 2225-2236.

Saeys, Y., Inza, I., & Larrañaga, P., (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.

Schölkopf, B., Sung, K.K., Burges, Chris J.C., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V., (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11), 2758-2765.

Shannon, C. E., (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*. 27, 379-432, 623-656.

Srivastava, S., Joshi, N., & Gaur, M., (2013). A Review Paper on Feature Selection Methodologies and Their Applications. *International Journal of Engineering*

DOI:10.6814/NCCU201900996

Research and Development, 7(6), 57-61.

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

Xu, X., (2006). Adaptive Intrusion Detection Based on Machine Learning: Feature Extraction, Classifier Construction and Sequential Pattern Prediction. *International Journal of Web Services Practices*, 2(1-2), 49-58.

Yu, L., & Liu, H., (2003). Feature selection for high-dimensional data: A fast correlation based filter solution. *Department of Computer Science & Engineering, Arizona State University.*

國立政治大學
National Chengchi University

DOI:10.6814/NCCU201900996