Machine Learning Approaches to Breast Cancer Classification

Chen-Po Liao

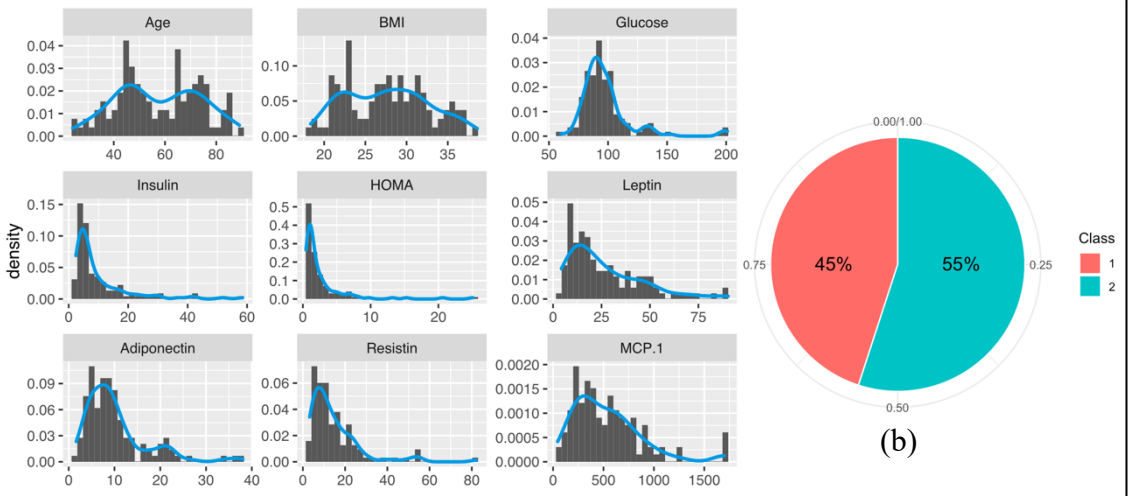
Institute of Epidemiology and Preventive Medicine, National Taiwan University

Introduction

乳癌(Breast cancer)不論在歐美或台灣婦女中的癌症發病率都已躍居於前三名，且屬於比較早期發病的癌症之一，一般來說透過正規治療後的五年存活率都可以達到 85%以上，甚至在前幾期乳癌的五年存活率可以達到 95%以上。因此，若能在癌症前期及早發現、儘早治療就會有很好的預後狀況。此研究目的在於利用機器學習中的分類模型針對健康檢查及血液分析的資料，找出適合的模型預測受試者是否有罹患乳癌[1]。

此研究使用來自科英布拉大學醫學中心婦女科(Gynaecology Department of the University Hospital Centre of Coimbra, CHUC)，在 2009 年至 2013 年所收集到的資料，共有 116 人，包含 64 名罹患癌症與 52 名健康的婦女[1, 2]。此外，該資料集共收集到 9 個變數，包含血糖(Glucose)、胰島素(Insulin)、HOMA、瘦體素(Leptin)、脂聯素(Adiponectin)、胰島素阻抗素(Resistin)、MCP-1、年齡(Age)與身體質量指數(BMI)。本研究將使用這些健康檢查以及血液分析的資料，建立機器學習的分類模型，例如邏輯斯迴歸(Logistic regression)、隨機森林(Random forest)與支持向量機(Support vector machine, SVM)。

首先，對此筆資料利用探索式資料分析(Exploratory data analysis)查看資料長相與特徵，如圖 1 為初步觀察資料的兩個統計圖表。圖 1(a)，為針對 9 個變數利用直方圖與密度曲線(density curve)畫出其資料的分布狀況，可以發現部分變數有右偏的情況，表示可能有離群值(outlier)出現，後續進一步分析需要考慮移除。圖 1(b)，以圓餅圖統計罹患乳癌與健康者的比例分佈情形，紅色為健康者，藍色標記為罹患乳癌者，兩者比例為 45%與 55%，應該不會出現資料不平衡的問題。

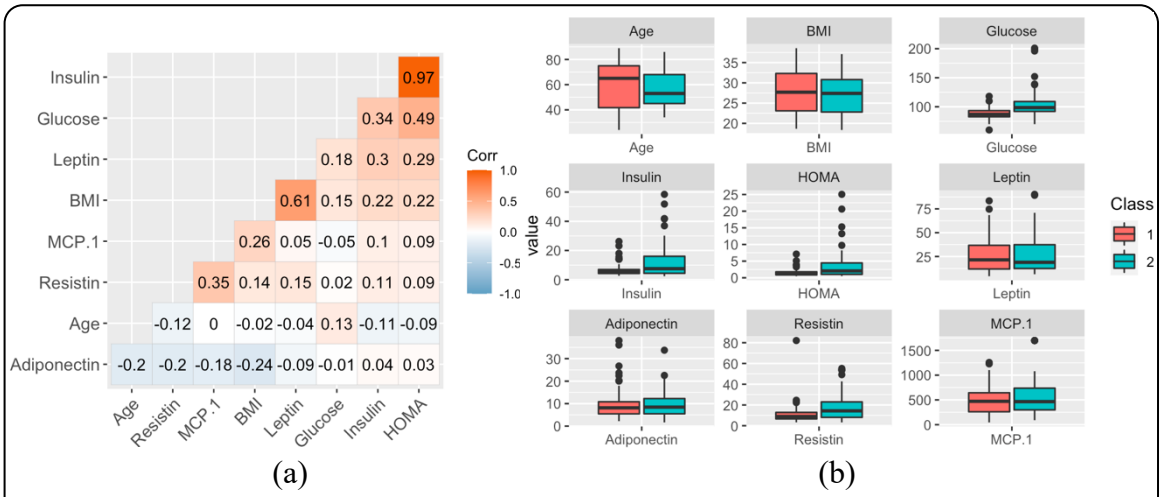


(a) (b)

圖 1. 利用統計圖表進行探索性資料分析。(a) 9 個變數分佈 (b) 受試者分佈

接著，觀察變數與變數之間的相關性，若存在高相關性，則可能不會選擇同時進入後續分析的模型，或需要使用能處理變數間高相關性的方法。此外，進一步考慮根據有無罹患乳癌將資料進行的分組，並且觀察各個變數在不同組別之間是否有些明顯的分佈差異存在，若有明顯的分佈差異表示此變數可能是進行分類的重要變數，如下圖 2。

圖 2(a)為根據變數之間相關性矩陣畫出的熱圖(heatmap)，橘色為正相關，藍色為負相關，可以發現只有胰島素(Insulin)與 HOMA 之間存在著高度相關，可能是後續分析需要注意的地方。圖 2(b)為根據有無罹患乳癌分組後 9 個變數的箱型圖，其中幾個變數在兩組間有明顯的差異，例如血糖(Glucose)、胰島素(Insulin)與胰島素阻抗素(Resistin)。另外，也可以發現某些變數在兩組間都有存在離群值，因為樣本數不多且有些離群值或許可以幫助區分出是否罹患乳癌，因此只針對胰島素阻抗素(Resistin)與脂聯素(Adiponectin)移除離群值，標準為超過平均值 3 倍標準差的數值，最後剩餘 111 人進入後續分析。



(a) (b)

圖 2. 進一步觀察變數間的相關性以及 9 個變數在有無乳癌兩組的分布差異 (a) 根據變數之間相關性矩陣所畫之熱圖 (b) 9 個變數在兩組之間的分佈圖

做分類可以先不用考慮outlier，因為他可能可以幫助分類

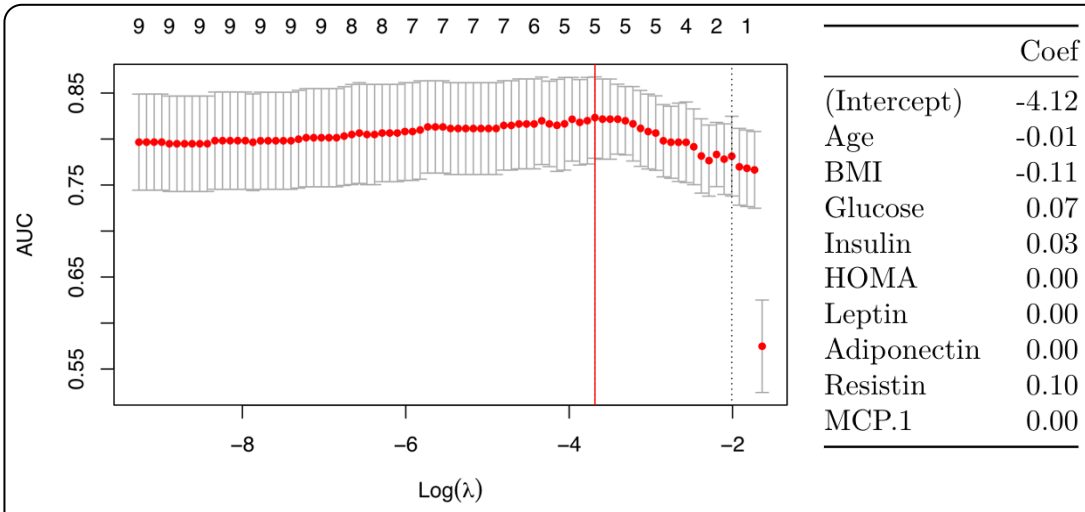
Methods

本研究將以邏輯斯迴歸(Logistic regression)、隨機森林(Random forest)與支持向量機(Support vector machine, SVM)，旨在探討與預測此筆資料的受試者是否罹患乳癌建立分類模型，並以交叉驗證的方式找出各個模型中適合的參數。

首先為建立 Logistic regression 的方式，由於資料中存在一組高度相關的變數，胰島素(Insulin)與 HOMA 相關係數高達 0.97，因此建立此迴歸模型時選擇加入 lasso 的懲罰項來處理。Lasso 針對此情況在模型的處理方式，將會保留幾個高度相關性的變數之中的一個變數，其中懲罰項的超參數(hyperplane)利用 5-fold cross validation 的方式找出表現較好的λ，以 AUC 較高的參數作為最後模型參數的選擇，$\lambda = 0.0251$，最後在模型中保留了 5 個變數，平均準確率為 73.87%，平均 AUC 為 0.8233，交叉驗證結果以及最後選擇模型中的迴歸係數如圖 3 與表 1。

可以考慮交互作用

畫兩個變數之間的散佈圖，可以將這些點以class標記顏色



	Coef
(Intercept)	-4.12
Age	-0.01
BMI	-0.11
Glucose	0.07
Insulin	0.03
HOMA	0.00
Leptin	0.00
Adiponectin	0.00
Resistin	0.10
MCP.1	0.00

圖 3. &表 1. 以交叉驗證選擇 Lasso 超參數結果與最後模型的迴歸係數

第二個使用的模型為 Random forest，共生成 5000 棵樹，並以 5-fold cross validation 的方式挑選模型中參數 mtry，此參數表示執行 Random forest 中以 bootstrap 方法抽取資料建立每顆樹時，每次抽取到的變數個數，挑選標準為以 AUC 較高的參數作為最後模型參數的選擇，最後挑擇每次以 7 個變數進入模型表現得最好，平均準確率為 75.4%，如圖 4(a)為選擇不同參數 mtry 各個驗證資料集中的結果，實線為平均值，虛線為最大值與最小值。此外也呈現了以 Random forest 計算各個的變數重要程度排序，如圖 4(b)。

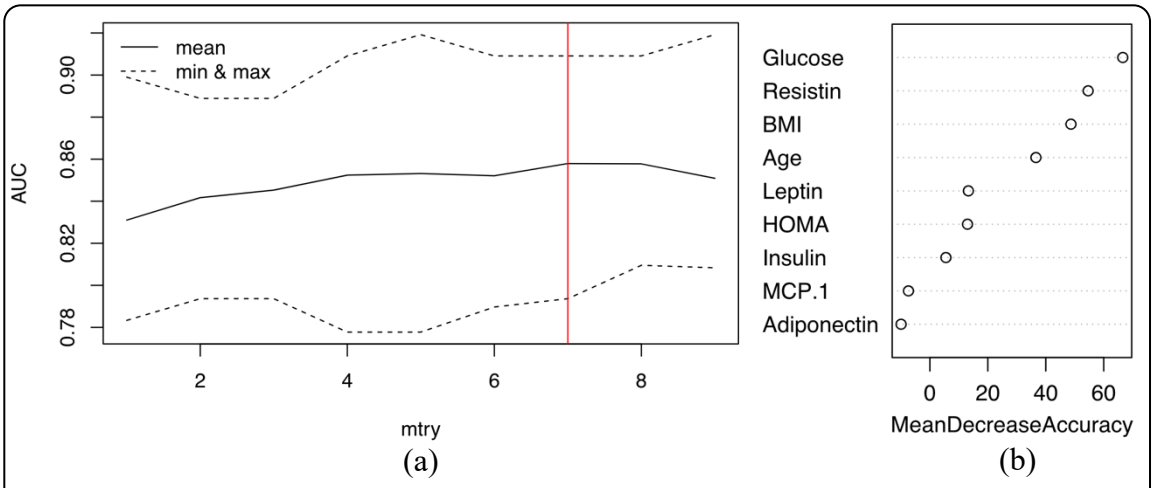
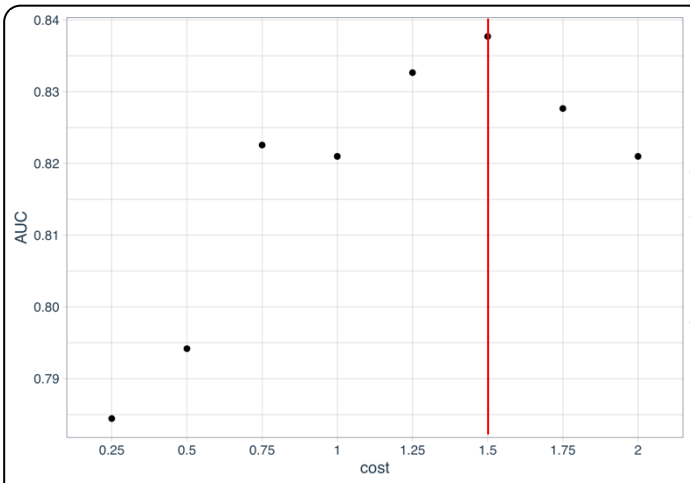


圖 4. (a)以交叉驗證的方式挑選 Random forest 參數 mtry (b) 各個變數重要性排序

第三個模型為支持向量機(Support vector machine, SVM)，使用 radial basis kernel，另外以 5-fold cross validation 的方式挑選 gamma 與 cost 參數。首先固定 cost 為 1 挑選 gamma，數值由 0.1 到 3 以 0.1 為間隔作挑選，以 AUC 較高的參數作為最後模型配飾的，最後選擇以 gamma 為 0.2 來建立模型。接著，固定以 gamma 為 0.2 挑選 cost 參數，最後選擇以 cost 等於 1.5 為最後模型的參數。如圖 5 為固定 gamma 為 0.2 挑選 cost 參數時，每個數值對應的 AUC 數值。

SVM經過挑選變數後是不是會表現更好？

3-fold表示training只有2/3，使用10-fold可能更像原始資料，有更多資料可以train參數



	AUC	Accuracy
Logistic Regression	0.8233	0.7387
Random Forest	0.8579	0.7493
SVM	0.8377	0.7743

圖 5. &表 2. 以交叉驗證的方式挑選 SVM 參數與三個模型表現結果

最後為三種模型的比較，表 2 為三種模型的 AUC 與準確率結果，以兩種標準衡量三個模型的排序都不太一樣，但看起來的差異也相差不大。不過在兩種指標下，邏輯斯迴歸模型都是表現最差的，原因大概是因為其他模型都可以一定程度處理非線性的分類關係。雖然邏輯斯迴歸模型表現不怎麼樣，但其中一個優勢為模型解釋力。以 lasso 挑選出的五個變數有四個變數為 Random forest 變數重要性排序的前四名，代表這四個變數確實是 9 個變數中，對於是否罹患乳癌的分類是較為重要的，包含年齡、BMI、Glucose 與 Resistin，其中幾個變數在圖 2(b)的箱型圖可以看出罹癌與健康者之間變數的分佈差異。

然而，由於樣本數數量只有 100 多筆，因此設定不同的種子在每個模型最後得到的結果差異還蠻大的。以 AUC 為標準來衡量時，Random forest 是最好的模型，也就是此模型的分類表現相對來說是比較好的，我想或許是因為有 bootstrap 抽取資料方法的關係使得 Random forest 有比較穩定的結果。

Further Investigations

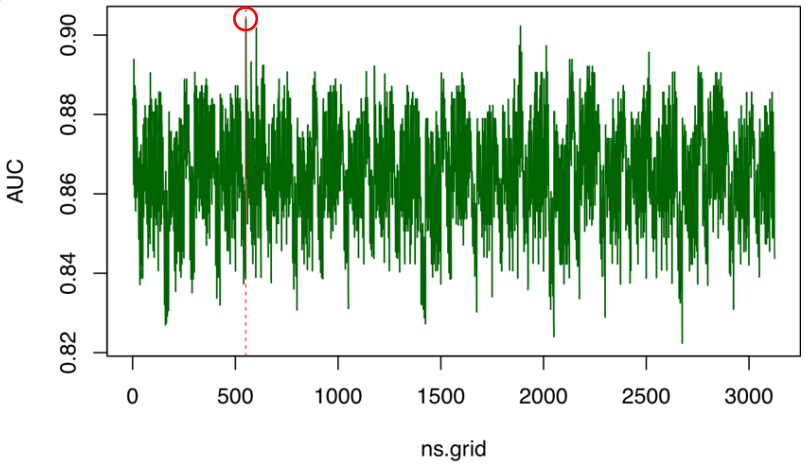
此研究想進一步探討的模型為邏輯斯迴歸模型，因為此模型較具有解釋力且可利用 basis expansion 的方式將模型延伸到處理非線性的關係。利用以上分析結果建立的邏輯斯迴歸模型如下：

$$\text{logit}\,P(Y = \text{乳癌}|X)$$
$$= -4.12 - 0.01\text{Age} - 0.11\text{BMI} + 0.07\text{Glucose} + 0.03\text{Insulin} + 0.1\text{Resistin}$$

如表 1 中呈現的迴歸係數，解釋如下：截距項為−4.12，表示在其他變數都為 0 之下，一個人罹患乳癌的機率為健康機率的 0.016 倍，而截距項本身並沒有任何意義；年齡係數為−0.01，表示其他變數固定下，年齡每增加一歲罹患乳癌的勝算(odds)變為原本的 0.99 倍，此結果與圖 2(b)年齡的分佈狀況相同，罹患乳癌者年齡的中位數明顯低於健康者；BMI 係數為−0.11，表示其他變數固定下，BMI 每增加一單位罹患乳癌的勝算(odds)變為原本的 0.89 倍；血糖(Glucose)係數

為 0.07，表示其他變數固定下，血糖每增加一單位罹患乳癌的勝算(odds)變為原本的 1.07 倍，此參數在圖 2(b)變數分佈也可以觀察到一樣的趨勢，且也被 Random forest 挑選為最重要的變數。Insulin 係數為 0.03，表示其他變數固定下，Insulin 每增加一單位罹患乳癌的勝算(odds)變為原本的 1.03 倍；Resistin 係數為 0.1，表示其他變數固定下，Resistin 每增加一單位罹患乳癌的勝算(odds)變為 1.11 倍。

接下來嘗試利用 nature cubic spline 將邏輯斯迴歸延伸到處理非線性的變數關係，希望能得到比原來模型更好的 AUC 與準確率。針對以 lasso 挑選出的五個變數尋找各自適合的 spline 自由度。每個變數自由度嘗試設定由 2 到 6，共有 3125 個組合，將每個組合都跑一次 5-fold cross validation 找出平均 AUC 表現最好的組合。結果如圖 6 與表 3，(Age, BMI, Glucose, Insulin, Resistin)最佳選擇的組合為(2, 2, 4, 6, 2)，平均 AUC 為 0.90，準確率為 0.78。



	GAM
AUC	0.9040
Accuracy	0.7838

圖 6. &表 3. 以 grid search 的方式挑選最適合的 Basis 組合與最佳表現結果

Summary

本研究旨在利用健康檢查及血液分析的資料，嘗試以三種模型找出最適合的分類模型預測受試者是否有罹患乳癌，結果達到 70%以上的準確率與 0.8 以上的 AUC，並不是非常傑出的表現。未來或許可以針對敏感度(sensitivity)的重要性來選擇分類機率的切點(cut point)，此標準下選擇的模型可以使真正罹患乳癌的病患有更高的機率被分類模型預測出來，因目前使用的標準僅僅以機率較高的類別作為最後的預測結果，也就是機率的切點選在 0.5。此外，經過 basis expansion 的邏輯斯迴歸模型，最後得到高達 0.9 的 AUC，且準確率也提高到將近八成，表示以非線性的方法來預測是否罹患乳癌更為合適。

Reference

1. Crisostomo, Joana, et al. "Hyperresistinemia and Metabolic Dysregulation: A Risky Crosstalk in Obese Breast Cancer." *Endocrine* (2016): 433-442.
2. Patrício, Miguel, et al. "Using Resistin, Glucose, Age and BMI to Predict the Presence of Breast Cancer." *BMC Cancer* 18.1 (2018): 1-8.