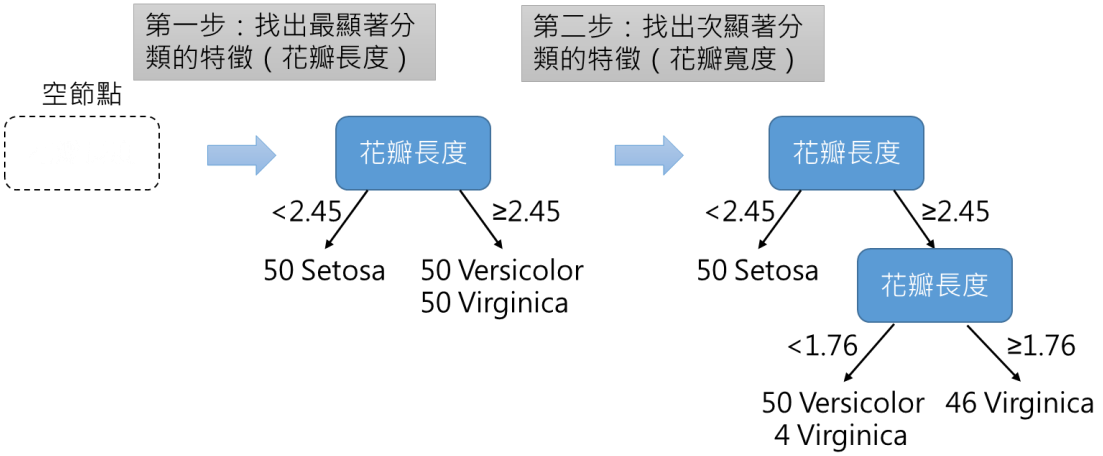
特徵值才是本體：以特徵選取為主要目的的機器學習模型建置

吳育瑋
臺北醫學大學

講到機器學習，許多人腦中或許會馬上浮現出「分類」二字。的確，一直以來機器學習一直都是以分類事物作為主要目的。比如說手寫辨識就是透過筆劃將手寫的鬼畫符分類成資料庫中的某個中英文字或數字，而物體辨識則是將整張照片或照片的某一隅分類成某種動植物或事物。醫院中建置的影像自動判讀系統最主要的目的，同樣也是判斷某張醫療影像是否有著那些特定的疾病（比如說有無不正常腫塊等等）。

事實上，分類本來就是各種機器學習演算法的主要目的。以最簡單的決策樹（Decision Tree）模型為例來說，其目的正是在於建出一棵有著數個至數十個分支節點的樹狀模型；只要照著分支節點的指示從根部「走」完這棵決策樹，即可得到手邊某筆未知資料最有可能的分類。以最簡單且經典的鳶尾花資料集為例來說，其建立出來的決策樹分支節點即可在使用者拿到某一朵未知的鳶尾花時，照著決策樹分支節點的指示（比如說花辦長度或寬度是否大於某個長度等等）很精確地分類這朵花。

讀到這裡，或許會有讀者開始產生疑問：「為什麼標題會寫特徵才是本體呢？」這是因為不管是那種機器學習模型，其分類的依據都還是「特徵」。唯有掌握特徵，才能建立出更精準的模型。再次以鳶尾花資料集為例來說明。雖然模型的目的仍然是分類，但是在建立模型分支節點的過程中我們必須判斷最能夠用來分類的特徵值作為節點才行。下圖一為鳶尾花資料集的決策樹建立步驟。在這個建立步驟中可以很清楚地看到分類效果最佳（最顯著）的特徵【花瓣長度】會被首先選取出來，並放在最靠近決策樹根部處；其次是選取出效果次佳的特徵【花瓣寬度】，以達到分類絕大部分資料的目的。在術語上我們會用重要性（Importance）這項評斷指標來給這些特徵值打分數，而分類效果較佳或較顯著的特徵就會有著較高的重要性分數。反之亦然。以圖一為例，之所以會將花瓣長度放置於根節點處，正是因為它的重要性（或者說是分類效果）最高的緣故。



圖一、鳶尾花資料集的決策樹建立步驟。

許多機器學習演算法都能夠進行程度不一的特徵值重要性判斷。其中最為人熟知的是決策樹以及決策森林系列的演算法，包含決策樹、隨機森林（Random Forest）、以及由隨機森林衍生出來的各種進階樹狀模型如 Gradient Boosting 或 eXtreme Gradient Boosting（XGBoost）等演算法。這些模型的共通點除了都與樹狀模型有關之外，還有就是都能夠計算並回報特徵值的重要性。同樣地，重要性越高，資料分類的效果越好。

在研究上，這類型的特徵值重要性判斷很常被拿來選取出「可能」具有某種意義的特徵值。比如說在數萬個基因中找出與疾病有著重要表現關聯性的數十或數百個基因，或是在一大堆醫療數據中找出與目標族群分類相關性最高者。一般來說，只透過這些重要性較高的特徵值訓練出來的演算法也會比用全部特徵值進行訓練的演算法有著更準確的分類效果。這類型的研究走到極致的其中一種應用就是生物標記的尋找，比如透過某種特徵選取演算法找出五六個能夠精確分類某種疾病的基因等等。這樣子的研究後續可進行的，包含 1) 透過挖掘出來的數個或數十個基因生物標記進行更精準的疾病判斷，以及 2) 針對找出來的基因生物標記進行更深入的研究或實驗，以了解未知的致病機制或因子。

我們實驗室最主要的研究方向之一，在於透過數千株細菌組成的泛基因體（pan-genome）研究細菌基因與抗藥性機制之間的關聯性。我們的研究方法即

為透過特徵選取演算法找出未知的抗藥性基因。概念也很簡單：在建立泛基因體後，我們透過決策森林系列的演算法針對每種特定基因進行重要性評分，並透過與已知抗藥性基因之間的交錯比對找出已知或未知的抗藥性基因。在我們的研究中我們不但發現大部分與抗藥性有著高度關聯性的基因並非已知的抗藥性基因，甚至透過基因功能性確認至少一半的基因為未知功能的基因。除此之外，我們也發現透過這些高度關聯性的基因特徵值建立出來的預測模型比起使用已知抗藥性基因建立的模型還要準確，顯示這些基因至少有一部分應該帶有著未知的抗藥性功能。

為了發掘生物標記，我們並發展出一套演算法，透過將資料集分割成子資料集，在每個子資料集中進行特徵選取，並交集所有子資料集特徵的方法，找出遠比現有特徵選取演算法還要少上許多，但卻不影響預測效能的特徵集合。舉例來說，針對某個細菌抗藥性預測，傳統的特徵選取或許可以在細菌泛基因體成千上萬個基因中找出數百個有著高度關聯性的基因；而我們的演算法則可以更進一步將找出的高度關聯性特徵值壓縮到數十個或甚至個位數，而且預測效能與傳統特徵選取演算法相當。這意味著我們找出的數十個或數個基因與抗藥性之間有著極高度的關聯性，而在基因數量相當少的情況下我們甚至可以針對這些基因進行深度挖掘，找出它們為什麼會與抗藥性高度相關。

當然，這類型的演算法再怎麼發展，原理都還是找出與預測目標（在我們的例子中即為細菌抗藥性）有著高度關聯的特徵值；而這些特徵選取演算法最大的限制，在於 1) 有可能會找出關聯性極高但其實與預測目標並無直接關聯性的特徵、以及 2) 無法說明找出的特徵值與預測目標之間的因果關係。再次以細菌抗藥性為例。我們發現部分透過特徵選取找出的基因與基因跳躍與插入有關。這些基因有可能會帶著真正的抗藥性基因一起插入基因體，因此同樣也與抗藥性有著高度關聯性；但是它們並不直接貢獻抗藥性，而是幫助細菌取得與傳播抗生素抗藥性。因此在透過這類型的方法進行研究時要特別小心。除非能夠進行實驗證實特徵值與預測目標之間的因果關係，要不然都只能描述關聯性，無法直接引申為因果關係。