應用 KNN 演算法於鑑別心臟病之分析研究

*楊偉修 戴良軒 彭冠傑 林柏豪
南開科技大學企業管理系

1. 研究背景與目的

心臟病於醫學上又稱為心肌梗塞(myocardial infarction)，通常是冠狀動脈疾病的結果。由於發病前較無明顯病徵所以常易被忽略，然一旦發作就會傷害心臟，並且在心臟留下結痂組織，嚴重時導致猝死。依據衛生福利部國民健康署公告 102 年國人死因統計結果顯示心臟疾病高居國人十大死因第二名，約占總死亡人數之 11.5%，且其中更以老年人占絕大多數（衛生福利部國民健康署，2014）。因此如何有效且準確的區分該疾病，及早發現及時治療，健康且成功的老化，則成為本研究的主要目的。

k 最近鄰演算法(k Nearest Neighbor Algorithm, KNN)是一種非常簡單且直覺化的樣式辨識(pattern classification)方法，由於該法操作簡單，早已廣泛且有效的應用於各種領域之辨識問題上。雖然此演算法屬於無參數分類法，但鄰近點的數量（k 值）選取、資料前處理方式以及距離測量的方法皆會影響計算結果。有鑑於此，本研究預先將資料以四種不同的資料正規化法處理，再分別以常用的歐幾里得距離(Euclidean distance)與曼哈頓距離(Manhattan distance)進行計算，分析每種情況下各種值的準確率變化情形，以期建立簡易使用且具高準確度的心臟病鑑別模型。

2. 研究方法

本研究所使用的實驗資料，是取自 UCI 資料庫(Lichman, 2013)所建立的心臟病資料集(Statlog (Heart) Data Set)。該資料集共有 270 筆資料，13 個診斷項目屬性（含 7 個類別型屬性與 6 個數值型屬性），與 1 個診斷結果(class)屬性。診斷結果分為有心臟病(class=2)與無心臟病(class=1)兩類，其中診斷為無心臟病者共有 155 筆，約佔全部資料之 55.56%。

KNN 是一種非常簡單又容易操作的分類演算法，本研究中先使用四種常用的資料正規化方法：極值正規化(min-max normalization)將值標準化為[0,1]範圍；Z 分數正規化(Z-score normalization)、最大正規化(max normalization)與十進位正規化(normalization by decimal scaling)，將某維度中的任一值 v 透過相關的正規化公式，將該值正規化 v'。再分別以歐幾里得距離(Euclidean distance)與曼哈頓距離(Manhattan distance)函數進行 KNN 計算，研究中依據訓練資料筆數詳盡計算所有可能 k 值，並平手時採以 NN（k=1）規則判別，以觀察每種 k 值的準確率變化情形，並尋找出最佳準確率的值。

3. 結果與討論

為與之前文獻的研究結果比較，本論文所有計算皆採用 5 組交互驗證方法(5-fold cross validation method)，也就是將資料隨機均分為五等份，其中 4/5 的資料作為訓練資料，剩下的 1/5 則為測試資料，總共計算 100 次，並將平均結果整理於後。

由圖 1 與圖 2 的結果顯示，未經正規化處理的心臟病資料以 KNN 演算法計算後所得之準確率最低，而將資料預先以四種不同的正規化方法處理後，每一種方法所得的準確率相對 k 值變化情形亦皆不相同，且其中以 Z-分數正規化所得之準確率最高，而未經任何正規化處理(original)的結果最差。此外，由圖中可發現當 k 值約增加至 200 時，準確率趨近於該資料集的原始準確率 55.56%。

*通訊作者：yws@nkut.edu.tw

最後將各種情況的最佳 k 值，以及相對應的最高平均準確率與標準差整理於表 1，由表中亦可得知當應用 KNN 演算法於心臟病資料並以 5 組交互驗證計算時，若將資料先以 Z-分數正規化處理後再以歐幾里得距離計算，當 k=39 時會得到最高平均準確率為 85.72%。

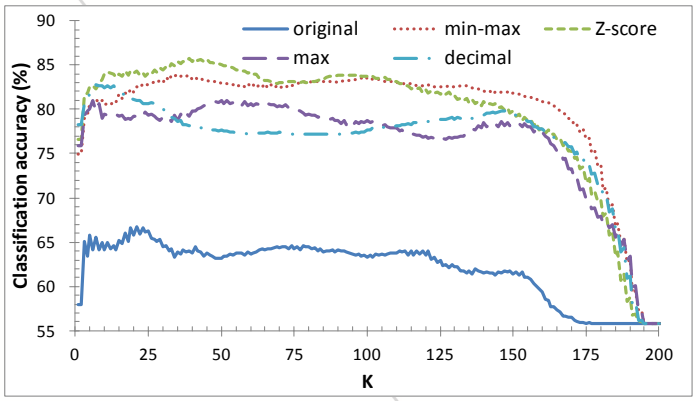

圖 1. 在五種不同資料正規化的情況下，各種 k 值之準確率變化圖(Euclidean distance)

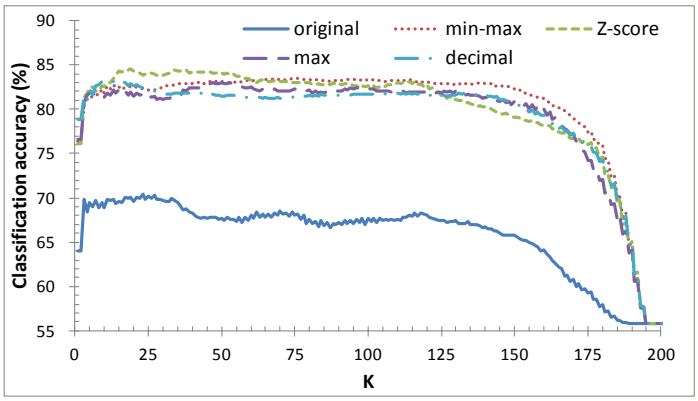

圖 2. 在五種不同資料正規化的情況下，各種 k 值之準確率變化圖(Manhattan distance)

此外，李御璽等人(2007)於未經屬性篩選前，以 Neural Network 計算此資料集所得之準確率為 85%，以 C5.0 計算所得之準確率為 78%，以 C&RT 方法計算所得之準確率為 80%。而翁政雄等人(2013)應用決策樹於心臟病預測之研究結果，所獲得之準確率為 83.70%。而本論文以非常簡單且易實現的 KNN 分類模型，預先將資料經由 Z 分數正規化處理後再以歐幾里得距離進行計算，所得之最佳 k 值分類準確率皆優於上述文獻。

表 1. 各種情況的最佳 *k* 值，以及相對應的最高平均準確率與標準差結果

	Euclidean distance			Manhattan distance		
Normalization	k	準確率	標準差	k	準確率	標準差
original data	21	66.70	6.09	23	70.38	5.99
min-max	35	83.94	4.54	75	83.51	4.63
Z-score	39	**85.72**	4.19	19	84.53	4.55
max	50	81.04	5.39	53	82.98	5.03
decimal scaling	7	82.74	4.55	15	83.40	3.98

參考文獻

1. 李御璽、顏秀珍、楊乃樺、廖晨涵、黃伯文、英家慶、賴郁菁(2007)。資料探勘在心臟病預測模型上之研究。*Journal of Informations & Electronics*，*2*(1)，19-28。
2. 翁政雄、洪令莊、呂培豪、陳學瀚、郭家佑、施博惟、謝孟哲(2013)。應用決策樹於心臟病預測之研究。第 19 屆資訊管理暨實務研討會。台中。
3. 衛生福利部國民健康署(2014)。102 年全國主要癌症死亡原因。取自 http://health99.hpa.gov.tw/Article/ArticleDetail.aspx?TopIcNo=846&DS=1-life。