建構預測乳癌之 KNN 鑑別模型

*楊偉修 黃敏誠 廖昇飛 張智恆 唐維志 朱宇凡
南開科技大學企業管理系

1. 研究背景與目的

依據衛生福利部國民健康署公告國人死因統計結果顯示（衛生福利部國民健康署，2014），102 年國人因惡性腫瘤死亡人數為 44,791 人，占總死亡人數之 29.0%，而女性乳癌則位居十大癌症死因之第四順位，占所有因癌症死亡人數之 16.8%，是目前國人不可忽視的重大疾病之一。然而乳癌的存活率平均達百分之六十，第一期的乳癌存活率約為百分之八十，甚至若能於乳癌零期時即發現，其存活率更可高達百分之百。因此如何有效且準確的區分該疾病，及早發現及時治療，讓所有婦女皆能免於此疾病的威脅，健康且成功的老化，則成為本研究的主要目的。

近年來已有許多的研究，成功的將資料探勘技術應用於醫療領域問題上，而在眾多的資料探勘技術中，k 最近鄰演算法(K Nearest Neighbor Algorithm, KNN)是一種最簡單且直覺化的方法。但 KNN 演算法從發展至今，一直存在兩個問題，一個就是必須儲存大量的訓練樣本，另一個則是無法準確的決定最佳值。為解決上述問題，已經有許多學者提出有效的樣本篩選方法，然對於如何決定最佳 值，至今仍無具體且有效的方法。但是 值的選取，絕對會影響 KNN 演算法的執行結果，分類過程中如能以真正的最佳 值進行計算，則可以建立高準確率的分類模型。此外，以不同的距離函數計算，亦會產生不同的準確率，因此亦有很多學者發展出各種不同的距離函數以提升 KNN 演算法的準確率。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乳癌鑑別模型。

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)所建立的乳癌 Original 資料集。該資料集共有 683 筆資料，9 個診斷項目屬性，與 1 個診斷結果(class)屬性。診斷屬性分為良性(class=2)與惡性(class=4)兩類，其中診斷為良性者共有 444 筆，約佔全部資料之 65.01%。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，將該值正規化為 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值，並採平手時以 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 k 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 10 組交互驗證方法(10-fold cross validation method)，並將計算 100 次之平均結果整理於後。經計算所有 k 值之準確率後，得知發生最佳準確率之 k 值皆不大，且當 k 值約增加至 430 時，準確率趨近於該資料集的原始準確率 65.01%，而為了清楚觀察準確率與 k 值之間關係，以下僅將 k=1~50 之準確率變化情形繪製於圖 1 與圖 2。由圖中的結果顯示，經過最大正規化處理的乳癌資料，以 KNN 演算法計算後所得之準確率最高，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

*通訊作者：yws@nkut.edu.tw

當應用 KNN 演算法於乳癌資料並以 10 組交互驗證計算時，若將資料先以最大正規化處理後再以歐幾里得距離計算，約於 k=5 時會得到最高平均準確率為 97.21%。

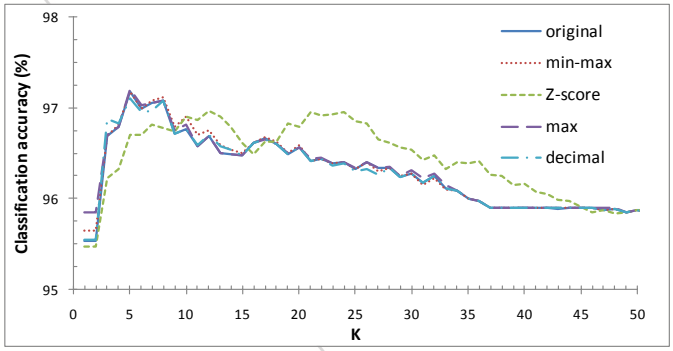

圖 1. 在五種不同資料正規化的情況下，各種 k 值之準確率變化圖(Euclidean distance)

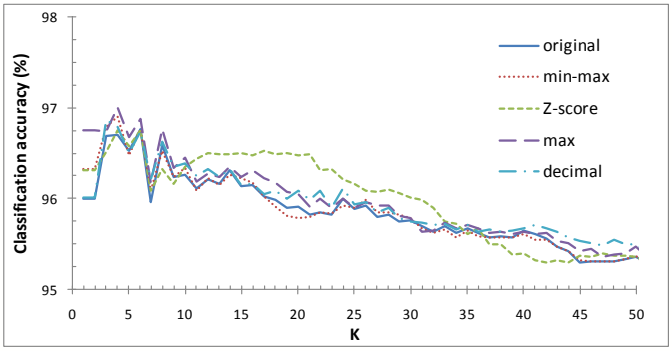

圖 2. 在五種不同資料正規化的情況下，各種 k 值之準確率變化圖(Manhattan distance)

此外，Quinlan(1996)以 C4.5 方法計算此資料集所得之準確率為 94.74%，Ster 和 Dobnikar(1996)以 LDA 方法計算所得之準確率為 96.80%，Abonyi 和 Szeifert(2003)使用 supervised fuzzy clustering 方法所獲得之準確率為 95.57%。而本論文以非常簡單且易實現的 KNN 分類模型，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

Normalization	Euclidean distance			Manhattan distance		
	k	準確率	標準差	k	準確率	標準差
original data	5	97.18	1.94	4	96.71	2.13
min-max	5	97.18	1.94	4	96.90	2.08
Z-score	12	96.97	1.95	6	96.75	1.99
max	5	97.21	1.96	4	97.00	2.04
decimal scaling	5	97.10	1.93	3	96.87	2.01

參考文獻

1. Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters*, *24*(14), 2195-2207.
2. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
3. Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. Journal of artificial intelligence research, 77-90.
4. Ster, B., & Dobnikar, A. (1996, June). Neural networks in medical diagnosis: Comparison with other methods. In *Proceedings of the International Conference EANN* (Vol. 96, pp. 427-430).
5. 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。