研究資料清理方法對 KNN 分類器鑑別含雜訊心臟病資料之影響

楊偉修
南開科技大學 企業管理系

1. 研究背景與目的

k 最近鄰演算法（k Nearest Neighbor Algorithm，簡稱 kNN）一種非常簡單且直覺化的樣式辨識方法，由於該法操作簡單，所以已廣泛且有效的應用於各種領域之辨識問題上。該法辨識未知類別測試資料的過程，是將測試資料與一組已知類別的訓練資料進行相似度比對，也就是先計算測試資料與所有訓練資料間的距離，然後取距離最接近測試資料的 k 個鄰近點，統計各類別的訓練點數，最後以得票數最多的類別作為測試資料之預測類別。所以選取不同的 k 值也會產生不同的分類結果，但如何選取最佳的 k 值，至今依然沒有適當的方法來預先決定，因此設定 k=1 是最簡單且常見的分類規則（簡稱 1NN），可是當訓練資料中含有誤判案例（又稱為類別雜訊）時，1NN 分類器亦是最易受類別雜訊影響的分類器。

為了獲得更高品質的訓練資料，已有許多研究發展出各種雜訊清理或資料編輯方法，而以 kNN 為基礎的編輯方法，則以 ENN (Edited Nearest Neighbour)、RENN (Repeated ENN)與 AENN (All-k-NN)三種方法，屬於最簡單、有效且常被使用的雜訊清理方法，三種方法在使用時，都必須先指定一個 k 值作為判定標準。ENN 清理方法(Wilson, 1972)是在訓練資料中隨機挑出一筆資料當作測試資料，並對剩下的其他訓練資料以 kNN 演算法計算，如果挑出的資料無法按照預先指定的 k 值成功分類，則將該筆資料視為雜訊並予以刪除。Tomek 於 1976 年提出 RENN 與 AENN 清理方法(Tomek, 1976)，其中 RENN 方法則是重複執行 ENN 清理方法，直到所有訓練點都能以剩下的訓練點成功分類為止。AENN 清理方法則是按照指定的 k 值，刪除訓練資料中無法同時以 1NN、2NN、…、kNN 演算法成功分類的資料，若 AENN 清理方法指定的 k=1，則清理結果與 ENN(k=1)相同。

心臟病是中老年人中常見的一種慢性疾病，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。而心血管疾病已經連續蟬聯國人十大死因第二或第三名，相較於癌症，心血管疾病較無明顯病徵所以常被忽略，而發病時是較猛烈且易致命的，因此是中老年人不可忽視的重大疾病之一。在資料探勘領域中，近年來已有許多以心臟病為主題的研究，但這些研究中鮮少考慮雜訊資料對於判別此疾病的影響。而當所獲得之資料品質不佳，並以此含有雜訊的訓練資料建立分類器，一定會減低分類器的功能。有鑑於此，如何應用有效的雜訊清理方法，建立高品質的訓練資料，進而改進分類器的雜訊容忍度與鑑別準確率，則成為本研究的主要目的。

2. 研究方法

本研究所使用的實驗資料，是取自 UCI 資料庫(UCI Machine Learning Repository)所建立的 Statlog (Heart)心臟病資料集。該資料集共有 270 筆資料，13 個診斷屬性（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。診斷屬性分為有心臟病與無心臟病兩類，當 class=1 時表示無心臟病者共有 155 筆，約佔全部資料之 55.56%。

依據之前與本論文相同的心臟疾病資料研究（楊偉修等，2015），經由不同的正規化方法處理後，再使用 kNN 演算法進行鑑別分析結果，顯示若預先將資料經由 Z-分數正規化處理後再以歐幾里得距離進行計算，可獲得最佳的分類準確率。因此本研究於計算過程，亦是先將資料以 Z-分數正規化處理後，再以歐幾里得距離進行計算。

*通訊作者：yws@nkut.edu.tw

本研究之計算過程概略可分為四個階段，第一階段是將資料以 10 組交互驗證方法(10-fold cross validation method)分割為訓練資料與測試資料；第二階段則於訓練資料中，分別加入 7 種不同程度的類別雜訊比（0%、5%、10%、15%、20%、25%與 30%）；第三階段分別以 ENN、RENN 與 AENN 三種方法對訓練資料進行清理工作，而三種方法所設定的 k 值皆為 1、3、5、7、9；第四階段則以未加入雜訊的測試資料對清理過的訓練資料進行 1NN 規則驗證，最後計算出各種方法的平均 1NN 準確率進行分析比較。

3. 結果與討論

各種資料清理方法在加入不同雜訊比的情況下，所得之平均 1NN 準確率整理於表 1 中，由計算結果得知，本研究所使用的心臟病資料，原始資料（雜訊比為 0%）的 1NN 準確率為 75.56%，而在使用三種資料清理方法後，由各種清理 k 值所得之訓練資料，皆可明顯提升 1NN 準確率，其中尤以 RENN (k=5)方法清理後之準確率 84.07%最佳。

此外，在加入雜訊後，整體的平均準確率以 RENN (k=9)所得之 82.12%最高，而 RENN（k=5 與 k=7）所得之準確率 81.59%次之。最後由以上研究結果得知，以 RENN 方法清理此心臟病資料具有相當好的效果，因此建議若以 RENN 方法清理此心臟病資料，可以設定 k=5、k=7 或 k=9 作為清理標準。

表 1. 各種資料清理方法在不同雜訊比的情況下所得之平均 1NN 準確率

	0%	5%	10%	15%	20%	25%	30%	平均值
未清理	75.56	73.33	71.48	68.15	65.56	59.26	64.07	68.20
ENN (k=1)	80.00	78.89	77.78	74.44	72.59	68.15	70.00	74.55
ENN (k=3)	82.96	82.22	83.33	80.37	78.15	75.56	74.44	79.58
ENN (k=5)	83.33	83.33	83.70	84.44	78.52	78.15	75.56	81.01
ENN (k=7)	83.70	82.96	83.70	84.07	81.48	75.93	75.56	81.06
ENN (k=9)	83.33	82.96	83.33	83.70	80.37	77.04	77.78	81.22
RENN (k=1)	80.74	79.63	78.52	76.67	74.07	69.26	71.11	75.71
RENN (k=3)	83.33	82.96	83.70	82.22	80.37	76.30	74.81	80.53
RENN (k=5)	84.07	83.70	82.59	85.56	79.26	77.78	78.15	81.59
RENN (k=7)	83.33	82.59	82.96	84.81	80.74	79.26	77.41	81.59
RENN (k=9)	83.33	83.70	82.96	84.44	80.74	80.00	79.63	82.12
AENN (k=3)	82.59	81.11	81.85	81.11	78.52	74.44	73.33	78.99
AENN (k=5)	82.22	81.85	82.22	82.22	81.85	79.26	74.07	80.53
AENN (k=7)	82.22	81.48	82.22	81.85	82.22	80.74	76.30	81.01
AENN (k=9)	82.22	81.48	82.22	81.85	81.85	79.63	78.52	81.11

參考文獻

1. Tomek, I. (1976). An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics, 6*(6), 448-452.
2. UCI Machine Learning repository, website: http://archive.ics.uci.edu/ml/.
3. Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. Systems, Man and Cybernetics, *IEEE Transactions on, 2*(3), 408-421.
4. 楊偉修、戴良軒、彭冠傑、林柏豪(2015)。應用 KNN 演算法於鑑別心臟病之分析研究。*福祉科技與服務管理學刊，3*(3)。