調整鄰近點距離以增進 kNN 演算法鑑別帕金森氏疾病的雜訊容忍度

楊偉修
南開科技大學 企業管理系

1. 研究背景與目的

帕金森氏病(Parkinson's Disease)是老年人中常見的神經系統退化性疾病，症狀表現相當複雜且與其他多種疾病的症狀相似。在資料探勘領域中，近年來已有許多以帕金森氏病為主題的研究，但這些研究中鮮少考慮當訓練資料中含有誤判案例（又稱為類別雜訊）時，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，因此如何增進 kNN 演算法鑑別帕金森氏疾病的雜訊容忍度，則成為本研究的主要目的。k 最近鄰演算法（k Nearest Neighbor Algorithm，簡稱 kNN）是一種非常簡單且常用的機器學習方法。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雜訊時，1NN 分類器亦是最易受雜訊影響的分類器。在應用 kNN 演算法的分類過程中，使用不同的距離計算方法，會產生不同的分類結果，因此有許多的學者致力於發展各種新的距離計算方式，期以提高 kNN 演算法的分類準確率。Wang et al. (2007)提出 A-kNN 演算法，該法是先計算每個訓練點與不同類別訓練點間的最近距離（邊界距離），再將測試點對所有的訓練點距離，除以各相對訓練點的邊界距離，以此簡單的邊界距離調整鄰近點的順序，可以有效提升 kNN 的分類效能，但目前亦無研究探討此演算法之雜訊容忍度。

綜合上述說明，本研究將採用 A-kNN 演算法計算帕金森氏病資料，並對於此資料加入各種不同程度的類別雜訊(class noise)，進一步與 kNN 比較鑑別結果受雜訊的影響程度。此外，預先將資料以不同的正規化方法進行前置處理，以及使用不同的距離函數計算鄰近點之間的距離，皆會產生不同的 kNN 分類準確率(Ma et al., 2014)，因此本研究中亦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)以 A-kNN 方法計算，以分析各種情況下 A-kNN 方法的雜訊容忍度，並以分析結果提出具有高雜訊容忍度與準確率的帕金森氏病分類模型。

2. 研究方法

本研究所使用的實驗資料，是取自 Little(Little, 2009)所建立的帕金森氏病資料集。該資料集是針對 31 位 46～85 歲的病人，花費約 28 年時間，針對每位病人進行約六次的發音測試，並紀錄測試結果而得。資料集中共有 195 筆記錄，22 個輸入屬性為連續性資料，與一個類別標記屬性 status，status 的值有 0 與 1 兩種，當 status=1 時表示為確定病例共有 147 筆，約佔全部資料之 75.38%。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 透過相關的正規化公式，將該值正規化為 。再分別以歐幾里得距離(Euclidean distance)與曼哈頓距離(Manhattan distance)距離函數進行計算，以觀察 1NN 準確率受雜訊的影響情形。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%）；第三階段分別以上述

*通訊作者：yws@nkut.edu.tw

之資料正規化方法處理屬性值；第四階段則以兩種不同的距離計算 1NN 分類器與 A-1NN 分類器的平均準確率。

3. 結果與討論

各種資料正規化方法在加入不同雜訊比的情況下，以歐幾里得距離計算之平均 1NN 準確率整理於表 1 中，而以曼哈頓距離計算的結果則整理於表 2。由計算結果得知，當雜訊比增加時，兩種分類器皆受雜訊影響致使平均準確率下降。但以邊界距離重新調整鄰近點順序的 A-kNN 演算法，其 A-1NN 結果在各種雜訊比下，皆明顯優於原 1NN 的準確率，此表示 A-1NN 具有較高的雜訊容忍度。此外，以最大正規化進行資料預處理並以曼哈頓距離計算時，A-1NN 所得之平均準確率 84.66%最高。而若以歐幾里得距離計算，以極值正規化所得結果 84.62%次之。

表 1. 各種資料正規化方法在不同雜訊比的情況下，所得之平均 1NN 準確率（歐幾里得距離）

	無正規化		極值正規化		Z-分數正規化		最大正規化		十進位正規化	
雜訊比	1NN	A-1NN	1NN	A-1NN	1NN	A-1NN	1NN	A-1NN	1NN	A-1NN
0%	83.61	87.18	95.39	94.37	94.37	93.84	93.87	92.29	95.87	93.29
5%	82.11	85.66	91.34	93.87	88.74	91.29	90.82	91.82	91.74	92.76
10%	74.34	77.97	85.61	89.16	86.63	90.18	87.71	88.13	85.58	85.63
15%	70.18	74.76	80.05	87.26	82.11	86.79	80.08	85.18	77.47	86.18
20%	74.32	77.89	82.05	84.08	80.05	85.13	82.61	83.03	83.03	85.63
25%	63.71	72.18	73.97	77.03	72.45	77.53	72.92	75.47	69.24	74.84
30%	67.29	67.66	59.00	66.61	61.53	66.66	62.53	67.61	65.16	67.76
平均值	73.65	77.62	81.06	84.62	80.84	84.49	81.50	83.36	81.15	83.73

表 2. 各種資料正規化方法在不同雜訊比的情況下，所得之平均 1NN 準確率（曼哈頓距離）

	無正規化		極值正規化		Z-分數正規化		最大正規化		十進位正規化	
雜訊比	1NN	A-1NN	1NN	A-1NN	1NN	A-1NN	1NN	A-1NN	1NN	A-1NN
0%	84.68	87.71	93.87	93.89	93.89	93.89	93.37	93.92	95.39	94.89
5%	83.16	88.18	89.74	93.89	89.76	93.39	89.76	93.95	91.82	93.87
10%	74.39	80.55	83.03	86.61	83.00	85.55	85.63	85.05	85.16	90.24
15%	72.32	74.76	81.58	87.29	80.53	87.29	80.55	89.29	80.05	85.16
20%	71.79	79.00	79.05	83.68	80.11	82.16	78.03	82.63	80.55	83.61
25%	67.39	73.26	74.45	75.53	73.42	78.05	71.87	78.58	72.89	76.95
30%	67.74	67.71	64.11	66.21	64.66	68.18	63.58	69.21	63.61	66.63
平均值	74.50	78.74	80.83	83.87	80.77	84.08	80.40	84.66	81.35	84.48

參考文獻

1. Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Biomedical Engineering, IEEE Transactions on, 56*(4), 1015-1022.
2. Ma, C. M., Yang, W. S., & Cheng, B. W. (2014). How the Parameters of K-nearest Neighbor Algorithm Impact on the Best Classification Accuracy: In Case of Parkinson Dataset. *Journal of Applied Sciences, 14*(2), 171-176.
3. Wang, J., Neskovic, P., & Cooper, L. N. (2007). Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recognition Letters, 28*(2), 207-213.