BPN、SVM 與 NCPLS-DA 之比較分析

目錄

第一章 緒論

1.1 研究背景與動機

PLS-DA(Partial Least Squares Discriminant Analysis)為一種多變量的統計分析方法，經常使用於預測類別。特別的是，PLS-DA 在處理共線性和高維度的資料時，仍有良好的表現。但 PLS-DA 面對多模的資料時，卻時常表現不佳，導致低正確率的情況。為改善這樣的情況，提出了新的方法 NCPLS-DA(Nearest clusters based PLS-DA)，以原始的 PLS-DA 方法為基礎，加入階層式分群法(Hierarchical clustering)進行改良。而 Dewan et al.(2018)於實驗中，將 NCPLS-DA 和原始 PLS-DA 進行比較。結果顯示，NCPLS-DA 的正確率明顯優於其他兩者。但此實驗所使用的方法皆為統計分析方法，若與類神經網路方法與其比較，結果不得而知。

1.2 研究目的

本研究以 Dewan et al.(2018)所提出 NCPLS-DA 方法的論文做為參考文獻，使用類神經網路方法與此文獻中的方法進行比較。因參考文獻所提出的 NCPLS-DA，以及實驗中所比較的方法，皆為一般統計方法。而類神經網路方法又和一般統計分析方法不同，不需嚴格的假設限制，且擁有處理非線性問題的能力，在預測分類上有良好的表現。故本研究為比較類神經網路方法和統計分析方法的表現，將在多個資料檔上，使用 BPN(Back Propagation Network)和 SVM (Support Vector Machine)此兩種類神經網路方法，評估其正確率，和參考文獻中的實驗結果進行比較。

1.3 NCPLS-DA 介紹

NCPLS-DA 是以原始的 PLS-DA 為基礎，加入階層式分群法進行改良的方法。一開始，和 PLS-DA 相同，先將訓練資料進行主成分分析(Principal Component Analysis；PCA)，選取重要性最高的兩個主成分作為屬性，加入每筆資料資料的分數和類別，形成新的訓練資料。以新的訓練資料進行階層式分群，而計算群心的方式如公式一：

$$\boldsymbol{\mu} = \frac{1}{c}\sum_{i=1}^{c}\boldsymbol{x}_{(i)} \tag{1}$$

在公式一中，c 為此群集的資料數量，x_i為群集內第 i 個樣本資料，而μ 即為此群集的群心。而加入的新一筆資料將會比較，與群心的 Euclidean 距離，並選取距離較近的幾個群集，進行後續分析。若這些群集皆屬與相同類別，則預測新資料為此一類別，否則，使用原始 PLS-DA 方法預測類別。

其中，有兩個參數必須事先決定，一個為資料要分割的群集數，以下以 CN (clustering number)表示，另一個為選擇最近群集的數量，以下以 NC (nearest cluster)表示。CN 太小，將會導致一些有用的資訊無法幫助預測，CN 太小，將會不必要的劃分類似的資料。為求出適當的 CN 值，將使用平均 Euclidean 距離方法，評估某範圍內最佳的 CN 值，如公式二：

$$AED_{\mathrm{CN}} = \frac{1}{\mathrm{CN}}\sum_{i=1}^{\mathrm{CN}}\left|\left|\overline{\boldsymbol{X}} - \boldsymbol{\mu}_i\right|\right|_2. \tag{2}$$

在公式二中，$\bar{X}$為樣本平均，μ_i代表第 i 個群集的群心。而 NC 的部分，若數量太小，也會導致一些資訊無法幫助預測，若太大將會趨近於全域的 PLS-DA，失去了改良的效果。而此 NC 將依照經驗設定，初始值將為接近 CN / 3 的整數，依情況再進行調整。以上這些參數並非在所有情況下，都能產生最佳效果，但幾乎在所有情況下，效果皆優於原始 PLS-DA。

第二章 實證研究

本研究使用 SVM 和 BPN 方法，在 12 個資料檔上進行實驗。評估方法與參考文獻相同，以重複 10 次的 10 等分交叉驗證求得平均正確率。

2.1 資料檔介紹

參考文獻總共使用 17 個資料檔進行實驗，其中，有 5 個資料檔為參考文獻自行模擬的光譜資料。因無法取得其光譜資料，故使用其餘 12 個來自 UCI 機器學習資料庫的資料檔進行實驗。本研究所使用的資料檔，一部分包含高維度、多類別的特性。其中，Arcene 資料檔包含了數千個屬性，參考文獻建使用了 ReliefF 演算法，從中選取 1000 個屬性，以達成特徵選取的效果。而本研究為求實驗結果公正，亦會採取相同作法。

下表為 12 個資料檔的資訊：

資料檔名稱	樣本筆數	屬性個數	類別個數
Arcene	200	1000	2
Breast Tissue	106	9	4
Ecoli	336	7	8
Forest Types	523	27	4
Glass	214	9	6
Ionosphere	351	34	2
Iris	150	4	3
Parkinsons	195	22	2
QSAR-biodeg	1055	41	2
Sonar	208	60	2
SPECTF heart	267	44	2
Zoo	101	16	7

表 2.1 資料檔資訊

2.2 實驗過程與參數設定

2.2.1 資料前置處理

當我們把資料輸入進模型時，常常需要對資料先做一些前處理，藉此提升模型的預測效果。首先檢查是否有遺失值，其原因是如果有缺值，就沒辦法在空間中表示出位置，而檢查結果並無發現遺失值，接著進行標準化(Standardization)，以使離群值(Outlier)對整個模型的影響大大減低。

2.2.2 BPN

a.方法設計

使用了倒傳遞類神經網路(Backpropagation Network)，並搭配慣性項(momentum)加速神經網路訓練，以及正規化(regularization)來防止過度合適（overfitting)的問題，而實作工具為 Python 的 tensorflow 套件。

b.結構與參數

在使用不同的資料檔時，除了學習率與慣性向比率皆相同外，神經網路結構與正規化比率會被多次挑出不同的測試組合，進行十等分交叉驗證，並將驗證出最高正確率之神經網路結構與正規化比率留下，作為目前使用資料檔之參數設定。學習率為 0.02，慣性項比率為 0.3，停止訓練條件為使用訓練資料進行回測，即將訓練資料放入模型中進行分類測試，當分類錯誤率不再大量變動時停止訓練，並經過結果觀察，沒有出現因為學習率過大造成錯誤率無法收斂之問題。

神經網路結構的選擇為使用一層隱藏層或兩層隱藏層，為了減少最佳結構搜尋時間，在使用兩層隱藏層時，各層皆使用相同的神經元數目，而每個隱藏層的最佳神經元數目搜尋範圍，介於 3 到 20 個神經元數目，搭配隱藏層共有 36 種神經網路結構組合。正規化比率範圍為3×10^{-i}與1×10^{-i}，其中 $\mathrm{i} = 2,3,4,5,6,7,8$，共使用 14 種大小做為測試，。

所以每個資料檔總共會以 504 次不同的神經網路結構與正規化比率之組合，進行十等分交叉驗證。

c.BPN 實驗結果

資料名稱	正確率(%)	正確率之標準差	隱藏層數目	每隱藏層之神經元數目	正規化比率
Arcene	90.00	1.24	1	12	0.00001
BreastTissue	87.37	1.26	1	13	0.00003
Ecoli	85.57	0.83	1	6	0.0003
ForestTypes	90.00	0.27	2	7	0.0001
Glass	70.32	1.82	2	13	0.00003
Ionosphere	89.88	0.83	2	5	0.00003
Iris	97.20	0.40	1	7	0.0001
Parkinsons	92.15	1.26	2	6	0.0003
QSAR biodegradation	87.66	0.45	1	3	0.0003
Sonar	80.11	1.75	2	4	0.00003
SPECTF heart	79.86	0.95	1	13	0.003
Zoo	94.40	1.21	1	14	0.000003
平均正確率	87.04				

表 2.2 BPN 實驗結果與參數設定

2.2.3 SVM

a.方法設計

使用支援向量機(Support Vector Machine)，並將資料標準化(Standardization) 經過標準化後，資料會符合常態分佈，不會有偏單邊的形況，同時使離群值(Outlier)對整個模型的影響大大減低。而實際操作工具為 Python 中 SVC，SVC 是 SVM 用 C++語言實作的版本，背後用的是台灣大學林智仁教授所開發的 libsvm。

b.參數設定

參數部分考慮 Kernel 與懲罰值 C，首先 Kernel 部分考慮 linear 與 rbf，而其 gamma 值是以$\frac{1}{特徵數}$設定，並找出適合各筆資料集的 Kernel，接著 C 值則先選用 0.01 到 1000，以 10 倍作為一次挑選，例如 0.01 後依續為 0.1、1、10 等等，最後在 1000 停止，在初次模擬結果跑出後，我們發現在 C 值為 10 時，有較佳的平均正確率，後續限縮在 10 附近搜尋更佳的 C 值，並在各筆資料集裡找出其正確率為最大的 C 值，接著進行十等分交叉驗證，求得平均正確率。

c. SVM 實驗結果

資料名稱	正確率(%)	正確率之標準差	Kernal	C 值
Arcene	90.05	(0.72)	linear	15.5405
BreastTissue	88.22	(1.06)	rbf	11.2162
Ecoli	86.54	(0.46)	rbf	15.5586
ForestTypes	89.49	(0.83)	linear	11.4144
Glass	69.02	(1.16)	rbf	16.3784
Ionosphere	94.84	(0.43)	rbf	15.7568
Iris	96.00	(0.60)	linear	14.8829
Parkinsons	91.57	(1.29)	rbf	15.7297
QSAR biodegradation	88.66	(0.29)	rbf	15.1171
Sonar	88.46	(0.89)	rbf	13.7297
SPECTF heart	79.71	(1.77)	rbf	12.3423
Zoo	96.53	(1.11)	linear	12.7840
平均正確率	88.26			

表 2.3 SVM 實驗結果與參數設定

2.3 實驗結果

以 12 個資料檔下的正確率和標準差進行比較，結果如下表所示：

	PLS-DA	NCPLS-DA	BPN	SVM
資料檔	正確率 (標準差)	正確率 (標準差)	正確率 (標準差)	正確率 (標準差)
Arcene	84.65 (0.97)	91.95 (1.09)	90.00 (1.24)	90.05 (0.72)
BreastTissue	84.62 (0.44)	89.60 (1.54)	87.37 (1.26)	88.22 (1.06)
Ecoli	84.52 (0.97)	86.83 (0.68)	85.57 (0.83)	86.54 (0.46)
ForestTypes	86.07 (0.32)	90.56 (0.46)	90.00 (0.27)	89.49 (0.83)
Glass	59.19 (2.01)	68.67 (1.16)	70.32 (1.82)	69.02 (1.16)
Ionosphere	86.77 (0.24)	95.41 (0.51)	89.88 (0.83)	94.84 (0.43)
Iris	82.40 (0.47)	98.20 (0.45)	97.20 (0.40)	96.00 (0.60)
Parkinsons	85.66 (0.70)	88.14 (1.07)	92.15 (1.26)	91.57 (1.29)
QSAR biodegradation	83.33 (0.37)	85.05 (0.39)	87.66 (0.45)	88.66 (0.29)
Sonar	76.33 (1.42)	84.12 (1.16)	80.11 (1.75)	88.46 (0.89)
SPECTF heart	78.28 (0.43)	81.26 (0.61)	79.86 (0.95)	79.71 (1.77)
Zoo	93.93 (0.88)	97.53 (0.70)	94.40 (1.21)	96.53 (1.11)
平均正確率	82.15	88.11	87.04	88.26

表 2.4 實驗結果比較

a. BPN

BPN 在 12 個資料檔上的正確率皆高於 PLS-DA 方法，整體而言，高於 PLS-DA 方法的平均正確率 4.89。與 NCPLS-DA 相比，BPN 只在 3 個資料檔上的正確率高於 NCPLS-DA 方法，此 3 個資料檔分別為 Glass、Parkinsons 和 QSAR biodegradation，而平均正確率低於 NCPLS-DA 方法 1.07。

b. SVM

SVM 在 12 個資料檔上的正確率高於 PLS-DA 方法，整體而言，高於 PLS-DA 方法的平均正確率 6.11。與 NCPLS-DA 相比，SVM 在 4 個資料檔上的正確率高於 NCPLS-DA 方法，此 4 個資料檔分別為 Glass、Parkinsons、Sonar 和 QSAR biodegradation，而平均正確率高於 NCPLS-DA 方法 0.15。SVM 在 6 個資料檔上的正確率皆優於 BPN，其平均提高正確率 1.22。

第三章 結論與建議

3.1 結論

實驗結果中，在多個資料檔上，SVM 總體正確率總和之平均高於 PLS-DA 與 NCPLS-DA 方法，BPN 總體正確率總和之平均高於 PLS-DA 方法，推測 SVM 方法比起一般統計分析方法，更能有效預測分類；而在部分資料檔上 SVM 與 BPN 的表現不及 NCPLS-DA，推測原因可能是參考文獻所提出的 NCPLS-DA，不但使用了 PCA 方法，達到降低維度的效果，又以階層式分群方法，只以類似的資料進行處理。相對來說，類神經方法卻使用了所有的屬性和資料。而觀察多數的資料檔，可發現部分的資料檔，相對於變數數量來說，樣本數量偏少，因此在與相同變數數量的維度中，過少的樣本可能無法適當的配適資料的母體分配。以上這些原因，皆有可能導致 BPN 和 SVM 分類效果不及 NCPLS-DA，但整體而言，從 12 個資料檔的平均正確率之總和平均數來看，SVM 與 NCPLS-DA 正確率並無明顯落差。

3.2 建議

在評估方法中，為求實驗結果公正，本研究依照參考論文，使用了重複 10 次的 10 等分交叉驗證法，求得平均正確率。但此種方法可能導致正確率的高估或低估，因為 k 等分交叉驗證法，是將資料隨機切割為 k 個獨立的等分，而每等分都只有一次作為測試資料的機會。但若重複使用 k 等分交叉驗證，在多次的切割等分中，將會造成測試資料的重疊，測試資料可能彼此相依，求出的正確率也可能相依。假設資料檔中包含對正確率有幫助的資料，而這些資料又重複出現於測試資料，以這些相依的測試資料求得正確率，再由相依的正確率求出平均正確率，可能導致此平均正確率過於樂觀。

參考文獻

Weiran Song , Hui Wang , Paul Maguire & Omar Nibouche.(2018).Nearest clusters based partial least squares discriminant analysis for the classification of spectral data. *Analytica Chimica Acta* ,1009,27-38.