改進支援向量資料描述中特徵分類準確度之研究

王啟凱*　林依亭**　游雅茹**　陳之寧**
中山科學研究院第二研究所*
龍華科技大學資訊管理系**
olm83tf2303@hotmail.com

摘要

多數的樣本識別研究多著重於分類與推論領域，但是事實上資料描述分類的重要性並不亞於前述兩項工作。資料描述的本質不在於如何有效分割重疊或混雜在一起的資料，而是如何明確地判斷出哪些資料是屬於同一個群體(Group)或類別(Class)，也就是資料描述或識別的意義在於離群值(Outlier)或邊界值(Boundary)的判別，離群值或邊界值愈少愈明確，有助提升影像判讀的準確度。支持向量資料描述(Support Vector Data Description, SVDD)脫胎於支援向量機(Support Vector Machine，SVM)，能提供有效之資料描述識別結果。但是常囿於樣本資料有限，影響識別結果之準確度。本研究提出利用極值區間法(Max-min range method)作為人工離群值取樣之計算基礎，產生圍繞於目標資料集(Target set)的離群樣本，並利用 N 摺 M 次交叉驗證(N-fold M times cross validation)方法，我們可以得到懲罰因子參數 C 與間距寬度參數 S 的最佳(C, S)組合結果。最後，我們以 UCI 標準測試資料庫對本研究所提出之極值區間法進行測試，以驗證其有效性。

關鍵詞：支持向量資料描述、交叉驗證方法、極值區間法、特徵分類

Abstract

Most pattern recognition tasks deal with classification and regression problems. But the data domain description problem is as important as them. In domain description, the task is not to part with overlapped or mixed objects, but to judge them into the same group or class. It means if we can find the boundary around the target data closely, we can get better accuracy of image judgment. Support Vector Data Description (SVDD) is inspired by the Support Vector Machine (SVM), and can provide an effect accuracy of data domain description. But the accuracy is blundered by the amount of samples. In this research, we utilize max-min range method to generate outlier objects around the target data artificially. By N-fold M times cross validation method, we can get the best (C, S) combination. At last, we use the UCI Machine Learning Dataset Repository to validate the effect of two methods.

Keywords: Support Vector Data Description (SVDD), cross validation method, max-min range method, feature classification

1. 前言

多數的樣本識別研究大多著重於分類與推論領域，而事實上資料描述分類的重要性不亞於前述兩項工作。資料描述的本質不在於如何有效分割重疊或混雜在一起的資料，而是如何明確地判斷出哪些資料是屬於同一個群體(Group)或類別(Class)。從某個角度來說，資料描述或識別的意義在於離群值(Outlier)或邊界值(Boundary)的判別，離群值或邊界值愈少或愈明確，意味著對於影像的判讀能更加準確。Moya、Koch 與 Hostetler 於 1993 年提出的以目標類別密度量測為核心的單類別分類法(One-Class Classification)[6]，或是 Vapnik 於 1998 年所提出的資料支援向量分類法[17]，都是近年來逐漸被學者重要及使用的重要影像特徵分類技術之一。

支持向量資料描述(Support Vector Data Discription, SVDD)[8][9][10]脫胎於支援向量機(Support Vector Machine，SVM)，有別於 SVM 尋找一超平面(Hyperplane)來分隔兩個類別資料，SVDD 則是試圖找出一個超球體(Hypersphere)使能儘量包圍住所有或最多的欲訓練樣本資料，當然也有許多降低其分類錯誤率的研究[1][2][6][7]。SVDD 方法已廣泛應用於許多實際領域，如影像特徵分類、機械故障

檢測、語音識別等，均有顯著的實用成效。在本研究中，我們利用兩摺式交叉驗證法(Two-Fold Cross Validation Method)來描述與分析影像資料，此方法主要是利用單類別分類的特性，將影像資料分成目標類別（目標類別）與離群類別（離群類別），利用類別資料集隨機平分後，再利用公式判斷與計算錯誤率，詳細的內容將於第二小節及第三小節中敘述。第四小節則針對實驗的數據結果進行分析探討，最後並提出一簡單的結論。

2. 方法簡介

2.1 SVDD

SVDD 方法是從支持向量機方法得到啟發的，其目的是去找尋能夠涵蓋所有的訓練資料且擁有最小體積(或是最小半徑)的最佳超球體(Optimize Sphere Hypersphere, OSH)，環繞著資料集合的球狀決策邊界(Decision boundary)是由可描述超球體邊界的支持向量集合所建構。而且可將原資料轉換至新的特徵空間而不需要太多經過額外的計算，經由轉換過的資料，SVDD 可以得到更合適、精確的資料描述。

當一個包含個資料的資料集合(目標類別資料集合) $\{x_i, i=1.....N\}$ 需要被描述，我們試著去找到能包含所有(或是多數)資料且擁有最小體積的超球體，但是，如果要包含少數較偏離的訓練資料，此一超球體將會變大且不能將資料表示的很精準，為了求出適當的精準結果，我們允許一些較偏離的資料在這個超球體之外，於是引入彈性變數(slack variables) ξ_i。

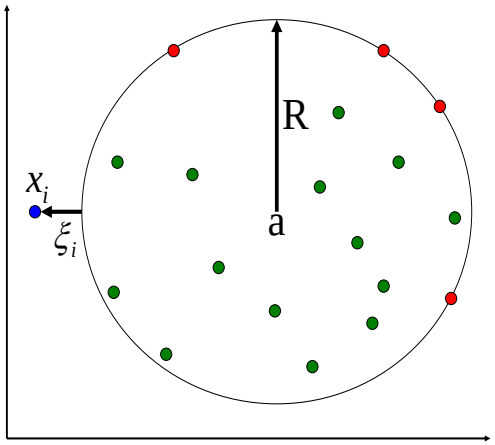


圖 1 SVDD 示意圖

為了描述一個中心為 a，半徑為 R 的超球體，我們試著最小化半徑：

$$F(R,a,\xi_i) = R^2 + C\sum_{i=1}^{N}\xi_i \quad (1)$$

其中 C 是資料的錯誤權重值(會影響超球體的體積或是目標類別被捨棄的個數)。而最小化(1)式的限制條件為(2)式：

$$(x_i - a)^T(x_i - a) \le R^2 + \xi_i$$
$$,\xi_i \ge 0 \quad i = 1.....N \quad (2)$$

此時導入 Lagrangian 可寫成：

$$L(R,a,\alpha_i,\xi_i) = R^2 + C\sum_{i=1}^{N}\xi_i -$$
$$\sum_{i=1}^{N}\alpha_i\left\{R^2 + \xi_i - (x_i^2 - 2ax_i + a^2) - \sum_{i=1}^{N}\gamma_i\xi_i\right\} \quad (3)$$

其中，Lagrangian 乘子 $\alpha_i \ge 0$，$\gamma_i \ge 0$。將(3)式對變數做偏微分並令其值等於零：

$$\left.\begin{aligned}\frac{\partial L}{\partial R} = 0 \Rightarrow \sum_{i=1}^{N}\partial_i = 1\\ \frac{\partial L}{\partial a} = 0 \Rightarrow a = \sum_{i=1}^{N}\partial_i x_i\\ \frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \gamma_i = 0\end{aligned}\right\} \quad (4)$$

於是，原問題的對偶問題可轉變成求取最大化問題：

$$L = \sum_{i=1}^{N}\alpha_i(x_i \cdot x_i) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(x_i \cdot x_j) \quad (5)$$

限制條件為：

$$\sum_{i=1}^{N}\alpha_i = 1, 0 \le \alpha_i \le C \quad i = 1,.....N \quad (6)$$

此時可以發現，超球體的中心 a 是資料點的線性組合，而擁有非零的 α 值的資料點為支持向量，半徑 R 由符合 $0 < \alpha < C$ 的資料點求得。對於目標類別的訓練資料而言，當 $\alpha_i = 0$ 時，資料點落在超球體邊界之內，判定接受；當 $0 < \alpha_i < C$ 時，資料點落在超球體邊界上，判定接受；$\alpha_i = C$ 時，資料點落於超球體邊界之外，為不被接受的資料點，稱做被捨棄的目標類別資料(Target rejected data)。

決定一個新的測試資料 Z 是否在超球體中，是去計算它與超球體中心 a 的距離，當它與 a 的距離小於超球體半徑 R，即 $(z-a)^T(z-a) \le R^2$，則接受。由支持向量表示時為(7)式：

$$(z\cdot z)-2\sum_{i=1}^{N}\alpha_i(z\cdot x_i)+\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(x_i\cdot x_j)\le R^2 \quad (7)$$

當原資料在輸入空間的分佈不為球狀時(即使一些離群值已經忽略)，我們無法由此方法找到適合且嚴謹的資料描述，於是引入核心函數(Kernel Function)K，兩資料向量的內積$(x_i\cdot x_j)$可以被$K(x_i\cdot x_j)$，則 Lagrangian 變成(8)式：

$$L=\sum_{i=1}^{N}\alpha_i K(x_i\cdot x_j)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K(x_i\cdot x_j) \quad (8)$$

限制條件為(9)式：

$$\sum_{i=1}^{N}\alpha_i=1 \quad ,\ 0\le\alpha_i\le C \quad i=1.....N \quad (9)$$

於是，一個新的測試資料可以被接受的情況如(10)式所示：

$$K(z\cdot z)-2\sum_{i=1}^{N}\alpha_i K(z\cdot x_i)+\sum_{i=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_j K(x_i\cdot x_j)\le R^2 \quad (10)$$

不同的核心函數 K 會在原輸入空間得到不同的描述邊界，所以問題在於找到最適合的核心函數。

2.2 SVDD 之參數分析

SVDD 是一種判斷邊界的方法，因為在目標資料周圍可以找出一條封閉的邊界，相較於 SVDD，SVM 則沒有邊界限制，取而代之的是以找出最大間距的超平面方式加以切割。根據[7]的證明，不論是 SVM 或是 SVDD，其訓練方法與訓練成果均與參數有關。因此在本小節將探討兩個參數，一個跟核心函數有關的參數，稱之為間距寬度參數S。另一個則是與單類別學習有關的參數，懲罰因子參數C。

從前面可以知道，SVDD 的做法是將原始資料利用核心函數映射一個高維特徵空間。此核心函數由一線性學習機(Linear Learning Machine)產生，藉以將資料轉換到特徵空間中並使得這些資料得以被分類。通常一個核心函數包含一個內積的運算，例如：

$$x,z\in X, K(x,z)=(\phi(x)\cdot\phi(z))$$

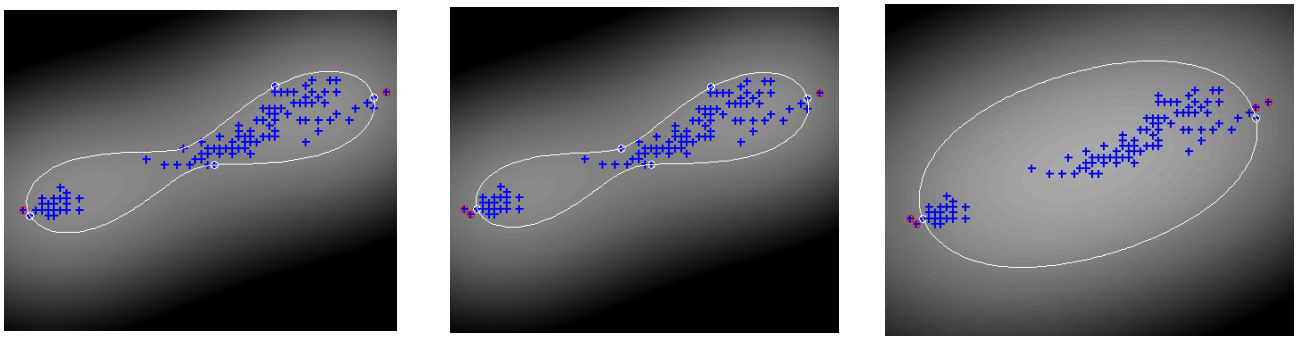
其中ϕ即為X映射到特徵空間F的映射。常見的 SVDD 核心函數有：線性函數、多項式函數、高斯函數等，其中高斯函數：

$$k(x,y)=e^{-\|x-y\|^2/s}$$

為最常用的核心函數，本研究亦採用高斯函數進行處理與計算。

在 SVDD 中，間距寬度參數S的大小影響著支援向量的個數。當S值變大時，則所產生的超球體體積會愈大，此時支援向量的個數就會減少。而 SVDD 的目的當然是希望降低超球體的體積，然而當S值減小時，則所產生的超球體體積會愈小，但相對會需要大量運算時間（如圖 2 的(b)、(c)）。因此如何在效果與運算效率上取得平衡，將由間距寬度參數S決定。

另外一個重要的參數稱為懲罰因子參數C，它可以視為是測試資料是否被拒絕的門檻。當C值愈小時，表示更多的資料會被判定位於超球體之外，也就是此時被拒絕的比率會增加，這個參數影響了容忍區域的大小，若要降低被拒絕資料的比率，相對的C值就必須愈大（如圖 2 的(a)、(b)）。因此最理想的狀況就是以最小的容忍區域去包含最多的訓練資料。當然這個問題如同S值一樣，C值的選擇受到效果與運算效率兩個因素的限制。



(a) C=0.3, s=3 **(b) C=0.2, s=3** **(c) C=0.2, s=5**

圖 2 不同的C與S值對容許範圍程度的影響

2.3 交叉驗證法

在單一類別之問題中，可供訓練的樣本資料是有限的，然而當樣本資料的數量較少時，將會導致訓練績效之正確性降低。為使資料得到較佳之正確性，目前常用的方式之一就是交叉驗證法。交叉驗證法之基本概念乃將樣本資料正確之部分視為接受培訓的訓練集，而其餘剩下之部分則視為測試績效的測試集。之後將訓練集與測試集組之資料互換，並將兩者的結果予以平均，如此將可使原先訓練集與測試集間取樣不均衡之影響程度降低，使運算後之結果接近更加真實性。依據前面內容所述，我們可以知道在單類別分類方法中，懲罰因子參數 C 與間距寬度參數 S 兩個參數為影響分類結果的重要因素，因此如何從已知的目標資料中找出最佳的(C, S)參數組合，將決定分類結果的精確程度。一般常見的交叉驗證法有兩類：

1. 保留式交叉驗證(Holdout cross validation)：保留式交叉驗證是最簡單的交叉驗證方法。原始的資料集被隨機劃分成兩個獨立的集合，也就是所謂的訓練集與測試集。只使用訓練集的資料來產生模型，接著對測試集預測輸出值，再利用原本已知的答案來評估正確性，如圖 3 所示。這個方法的優點是提供模型實地應用的表現成效，然而資料的分割方法可能會影響到評估的結果，如果將訓練集與測試集資料重新劃分，評估效能有可能會大不相同。一般來說，作為測試的資料量不會超過原始樣本的三分之一。
2. k 摺交叉驗證(k-fold cross validation)：由於使用保持交叉驗證方法建立模型時只用了一部份的初始資料，因此評估的結果較為保守，k 摺交叉檢定可以用來改進保持交叉驗證。將樣本資料分為 k 個子集，然後重複 k 次保持交叉驗證，在每一次進行中，選 k 個子集的其中一個作為測試集，其餘的 $(k-1)$ 個子集作為訓練集，取 k 次結果的平均作為整體評估。這個方法的優點是比較不會受到資料分割方式的影響，每一筆資料都當了一次的測試資料以及 $(k-1)$ 次的訓練資料，當 k 值越大則結果變異度越小。而這個方法的缺點就是演算法必須重複執行 k 次，也就得花上 k 倍的計算量。此外這個方法還有一個變形：隨機將資料分為訓練與測試資料 k 次，這個方法的優點是使用者可以自由的將訓練集大小與測試次數分開來考慮。

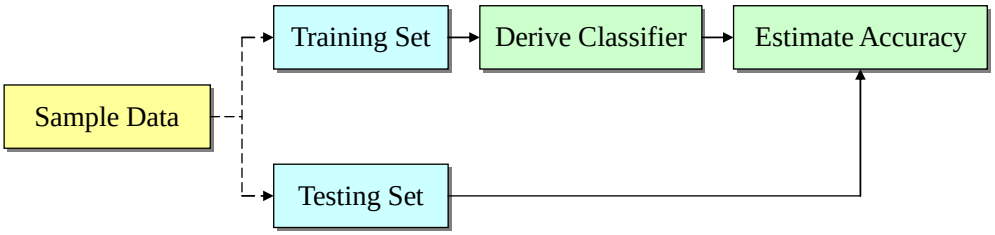


圖 3 保留式交叉驗證流程

3.研究方法

3.1 產生人造離群值資料

單類別分類法的優點在於不需要大量的原始資料進行模型訓練，然而在訓練過程中，較多的測試資料卻反而能提高模型訓練的成效，使訓練出來的模型具有較高的分類準確率。為了彌補原始資料的不足造成訓練結果錯誤率偏高的問題，本研究提出利用內插法方式，從已知測試資料中的離群資料，衍生創造出若干個離群資料，藉以增加測試資料的數量，同時不影響原有的(C, S)最佳參數組合結果。

由圖 2 的例子中可以知道，若能得到足夠數量的離群值資料圍繞在目標資料集外圍，便能求得更為精確的邊界值。為了達到這個目的，則必須透過人工產生的方式增加離群值資料的數量。一般而言單類別分類的優點及在於僅需要較少的資料即可進行分類，但在另外一方面意味著必須透過其他方式增加測試資料的數量，以提高訓練結果的準確率。在本研究提出極值區間法(Max-min range method)用以產生能圍繞目標資料集的人工離群值資料。假

設已知的資料集包含 M ($M \geq 1$) 個資料 $\{x_i, i = 1...M\}$，每個資料均包含 N ($N \geq 1$) 個屬性值 $\{a_i, i = 1...N\}$。此時可以將每個資料看成是位於 N 維特徵空間中的向量。而極值區間法的主要精神，就是在每一維度中找出適當的值，使其剛好落於目標資料集的外圍，藉以找出包圍於目標資料集的人工離群值的「向量值」。假設要產生 j 個人工離群值，則產生的步驟如下：

1. 隨機選取已知樣本尚未選取過的任意兩個維度（分別稱為 a_x 和 a_y，$(x \neq y)$），分別將所有樣本中這兩個維度的值進行升冪排序的程序。此時，a_x 和 a_y 分別表示如下：

$$a_x = Dim(V_x) = \{V_{x_1}, V_{x_2}, ..., V_{x_M}\} \quad (13)$$

$$a_y = Dim(V_y) = \{V_{y_1}, V_{y_2}, ..., V_{y_M}\} \quad (14)$$

2. 從 a_x 和 a_y 分別找出其最大值與最小值：a_x^{Max} 、 a_x^{Min} 、 a_y^{Max} 、 a_y^{Min} 。
3. 從已知樣本資料中隨機選取出一個，接著從 a_x 維度中隨機選取 a_x^{Max} 或者 a_x^{Min} ，然後將剛才隨機選取出的樣本資料的同一維度的數值置換成 a_x^{Max} 或者 a_x^{Min} ，其他維度的數值則保持不變。同理，從 a_y 維度中隨機選取 a_y^{Max} 或者 a_y^{Min} ，然後將剛才樣本資料的同一維度的數值置換成 a_y^{Max} 或者 a_y^{Min} ，其他維度的數值亦保持不變。
4. 將上述的結果儲存起來，並且視為一個新的離群值資料。
5. 重複步驟 1 到步驟 4 的程序 j 次，則可得到要產生的 j 個人工離群值值資料。

3.2 N 摺 M 次交叉驗證

利用均值變異數法以原始資料中的資料為基礎，產生新的人工離群值資料之後，藉由增加訓練次數來提高模型的分類準確能力。其演算法如下：

1. 將原始資料分成目標資料與離群資料。
2. 隨機將目標資料等分成 N 等分。
3. 以極值區間法產生人工離群值（產生方法如前一小節所述），並將其納入測試資料中。
4. 將目標資料的其中一等分加入測試資料集合中，以剩餘的 $(N-1)$ 等分的目標資料做為訓練資料，利用 SVDD 方法計算出錯誤率，錯誤率計算公式為：

$$ER = \frac{FP + FN}{TP + FP + FN + TN} \times 100\%$$

其中：

TP ：被判斷正確的訓練資料個數。

FP ：被判斷錯誤的測試資料個數。

FN ：被判斷錯誤的訓練資料個數。

TN ：被判斷正確的測試資料個數。

5. 將 N 個錯誤率值加總平均，得出 N 摺平均錯誤率 E_{aveage} 。
6. 重覆步驟 2 到步驟 5 的過程 M 次，並將所得之 M 個平均錯誤率 E_{aveage} 加總平均，得出總體平均錯誤率 E_{total} 。
7. 以網格(Grid)方式調整 (C, S) 參數組合重覆步驟 6 的過程，比較後可找出最小總體平均錯誤率 $Min(E_{total})$，藉此可得出原始資料之最佳 (C, S) 參數組合結果。

4. 實驗結果與分析

為測試極值區間法的效果，本研究使用 UCI 機器學習實驗室所提供之標準資料集[5]中的 Iris、Wine、Balance-scale 三個資料集做為測試對象。事實上，一般進行相關研究之測試，大多採用 UCI 機器學習實驗室之標準資料集，藉以比較準確率及效果。三個資料集基本資料如表 1 所示：

表 1 測試資料集

資料集名稱	類別數	資料數	特徵數
Iris	3	150	4
Wine	3	178	13
Balance-scale	3	625	4

經過 N 摺 M 次交叉驗證法結合內插法進行分類訓練，所得到的 (C, S) 最佳參數組合及最低錯誤率結果如表 2 所示：

表 2 測試結果

資料集	原始離群值個數	新增的人工離群值個數		
		產生 50 個	產生 100 個	產生 200 個
Iris				
類別 1(50)	(0.6, 4.2)2.62%	(0.6, 5.1)1.74%	(0.6, 4)1.23%	(0.6, 3.9)0.77%
類別 2(50)	(0.57, 2.7)8.17%	(0.55, 2.9)4.78%	(0.56, 2.6)3.43%	(0.6, 3)2.18%
類別 3(50)	(0.38, 3)8.32%	(0.39, 2.8)4.8%	(0.36, 2.7)3.46%	(0.37, 2.9)2.26%
Wine				
類別 1 (59)	(0.1, 9.2)32.82%	(0.1, 9.4)21.54%	(0.1, 9.2)15.45%	(0.1, 9.4)10.1%
類別 2 (71)	(0.1, 10)39.23%	(0.1, 10)25.12%	(0.1, 9.9)18.47%	(0.1, 10)12.07%
類別 3 (48)	(0.1, 1)26.97%	(0.1, 1)17.27%	(0.1, 1)12.7%	(0.1, 1)8.3%
Balance-scale				
類別 1 (288)	(0.1, 2)17.81%	(0.1, 2)12.79%	(0.1, 2)10.92%	(0.1, 2)9.26%
類別 2 (49)	(0.1, 1)7.84%	(0.1, 1)6.76%	(0.1, 1)5.8%	(0.1, 1)4.78%
類別 3 (288)	(0.1, 2)17.94%	(0.1, 2)12.91%	(0.1, 2)10.86%	(0.1, 2)9.11%

如同前面內容所描述，當其中一種類別被當成是目標類別時，其他的類別則都被視為是離群類別（非目標類別）。在本研究中，我們採用 2 摺 10 次的交叉驗證方法來求取(C, s)的最佳組合值。至於產生人工離群值的方法則是採用前一小節所描述的極值區間法。至於要產生多少人工離群值，則可視需要自行決定。

在表 2 中，左邊的數字組合表示所得到的最佳(C, s)組合值，而右邊的數字則是代表所獲得的最小錯誤率。舉例來說，在 Iris 資料集中，若以類別 1 當成目標類別，類別 2 與類別 3 當成離群類別，則利用經過 2 摺 10 次的交叉驗證之後，可以得到(0.6, 4.2)為最佳組合值，此時的最小錯誤率則為 2.62%。

從 Iris 資料集來觀察，當人工離群資料增加 50、100、200 個之後，可以看出所計算出的(C, s)組合約在(0.6, 4)附近，而最小錯誤率則隨著人工離群資料數量的增加而降低，可見所產生的人工離群資料密集圍繞在目標資料集，確實使訓練出來的模型的錯誤率有降低的趨勢。再從 Wine 資料集與 Balance-scale 資料集來看，不論是哪一個類別當成目標類別，在利用極值區間法產生人工離群資料之後，其錯誤率亦均有降低趨勢。由此可見利用極值區間法所產生確實沒有在超空間中發散，而是有效圍繞在超球體的外圍附近。表 2 中錯誤率與人工離群值增加個數的關係如圖 3 所示。所有曲線均呈現下降趨勢。

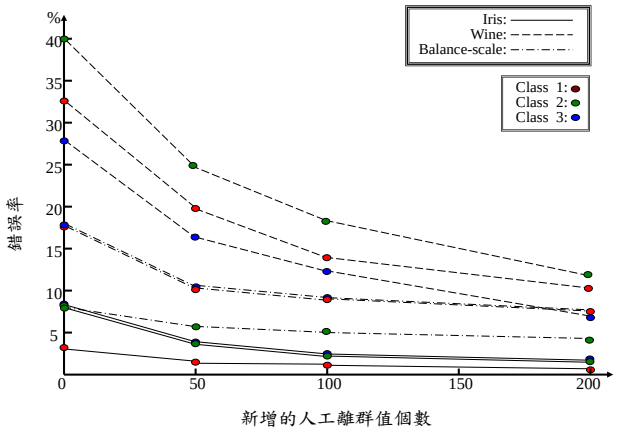


圖 3 錯誤率與人工離群值個數

另外值得我們注意的是，當測試資料與訓練資料之分布有重疊現象時（如 Iris 資料庫中類別 2 與類別 3），所得之最低錯誤率相較有偏高狀況，這是因為重疊區域的資料原本就有不易分類的難度，但是加入本研究所提出之方法產生之人工離群值進行訓練後，透過增加測試資料的方式加強模型的訓練次數，確實達到降低錯誤率的目的。

就理論上而言，持續增加離群資料的資料數量應可使錯誤率持續降低，但人工離群值之產生除了可能落於目標資料集之外，亦有可能落於目標資料集之內，這將造成錯誤率之增加，故人工離群值之產生個數不會因其數量持續增加而使錯誤率無限降低，而是趨近一極限值（不為零）。

5.結論與未來研究方向

單類別分類法的優點就是可利用較少的已知樣本來進行訓練模型推估，這個特性對於一般較難取得實際資料的情況來說，確實佔有優勢。然而從另外一方面來說，這也造成訓練模型推估時資料不足的困擾，使得錯誤率無法有效降低。SVDD 脫胎於 SVM 方法，利用超球體的觀念取代超平面區隔，具有較少樣本資料需求的優點。在本研究中，我們透過增加人工離群值資料並且讓它們環繞在目標資料集外圍的方式，一方面讓合理的樣本資料增加，同時又讓這些資料環繞在目標資料集外圍，以便進行訓練模型推估時讓推導結果更加緊密地趨近目標資料集。利用 UCI 機器學習資料庫中標準資料集做為測試對象，可以驗證此方式確實能降低分類判斷時的錯誤率。

未來主要研究方向將是持續尋找可降低錯誤率之方法或模型，諸如：基因演算法或類神經網路的應用或結合都是可能引入的發展方向。

誌謝

本研究承蒙經濟部「「陸用智慧型載台關鍵技術研發」學研聯合研究計畫(計畫編號：98-EC-17-A-05-S2-0012)贊助，特此致謝。

參考文獻

[1] 王啟凱、董敏寶、許昭燦，"應用兩摺式交叉驗證法於影像特徵分類之研究"，***第 16 屆國防管理學術暨實務研討會論文集***，pp. IM82 – IM89，2008。

[2] 王啟凱，"一種新的用於 SVDD 中的產生離群值方法"，***中山科學研究院 40 周年院慶暨科專計畫成果發表研討會論文集***，pp. 272–283，2009。

[3] 赵英刚、刘仰光、何钦铭，"一种约减型支持向量域数据描述算法"，*Proceeding of the 25th Chinese Control Conference*，中国宁波，pp. 830–833，2006。

[4] Banerjee. A, Burlina. P, and Diehl. C, "A support vector method for anomaly detection in hyperspectral imagery, " *Geoscience and Remote Sensing*, Vol. 44, Issue 8, pp. 2282–2291, 2006.

[5] Blake, C., Keogh, E., Merz, C., UCI repository of Machine Learning Datasets. http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Irvine, Department of Information and Computer Sciences, 1998.

[6] Chi-Kai Wang, Yung Ting, and Yi-Hung Liu, "A Novel Approach of Feature Classification using Support Vector Data Description combined with Interpolation Method," *the 34th Annual Conference of the IEEE Industrial Electronics Society (IECON 2008)*, pp.1828-1832, 2008.

[7] Chi-Kai Wang, Yung Ting, Yi-Hung Liu, and Gunawan Hariyanto, "A Novel Approach to Generate Artificial Outliers for Support Vector Data Description," *IEEE International Symposium on Industrial Electronics (ISIE 2009)*, pp. 2202-2207, Jul. 2009.

[8] David M. J. Tax, and Robert P. W. Duin, "Support Vector Domain Description", *Pattern Recognition Letters*, No. 20(11-13), pp. 1191-1199, December 1999.

[9] David M. J. Tax, and Robert P. W. Duin, "Uniform Object Generation for Optimizing One-class Classifiers," *Journal of Machine Learning Research*, Vol. 2, No. 2, pp. 155-173, 2001.

[10] David M. J. Tax, and Robert P. W. Duin, "Support Vector Data description," *Machine Learning*, pp. 45-66, 2004.

[11] Defeng Wang, Daniel S. Yeung, Eric C. C., . "Structured One-Class Classification," *IEEE transaction on systems*, Vol. 36, No. 6, pp. 1283–1295, 2006.

[12] Ki-Young Lee, Dae-Won Kim, Lee, K.H, and Doheon Lee, "Density-Induced Support Vector Data Description," *IEEE Transactions on Neural Networks*, Vol. 18, pp. 284–289, 2007.

[13] Ling Zhuang, and Honghua Dai, "Parameter Optimization of Kernel-based One-class Classifier on Imbalance Learning", *Journal of Computers*, Vol. 1, No. 7, pp. 32-40, 2006.

[14] M.R. Moya, M.W. Koch, and L.D. Hostetler, "One-class classifier networks for target recognition allocations", *In Proceedings world congress on neural networks*, pp. 797-801, 1993.

[15] Rui-Rui Ji, Ding Liu, Min Wu, and Jing Liu, "The Application of SVDD in Gene Expression Data Clustering," *The 2nd*

International Conference of Bioinformatics and Biomedical Engineering (ICBBE 2008), pp. 371–374, 2008.

[16] Tao Xin-min, Chen Wan-Hai, Du Bao-Xiang, XuYoung, and Dong Han-Guang, "A Novel Approach to Intrusion Detection Based on Support Vector Data Description," *the 2nd IEEE Conference on Industrial Electronics and Applications (ICIEA 2007)*, pp. 802–807, 2007.

[17] V. Vapnik. , Statistical Learning Theory, New York:Wiley, 1998.

[18] Wei-min Huang, and Le-ping Shen, "Weighted Support Vector Regression Algorithm Based on Data Description," *Computing, Communication, Control, and Management (CCCM 2008)*, Vol. 1, pp. 250–254, 2008.Yi-Meng Lin, Xuan Wang, Wing W.Y. NG, Qun Chang, Daniel S. Yeung, and Xiao-Long Wang, "Sphere Classification for Ambiguous Data," in *Proceedings of 5th International Conference on Machine Learning and Cybernetics (ICMLC 2006)*, pp. 2571–2574, 2006.