結合馬氏距離和支援向量機之兩階段分類方法以解決類別不平衡問題

陳隆昇
朝陽科技大學資訊管理系
助理教授
e-mail：lschen@cyut.edu.tw

張毓珊
朝陽科技大學資訊管理系
研究生
e-mail：s9614617@cyut.edu.tw

摘要

在分類問題中，類別不平衡問題會使分類器在訓練時產生偏誤，導致其對少數類別有相當低的預測正確率。這個問題是因為不平衡的資料所造成，在此種型態資料中，一個類別的樣本數會遠超過其它類別的樣本數，使類別樣本的分佈呈現偏斜狀況，而相較於多數類別樣本，少數樣本通常是較有趣的類別。例如，醫學診斷資料的少見疾病、監測資料中的錯誤資料等。為了解決類別不均問題，本研究目的有(1)從決策樹，邏輯式迴歸，馬氏距離與支撐向量機，找出較穩健的分類器。(2)提出一種新方法，稱為「馬氏距離與支撐向量機之兩階段分類法」(MD-SVM)。實驗結果顯示，所提的MD-SVM，相較於傳統處理不平衡資料的方法如調整錯誤分類成本法、隨機減少多數法、分群抽樣法等，有較佳的績效表現。

關鍵詞：類別不平衡問題、分類、馬氏距離、資料探勘、不平衡資料。

Abstract

In classification problems, the class imbalance problem would cause a bias on the training of classifiers and result in a low predictive accuracy over the minority class examples. This problem is caused by imbalanced data in which almost all examples belong to one class and far fewer instances belong to others. Compared with the majority examples, the minority examples are usually more interesting class, such as rare diseases in medical diagnosis data, failures in inspection data, and so on. In order to tackle the class imbalance problem, this study aims to (1) find a robust classifier from different candidates including Decision Tree (DT), Logistic Regression (LR), Mahalanobis Distance (MD), and Support Vector Machines (SVM); (2) propose one novel method called MD-SVM (a new two-phase learning scheme). Experimental results indicated our proposed MD-SVM has better performance in identifying the minority class examples compared with traditional techniques such as under-sampling, cost adjusting, and cluster based sampling.

Keywords: **Class Imbalance Problem, Classification, Mahalanobis Distance, Data Mining, Imbalanced Data.**

1. 前言

近年來，在現實世界資料分類的應用中，通常會遇到類別不平衡問題(Class Imbalance Problems)。這個問題是因為不平衡的資料所造成，在此種型態資料中，一個類別的樣本數會遠超過其它類別的樣本數，使類別樣本的分佈呈現偏斜狀況(skewed class distribution) (Barandela et al, 2003)。當從不平衡資料萃取知識時，傳統的資料探勘方法，會使學習過程產生偏差，而減低判別少數類別的精確度，進而導致分類方法的可靠度降低(Su et al., 2006a & 2006b; Zhou and Liu, 2006; Xu and Chow, 2006; An and Wang, 2001)。許多分類技術的應用都會面臨到這個問題，例如，偵測流失客戶、罹患少數疾病的診斷(Su et al., 2006a)、檢驗資料中對不良品之預測(Su et al., 2006b; Liao, 2008)、保險業上詐領保費案件的偵測(Chae et al., 2001)等等。近年來這個問題逐漸受到學界的關注。分別在 2000 年的 AAAI(American Association for Artificial Intelligence)與 2003 年的 ICML(International Conference on Machine Learning)研討會裡中，被提出討論。另外 ACM 也以特別專刊來討論相關議題(June 2004)。由此可見，在機器學習領域裡，類別不平衡的問題，已經受到很多研究人員的關注(Xie and Qiu, 2007; Su and Hsiao, 2007; Chen et al., 2008;

Liao, 2008)。

造成類別不均問題的原因，主要是因為傳統的分類器，在一個訓練的樣本上，只為了尋找一個整體準確度最佳之績效，所以它們並不適合用來處理類別不平衡的工作(Batista et al., 2004; Chawla et al., 2004; Guo and Viktor, 2004; Japkowicz and Stephen, 2002)。

通常解決類別不平衡的問題，有許多種方法(Weiss, 2004)。其中最基本的方法就是 Kubat et al. (1997)學者所提出的「隨機減少多數法(Random Under-sampling, RU)」和 Lewis 和 Catlett (1994)學者所提出的「隨機增加少數法(Random Over-sampling, RO)」。RU 就是減少多數類別的樣本，而 RO 則是增加少數類別的樣本，它們的主要目的就是，修改訓練資料(Training data)二個類別的分配，讓兩個類別分佈平衡。但是，這兩個方法都有其缺點。例如：RU 方法可能在刪減資料的同時，也會把有用的資料刪除，所以有可能會失去有用的資訊，因而可能降低分類器的性能。而 RO 方法，會引入額外的雜訊，造成在建立分類器所需的時間變長，亦會導致過度訓練。研究指出，就整體的績效而言，RU 會比 RO 的方法來的好(Liao, 2008；張琦等人，2005)。

因為以上的二種基本的抽樣方法都有其缺點，所以又有學者提出了另外一種「分群抽樣法(Cluster based Sampling)」。這種方法就是先把訓練資料做分群的動作，然後在從每一群裡面挑選出具有代表性的樣本，來減少失去有用資訊的可能性(Altincay and Ergun, 2004)。另外 Weiss and Provost (2001)也提出一種「調整錯誤分類成本法(Adjust misclassification Cost)」，也可稱為「成本敏感度訓練」。這種方法就是因為誤判少數類別樣本的成本通常會比誤判多數類別樣本的成本高，如同誤判病人為無病與誤判無病的人為有病一樣，其成本是不相同的。所以藉著調高誤判少數類別之成本，可使分類器正確判別出少數類別。但這兩種方法，還是有不足之處。例如，在分群抽樣的方法中，我們必須決定適當的群數，且分完群之後再進行抽樣，還是有可能會失去有用的資訊。而調整錯誤分類成本的方法，它在判斷指定成本的訊息不易得到，換言之，在每案例中，它們必須透過試誤法才會決定錯誤分類之成本比例(張琦等人，2005)。

針對以上所提不足之處，本研究將提出一種新方法，稱之為「馬氏距離與支撐向量機之兩階段分類法」(Mahalanobis Distance-Support Vector Machines, MD-SVM)，用來解決分類預測模式偏向將未知類別的資料預測為多數類別的預測偏誤，以便提高預測模式對屬於少數類別的目標資料的預測能力。另外，也評估各種分類器，包含支援向量機 (Support Vector Machines, SVM)、決策樹(Decision Tree, DT)、邏輯斯迴歸(Logistic Regression, LR)、馬氏距離(Mahalanobis distance, MD)，這些方法在處理不平衡資料之穩健性(Robustness)。

實驗結果顯示，所提的 MD-SVM 方法，相較於傳統處理不平衡資料的方法如調整錯誤分類成本法、隨機減少多數法、分群抽樣法等，在偵測少數類別範例上有較佳的績效表現。

2. 文獻探討

一般而言，為了解決類別不平衡的問題，相關之研究主要可區分為兩類，分別為：

■ **演算法層級**

這種解決方法，主要是以提出新的演算法，或修改演算法架構來解決此類方法中，包含了單一類別學習法(one-class learning)，SVM 等等。其中單一類別學習法，僅用單一類別，(Raskutti & Kowalczyk, 2004)來代替二個類別(two-class)的學習。在這個部份，我們提出了幾個可能解決類別不平衡資料的學習演算法包括了，SVM、DT、MD、LR 並且，SVM 通常被用來處理類別不平衡問題(Wu & Chang, 2005)。

■ **資料操弄技術**

這種解決方法，主要是平衡兩個類別樣本。

在研究分類的觀點上，我們能發現一些資料適合重新抽樣技術(Weiss & Provost, 2003; Estabrooks et al., 2004)以及選擇適當的方法(Wilson & Martinez, 2000)去處理類別不平衡資料。它已經證實應用預先處理步驟是為了去平衡類別分配，也確定能解決類別不平衡問題(Batista et al., 2004)。而且這些技術主要的優勢，是他們不依賴使用分類器。因此在我們的研究裡，我們使用減少多數法、增加少數法、調整錯誤分類成本法、分群抽樣法，為目前最基本之解決類別不平衡之方法，以下為四種方法的介紹：

■ **減少多數法**

這個方法是以減少多數類別樣本來平衡

一個訓練樣本。具體來說，就是減少多數類別樣本，直到多數類別樣本的大小等於少數類別樣本。一些研究顯示減少多數類別法比增加少數類別法對於學習不平衡資料好(Drummond & Holte, 2003)。不過這種方法，可能會刪減一些潛在有用的訓練樣本，並且會降低分類器的表現(Batista et al., 2004)。

- **增加少數法**

這個方法是，增加少數類別樣本來平衡一個訓練樣本。具體來說，就是增加少數類別樣本直到少數類別樣本的大小等於多數類別樣本。增加少數法是解決類別不平衡問題的一種受歡迎的方法，並且研究已經顯示，增加少數法對於類別不平衡學習是有效的(Japkowicz & Stephen, 2002)，但是，研究也指出，因為它引進一些精確的樣本來培養訓練用資料，它通常會增加訓練時間並且可能導致過度訓練(Drummond & Holte, 2003)。

- **分群抽樣法**

在這個想去後面的直覺是建立一個過濾器，去過濾掉多數類別樣本，而不會失去少數類別樣本，讓我們減少資料不平衡的情況，使得學習工作更容易處理。在處理不平衡資料時，有一個非常重要的關鍵，就是少數樣本是缺少的，因此在處理不平衡資料時，應該嘗試不要刪減任何的少數類別樣本。要達到這個目標，應將多數類別樣本，分為許多群組。尤其，我們將找出多數遠離目標邊界的樣本(並且因此降低不平衡資料的情況)，以便我們能專注於區分更難的邊界樣本。因此，使群組的資料有更高的純度(purity)。

- **調整錯誤分類成本法**

Weiss and Provost (2001)提出一種「調整錯誤分類成本法」，也可稱為「成本敏感度訓練」。這種技術，則是為少數類別增加錯誤分類成本。傳統的分類效能指標認為，多數類別樣本和少數類別樣本，它們分類錯誤成本是相等的。因此導致分類結果傾向於多數類別樣本，而忽略少數類別樣本。所以它提出一種方法，強調二個類別要給予不同的錯誤分類成本，以減少整體分類所花費的成本(Zadrozny & Elkan, 2001)。也就是當類別之間數量呈現非對稱分配時，不採取增加或減少類別資料的方式，而是透過對於預測錯誤成本的調整，提升分類預測的效能，此種方法最大的缺點在於使用者須自行定義不同目標類別的錯誤分類方本，當使用者對於該領域的知識不足時，並無法正確的定義出適當的錯誤成本，而且若類別之間已呈對稱性分配時，成本差異提升並無法有效提升預測能力(Witten & Frank, 2002)。

3.研究方法

本研究，提出了一個以馬氏距離(Mahalanobis Distance, MD)為基礎的二階段分類法，稱之為 MD-SVM(Mahalanobis Distance-Support Vector Machines)，以提升分類器對不平衡資料之績效。這個方法大致上可分為二個階段，首先是先建立一個以 MD 為基礎的過濾器，初步地刪去那些有十足把握一定是屬於多數類別的樣本，改善類別失衡的狀況。第二階段，再以 SVM 來做為主要的學習演算法，來攫取知識。

3.1 MD-SVM 方法與流程

本節將介紹，我們所提出 MDS-SVM 法，用來改善類別不平衡的問題。我們主要的想法是，利用簡單的 MD 分類器，先行濾掉一些可以確定為多數類別的樣本，再以 SVM 進行學習。另在過濾樣本階段中，我們希望將，少數類別樣本被分配到多數類別樣本的錯誤機率(β)降至最小，而不是考慮多數類別樣本被分配到少數類別樣本的錯誤機率(α)。

這方法主要分為二個階段，第一階段是「過濾多數樣本(Screening)」，第二階段是「SVM 學習(Learning)」階段。這方法總共有五個步驟，流程如圖 1，詳細步驟介紹如下：

第一階段:過濾多數類樣本

步驟一：建立馬氏空間

利用正常資料(多數類別樣本)，計算「馬氏距離」，建立一個馬氏空間。

步驟二：決定一個門檻值(threshold)

這個步驟主要是要設一個門檻值來區分屬性的類別。在建立門檻值時，我們希望將錯分少數類別的機率(β)到最小。

步驟三：改善類別不平衡情況

利用這個分界線以刪除十足把握為多數類別樣本，改善類別不平衡的問題。最後把其餘的多數類別樣本和少數數別樣本結合，成為訓練資料(training set)(方法 1&2)。在此步驟中，我們也嘗試了將刪除資料回補之策略(方法 3&4)。

第二階段:學習

步驟四：建立分類器

我們嘗試去找一個，最不受不平衡資料影響的學習演算法(SVM、DT、LR、MD)，來建立一個分類器。

步驟五：評估效能

找到最好的學習演算法之後，利用階段一所得到修正後的訓練資料，來評估這個方法(MD-SVM)的效能。最後，在使用幾個評估不平衡資料的標準(PA、NA、OA、GM、F1)來做為方法的這些方法的評估指標，並其他傳統的方法如，調整錯誤分類成本法、隨機減少多數法、分群抽樣法做比較。

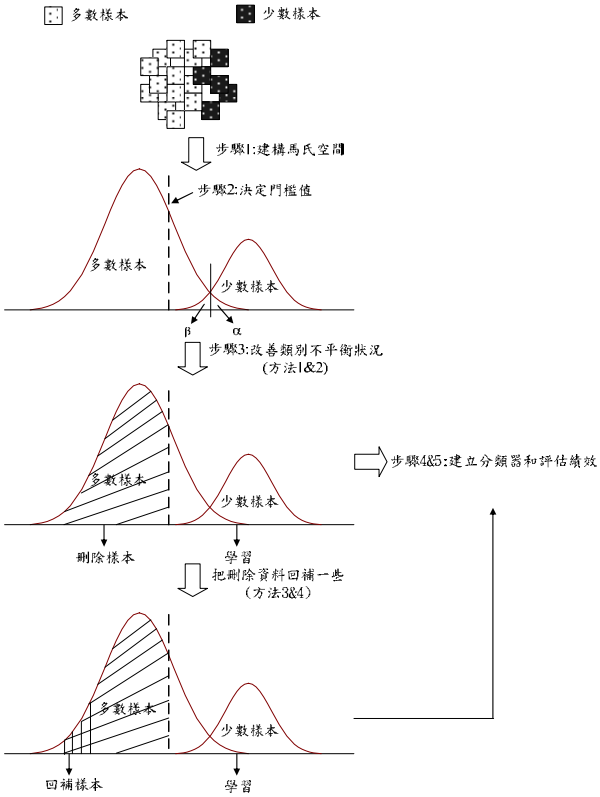

圖 1 MD-SVM 方法流程圖

3.2 簡單說明例

接下來本研究以一個乳癌例子來詳細的說明本方法，資料筆數共 306，屬性數目為 4(包含 1 個類別屬性)，欄位說明如下：

(1)正在手術的患者的年齡。

(2)患者手術的年度。

(3)腋窩腫瘤檢驗陽性程度。

(4)患者存活狀況(類別屬性)：

- "1"代表患者存活(Survived)超過 5 年或更長。
- "2"代表患者死(Died)於 5 年內。

第一階段:過濾多數類樣本

步驟一：利用正常資料(多數類樣本)，計算「馬氏距離」，建立一個馬氏空間。

步驟1.1: 選取 Survived 樣本的所有特微值。

步驟1.2: 計算每個特性變數的平均值與標準差，將每個數據做標準化。

步驟1.3: 計算每個特性變數間的相關系數及其反矩陣 C^{-1}

步驟1.4: 計算馬氏距離。(馬氏距離公式，如 3.3 節)。

步驟二：決定一個門檻值(threshold)，來決定一個決策分界線。

這個步驟主要是要設一個門檻值來區分 Survived 與 Died 的病患。並且要降低 β 發生的機率。本研究利用試誤法的方式來建立一個可以接受的門檻值，本文所採取建立門檻值之方法，共有兩種，分述如下：

方法 1：

利用多數類別(Survived)的樣本，來建立門檻值，如圖 2。然後再利用試誤法決定一個門檻值(例如：取 Survived 樣本前 50%)，使用這個決策門檻值(0.847293)來判斷是否做刪除。

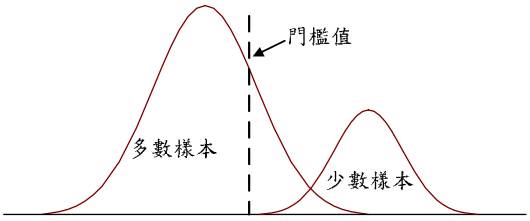

圖 2 方法 1 門檻值示意圖

方法 2：

利用少數類別樣本來建立決策門檻值，如圖 3 所示。然後再利用試誤法決定一個門檻值(例如：取 Died 樣本前 20%)，使用這個決策門檻值(1.851107)來判斷是否做刪除。

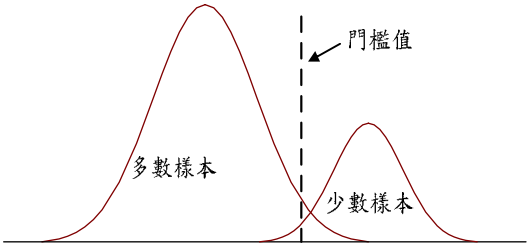

圖 3 方法 2 門檻值示意圖

步驟三：改善類別不平衡情況

接續以上的方法 1 與方法 2，刪除決策門檻值以內的多數樣本，改善類別不平衡的問

題。最後把其餘的多數類別樣本和少數類別樣本結合，成為訓練資料(training set)。

方法 1：

大於這個值設為 Survived，小於這個值設為 Died，因此我們把 Survived 刪除，如圖 3-4 所示，把 Died 保留並與少數類別樣本(Died)組成一個訓練樣本。

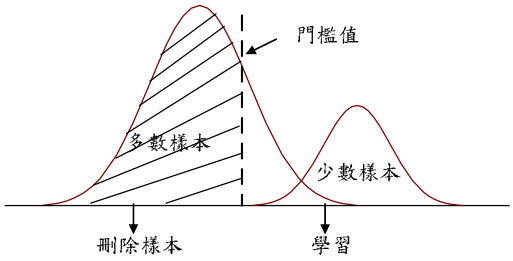

圖 4 方法 1 示意圖

方法 2：

大於這個值設為 Died，小於這個值設為 Survived，因此我們把 Survived 刪除，如圖 3-5 把 Died 保留並與少數類別樣本(Died)組成一個訓練樣本。

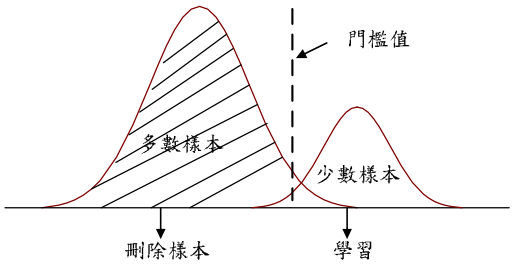

圖 5 方法 2 示意圖

另外，我們分別以方法 1 與方法 2 為基礎，提出刪除資料，回補策略稱之為方法 3&4。其中方法 3 與方法 4，是利用方法 1 與方法 2 的資料做回補，主要是因為在刪減資料的同時，有可能會刪到重要的資料，而馬氏距離考慮了變數的相關性，所以它們的變數是具有相關性的，因此在這裡要把刪除的資料回補一點點。

方法 3：

把方法 1 刪除的資料做回補。因為馬氏距離考慮了變數的相關性，所以它們的變數是具有相關性的，因此在這裡要把刪除的資料回補一點點。首先，因為我們剛剛是取 Survived 樣本前的 50%，因此我們刪減了大概 50%的 Survived 樣本，接下來我們要從刪除的這 50%裡，拿回 5%的資料，使用這個決策門檻值(0.584986)來判斷是否做回補。大於這個值設為 Died，小於這個值設為 Survived，因此我們把 Died 刪除，把 Survived 回補並與少數類別樣本(Died)組成一個訓練樣本。如圖 6 所示。

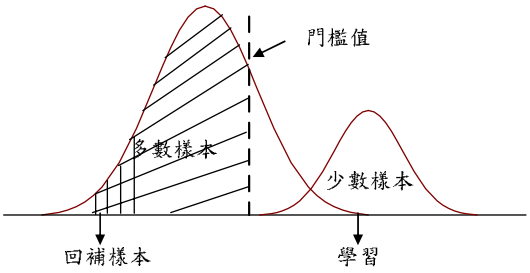

圖 6 方法 3 示意圖

方法 4：

把方法 2 刪除的資料做回補。因為馬氏距離考慮了變數的相關性，所以它們的變數是具有相關性的，因此在這裡要把刪除的資料回補一點點。首先，因為我們剛剛是利用 Died 樣本來判斷 Survived 樣本，因此我們刪減了大概 20%的 Survived 樣本，接下來我們要從刪除的這 20%裡，拿回 5%的資料，使用這個決策門檻值(0.793443)來判斷是否做回補。大於這個值設為 Died，小於這個值設為 Survived，因此我們把 Died 刪除，把 Survived 回補並與少數類別樣本(Died)組成一個訓練樣本。如圖 7 所示。

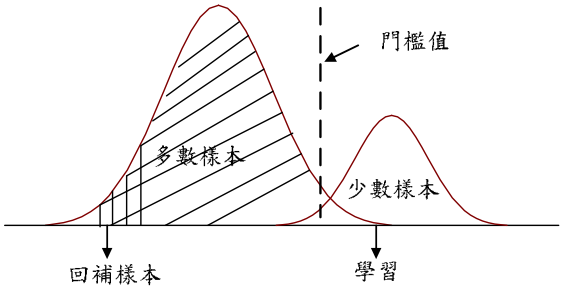

圖 7 方法 4 示意圖

第二階段(學習)

步驟五：建立分類器。

我們嘗試去找一個，最不受不平衡資料影響的學習演算法(SVM、DT、LR、MD)，並於所提方法(MD-SVM)所建立出來的訓練樣本，來建立一個分類器。

步驟六：評估效能。

找到最好的學習演算法之後，利用階段一所得到修正後的訓練資料，來評估這個方法(MD-SVM)的效能。最後，在使用幾個評估不平衡資料的標準(PA、NA、OA、GM、F1)來做為方法的這些方法的評估指標，並其他傳統的方法如，調整錯誤分類成本法、隨機減少多數

法、分群抽樣法做比較。

3.3 馬氏距離

馬氏距離是一個有名的學習方法，它是由印度統計學家馬哈拉諾比斯(P. C. Mahalanobis)提出的，表示數據的協方差距離。它是一種有效的計算兩個未知樣本集的相似度的方法。與歐式距離不同的是它考慮到各種特性之間的關聯(例如：一條關於身高的訊息會帶來一條關於體重的訊息，因為兩者是有關聯的)。當它在學習一個訓練資料時，馬氏距離不考慮二個類別的大小，所以這是為什麼我們選擇馬氏距離，來做為不平衡資料的最佳方法的最主要原因。下列為馬氏距離簡短的介紹：

馬氏距離應用的第一步驟是，先定義出正常樣本(Normal sample)，並從中選擇做為正常群的參考，來建立一個馬氏空間(Mahalanobis space，MS)。接下來建立一個測量尺度，我們需要收集一組正常的資料與標準化的值，來計馬氏距離。

以下為馬氏距離的公式：

$$MD_j = D_j^2 = Z_{ij} C^{-1} Z_{ij}^T \quad (1)$$

其中 Zi=標準化值之標準向量，

Xi=(i=1,2⋯k)

$$Z_{ij} = (X_{ij} - \overline{X}_i)/S_i, i = 1,2,\ldots,k, j = 1,2,\ldots,n$$

$$\overline{X}_i = \frac{\sum_{j=1}^{n} X_{ij}}{n}$$

其中 Xij=第 j 個樣本的第 i 個特微變數值；

$$S_i = \sqrt{\frac{\sum_{j=1}^{n} (X_{ij} - \overline{X}_i)^2}{n-1}}$$

其中 Si 為第 i 個特微變數之標準差

C-1=相關反矩陣

k=特徵變數的各數；n=觀測變數的各數

T=轉置向量

之後，我們建立一個馬氏空間，然後我們要決定一個門檻值來分類正常與異常的資料，而這個門檻值，正好跟 SVM 的超平面一樣，並且也能被當成一個分類器。如圖 8 為一個馬氏距離的例子。

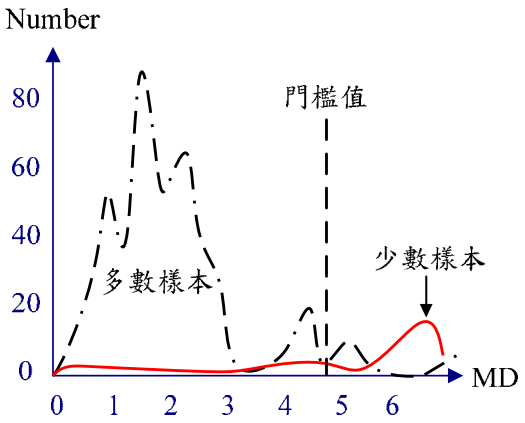

圖 8 馬氏距離的例子

3.4 學習演算法

除了在 3.3 介紹的 MD 演算法之外，在本研究另外還有使用到 SVM、DT、LR 這三個演算法，因此在這裡分別做介紹：

3.3.1 支援向量機

支援向量機(Support Vector Machines；SVM)是由Vapnik在1995年和AT＆T實驗室團隊所提出的一個新方法，其主要的理論是來自統計學習理論中結構化風險最小原理(Structural Risk Minimization, SRM)(H.T. Lin & C.J. Lin, 2003)。支援向量機最主要是利用區分超平面(Separating Hyperplane)來分隔兩個或多個不同類別(Class)的資料，處理資料探勘中分類(Classification)的問題。它不僅被應用到分類，也被廣泛的應用到特徵選取(feature selection)和迴歸分析(regression)，而且許多研究報告也顯示，SVM在處理類別不平衡的資料，也有不錯的效能。所以我們嘗試去測試它的優點與限制。然而在現實生活中，資料並非容易分類，當遇到非線性分割的資料時，我們就需要靠核心函數，來幫助我們做分類。簡單來說，我將一個訓練資料的二個類別，輸入到一個低維度的空間裡，然後透過核心函數(Kernel functions)轉換到一個高維度的空間，又稱為特徵空間(feature space)，之後找到一個最小的邊界超平面(Maximal margin hyperplane)。最後這個超平面(Hyperplane)也可以把他當成一個分類器。圖9顯示SVM的基本概念。

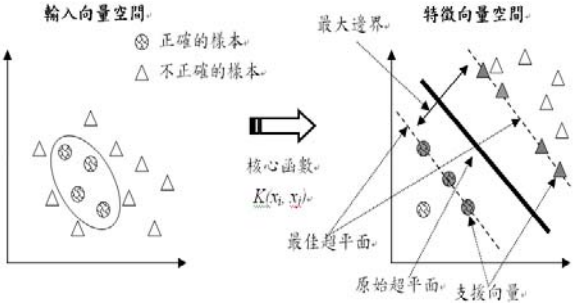

圖 9 SVM 基本概念

(資料來源：修改自郭琇靜，2007)

接下來讓我們簡單的介紹 SVM，例如以二維的例子來說，如上圖，我們希望能找出一條線能夠將圓點和方形點分開，而且我們還希望這條線距離這兩個集合的邊界(margin)越大越好，這樣我們才能夠很明確的分辨這個點是屬於那個集合。以下用「數學」的方法來描述這個問題：當我們給一對 (x_i, y_i) 訓練資料，i=1,……,m 當 $x_i \in R^n$ 和 $y \in \{1,-1\}^m$，x 是屬性和 y 是類別，需要利用 SVM 來解決最佳化的問題。因此它的目標函數為：

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i \qquad (2)$$

Subject to

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \qquad (3)$$

$$\xi_i \geq 0$$

這裡的訓練向量 x_i 被函數 ϕ 轉換到一個高維度的空間。然後SVM在這高維度空間裡，找到一條線性分隔的超平面與最大邊界。C > 0是損失參數(penalty parameter)，另外，核心函數是 $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$。在本研究裡，我們使用了，林智仁老師所開發的 LIBSVM(版本為 2.8)，libsvm 工具可在這個網址取得 http://www.csie.ntu.edu.tw/~cjlin/libsvm/。我們使用這個演算法標準的參數，在這個工具裡，參數設定的方式，它可以自動找出最佳的參數，另外，根據 Hus et al.建議，選擇一個核心函數時，應優先考慮放射型(radial basis function，RBF)，因為它具有以下優點：①它能分類非線性與高維度的資料②它只需調整二個參數，c 和 g，使得操作上變的簡單，而且也能達到高的預測能力③輸入資料限在 0-1 之間，減少運算時間，因此，我們使用的核心函數為 RBF。

3.3.2 決策樹

決策樹是功能強大且相當受歡迎的分類和預測工具，並且能成功的應用到許多領域。它是以樹狀資料結構為基礎的分類分析方法，主要是由，根節點、子節點、葉節點所構成，每一個結點代表不同的特徵(feature)，樹枝為特徵的值，而樹葉則是指不同的分類類別(class label)。這種方法是先找一個最佳的特徵作為根節點，所有的資料以此根節點為判斷根據，進行分類，分類在每一個分支的資料再選出最佳的特徵作為根節點，再進行分類，形成一棵子樹，如此的過程一直重複，直到在這一個分支內的所有資料都屬於同一個類別，推導過程才算結束，這樣就會形成一棵決策樹，如圖 10，預估「是否會使用電腦」的決策樹。

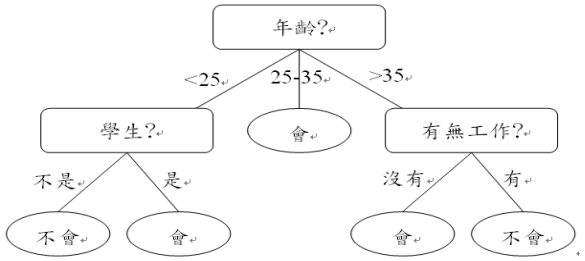

圖 10 決策樹

決策樹分析模式主要演算法包括 ID3 (Interactive Dichotomimizer 3)與 C4.5(緣起於人工智慧)、CART(Classification And Regression Tree)分類樹及 CHAID(Chi-squared Automatic Interaction Detection)卡方自動互動偵測(緣起於統計)。而在決策樹的研究領域裡主要的焦點是在 ID3 與它的延伸 C4.5 演算法(Quinlan, 1986; Quinlan, 1993)。然而 C4.5 是建構決策樹，最受歡迎的演算法(凌士雄，2002)。

決策樹歸納法的基本演算步驟如下：

步驟1: 將訓練樣本的原始資料放入決策樹的樹根。

步驟2: 將原始資料分成兩組，一部份為訓練組資料，另一部份為測試組資料。

步驟3: 使用訓練資料來建立決策樹，而在每一個內部節點，則依據資訊理論(Information Theory)來評估選擇哪個屬性繼續做分支的依據。

步驟4: 使用訓練資料來建立決策樹，而在每一個內部節點，則依據資訊理論(Information Theory)來評估選擇哪個屬性繼續做分支的依據。

步驟5: 將步驟 1-步驟 4 不斷重復進行，直到所有的新內部節點都是樹葉節點為止。

步驟6: 重復進行之，它的停止條件為：

① 在這群資料中，每一筆資料都已經被分在同一類別下面。

② 在這群資料中，已經沒有辦法再找到新的屬性來進行節點分割。

③ 在這群資料中，已經沒有資料可以處理了。

根據相關的文獻，決策樹的優點有(Berry and Linoff, 1997)：①能產生讓人易懂的規則②不需要復雜的運算③容易計算分類時間④能夠處理連續和分類之變數⑤不管在預測和分類上，它都能清楚的提出一個解釋。因此，我們把決策樹做為我們學習演算法之一。

3.3.3 邏輯式迴歸

迴歸分析是描述一個應變數與一個或多個預測變數之間的關係式，它是資料分析最重要的工具。它在統計分析應用也已經有很多年，但是從1967年以後，羅吉斯迴歸才變普遍，現在對於二元的離散資料尤其是在醫學健康方面使用的很廣泛。所以當我們探討結果的應變數是離散型，其分類只有二類或少數幾類時，例如「成功」或「失敗」，以邏輯斯迴歸作分析，在很多領域已變成是最標準的分析方法。因此邏輯斯迴歸就是針對二元因變數，即是1或0。在 Logistic Curve 中有一個臨界遞增的 S 型函數，如圖11，適用於分析機率模型，而根據分類性變數，產生輸出變數，其值可為0或1。在統計學上不同於線性判別分析，邏輯斯迴歸不需要滿足常態分配假設(Press and Wilson, 1978; Desai et al., 1996)，許多學者認為邏輯斯迴歸的優點，主要能處理依變項有兩個類別的名目變項，用以預測事件發生的勝算比(Odds Ratio) ，它可解決了傳統線性迴歸模式中，不能處理依變項是兩個類別的名目變項的缺點。因此，在傳統的統計分類技術，在這個研究，我們選擇使用邏輯斯迴歸來做為我們的學習演算法之一。接下來為邏輯斯迴歸的基本模式：

$$p = \frac{e^{\alpha+\beta_i X_i}}{1+e^{\alpha+\beta_i X_i}} \tag{4}$$

p：機率

X_i：自變數

α & β：迴歸係數

之後把公式2做轉換，定義如下：

$$e^{\alpha+\beta_i X_i} = \frac{p}{1-p} \tag{5}$$

當時，邏輯斯迴歸模示的公式表示為

$$\ln(\frac{p}{1-p}) = \alpha + \beta_i X_i \tag{6}$$

其中 $\frac{p}{1-p}$ 稱為勝算比(Odds Ratio)，P為事件發生的或機率，勝算比的定義是一件事情會發生的或然率除以不會發生的或然率，若以或然率P(Y)=0.5為判別值(Cut Value)，將0.5以上判別為1，0.5以下判別為0，則利用邏輯迴歸，便可進行類別預測。

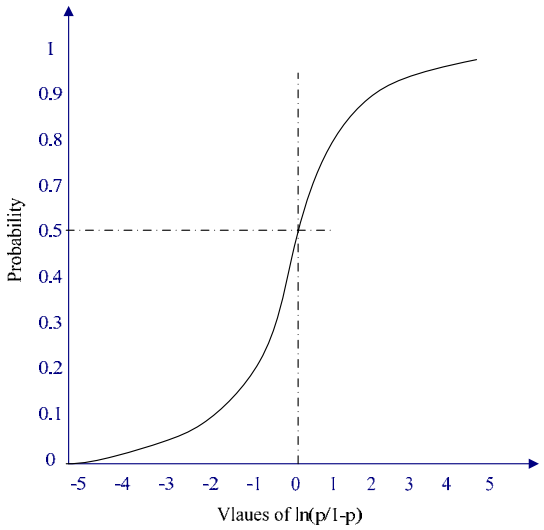

圖 11 邏輯斯曲線

4. 實驗

本研究將從 UCI 機器學習資料庫裡，收集十個具有不平衡資料的集合，來評估所提方法之有效性，並與目前用來處理不平衡資料的抽樣方法做比較。

4.1 實驗資料與前置處理

本研究將從 UCI 機器學習資料庫(UCI Machine Learning Repository，http://www.ics.uci.edu/~mlearn/MLSummary.html)中挑選出包括 Diabetes、Car Evaluation、Pima Indians Diabetes、Contraceptive Method Choice、Spambase、Ionosphere、Teaching Assistion Evaluation，等七個具有高度類別不平衡的資料集合，總共七份資料集合作為本研究的實驗資料集合。它們的特徵如下表：

表 1 本研究使用的 UCI 資料集基本資料

資料集合	樣本數	屬性數目	屬性型態	類別分配
Diabetes	759	9	整數	Healthy: 66% Diabetic*: 34%
Car Evaluation	1785	6	文字	Unacc: 70.023 % Acc: 22.222 % Good: 3.993 % v-good*: 3.762 %
Pima Indians Diabetes (PIMA)	768	8	整數	Healthy: 65% Diabetic*: 35%
Contraceptive Method Choice (CMC)	1473	9	整數	No-use*: 43% Long-term: 22% Short-term: 35%
Spambase	4601	57	整數	Spam*: 39.4% non-spam: 60.6%
Ionosphere	351	34	整數	Good: 69% Bad*: 31%
Teaching Assistion Evaluation (TAE)	151	5	整數	Low*: 32.5% Med: 33.1% High: 34.4%

注意：「*」表示少數類別

本研究，只針對兩個類別的資料集合進行分類效能的評估，所以對蒐集到的資料集合，若資料類別的種類超過二個以上，將會把資料中的少數類別合併成一個類別，來產生資料類別不平衡的特性，並且使用 holdout 方法來評估分類器的準確性，這個方法是利用隨機抽樣方式來切割所給予的學習資料，主要將學習資料分成兩大部份，training set(90%)和 testing set(10%)，其中將 90%的資料作為訓練集合的物件用以建立模型，然後剩下 10%的資料作為測試集合的案例來評估這個模型的分類準確性。

4.2 評估指標

當要學習一個不平衡資料時，我們因該討論一下，在分類預測上的一些評估指標對於處理不平衡資料的有效性，因此，通常在評估一個分類器的好壞，都會利用 Confusion matrix 來評估分類器的表現，如表 2。

表 2 Confusion matrix

Actual \ Predicted	**Positive**	**Negative**
Positive	TP	FN
Negative	FP	TN

在 Confusion matrix 表中的 TP、FP、FN、TN 這四種指標，分別定義如下：

(1) TP 指目標樣本被分類正確的樣本數目。

(2) FP 指實際目標樣本為異常，但預測分類結果是正常的分類錯誤樣本數目。

(3) FN 指實際目標樣本為正常，但預測分類結果是異常的分類錯誤樣本數目。

(4) TN 指非目標樣本被分類正確的樣本數目。

因此，TP 與 TN 為被正確分類的資料個數，而 FN 與 TN 則是被分類錯誤的資料個數。

在傳統中，要評估一個分類器的好壞，通常都是使用 Overall Accuracy(Provost and Fawcett, 1997)，它是用來計算分類系統對整體資料，分類正確的比率。公式如下：

$$Overall\ Accuracy = \frac{TP + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{7}$$

但是研究顯示，當資料是不平衡時，已不能只用 Overall Accuracy 來當評估指標(Su et al., 2006 a&b; Chen et al., 2008)。另外，研究也顯示，評估指標對於考慮不同的分類錯誤也相當重要。因此在這個研究，PA 和 NA 分別表示，正確判斷多數類別樣本和少數類別樣本的能力。公式如下：

$$PA = \frac{TP}{TP + FN} \tag{8}$$

$$NA = \frac{TN}{FP + TN} \tag{9}$$

在處理不平衡資料時，有學者指出 G-mean、F-measure 是適當的評估指標。因為 Kubat 和 Matwin(1997)提議使用二個準確度的幾何平均數的公式，當 PA 和 NA 準確率都是高的，那麼幾何平均數就是高的。公式如下：

$$G - mean = \sqrt{PA * NA} \tag{10}$$

另一個 F-Measure 是參數化的，可以調整 Recall 與 precision 二者的相對權重（F1 是按二者權重值相等計算、F2 是以 Precision 乘以 2 計算），這個指標是為了，若我們想刻意提高 Precision，則勢必會導致 Recall 下降，反之亦然，因此才會有 F-Measure 的評估方式，希望同時兼顧兩種指標。公式如下：

$$\mathrm{Pr}ecision = \frac{TP}{TP + FN} \tag{11}$$

$$\mathrm{Re}all = \frac{TP}{TP + FP} \tag{12}$$

$$F1 = \frac{2 * \mathrm{Pr}ecision * \mathrm{Re}call}{(\mathrm{Pr}ecision + \mathrm{Re}call)} \tag{13}$$

另外也有研究顯示，使用 Overall Accuracy、G-mean、F1，這三個評估指標，來當不平衡資料的評估標準是適當的(Kubat and Matwin, 1997; Maloof, 2003; Chawla et al.,

2003)。因此本研究共使用了，五個評估指標，分別為：Positive Accuracy、Negative Accuracy、Overall Accuracy、G-mean、F1。

4.3 學習演算法穩健性之評估

表 3 為 SVM、DT、MD、LR 四個學習演算法的分類結果比較，PA 與 NA 則表示，判斷多數樣本與少數樣本的能力，然而 MD 有高的 PA(96.69%)，但它也有低的 NA(3.10%)，這個結果表示，MD 這個方法明顯受到不平衡資料的影響，然後是 LR，之後是 DT，然而最不受不平衡資料影響的是 SVM，因為他有高的 PA(88.80%)與高的 NA(80.03%)，這表示 SVM 這個方法，最不受不平衡資料的影響。在圖 12 也看的出 SVM 無論在 OA 或 G-mean 及 F1 這三個評估標準，和其他三個學習演算法比較，它都顯示為最好的表現，因此，我們決定使用 SVM 來做為我們提出方法(MD-SVM)以及其他目前處理類別不平衡資料技術的學習演算法。

表 3 四個學習演算法的分類結果

方法 / 指標	DT		SVM		MD		LR	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
PA	84.74	11.17	88.80	9.09	96.69	2.57	82.26	12.25
NA	75.49	17.51	80.03	16.40	3.10	6.51	67.41	25.68
OA	80.94	13.34	84.91	8.95	62.87	7.78	77.87	13.91
G-mean	79.80	14.07	83.76	10.21	8.73	15.98	72.70	18.67
F1	84.62	11.27	88.05	7.07	77.10	5.14	79.72	13.50

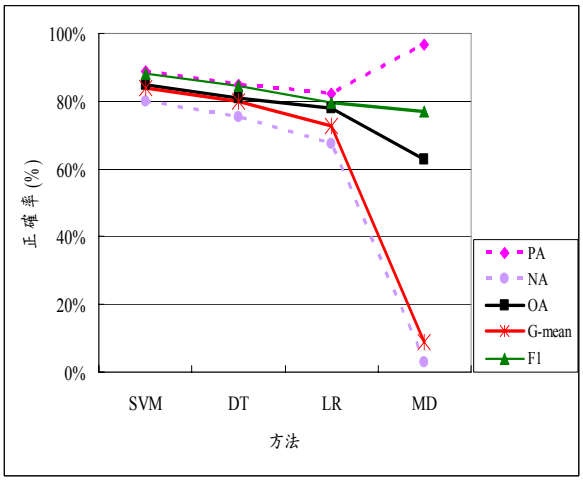

圖 12 四個學習演算法的分類結果比較

4.4 MD-SVM 四個方法的比較

表 4 為本研究所提之方法的分類結果比較，總共分為四種 MD-SVM 的方法，如果考慮 OA，MD-SVM 方法 1 有比較好的表現(86.24%)，和低的標準差(8.37%)，與 MD-SVM 方法 2(Mean/SD: 85.47%/8.11%)、MD-SVM 方法 3 (Mean/SD: 85.99%/8.59%)和 MD-SVM 方法 4(Mean/SD: 84.93%/9.39%)，由這個結果來看，MD-SVM 分類的能力，比其他方法更穩定。如果也把 G-mean 與 F1 考慮進去的結果來看，雖然 MD-SVM 方法 1，它的 F1 比 MD-SVM 方法 3 好一點點，但是 MD-SVM 方法 3，它 G-mean 的結果是比其他三個方法好很多的，因此，MD-SVM 方法 3 這個方法對於處理不平衡資料的能力是值得探討的。

表 4 四種 MD-SVM 方法的分類結果

方法 / 指標	MD-SVM 方法 1		MD-SVM 方法 2		MD-SVM 方法 3		MD-SVM 方法 4	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
PA	91.60	7.29	90.76	5.41	91.80	7.41	89.16	7.35
NA	78.40	14.14	77.50	14.23	75.61	17.40	77.63	17.51
OA	86.24	8.37	85.47	8.11	85.99	8.59	84.93	9.39
G-mean	84.45	9.31	83.65	9.42	86.53	7.46	82.81	11.59
F1	89.30	6.64	88.73	6.37	89.10	6.94	88.18	7.25

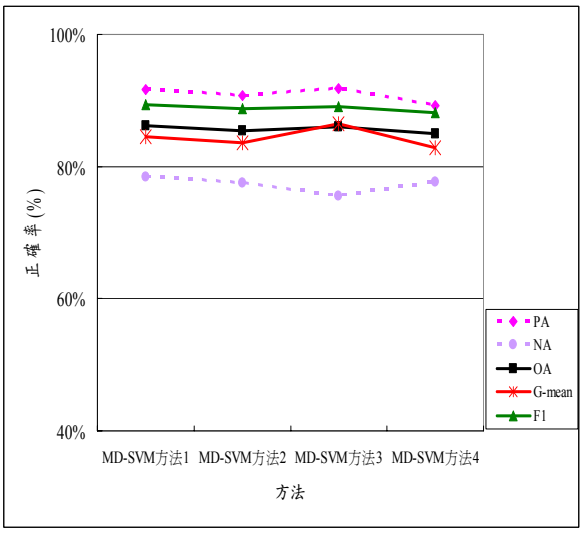

圖 13 四種 MD-SVM 方法的分類結果比較

4.5 目前處理不平衡資料技術的比較

表 5 為 MD-SVM 方法 3 與其他技術的比較結果，如果考慮 OA，MD-SVM 方法 3 是有比較好的表現(85.99)，和低的標準差(8.59)，與

沙少多數法 (Mean/SD: 80.36%/12.99%)、調成錯誤分類成本法(Mean/SD: 74.81%/11.91%)和分群抽樣法(Mean/SD: 68.89%/29.06%)，由這個結果來看，MD-SVM 方法 3 分類的能力，比其他方法更穩定。如果也把 G-mean 與 F1 考慮進去的結果來看，MD-SVM 方法 3 還是比其他的方法好，因此，MD-SVM 方法 3 這個方法對於處理不平衡資料的能力是非常有義意的。

表 5 MD-SVM 方法 3 與其他技術比較的分類結果

方法 / 指標	MD-SVM 方法 3		減少多數法		調整錯誤分類成本法		分群抽樣法	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
PA	91.80	7.41	86.41	15.61	71.43	15.01	68.10	36.71
NA	75.61	17.40	69.51	33.71	81.41	12.91	72.94	33.47
OA	85.99	8.59	80.36	12.99	74.81	11.91	68.89	29.06
G-mean	86.53	7.46	70.41	33.14	75.94	11.78	62.15	39.37
F1	89.10	6.94	84.46	10.65	77.40	12.29	68.21	35.61

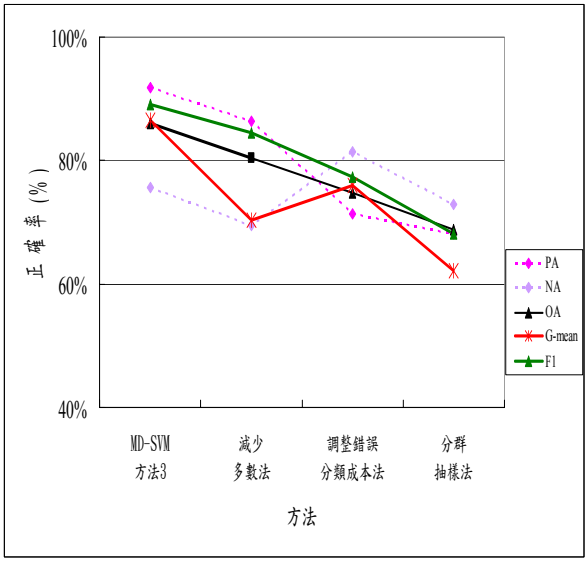

圖 14 MD-SVM 方法 3 與其他技術比較的分類結果比較

5. 結論

在本研究裡，我們提出一個新的方法稱為 MD-SVM，來處理類別不平衡之問題。實驗結果顯示，所提的 MD-SVM，相較於傳統處理不平衡資料的方法如調整錯誤分類成本法、隨機減少多數法、分群抽樣法等，在偵測少數類別範例上有較佳的績效表現。然而，其它傳統技術確實也能改善類別不平衡的情況，但是，實驗結果顯示，它們的表現是不穩定的。因此，這根本的原因是，它們缺少一個有系統刪減資料的方法。例如：分群抽樣法，它很難去決定適當的群數，另外，調整錯誤分類成本法，它在判斷指定成本的訊息不易得到，所以它需要一個標準，去決定分類錯誤成本。這些傳統技術的缺點，也許能做為，我們未來的研究方向。

6. 致謝

本研究受到國科會計畫(契約編號 NSC 96-2416-H-324 -003 -MY2)部分贊助，作者在此表達感謝之意。

參考文獻

[1] 張琦、吳斌、王柏，“非平衡數據訓練方法概述”，***計算機科學***，第三二卷，第十期，pp. 181-186，2005。

[2] 凌士雄，“非對稱性分類分析解決策略之效能比較”，***碩士論文***，國立中山大學資訊管理學系，2004。

[3] 郭琇靜，“應用支援向量機與製程統計特徵於線上偵測製程異常之研究”，***碩士論文***，國立虎尾科技大學工業工程與管理研究所，2007。

[4] Altincay, H. and Ergun, C., “Clustering based undersampling for improving speaker verification decisions using AdaBoost,” ***Lecture Notes in Computer Science***, Vol. 3138, pp. 698- 706, 2004.

[5] An, A. and Wang, Y. (2001), “Comparisons of classification methods for screening potential compounds,” *Proceedings of the 2001 IEEE International Conference on Data Mining*, pp. 11-18, 2001.

[6] B. Raskutti, A. Kowalczyk, “Extreme rebalancing for SVMs: a case study,” ***SIGKDD Explorations***, Vol. 6 No. 1, pp. 60–69.

[7] B. Zadrozny and C. Elkan, “Learning and Making Decisions When Costs and Probilities are Both Unknown,” ***In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining***, pp. 204-213, 2001.

[8] Barandela, R., Sanchez, J. S., Garcia, V. and Rangel, E., “Strategies for learning in class imbalance problems,” ***Pattern Recognition***, Vol. 36, No. 3, pp. 849-851, 2003.

[9] Batista, G., Prati, R.C., and Monard, M.C., “A study of the behavior of several methods

for balancing machine learning training data," ***SIGKDD Explorations***, Vol. 6, No. 1, pp. 20-29, 2004.

[10] Berry, M. J. A. and Linoff, G., "Data Mining Techniques: For Marketing Sale and Customer Support," John Wiley & Sons, Inc., 2007.

[11] C. Drummond, R.C. Holte, "C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," ***Workshop on Learning from Imbalanced Datasets***, NRC 47381., 2003.

[12] Chawla, N. V., Japkowicz, N. and Kolcz, A., "Editorial: special issue on learning from imbalanced data sets," ***SIGKDD Explorations***, Vol. 6, No. 1, pp. 1-6, 2004.

[13] Chen, M.-C., Chen, L.-S., C.-C., Hsu, and Zeng, W.-R., "An information granulation based data mining approach for classifying imbalanced data," ***Information Sciences***, Vol. 178, No. 16, pp. 3214-3227, 2008.

[14] D.R. Wilson, T.R. Martinez, "Reduction techniques for instance-based learning algorithms," ***Mach. Learning***, Vol. 38, No.3, pp. 257-286, 2000.

[15] Desai, V.S., Crook, J. N., and Overstreet, G. A., "A comparison of neural networks and linear scoring models in the credit union environment" ***European Journal of Operation Research***, Vol. 95, pp. 24-37, 1996.

[16] Estabrooks, T. Jo, N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," ***Comput. Intelligence***, Vol. 20, No. 1, pp. 18-36, 2004.

[17] G. Weiss, F. Provost, "Learning when training data are costly: the effect of class distribution on tree induction," ***Journal of Artificial Intelligence Research***, No. 19, pp. 315-354, 2003.

[18] Guo, H. and Viktor, H. L., "Learning from imbalanced data sets with boosting and data generation: the DataBoost- IM approach," ***SIGKDD Explorations***, Vol. 6, No. 1, pp. 30-39, 2004.

[19] H.T. Lin, C.J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," Technical report, Department of Computer Science & Information Engineering, National Taiwan University, 2003.

[20] Hsu, Chih-Wei, Chang, Chin-Chung, and Lin, Chih-Jen, "A Practical Guide to Support Vector Classification," Available at http://www.csie.ntu.edu.tw/~cjlin/papers/papers/guide/guide.pdf, 2003.

[21] Japkowicz, N. and Stephen, S., "The class imbalance problem: a systematic study," ***Intelligent Data Analysis***, Vol. 6, No. 5, pp. 429-449, 2002.

[22] Kubat, M., Holte, R., Matwin, S., "Learning when negative examples abound," ***Proceedings of Europeanm Conference on Machine Learning***, pp. 146-153, 1997.

[23] Lewis, D. and Catlett, J., "Heterogeneous Uncertainty Sampling for Supervised Learning," ***Proceedings of the 11th International Conference on Machine Learning***, pp. 144-156, 1994.

[24] Liao, T. W, "Classification of weld flaws with imbalanced class data," ***Expert Systems with Applications***, Vol. 35, No. 3, pp. 1041-1052, 2008.

[25] M. Kubat, S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," ***Machine Learning***, pp. 179-186, 1997.

[26] M.A. Maloof., "Learning when data sets are Imbalanced and when costs are unequal and unknown," ICML-2003 Workshop on Learning from Imbalanced Data Sets, 2003.

[27] N. Chawla, A. Lazarevic, L. Hall and K. Bowyer., "SMOTEBoost: improving prediction of the minority class in boosting," 7th uropean Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia , pp. 107-119, 2003.

[28] Press, S.J. & S. Willson, "Choosing Between Logistic Regression and Discriminant Analysis," ***Journal of the American Statistical Association***, pp. 699-705, 1978.

[29] Quinlan, J.R., "Induction of Decision Tree," ***Machine Learning***, Vol. 1, No. 1, pp.81-106, 1986.

[30] Quinlan, J.R., "C4.5:Programs for Machine Learning," Morgankaufmann, San Mateo, CA, 1993.

[31] Su, C.-T. and Hsiao, Y.-H., "An Evaluation of the Robustness of MTS for Imbalanced Data," ***IEEE Transactions on Knowledge and Data Engineering***, Vol. 19, No. 10, pp. 1321-1332, 2007.

[32] Su, C.-T., Chen, L.-S. and Yih, Y., " Knowledge acquisition through information granulation for imbalanced

data," ***Expert System with Applications***, Vol. 31, No. 3, pp. 531-541, 2006a.

[33] Su, C.-T., Chen, L.-S., and Chiang, T.-L., "A neural network based information granulation approach to shorten the cellular phone test process," ***Computers In Industry***, Vol. 57, No. 5, pp. 412-423, 2006b.

[34] Weiss G. M., and Provost F., "The Effect of Class Distribution on Classifier Learning," Technical Report, MLTR43, Department of Computer Science, Rutgers University, 2001.

[35] Weiss, G. M., "Mining with rarity: a unifying framework," ***SIGKDD Exploration***, Vol. 6, No. 1, pp. 7-19, 2004.

[36] Witten, I. H. and Frank, E., "Data Mining: Practical machine learning tools with Java implementations," Morgan Kaufmann, San Francisco, 2002.

[37] Wu, G. and Chang, E. Y., "KBA: kernel boundary alignment considering imbalanced data distribution," ***IEEE Transactions on Knowledge and Data Engineering***, Vol. 17, Vol. 6, pp. 786-795, 2005.

[38] Xie, J.G., and Qiu, Z.D., "The effect of imbalanced data sets on LDA: a theoretical and empirical analysis," ***Pattern Recognition***, Vol. 40, No. 2, pp. 557–562, 2007.

[39] Xu, L. and Chow, M.-Y., "A classification approach for power distribution systems fault cause identification," ***IEEE Transactions on Power Systems***, Vol. 21, No. 1, pp. 53-60, 2006.

[40] Zhou, Z.-H. and Liu, X.-Y., "Training cost-sensitive neural networks with methods addressing the class imbalance problem," ***IEEE Transactions on Knowledge and Data Engineering***, Vol. 18, No. 1, pp. 63-77, 2006.