模糊支援向量機應用於多類別的文件分類

王泰裕*、江蕙民

成功大學工業與資訊管理系

一、緒論

傳統上，文件分類的工作都是仰賴人力來處理。然而，隨著資訊時代的來臨，各種電子文件相繼問世，再加上個人電腦的日益普及和寬頻網路的迅速發展，人們所能 接收到的文件不再僅限於傳統的書籍、雜誌、報紙等等。藉由網際網路的傳遞，各種電子刊物、網頁以及電子郵件成爲人們不用出門也能獲得知識的來源。這些邁入 資訊社會所伴隨而來的改變，使得欲分類處理的文件種類，從傳統的紙本資料演進成各種數位化形式的檔案;文件數量也比以往龐大數千甚至數萬倍，而這些改變也 使得文件分類的工作，不可能再利用龐大的人力和時間來加以處理。爲因應這些情勢的改變，建立一個以電腦來自動處理文件分類的方法，已成爲一個重要的課題。

文件分類是依文件「內容主旨」給定文件類別的意思。例如，新聞文件可按其報導的內容，給予「政治」、「外交」、「娛樂」、「運動」等類別，而這些類別通常 是預定的，由於在網際網路上的文件如雨後春筍的出現，因此很難每一次的分類都靠領域專家以人工來分類,必要藉助自動分類系統，來協助此一工作。

就分類器而言：所謂「二元分類」主要是解決兩個類別(binary classification)的分類問題，例如：一個須檢驗的產品分成良品與不良品兩個類別，實驗的目的找出一決策函數可以區分出良品與不良品，藉以提 供給檢驗人員參考。而解決二元分類基本問題上，Vapnik (1998) 提出了線性支援向量機 (Linear Support Vector Machines) 就是用於處理兩類別的分類問題，近來學者Joachims (1998) 對支援向量機 (Support Vector Machine,SVM) 應用於文件分類進行研究，證明SVM在文件分類上有非常好的效用。在其簡單的基本線性模式中，SVM分類器乃是發現一組能區別正類別範例與負類別範例的一 個有最大邊界的超平面。應用SVM分類器就是找出一個線性可分割的超平面以解決二元分類的問題。但若是資料有偏離值的問題或是資料在本身類別權重不一時， 則可將資料給予不同的權重。於是模糊支援向量機(Fuzzy Support Vector Machine, FSVM) 被提出用以解決偏離值的問題。透過不同的權重使得最大邊界的超平面隨之改變，因此可得到更好的分類效率。但是之前的研究者皆著重於兩類別(binary class)模糊化的問題，對於「多類別」(multi-class)模糊化問題則較少著墨。

二、研究目的

本研究之研究目的即在研發出符合多類別文件分類系統。具體而言，本研究之研究目的爲：

1. 利用模糊支援向量機的概念發展一套多類別文件分類系統，以建立文件多類別分類的準則，方便文件的檢索，達到文件的預先分類機制，減少人爲分類差異，以達到加快檢索的速度及提高檢索的正確率。

2. 爲凸顯關鍵字詞在類別與類別之間的鑑別度以及降低輸入資料的維度，在關鍵字詞的選取上，加入集中度與廣度的概念。

三、研究方法

本研究所擬開發之多類別文件分類系統包含兩個模組一爲文件處理模組(document processing module)另一爲分類模組(classifying module)。文件處理模組包含文件預處理過程(data processing)、特徵選取過程(feature selection)、文件向量化表示過程(document frequency representation)、以及每一字詞權重計算方式(term's weight)等。處理順序爲從可用的資料集中收集類別文件，因可用的資料集爲SGML格式，所

以文件預處理工作爲移除SGML標籤及刪除一些無用的字 等。經過了預處理後文件包含許多單字，這些單字並不一定是對類別有區別力的因此依據兩個指標Uni及ICF萃取關鍵字，稱爲特徵擷取過程。這些關鍵字可以 向量空間模式Vector Space Model (VSM)表示，向量內的元素稱爲TF_IDF 權重，爲其對應分類器的輸入資料。

而分類模組基本上是先訓練分類器，接著測試分類器的好壞。在資料處理部分，資料先分成n個folds，其中一個fold 當做測試資料，n-1個folds當作訓練資料。爲了評估分類器的正確性，模糊支援向量機發現最佳的支援向量參數，隨後利用OAA(one- against-all)分類器找到多個最大邊界的超平面，並決定文件要歸屬的類別。

而在分類器目標值表示方面，假設k類問題必須使用k bipolar 目標函數去代表k個類別。一般而言當有n個文件k個類別需分類時, vij代表第i個文件且屬於第j類，其數學式可表示如下式

$$v_{ij} = \begin{cases} 1 & \text{if the } i^{th} \text{ document belongs to the } j^{th} \text{ class , for } i = 1,\ldots\ldots,n,\ j = 1,\ldots\ldots,k. \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

where

n : the number of documents

k : the number of classes

矩陣V代表vij的聯集。

另外同時定義OAA-FSVM ($ssij^+$,$ssij^-$) 當作OAA-FSVM 分類器的模糊歸屬函數，其中 $ssij^+$ 代表文件i 屬於第 j類的權重，相反的 $ssij^-$ 代表文件 i 不屬於第 j 類的權重。每一個輸入矩陣包含正範例與負範例，我們可給正範例較高的權重稱爲 $ssij^+$，相對的我們給於負範例較低的權重稱爲 $ssij^+$，可以數學式表之如下式

$$ss_{ij} = \begin{cases} ss_{ij}^{+} \text{ if } v_{ij} = 1, \\ ss_{ij}^{-} \text{ if } v_{ij} = -1 \end{cases} \quad (2)$$

where

$$0 \le ss_{ij}^{+} \le 1, 0 \le ss_{ij}^{-} \le 1,\ i = 1\ldots n,\ j = 1\ldots k$$

最後我們以實驗來驗證所提之理論，定義OAA-FSVM (1, h) 其中1爲正類別的權重，負類別的權重爲h ，h 實驗範圍從0.1 至1，結果顯示OAA-FSVM (1, 0.5) ，OAA-FSVM(1, 0.6) 表現較好，若以McNemar's 測試，其結果如Table 1，此結果顯示OAA-FSVM (1, 0.5) ，OAA-FSVM(1, 0.6) 與OAA-SVM 有統計上顯著差異。

Table 1
Results of McNemar's test for comparing the OAA-FSVM with the OAA-SVM algorithms

	4-fold			3-fold	
Classifier	*Z*	*p*-value		*Z*	*p*-value
OAA-FSVM (1,0.5) vs OAA-SVM	2.324**	0.0204		0.927	0.3524
OAA-FSVM (1,0.6) vs OAA-SVM	2.832**	0.0046		2.030*	0.0424

四、結論

本研究所所提之文件分類系統，包含萃取關鍵字及分類器的改善，分類器改善方面將資料模糊化的方法加入分類器運作中，所提之OAA-FSVM 方法可依據正類別範例與負類別範例在其所在類別的權重，使得正類別與負類別產生最大的區分邊界，更能區分正類別範例與負類別範例，經由實驗結果證明其有好 的分類效率。

< 上一篇 下一篇 >