集成分類器結合特徵選取與多字詞判斷疾病分類碼

許中川　　鄭紹偉　　王江山
國立雲林科技大學　　國立雲林科技大學　　國立雲林科技大學
hsucc@yuntech.edu.tw　g9623705@yuntech.edu.tw　g9823703@yuntech.edu.tw

摘要

全民健保實施後，健保局規定醫療單位申請補助費用時，必須一併呈報國際疾病分類碼。若不依此規定則不予以補助，尤其是疾病分類碼的編碼錯誤或不完整時，將被刪減或放大比例扣減醫療補助費用。目前大部份的醫療單位,在病歷摘要分類上，必須藉由人工作業完成，因此造成人力與資源的浪費。過去研究處理病歷摘要時，往往只考慮單字詞而忽略多字詞的重要性，造成資訊不完整影響分類績效；此外疾病分類碼普遍存在不均衡資料集的情況，而這些資料集往往極為相似，容易造成分類誤判。本研究透過醫療領域的語料庫擷取病歷摘要中的多字詞，並進行概念字詞擴充來增強病歷摘要的資訊，再搭配卡方檢定選取關鍵字以分辨相似類別，增進支援向量機與貝式分類器搭配多模激發法的績效。實驗透過 50 類的疾病分類碼進行驗證，發現卡方檢定能選取出品質較佳的關鍵字，且多字詞與擴詞能更完整的保留文件向量資訊，分類器集成法能大幅提升貝式分類器的分類績效。

關鍵字：國際疾病分類碼、概念字詞擴詞、卡方檢定、貝式分類器、支援向量機、多模激發法

1. 緒論

全民健保制度的實行，醫療單位必須藉由向健保局申報國際疾病分類碼(International Classification of Disease, 9th Revision, Clinical Modification, ICD9-CM)以獲得醫療費用的補助。由於採按件計酬的方式，一旦疾病分類碼編碼錯誤或不完整時，將會被刪減或被放大比例扣減醫療費用補助，故疾病碼判斷的正確性對於醫療體系是非常重要。

傳統做法，判斷病歷摘要的疾病分類碼須由人工完成，因此在執行編碼時容易產生許多人為上的疏忽。疾病分類碼繁雜且數量龐大，作業人員無法記憶所有疾病碼，較少出現的疾病容易出現編碼錯誤的案例，或先給予一般性的疾病分類碼，待日後再由人員覆核。除此之外，醫護人員難免不慎鍵入錯誤的分類碼，這不僅影響疾病分類的正確性，更可能因此而增加健保核減比例。

在醫療文件分類領域，部份研究提出將病歷摘要用向量空間表示，然後訓練分類器來預測疾病編碼。林偉彥、黃一平、楊家韋與林昕彥(2006)以貝式分類法、k 個最近鄰居分類法(k-nearest-neighbor, KNN)與支援向量機(Support Vector Machines, SVMs)三種學習演算法，分別對於大量類別的病歷摘要進行文件分類，透過實驗後發現，貝式分類法與 k 個最近鄰居分類法的正確率大幅下降，表示類別數量的多寡會影響其分類績效；而謝玉珊(2007)以創新之分類技術，由病歷摘要來預測疾病分類碼，針對多達 50 個疾病分類碼的 6664 筆病歷摘要進行實驗來解決以上問題，經實驗結果證明，採用支援向量機結合多模激發法可提升單一分類器之正確率。

陳紹佩(2008)更以領域知識強化病歷摘要文件向量的代表性來提升分類績效，實驗結果顯示透過醫療領域的語料庫進行概念字詞擴充時，同時考慮相關字詞出現在文件的頻率，並將相關字詞加入病歷摘要文件向量中，所能提升分類績效的影響最為顯著。廖艾貞(2008)以「局部分群分類法」改善疾病類別分佈不均的問題，實驗結果發現使用局部分

群演算法可以提昇小類別之分類準確率，進而提昇整體分類績效。但整體而言，於醫療文件分類領域中含有下列問題：

1. 類別數量將會影響分類績效。疾病分類碼繁雜且數量龐大，疾病分類的難度相對提升。
2. 國際疾病分類碼屬於階層式架構，越下層的類別彼此的相似性越高，因此判斷子類別疾病碼時，由於相似類別之間的關鍵字大部份都相同，無法利用重複性高的關鍵字來區分類別。
3. 判斷相似性高的類別時，若類別的文件數量差異大，訓練資料量少的疾病分類碼容易被資料量多的疾病分類碼所影響，進而影響分類器判斷能力。
4. 病歷摘要經由前置處理轉為文件向量時，往往只保留單字詞，而忽略多字詞，導致前置處理後的文件向量所保留資訊並不完整。
5. 透過醫療領域的語料庫進行概念字詞擴充時，只考慮了單字詞卻忽略了二字詞、三字詞等多字詞對文件向量的影響。

本研究目的希望能將非結構化的病歷摘要，自動判斷相對應之疾病分類碼，以目前新進的機器學習及資訊檢索技術改進先前研究之缺點，以協助人工編碼作業。具體而言，本研究探討下列議題：

1. 探討不同關鍵字選取方法及不同關鍵字權重設定方法，是否影響疾病碼判斷之績效。
2. 探討結合領域知識協助擷取病歷摘要中的多字詞，保留病歷摘要資訊的完整性，並進一步改良領域知識擴充用詞的做法，增強多字詞資訊對文件向量的影響，是否能提升疾病碼判斷之績效。
3. 探討分類器集成方法是否能提升疾病碼判斷之績效。

本研究使用中南部某大醫院所提供的病歷摘要做為主要研究的資料集，其來源由醫師撰寫看診病人所屬的病歷資料後，轉交至專業的醫療人員進行病摘編碼，再經確認後送交健保局以申報其醫療費用補助，因此資料來源具可靠性，並假設所取得之醫療文件編碼皆為正確。

2. 文獻探討

本節介紹本研究使用之語料庫 MeSH、相關的機器學習及資料檢索演算法以及相關研究，包括文件分類、領域知識擴充、兩階段分類法等較新的資訊技術。

2.1 醫學標題表

美國國家醫學圖書館(National Library of Medicine, NLM)於 1954 年首次正式出版標題表(Subject Heading Authority List)，1960 年由於 Index Medicus 的發行，全新的醫學標題表(Medical Subject Headings, MeSH)也隨之出版。MeSH 內含有豐富的醫學領域知識，可從不同角度加以剖析，並供給不同用途的使用者加以分析使用，到 2007 年為止，MeSH 包含了 24,357 項標題表，也稱為主標目(Descriptor, 底下皆以英文表示)，而本研究根據陳紹佩(2008)的研究將 MeSH 視為領域知識的來源。

目前 MeSH 透過延伸標記語言(Extensible Markup Language, XML)做為表達階層關係的標準格式，其中“ConceptRelation”表達此 Descriptor 與其下層 Concept(S)的關係，包含廣義(broader, BRD)、狹義(narrower, NRW)、有關但非廣義也非狹義(related but not broader or narrower, REL)，除此之外還具有“ScopeNote”說明此 Concept 的範圍。因此 MeSH 可表達 Descriptor 與其下層 Concept(s)的關係與詞類變化清單，故可被視為處理醫療文件分類時所用的語料庫(thesaurus)。

2.2 相關演算法

下列介紹與本研究相關之演算法，範圍包含特徵選取(feature selection)常使用的卡方檢定(chi-square test)，分類演算法中常見的支援向量機(Support Vector Machines, SVM)，以及屬於分類器

集成法(Ensemble)的多模激發法(Adaptive Boosting, AdaBoost)。

2.2.1 卡方檢定

卡方檢定(chi-square test)是一種常見的特徵選取(Feature selection)方法(March,A. D., Laur ía,E. J. M. and Lantos,J., 2004)，Yiming Yang 與 Jan O. Pedersen (1997)兩位學者也提到卡方檢定在特徵選取的領域是相當有效率，並且不降低整體準確率的一種做法，故本研究也使用卡方檢定處理關鍵字選取的動作。在統計上，卡方值用以檢定兩事件間的獨立性；在特徵選取的領域，可以評估"關鍵字的出現"與否與"類別的出現與否"，並可依據評估的數值對關鍵字進行排序，其評估公式如下：

$$X^2(D,t,c)=\sum_{e_t\in\{0,1\}}\sum_{e_c\in\{0,1\}}\frac{(N_{e_te_c}-E_{e_te_c})^2}{E_{e_te_c}} \quad (1)$$

其中 D 表整體文件，N 表在 D 中觀察到的 e_te_c 出現次數，E 表期望的 e_te_c 出現次數，e_t 表示文件是否包含關鍵字 t，$e_t = 1$ 表包含，$et = 0$ 表不包含；e_c 表文件是否屬於某類別 c，$e_c = 1$ 表屬於，$e_c = 0$ 表不屬於。

公式 1 可經由數學推導簡化如下式：

$$X^2(D,t,c)=\frac{(N_{11}+N_{10}+N_{01}+N_{00})\times(N_{11}N_{00}-N_{10}N_{01})^2}{(N_{11}+N_{01})\times(N_{11}+N_{10})\times(N_{10}+N_{00})\times(N_{01}+N_{00})} \quad (2)$$

2.2.2 支援向量機

近幾年，由於支援向量機(Support Vector Machines, SVM)夠成功的解決許多資料檢索上的問題，如：離群值偵測、文件分類、影像辨識…等，因此在資訊技術領域中逐漸被廣泛的被應用。

支援向量機由 Vapnik 在 1995 年與 AT&T 實驗室團隊共同提出的方法，依據統計學習理論中結構化風險最小化(Structural Risk Minimization, SRM)的原則(Vapnik,1995)，主要利用超平面(Hyper plane)區分兩個或多個不同類別(class)的資料，以此處理分類(Classification)問題。分類資料定義如下：

$$(x_1,y_1),\dots,(x_m,y_m),x_i\in X,y_i\{+1,-1\},i=1,2,\dots,m$$

在向量空間 X 中的所有資料$\{x_1,\dots,x_m\}$皆以向量的方式表示。y_i 為資料的標記(Label)或目標(Target)，若資料為兩個類別時，通常用{±1}表示，+1 和-1 表示兩個不同的類別。依照 SVM 的處理方式可分為線性支援向量機與非線性支援向量機。

2.2.3 貝式分類器

貝氏分類法是運用貝式定理(Bayes' theorem)的概念，在一組有限的類別集合中，計算未知的類別文件落入集合中各類別的機率，然後指定可能性最大的類別為此文件的類別。

演算法主要去計算 $P(C_i|X)$，判斷文件 X 屬於 C_i 類別的機率，C_i 存在於一個有限的類別集合 C。依據貝氏定理，為了達到最高的分類正確率，文件 X 將分類至 $P(C_i|X)$的最高機率之類別。

$$X\in C_i\equiv \arg\max_{C_i\in C}P(C_i|X) \quad (3)$$

依照貝氏定理可將 $P(Ci|X)$分為兩個部分

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{\sum_{C'\in C}P(C|C')P(C')} \quad (4)$$

$P(Ci)$為一篇文件落入類別 Ci 的事前可能性(prior probability)，$P(Ci|X)$則是在類別 Ci 下，文件 X 落入 Ci 的條件可能性(conditional probability)。$P(Ci)$為類別 Ci 占訓練資料的文件比例。

$$P(C_i)=\frac{|C_i|}{\sum_{C'\in C}|C'|} \quad (5)$$

由於資料集中存在大量的相異文件，使得計算 $P(C_i|X)$變得較為困難。我們不可能收集到足夠的訓練資料來計算 $P(X|C_i)$之值，因此我們假定組成文件的詞彙與文件的類別相依，但文件中的詞彙彼此之間互相獨立，所以 $P(X|C_i)$即可表示為：

$$P(X|C_i)=\prod_{k=1}^{|X|}P(w_k|C_i) \quad (6)$$

$|X|$為文件 X 內詞彙的個數，在計算 $P(w_k|C_i)$ 的過程中，我們必須考量到詞彙 w_k 的詞頻所帶來的資訊，故藉由 $P(w_k|C_i)$即可得知：

$$P(w_k|C_i)=\frac{1+TF(w_k,C_i)}{|vocabulary|+\sum_{w'\in vocabulary}TF(w',C_i)} \quad (7)$$

$TF(w_k, C_i)$是指詞彙 w_k 在類別 C_i 出現的次數，$|vocabulary|$是指整個訓練資料中不重覆的詞彙個

數， $\sum_{w' \in vocabulary} TF(w', C_i)$是指所有在 vocabulary 集合中的詞彙，出現於類別 C_i 的次數加總。

最後，我們將(4)、(5)和(7)整合成為(8)，又由於(8)中分母對於最後結果不會產生影響，因此將分母部分省略推導出公式(9)，以此做為文件分類的基礎：

$$X \in C_i \equiv \arg\max_{C_i \in C} \frac{P(C_i) \cdot \prod_{k=1}^{|X|} P(w_k|C_i)}{\sum_{C' \in C} P(C') \cdot \prod_{k=1}^{|X|} P(w_k|C')} \quad (8)$$

$$\approx \arg\max_{C_i \in C} P(C_i) \cdot \prod_{k=1}^{|X|} P(w_k|C_i) \quad (9)$$

2.2.4 多模激發法

多模激發法(Boosting)為分類器集成法(Ensemble)的一種，在訓練階段時，主要利用抽樣的方式產生訓練資料集，再利用學習演算法來建立分類器。經過 T 回合的訓練後，則可得到 t 個分類器。因此，當判斷測試資料的類別時，我們結合 t 個分類器所得的分類結果，來決定資料的類別。

Freund and Schapire(1995)兩位學者提出另一種更有效率的多模激發法－調適性多模激發法(Adaptive Boosting, AdaBoost)，與原本的計算方式不相同，但效率一樣。多模激發法執行時，必須預先知道假設(Hypothesis)的錯誤率下限；但調適性多模激發法，則是計算在抽樣訓練資料中，將預測錯誤的資料權重加總，即為此假設的錯誤率。利用此方式調整假設的錯誤率，因此不需事先了解。

本研究為 50 個類別的實驗資料集，因此無法使用一般的調適性多模激發法。在此我們採用調適性多模激發法 M1 型，來處理多類別的分類問題，以下皆以 AdaBoost 稱之。圖為 AdaBoost 之演算法虛擬碼(Freund & Schapire,1996)。

Input: sequence of N examples $< (x_1, y_1), \ldots, (x_N, y_N) >$
with label $y_i \in Y = \{1, \ldots, k\}$
weak learning algorithm WeakLearn
integer T specifying number of iteations
Initialize the weight vector: $w_i^1 = D(i)$ for $i = 1, \ldots, N$
Do for $t = 1, 2, \ldots, T$
1.Set

$$p^t = \frac{w^t}{\sum_{i=1}^{N} w_i^t}$$

2.Call WeakLearn, provideing it with the distribution p^t; get back a hypothesish$_t$: $X \to Y$
3.Calculate the error of ht : $\varepsilon_t = \sum_{i:h_t(x_i) \neq y_i} w_i^t$
if $\varepsilon_t > 1/2$, then set T = t -1 and abort loop
4.set $\beta_t = \frac{1}{2} \ln(\frac{1-\varepsilon_t}{\varepsilon_t})$
5.Updatedistribution D_t :

$$D_{t+1}(i) = D_t(i) \times \begin{cases} e^{-\beta_t}, \text{if } h_t(x_i) = y_i \\ e^{\beta_t}, \text{if } h_t(x_i) \neq y_i \end{cases}$$

where Z_t is a normalization constant (chosen so that D_{t+1} will be a distribution)

Output thefinal hypothesis:

$$h_{fin}(x) = \arg\max_{y \in Y} \sum_{t:h_t(x)=y} \beta_t$$

圖 1 調適性多模激發法 M1 型

我們將訓練資料定義為$\{x_1,\ldots,x_N\}$，每筆資料可分類至$\{y_1,\ldots,y_N\}$中，$y_i \in \{1,\ldots,k\}$。另外，在多模激發法中，並未規定必須使用的學習演算法，因此將它稱之為簡單學習演算法(weak learn)。

在訓練階段中，AdaBoost 必須連續地執行簡單學習法，直到產生 T 個分類器為止。我們針對抽樣所產生的訓練資料，利用簡單學習法來產生分類器或假設 ht：X→Y，以此評估訓練資料集的錯誤率。然而，簡單學習法最重要的目的就是找到一個假說，可以使得訓練的錯誤率達到最小化，計算方式如(10)，

$$\varepsilon_t = Pr_i \sim D_t[h_t(x_i) \neq y_i] \quad (10)$$

在計算訓練錯誤率時，我們利用簡單假說(Weak hypothesis)判斷抽樣訓練資料集，將預測錯誤的資料權重加總，即為此回合的錯誤率。這樣的處理流程需要持續 T 回合，最後將產生的所有簡單假說 $h_1, h_2, \ldots, h_T$ 整合成為一個最後假說 h_{fin}。

在上述的說明中，仍有尚未被定義的問題：首先，每回合的訓練中，D_t 應該如何被計算。其次是最後假說 h_{fin} 應該如何決定。

在本研究中，我們使用最簡單的方式來調整資料的權重值 D_t，並且設定抽樣訓練資料的初始權重值為 1。在每次訓練中，必須正規化所有抽樣訓練

資料的權重。我們以 D_t 與上一回中所評估的錯誤率來計算 D_{t+1} 的權重值，調整資料的權重值可採用(11)(12)的方式。當預測的類別是正確時，以(11)的方式來調整資料的權重，否則採用(12)的方式來調整。藉此將訓練中容易分類正確的資料，降低其權重值；較難被分類正確的資料，則提高權重值。因此，在下一回合時，權重大的資料被抽樣的機率也相對提高。AdaBoost 可以逐漸地將訓練重點，放在不易被分類正確的資料上，相對也提升分類器的訓練能力。

$$D_{t+1}(i) = D_t(i) \times e^{-\beta_t}, if\ h_t(x_i) = y_i \quad (11)$$

$$D_{t+1}(i) = D_t(i) \times e^{\beta_t}, if\ h_t(x_i) \neq y_i \quad (12)$$

我們以最後假說 h_{fin} 來決定測試資料的類別。主要利用簡單假說的權重 β_t 做為票選之依據，例如：在決定資料 x_i 的類別時，將 $h_1, h_2, ..., h_t$,所預測的類別，利用每一個簡單假說的權重加總，來計算出擁有最大權重的類別 y 即為 h_{fin} 的最後決策結果。權重 β_t 的計算方式可參考(13)，當訓練錯誤率越小時，則權重越大。

$$\beta_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \quad (13)$$

在此需特別注意，當評估簡單假說的訓練錯誤率已大於 0.5 時，將會造成 h_{fin} 的預測類別機率，以指數的方式掉到零以下。這表示此回合的訓練結果不佳，必須重新設定抽樣訓練資料以及其權重。

2.3 相關研究

在醫療文件分類的文獻上，已有許多學者提出各種方法來自動判斷病歷文件的分類碼。以下分為疾病碼預測(Ribeiro-Neto, Laender & Lima, 2001)、領域知識擴充、類別量不均衡之處理、兩階段分類法四部分加以討論。

2.3.1 疾病碼預測

在醫療文件的分類上，已有許多學者提出利用不同方法，自動判斷病歷摘要的疾病碼。Larkey and Croft(1996)使用 k 個最近鄰居分類法、相關回饋(relevance feedback)和貝式分類法，藉由個別與整合三個分類器的方式，分別針對大量資料集進行實驗，相較於單一分類器的實驗結果，結合不同的分類方法可明顯地改善分類正確率。實驗結果發現，結合三種分類方法來判斷文件類別時，由所提供的前十個候選分類碼中，高達 91.1%可正確預測疾病分類碼；但在直接決定文件的分類碼時，其正確率只有 46.5%。March, Lauria and Lantos(2004)利用貝式分類法自動判斷疾病碼的階層碼(section code)，分類正確率可達到 86%。但只使用少量病歷資料進行實驗，判斷結果為疾病分類碼的階層碼，無法提供正確的疾病碼。

2.3.2 領域知識擴充

過去有多位學者研究關於結合領域知識，以提升分類準確率的方法。Franz ,Zaiss, Schulz, Hahn and Klar(2000)將病歷摘要的分類方法分為索引(indexing)與檢索(retrieval)兩階段，索引階段使用了 Trigrams 與 SNOMED 編碼，其中 SNOMED 是醫學專用術語的資料庫，稱為人類與獸類醫學系統術語(the Systematic Nomenclature of Human and Veterinary Medicine, SNOMED)，搭配檢索階段的兩種方法，組合成 TGVS、MSVS 與 MSMS 如圖，其中 MSMS 結合領域知識進行編碼與檢索，實驗結果也顯示 MSMS 有較佳的結果，顯示了領域知識對於醫學文件分類的確有其重要性。Zhou, Yu, Torvik ,Smalheiser and Hong (2007)運用領域知識包含同義字、狹義詞(hyponyms)、廣義詞(hypernyms)、詞類變化(lexical variants)與隱含相關概念詞(implicitly related concepts)對查詢字詞進行擴充，結果也顯示此種做法能獲取更多相關的文件。

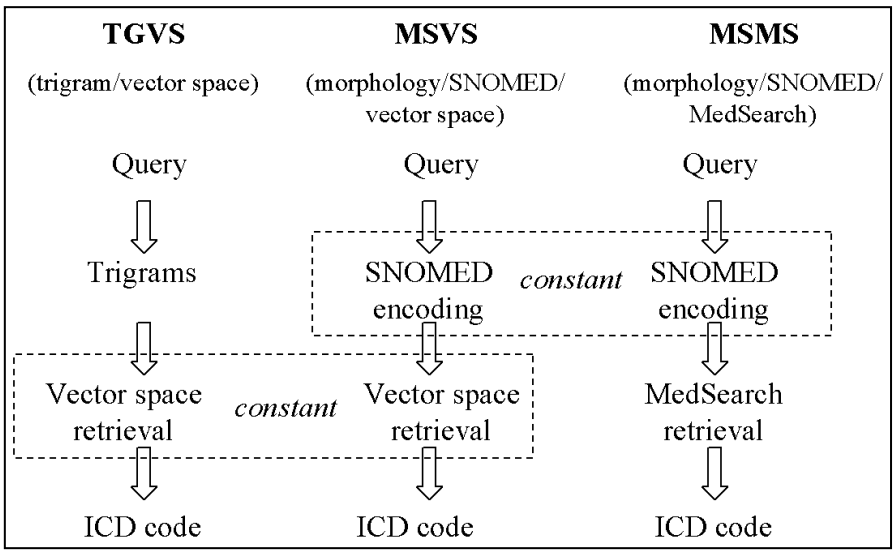

圖 2 Franz et al.(2000)所提 TGVS、MSVS 與 MSMS 方法

2.3.3 類別量不均衡之處理

類別不均衡問題處理，即探討如何降低兩類別資料數量的不均衡性、減少建構出的分類器對多數資料類別有預測偏向的問題，以避免分類出來的結果會偏向數量較多的類別。本研究將以小類別代表集合內數量稀少的類別，而以大類別代表擁有多數資料的類別。過去解決此問題的作法大致可分為兩種(廖艾貞, 2008)，分別是演算法層面與資料集層面，以下將分別介紹：

1. 演算法層面：透過調整不同類別的成本、調整決策門檻值等方式來達成。較有代表性之方法，包括了成本敏感的學習演算法(cost-sensitive learning)，以及辨識為基的學習演算法(recognition-based learning)。
2. 資料集層面：可分為減少多數法(under-sampling)、增加少數法(over-sampling)、多專家分類器(multi-classifier committee)與局部分群分類法(classification using local clustering)

除此之外，Forman(2003)認為處理不均衡資料時，特徵選取相對比分類演算法來的重要，Zheng ,Wu and Srihari (2004)也認同此觀點，並改以特徵選取的角度嘗試解決不均衡資料的問題，在不均衡資料集中挑選具有類別鑑別度的特徵，再進行分類。

2.3.4 兩階段分類法

過去執行分類任務時，有多位學者提出以兩階段甚至多階段的方式對資料進行處理，以提升績效。Tu, Chen, Wu and Chang (1998)針對高維度的遠端感應資料提出兩階段的快速分類法，於第一階段透過 BS(Band selection) 與 FSE(Feature Extraction/Selection)對資料進行降低維度的動作，再於第二階段使用 fast RMLC 執行分類，實驗結果顯示此法可大量降低運算時間。韓歆儀(2004)以兩階段分類法提升分類準確率，對人工資料與真實糖尿病資料(Pima Indians Diabetes Database)，先透過決策樹(decision tree)簡化資料，再以支援向量機(Support Vector Machine)進行第二階段分類，結果顯示此法可對複雜性較高的資料提升分類準確率。Bedingfield and Smith-Miles (2006)提出兩階段方式處理不平衡資料的分類法，結果顯示此法可以提升資料類別不平衡時的績效。整理以上結果發現由於資料具有不同特性，故適度於演算法中的不同階段對資料進行不同處理，可有效提升整體分類績效。

3. 研究方法

本研究所提出之研究方法，包含結合領域知識擷取多字詞與擴充文件向量、關鍵字選取與所使用之分類架構。本節說明如何將非結構化的原始病歷摘要，經由此過程預測國際疾病分類碼。

非結構化的原始病歷摘要存在許多雜訊資料，這些雜訊也會影響分類績效，故須經由前置處理對病摘進行過濾與篩選，以獲得正確病摘資訊；接著結合領域知識進行多字詞擷取與病摘資訊的擴充；並經由關鍵字選取階段，比較詞頻與反轉文件頻率(Term Frequency and Inverse Document Frequency, TF-IDF)與卡方檢定(Chi-square test)兩種方法，何者較能辨別出具有判別類別能力之關鍵字，再分別使用支援向量機結合多模激發法與貝式分類器結合拔靴集成法，訓練分類規則以預測類別。圖即為本研究所提方法之流程，並以灰底表示本研究所提出與過去不同做法之部分。

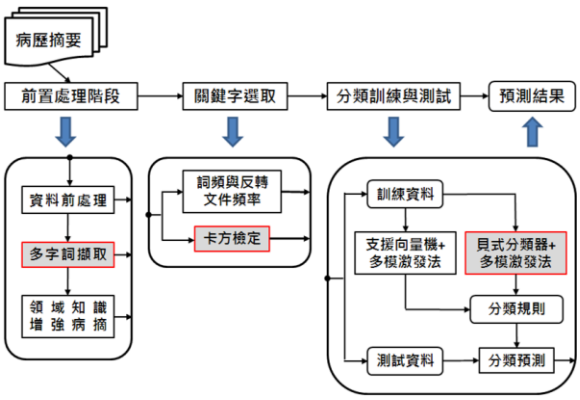

圖 3 本研究所提之預測疾病分類碼之處理流程

此流程分為前置處理，關鍵字選取以及分類訓練與測試三階段，最後可得到分類規則，對前置處理後的測試資料集進行國際疾病分類碼的預測。此章主要探討訓練階段內的各項步驟，包含對原始病歷摘要記錄進行前置處理，多字詞擷取與領域知識增強病摘資訊，接著介紹關鍵字選取方法與分類訓練與測試階段，最後獲得分類規則，以下就依序詳加說明。

3.1 前置處理

前置處理為本方法流程中的第一步，其重要性在於病歷摘要為非結構化的醫療文件，可能存在許多不正確的雜訊資料，或是內容不一致及不完整的問題，會影響後續使用的資料探勘技術所產生之結果。故前置處理的目的即希望透過整合(Integration)、轉換(Transformation)、過濾(Filtering)等方式提昇資料品質，在大量的資料中獲得有用的資訊，以改善執行的分類效率及準確性。文件處理常見的資料前處理可分為置換字母大小寫、去除數值型資料、更正拼字錯誤、還原基本字詞與刪除停用字共五步驟。

除了一般常見的資料前處理過程，本研究於此階段加入多字詞擷取與領域知識增強病摘兩步驟，希望透過多字詞的辨識，能更完整的保留病摘資訊，並且延續陳紹佩(2008)所提出藉由領域知識強化病歷摘要文件向量的代表性。

3.1.1 多字詞擷取

本研究希望透過多字詞的辨識，能更完整保留病摘資訊，本研究整理 MeSH 內所記錄的詞彙，對病歷摘要內的多字詞進行辨識與擷取，其單字詞與多字詞個數統計如下：單字詞 66711、二字詞 71984、三字詞 26243、四字詞 9746、其餘多字詞 7160，總計共有 181844 個詞彙。由此可發現 MeSH 中的詞彙有半數以上為多字詞，顯示多字詞的判斷的確有其必要性。

統計結果顯示多字詞個數超過十萬個，若病歷摘要中所有可能的多字詞都要進行比對，將耗費許多時間，在此本研究將病歷摘要中所擷取的單字詞，先行與 MeSH 中的多字詞比對，並保留含有這些單字詞為字首的多字詞，如此就大大的降低需要比對的多字詞個數。經過此步驟之後，MeSH 中的多字詞個數也從 115133 個降為 7359 個，比對次數也減少許多。

完成多字詞的擴充後，尚有下列兩點問題需解決：第一、形成多字詞的單字詞是否應保留，如多

字詞“benign neoplasm”中的單字詞“benign”、“neoplasm”，是否應做為文件向量的關鍵字。第二、多字詞為經由 MeSH 的詞彙辨識而得，可確定皆為專有名詞，而過去透過詞頻(term frequency, TF)評估單字詞在病摘中，是否有足夠的出現次數來做為關鍵字的做法，能否適用於多字詞。對於上述兩點的解決方法，有待取得實驗結果後才能確定何種做法較能提升實驗績效，不過在此可先做合理的推論，進一步觀察實驗結果是否與推論結果相符，並找出其原因。

關於問題一的處理，發現病摘中這些組成多字詞的單字詞，有極高比例在其餘病摘中以單字詞出現，如“benign neoplasm”此多字詞在病摘中出現被擷取後，“benign”與“neoplasm”在其他病摘中也有出現，但並非相鄰，而是具有各自所代表的意義，因此多字詞與單字詞須分開考量，也就是說“benign”與“neoplasm”若單獨出現，便可成為文件字詞，但若同時出現，則“benign”與“neoplasm”就不被當作文件字詞，其資訊改由“benign neoplasm”所取代，因此在計算詞頻與文件頻率時，“benign neoplasm”中的“benign”與“neoplasm”也不會納入計算，除非是單獨出現的情況。

對於問題二的處理，本研究認為多字詞應完全保留做為文件向量的一部分，因為這些多字詞皆有其獨特意義，也表示這些詞彙能表達該篇病摘的部分資訊，故不該因為詞頻較低而將其刪除，因為詞頻較低反而表達多字詞詞意的獨特性。

3.1.2 領域知識增強病摘

本研究以 MeSH 做為醫學語料庫，透過醫學標題表協助擷取病歷摘要中的二字詞、三字詞等多字詞，並接續陳紹佩(2008)提出以領域用詞擴充病歷摘要的做法，進一步改良過去研究中無法對多字詞進行擴充的問題，以期提升分類績效。以下介紹領域用詞擴充病歷摘要的做法。

病歷摘要為醫生臨床撰寫，對於疾病的描述與所用字詞有限，使用機器學習方式進行分類時，常因為資訊不充足而無法有效預測疾病碼，因此若能藉由醫學領域的語料庫對病歷摘要擴充相關資訊，讓病歷摘要形成的文件向量能蘊藏更多資訊，期望能進一步提升分類績效。

擴充字詞的方法是將前處理過的文件字詞，透過醫學標題表加以擴充，簡稱為擴詞，如圖所示，獲取相關的概念與字詞做為擴充的來源。使用詞頻與反轉文件頻率衡量字詞重要性時，除了計算該字詞頻率，也會將醫學標題表內所獲取的相關字納入考慮，透過領域知識的擴充，使文件向量能蘊藏更多資訊。

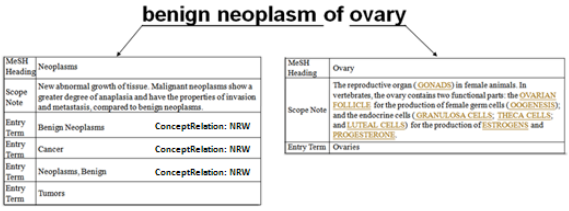

圖 4 benign neoplasm of ovary 的擴充結果

這些相關字詞在文件中所表達的意義將反映在文件頻率上，根據 Wei and Clement (2007)將字詞間的關係分為三種，而字詞間的不同關係透過參數設定整合如公式(14)，並使用權重 **β** 為不同的關係設定不同權重如下。(1)在醫學標題表內字詞屬於 t_k 的**同義字與狹義字時**，將 **β** 設為 1，並以 u_1 表示所有文件中出現此類字詞的文件數。(2)在醫學標題表內字詞屬於 t_k 的**廣義字時**，將 **β** 設為 0.95，並以 u_2 表示所有文件中出現此類字詞的文件數。(3)在醫學標題表內字詞與 t_k 屬於**其他關係時**，將 **β** 設為 0.9，並以 u_3 表示所有文件中出現此類字詞的文件數。並以公式(14)將三者結合，成為新的文件頻率值 df_{t_k}。

$$df_{t_k} = 1 * u_1 + 0.95 * u_2 + 0.9 * u_3 \qquad (14)$$

決定領域知識對於文件頻率的影響後，陳紹佩(2008) 提出了四種擴詞策略如圖所示，並與過去未擴詞(Non-extending, Non-E)的做法加以比較，四種策略描述如下：

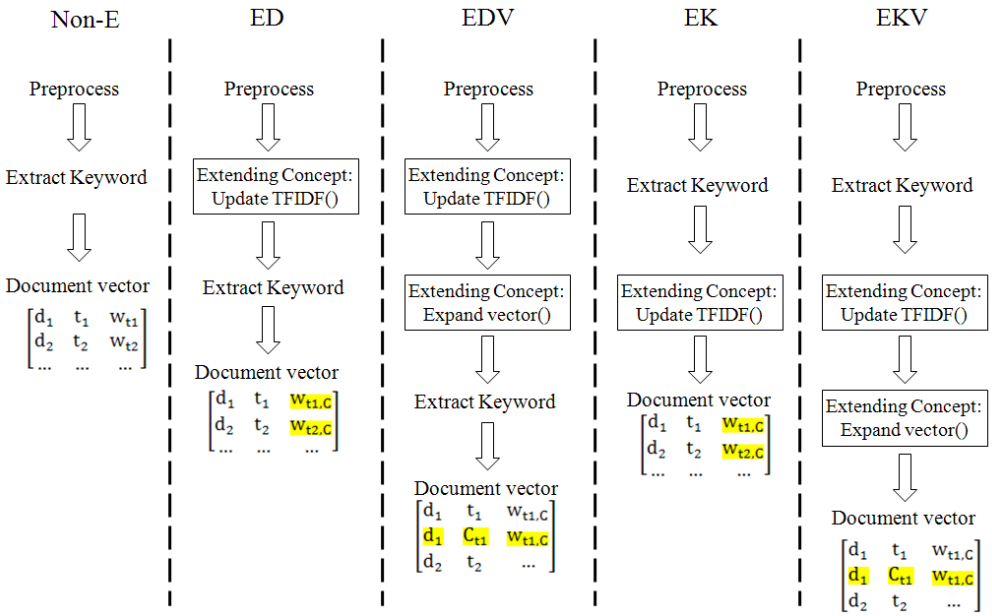

圖 5 未擴詞與四種擴詞做法

1. 對經過前處理後而未經篩選過的文件字詞(document terms)進行擴詞，以 ED 表示。
2. 對經過前處理後而未經篩選過的文件字詞進行擴詞，並將找到的詞加入到該篇病摘的文件向量中，以 EDV 表示。
3. 對篩選過後的關鍵字(keywords)進行擴詞，以 EK 表示。
4. 對篩選過後的關鍵字進行擴詞，並將找到的詞加入到該篇病摘的文件向量中，以 EKV 表示。

這四種做法均針對病歷摘要中的單一字詞個別擴充，並未考慮到醫生撰寫病歷摘要時可能出現二字詞、三字詞等多字詞的情形，因此績效提升並不顯著，在此本研究希望解決此項問題，加入針對多字詞擴充的做法，期望提升演算法的績效。故在此選擇 ED 與 EDV 的策略，因為不希望在擷取關鍵字過程中喪失了多字詞的組成字詞，因此針對文件字詞而不針對關鍵字擴充。

3.2 關鍵字選取

關鍵字選取的目的希望能適度的將文件字詞進行篩選，選取出具有辨別類別能力的關鍵字，過去謝玉珊(2007)使用詞頻與反轉文件頻率(TF-IDF) (Salton & Buckley,1988)當中的詞頻與文件頻率做為選取關鍵字的條件。本研究則使用卡方檢定進行關鍵字選取，並與 TF-IDF 比較，期望透過卡方檢定檢驗字詞與類別獨立程度的能力，挑選出品質較佳的關鍵字詞來提升績效。

以下將分別介紹 TF-IDF 以及卡方檢定的做法與概念，並進一步探討使用 TF-IDF 或卡方值的權重值可獲得較好的結果。

3.2.1 詞頻與反轉文件頻率

此方法主要利用統計與機率的概念，針對文件中的所有字詞與擴充的相關字詞進行權重計算，以字詞的權重來決定是否足以代表本篇文章，最後刪除與本文關聯性低的詞彙，故可形成一個維度較小且具有意義的新特徵空間(feature space)。

TF-IDF 值乃一種用於資訊檢索與文件探勘上的權重計算方法，時常被用於評估字詞在文件中的相關度，當詞彙的權重值越高時，則表示越能代表本文，其定義如下：

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log\left(\frac{\#Tr}{\#Tr(t_k)}\right) \quad (15)$$

其中，詞頻(term frequency, TF)以$\#(t_k, d_j)$來表示詞彙 t_k 在文件 d_j 中所出現的次數，當數值越大時則

代表詞彙 t_k 在文件 d_j 中越重要。此外將大寫的 TF 設為全體文件中詞彙 t_k 出現的次數，並設定一個門檻值 α，藉此來過濾總詞頻 TF 小於 α 的詞彙。文件頻率(document frequency, DF)以 $\#Tr(t_k)$ 來表示出現詞彙 t_k 的所有文件數量，當此數值越大時，即意指詞彙 t_k 時常出現於許多文件之中，無法以此當作文件的關鍵字，故刪除文件頻率大於所有文章數 β% 的詞彙，以此把集中度低的詞彙過濾掉。反轉文件頻率(inverse document frequency, IDF)以 $\log\frac{\#T_r}{\#T_r(t_k)}$ 來計算與文件頻率的反向關係，$\#T_r$ 為所有文件之數量。當數值越高時則表示詞彙 t_k 越具關鍵性，則可以利用它來區分文件。

將所有出現文件中的詞彙 t_k 之 TF-IDF 值都加以計算後，建立一個門檻值 γ，以此做為刪除小於此數值的所有詞彙之依據，來達到篩選出現頻率低或不具代表性的詞彙。在此本研究經過參數設定實驗後，於支援向量機的部分將 α 值設為 3、β 值設為 10%、γ 值為最小的 TF-IDF 值，而貝式分類器使用 α 值設為 1、β 值設為 12%、γ 值為 20 做為關鍵字擷取的參數依據。

3.2.2 卡方檢定

過去研究(謝玉珊, 2007)提到若相似類別的文件數量差異過大，在訓練階段文件數量少的類別容易受到數量多的類別的影響，導致測試階段時將文件數量少的類別文件，錯誤分類至文件數量多的類別，降低分類績效。相似類別是指疾病碼前 3 碼相同或相近的疾病碼，而文件數量差異過大也可稱為不平衡的資料(imbalance data)，過去研究指出本研究資料集具有這兩點特性。卡方檢定透過檢定關鍵字與類別的相依程度，當關鍵字與類別相依程度越高，所獲取卡方值也越高，也表示此關鍵字對該類別的重要程度，因此選取關鍵字時便將卡方值加以排序，從最高開始選擇一定比率的字詞作為關鍵字，而這些具有較高卡方值的關鍵字也表示代表類別的能力。在選擇上分為兩種策略。兩種做法均有其特點，在此無法評判優劣，於實驗階段將兩種不同做法進行比較。第一種為整體的策略：文獻中常見的做法(Yang and Pedersen, 1997)，將所有字詞依類別比例加以平均，排序後根據比例挑選表現值較高的字詞作為關鍵字。

$$X^2(t) = \sum_i^m P(C_i) \times X^2(D, t, C_i) \quad (16)$$

$$keywordset = \{t | X^2(t) \geq threshold\} \quad (17)$$

第二種為各類別的策略：本研究所提之做法，依相同比例在 m 個類別中各自挑選卡方值較高的字詞作為關鍵字，其概念為各類別擁有各自相依程度較高的關鍵字，故將這些關鍵字根據相同比例加以挑選。

$$keywordset = \{t | X^2(D, t, C_i) \geq threshold_i\}_{i=1}^m \quad (18)$$

3.2.3 關鍵字權重值

在關鍵字選取獲得關鍵字詞後，下一步便要評估這些關鍵字詞對病摘的表現權重值，過去常見的作法為使用直接 TF-IDF 值作為關鍵字的權重值，本研究為探討卡方值對實驗結果的影響，延續使用整體策略所獲的卡方平均值做為權重，雖然此整合後的一個卡方值代表多個類別的卡方值的平均值，無法突顯針對單一類別的辨別能力，然而一個文件向量有很高維度，不同關鍵字卡方值的交互影響，或許可以影響對單一類別的鑑別能力。故在此也提出用卡方值做為權重方式的兩種做法。第一種方法直接使用卡方值做為權重，若關鍵字詞在病摘中出現，則使用整體策略所獲的的卡方平均值做為權重。第二種方法將詞頻與卡方值做相乘，考慮關鍵字詞在病摘中的出現次數，在與卡方平均值相乘做為權重。

3.3 分類訓練與測試

本研究於分類訓練與測試階段延續謝玉珊(2007)所使用的多模激發法結合支援向量機與貝式分類器，其訓練與測試流程如圖所示。經過關鍵字選取與權重後所形成的文件向量，進入分類訓練與測試階段後，先分為訓練資料與測試資料，訓練資料分別用來訓練支援向量機+多模激發法以及貝式

分類器+多模激發法，建立分類規則，而測試資料使用這些分類規則進行分類預測產生預測結果，並檢驗分類規則的預測能力。其中支援向量機與貝式分類器均搭配多模激發法，希望追求最高績效。

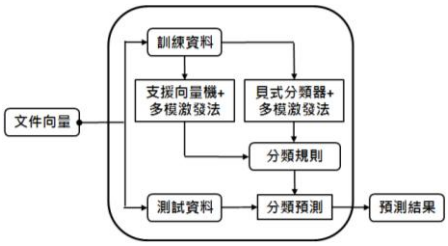

圖 6 分類訓練與測試階段

4. 實驗

實驗資料集由中南部某醫院所提供的病歷記錄，病歷資料庫中共有 17014 筆記錄，包含 1926 種 ICD9 疾病分類碼。由於部份類別的文件少於 50 篇，暫不納入本研究的實驗中。因此本研究取 50 種 ICD9 分類碼，大約百分之四十二的病歷摘要做為實驗資料集，其中包括 6664 份醫療病歷。我們將百分之八十的文件共 5379 篇做為訓練資料；其餘的百分之二十共 1285 篇做為測試資料。

過去陳紹佩(2008)提出以領域知識擴充病歷摘要的 4 種策略，與未擴充時的分類結果比較如錯誤! 找不到參照來源。，發現以支援向量機進行分類時先篩選關鍵字再擴詞，並擴充文件向量(EKV)的做法能擴充關鍵字至 3096 字，且獲得最佳績效 87.33%。為追求最高績效，結合過去研究(謝玉珊, 2007)提出能有效提升分類器績效的調適性多模激發法，如錯誤! 找不到參照來源。，發現績效提升不如想像中顯著，甚至還有下降情形，其中績效表現最佳為先擴詞再篩選關鍵字(ED)的做法，達 87.40%。故本研究希望以過去研究為基礎，提出新演算法並期望能持續提升分類器之績效。

表 1 以支援向量機之預測疾病碼之正確率

Strategy	No. of keywords	Accu.
Non-E	3083	87.09
ED	3083	86.70
EDV	3095	87.24
EK	3079	86.85
EKV	3096	87.33

表 2 以支援向量機結合調適性多模激發法之預測正確率

Strategy	Accu. of SVM	Best Accu. of AdaBoost
Non-E	87.01	87.56
ED	86.70	87.40
EDV	87.24	87.33
EK	86.85	87.25
EKV	87.33	87.01

為探討多字詞、擴詞、關鍵字選取以及分類器集成法對於績效的影響，以下分為四小節，首先探討關鍵字選取方法的比較，第二小節觀察多字詞對績效的影響程度，第三小節再加入擴詞的做法，最後將上列各小節所歸納而成的策略進行分類器集成，期望能獲得最佳的績效。

4.1 關鍵字選取方法之比較

關鍵字選取與權重值使用方式，包含卡方與 TF-IDF 的參數設定，以及後續的權重值組合實驗，最後會探討卡方與 TF-IDF 所選取關鍵字品質的比較。

首先為卡方參數之設定。過去研究使用 TF-IDF 作為關鍵字選取與權重的方法，在 TF>1 與 DF 前 10%的參數條件設定下，準確率最高可達 87.32%。本研究為探討不同關鍵字選取方法對分類績效之影響，加入了 chi-square 作為關鍵字選取與權重的方法，並透過不同的方式搭配，期望找出一組搭配獲得最佳績效。

本研究以單字詞搭配支援向量機做為基礎，分別搭配不同的關鍵字選取與權重值組合，來評估 TFIDF 與 chi-square 的優劣。首先針對 chi-square 的參數設定進行實驗，而 chi-square 有兩組不同的關鍵字選取做法如 3.2.2 節所述，分別為整體與各類別的策略，並透過不同百分比的設定選取關鍵字詞。整體的策略：針對單一關鍵字，將其在不同類別中的卡方值平均後，得到平均後的卡方值，如此計算整體關鍵字的卡方平均值，並加以排序挑選。各類別的策略：針對每一個類別的關鍵字挑選卡方值最高的關鍵字形成向量。

不同的卡方門檻值設定對準確率的影響如表。結果顯示用卡方選取關鍵字時，當關鍵字數量相近，績效也相近，但使用整體的策略所選取關鍵字品質較好。因此在關鍵字選取關於卡方的做法，採用整體的策略並挑選前 30%的關鍵字做為文件向量。

表 3 卡方檢定不同策略與門檻值比較

關鍵字選取方法	篩選門檻	向量長度	績效
CHI by class	top 5%	928	87.00
	top 10%	1911	**87.47**
	top 20%	4075	87.16
	top 100%	6136	87.39
CHI by total	top 5%	307	82.73
	top 10%	614	86.77
	top 20%	1227	87.63
	top 30%	1861	**87.86**

其次為詞頻與反轉文件頻率方法之參數設定。為了檢驗向量長度對績效的影響，我們進一步調整 TF-IDF 選取關鍵字的條件，將選取出的關鍵字詞數量降至 1900 個左右如表。結果發現在 TF>3 與 DF 前 10%的參數條件設定下，關鍵字數量降至 1956 個，準確率也提升至 87.47%，故之後的實驗便將 TFIDF 的參數設定為 TF>3 與 DF 前 10%。

表 4 不同參數設定對 TFIDF 所選取關鍵字的數量與績效影響

TFIDF 參數	TF 說明	在所有病摘中某詞彙出現的次數					
DF 說明	設定	TF>2		TF>3		TF>4	
擁有某詞彙的文章篇數少於全體文章數量的比例	DF 前 10%	2311	87.32	1956	87.47	1740	87.39
	DF 前 7%	2284	87.16	1929	87.08	1713	87.16
	DF 前 4%	2199	76.58	1844	76.81	1628	76.42
當 TF=1, DF=10%, 績效為 87.32%		向量長度	績效	向量長度	績效	向量長度	績效

我們進一步比較不同方式設定關鍵字權重值對預測績效之影響。實驗比較了 TFIDF 與卡方(CHI)兩種不同的權重方法。從表的結果可發現用卡方值(卡方平均值)也能獲得不錯的績效，但其表現不如直接使用 TFIDF 的做法，故本研究持續使用 TF-IDF 值作為文件向量中關鍵字的權重值。

表 5 使用不同關鍵字權重值對績效影響比較表

選詞方法	向量長度	權重值	績效
TFIDF	1956	tf*idf	87.47
		CHI	83.42
		tf*CHI	83.81
CHI	1861	tf*idf	87.78
		CHI	81.09
		tf*CHI	81.01

經過以上實驗觀察，可發現卡方所選取的關鍵字品質較好，因此能獲得較高的績效。我們針對卡方分析其績效提升的原因，並檢視卡方是否具備處理相似類別問題的能力。

首先使用 TF>3 與 DF 前 10%為門檻值的詞頻與反轉文件頻率搭配支援向量機進行實驗，並將分類錯誤率高於 50%的類別其錯誤矩陣列出如表，於此階段所得結果與謝玉珊(2007)年獲得結果類似，當相似類別的文件數量差異較大時，資料量較少的分類碼容易被資料量多的分類碼影響，因而將該類別的文件判斷為另一個類別。

表 6 TF-IDF 選取關鍵字搭配支援向量機預測 ICD9 分類錯誤率高於 50%的類別

ICD9	使用 TFIDF 搭配支援向量機預測 ICD9											錯誤篇數
	241.9	478.4	478.5	540	540.9	574	574.1	575	590.1	599	780.6	
241.1	10	0	0	0	0	0	0	0	0	0	0	10
574	0	0	0	0	0	1	11	4	0	0	0	15
540	0	0	0	2	15	0	0	0	0	0	0	15
599	0	0	0	0	2	0	0	0	4	5	1	7
478.5	0	6	6	0	0	0	0	0	0	0	0	6
575	0	0	0	0	1	1	2	5	0	0	1	5

表則為卡方選取關鍵字搭配支援向量機的錯誤矩陣，同樣選取分類錯誤率高於 50%的類別，發現這些錯誤率較高的類別與表的類別相同(最左欄)，而預測錯誤的類別與資料筆數也雷同，顯示當相似類別的文件數量差異較大，卡方的辨別能力並未高於詞頻與反轉文件頻率。

表 7 卡方選取關鍵字搭配支援向量機預測 ICD9 分類錯誤率高於 50%的類別

ICD9	使用卡方搭配支援向量機預測 ICD9											錯誤篇數
	155	241.9	478.4	478.5	540.9	574	574.1	575	590.1	599	780.6	
241.1	0	10	0	0	0	0	0	0	0	0	0	10
574	0	0	0	0	1	2	11	2	0	0	0	14
540	1	0	0	0	16	0	0	0	0	0	0	17
599	0	0	0	0	2	0	0	0	4	5	1	7
478.5	0	0	6	6	0	0	0	0	0	0	0	6
575	0	0	0	0	1	2	1	4	0	0	2	6

從上述的結論可發現雖然卡方辨別相似類別的能力並不突出，但最後的績效卻比 TF-IDF 來的好，表示卡方所選的關鍵字品質的確較好。但因為相似類別的關鍵字也相同，故無法在相似類別上獲得好的表現，只能選取出具有判斷差異類別能力的關鍵字。

4.2 加入多字詞之績效比較

此節探討多字詞對於支援向量機以及貝式分類器分類績效的影響，並根據前一節所得的參數設定繼續進行此節的實驗。以下分別對不同分類方法的結果進行討論。

首先使用支援向量機。根據 3.1.2 節的討論，認為應採用使用多字詞不保留單字詞的做法，而實驗結果如表,顯示當使用 TF-IDF 做為選詞方法時，多字詞不保留單字詞的作法的確有較佳的結果，與原先的推論相符，比多字詞+原單字詞的做法提升了 0.16%，但不用所有多字詞的做法績效更好，可達 87.32%，是多字詞四種組合中表現最好的，但仍不及只有單字詞時的 87.47%。

表 8 多字詞對支援向量機搭配詞頻與反轉文件頻率的影響

	多字詞個數	向量長度	績效
只有單字詞	0	1956	87.47
多字詞+原單字詞	175	2128	87.16
	494	2447	87.00
多字詞不保留單字詞	175	2040	87.32
	494	2359	87.16
註: Tfidf 選取條件：TF > 3, DF 前 10%			

使用卡方做為選詞方法的實驗結果如表，顯示了只有單字詞的績效是最好的，而多字詞的組合中，使用所有多字詞+原單字詞的做法可獲得較好的績效達到 87.63%，可發現不同的選詞方法，在多字詞組合的搭配上也應有所不同。

使用 TF-IDF 做為選詞方法時，透過詞頻與文件頻率來選擇關鍵字，發現加入選取過後的多字詞比加入所有多字詞的績效來的較好，顯示多字詞的重要性並不比單字詞來的高，也不會包含更多的資訊，所以越多的多字詞，績效並不會隨之增加，反而有下降的情況。

表 9 多字詞對支援向量機搭配卡方檢定的影響

	選詞方法	多字詞個數	向量長度	績效
只有單字詞	CHI	0	1861	87.86
多字詞+原單字詞	CHI	188	1987	87.47
	CHI	494	2293	87.63
多字詞不保留單字詞	CHI	192	1971	87.39
	CHI	494	2273	87.55
註：CHI 選取條件：卡方值前 30%				

使用卡方做為選詞的結果，在多字詞的部分呈現了向量長度越長，績效越高的情況，但也沒有比只有單字詞的結果來得好，不論是卡方或是 TFIDF 加入多字詞後，反而造成分類績效的降低。

其次，在貝式分類器的部分如表所示，發現使用卡方選詞時的績效普遍較差，而較好的多字詞組合變為多字詞+原單字詞的做法，但績效也只是與使用單字詞的 80.93%持平，並沒有提升，可見多字

詞的幫助相當有限。

表 10 多字詞對貝式分類器的影響

	選詞方法	多字詞個數	向量長度	績效
只有單字詞	TFIDF	0	1478	80.93
	CHI	0	1862	78.91
多字詞 不保留單字詞	TFIDF	132	1525	80.08
		494	1887	80.08
	CHI	192	1972	78.75
多字詞 +原單字詞	TFIDF	132	1607	80.93
		494	1969	80.86
	CHI	188	1988	78.75
註: TFIDF 參數設定：TF > 1, DF 前 12%, TF*IDF < 20 CHI：top 30% total				

多字詞之影響分析。本研究在此為了確認支援向量機績效下降的原因，進一步針對表中的結果進行分析，針對單字詞與多字詞不保留單字詞的策略，將測試資料中預測類別錯誤的資料進行探討，並整理成表，並將真實類別與預測類別兩者屬於相似類別的情況，用底線畫記。

表 11 單字詞與多字詞不保留單字詞分類錯誤資料

單字詞多分錯的資料	正確類別	誤判類別	多字詞多分錯的資料	正確類別	誤判類別
87053	733.42	715.35	117082	574.1	575
87219	575	574.1	101444	575	574.1
87281	475	478.4	115490	599	590.1
			116408	733.42	820.8
			107320	780.6	473.9

從表中發現加入多字詞的策略無論是相似類別或差異較大的類別，均多分錯了一篇，再深入分析資料中所含有的多字詞整理成表。

表 12 多字詞影響統計表

資料	多字詞	正確*	誤判**
107320	nasal obstruct	3	10
116408	femor neck	2	9
117082	acut cholecyst	1	8
101444	physic examin	2	10
	gall stone	1	21
115490	urinary tract	4	2
*正確類別中擁有此詞的文件數 **誤判類別中擁有此詞的文件數			

可發現多字詞辨識後的確會造成類別混淆的情況，由於這些多字詞被辨識後，原本所屬類別內的文件並沒有擁有相同的多字詞，此時該篇文件便會傾向於擁有較多此字詞的類別。因此雖然多字詞能更完整的表達文件向量所擁有的資訊，但也因為這些多字詞，造成分類的混淆進而影響績效。

4.3 加入擴詞之績效比較

本階段以 SVM 為主，延續上述績效表現較佳的四種組合，繼續進行擴詞部分實驗，觀察 EV 與 EDV 對於多字詞的影響如表，發現加入擴詞後績效並不穩定，尤其是 EDV 的策略，在組合一與四有大幅度的下降，但在組合二與三表現卻又比 ED 來的好，整體來說只有組合一的 ED 有提升績效的效果，但並不明顯。

表 13 擴詞策略 ED 與 EDV 對績效的影響

選詞	字詞策略	向量(擴詞)		績效	擴詞後績效
Tfidf	只有單字詞	1956	(1952)	87.47	87.55 (ED)
			(1974)		87.07 (EDV)
tfidf	多字詞 不留單字詞	2040	(2036)	87.32	87.16 (ED)
			(2071)		87.24 (EDV)
CHI	只有單字詞	1861	(1842)	87.86	87.63 (ED)
			(1974)		87.78 (EDV)
CHI	所有多字詞 +原單字詞	2293	(2301)	87.63	86.23 (ED)
			(2306)		83.11 (EDV)

擴詞之影響分析。本研究為了分析加入擴詞後導致績效下降的原因，針對單字詞與單字詞 EDV 的策略，將測試資料中預測類別錯誤的資料進行探討，並整理成表，並將真實類別與預測類別兩者屬於相似類別的情況，用底線畫記。從表中發現擴詞後相似類別的分類錯誤筆數多了 1 篇，差異較大的類別則多分錯了 5 篇，發現擴詞與多字詞同樣會造成相似類別的混淆，因為相似的類別所擴的字詞大多相同，並且擴詞還會造成差異類別的混淆情況，因此下節所進行實驗將不考慮擴詞的部分。

表 14 單字詞與擴詞後單字詞的分類錯誤資料

單字詞多分錯的資料	正確類別	誤判類別	擴詞後多分錯的資料	正確類別	誤判類別
119955	218.9	617	119047	220	617.1
107470	463	473.9	118222	550.9	590.1
109808	617.1	617	120610	574	574.1
118989	617.1	682.7	101444	575	155
			115490	599	590.1

			120681 116950 112172 107688	617 724.02 780.6 780.6	218.9 574.1 486 485

4.4 分類器集成法

總和上列各節所述，選出績效最好的組合，進行集成，觀察集成後的績效是否有所提升。表為支援向量機的集成結果，可發現除了 TFIDF+多字詞提升了將近1%的績效外，其餘都呈現下降的情況，但下降幅度亦不明顯，可顯示支援向量機的分類結果極為穩定，並且在集成之前就已經達到 87%以上的績效，集成所能進步的空間相當有限。將表集成 50 次的結果製成圖表如圖，可發現前 20 回合的績效震盪幅度較大，但高低也不會超過 2 個百分比，到了 20 次以後，績效也趨於穩定。

表 15 支援向量機集成 50 次前後之比較表

策略	集成 50 次	集成前	集成後	增減
TFIDF	26(50)	87.47	87.32	– 0.15
卡方	45(50)	87.86	87.62	– 0.14
TFIDF+多字詞	17(50)	87.32	88.17	+ 0.85
卡方+多字詞	12(50)	87.63	87.39	– 0.24

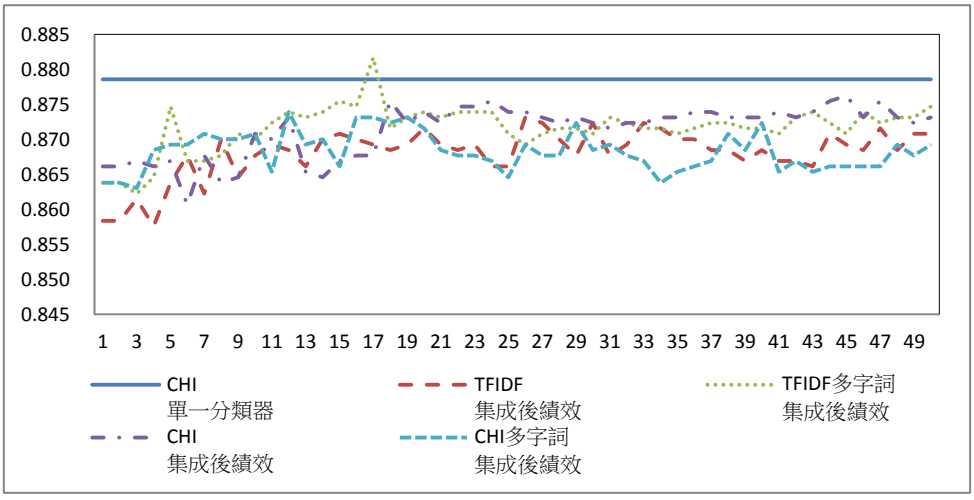

圖 7 支援向量機集成 50 回合之比較

接著針對貝式分類器進行集成，在此選擇表中加入多字詞策略後績效表現最佳的貝式分類器，並與使用原單字詞策略的貝式分類器進行 20 回合的集成，再與原本 80.93%的績效做比較如圖，可看出貝式分類器的集成效果比較顯著，搭配多字詞策略為多字詞+原單字詞的貝式分類器，績效可從原本未集成的 80.93%，於集成第 7 個分類器時提升至 84.20%。

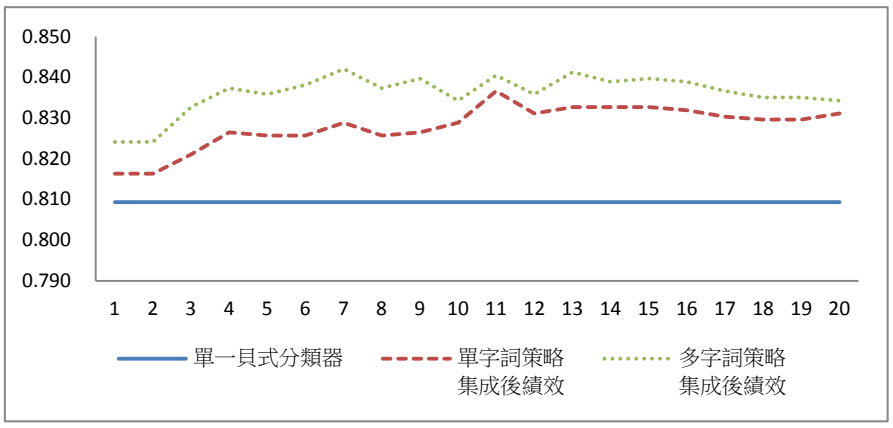

圖 8 貝式分類器集成 20 回合之比較

接著將支援向量機與貝式分類器中，單一分類器績效表現最佳的結果與集成後績效表現最佳的結果進行比較如圖。整體而言，支援向量機在績效上，表現都比貝式分類器還好，但貝式分類器可能因為進步空間較大的原因，導致集成後的績效提升較為顯著。

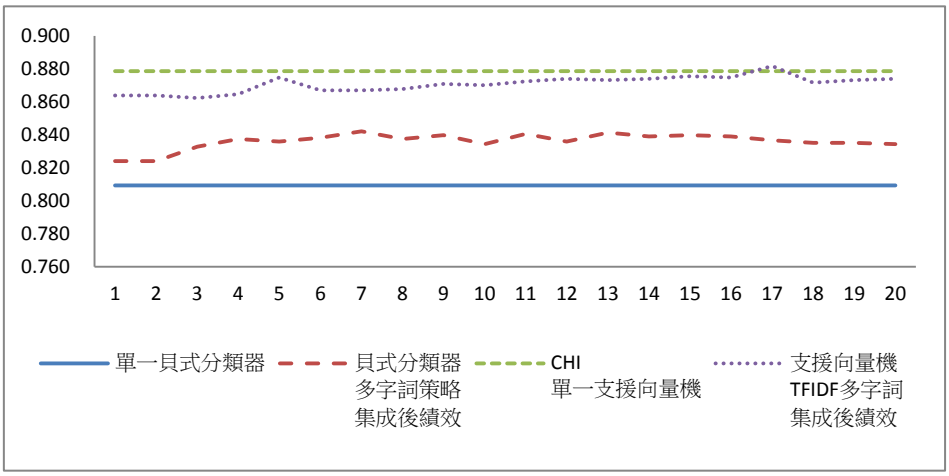

圖 9 支援向量機與貝式分類器集成之比較

集成結果之分析。分類器集成法的概念是針對此回合分類錯誤的資料於下一回合加強訓練，希望透過這樣的過程，能對容易混淆類別的資料建立正確的分類模型。以下針對前一節集成績效表現最突出的支援向量機搭配 TFIDF+多字詞的做法，分析集成過程中績效無法持續提升的原因。

表為類別分類錯誤率高於 50%之錯誤矩陣，其中 996.4 尚有 5 篇分類錯誤的文件各自分散在 5 個類別中(682,724.02,820.21,820.8,V54)，因表格篇幅關係未全部列出。

表 16 分類錯誤率高於 50%之錯誤矩陣

ICD9	支援向量機預測之 ICD9										錯誤篇數
	241.9	540	540.9	574	574.1	575.0	590.1	599.0	715.36	996.4	
241.1	9	0	0	0	0	0	0	0	0	0	9
540.0	0	3	14	0	0	0	0	0	0	0	14
575.0	0	0	1	6	0	2	0	0	0	0	7
599.0	0	0	2	0	0	0	3	5	0	0	5
574.00	0	0	0	7	8	1	0	0	0	0	9
996.4	0	0	0	0	0	0	0	0	3	8	3

觀察可發現大多的分類錯誤資料，均來自於相似類別，並且類別數量不平衡的情況，如 241.1 有 9 筆資料分類至 241.9，540.0 有 14 筆分類至 540.9，574 有 8 比分類至 574.1 等等，而進一步統計後發現這 160 筆的分類錯誤資料，在集成 50 回合的過程中被分類錯誤的次數也相當多，如表所示。

表 17 集成 50 回合分類錯誤次數統計表

次數	10~29	30~39	40~44	45~49	50	總數
筆數	6	5	10	18	121	160

從表可發現這些分類錯誤的資料有 121 筆從頭至尾就沒有分對過，其餘的分類錯誤資料也有超過一半的次數都無法被分對，並且這 121 筆沒有分對過的資料，大多來自於相似且不平衡的類別，這也突顯了分類器集成法對錯誤資料加強訓練以提升績效的功能，但卻也無法有效克服這些相似且不平衡的類別。

本研究觀察 241.1 中沒有分類正確過的文件，挑選其中兩筆觀察其關鍵字數的比較如表，並進一步探討這兩筆誤判的文件分別跟 241.1 與 241.9 中的預測正確的文件進行比較如表，發現這兩筆從 241.1 誤判至 241.9 的文件，與 241.9 中的文件所共同擁有的字詞數遠多於 241.1 的文件，而除了這兩筆文件外，其餘從沒有預測正確過的文件也有類似的情況。因此可推論這些關鍵字無法強調相似類別的差異，即使透過分類器集成法加強分類錯誤文件的訓練，仍是無法有效解決這些相似且不平衡的類別。

表 18 關鍵字數比較表

文件	關鍵字(文字數)	正確類別	預測類別
114669	36(41)	241.1	241.1
105797	16(36)	241.1	241.9
116880	16(35)	241.1	241.9
120690	16(31)	241.9	241.9

表 19 文件共同擁有字數比較表

兩文件共同擁有字詞數	誤判文件	
	105797	116880
241.1：114669	3	3
241.9：120690	12	12

本研究實驗結果匯整如表，統整各種不同策略對疾病碼預測正確率之影響。

表 20 本研究實驗結果統整

策略	文件資訊量	辨識相似類別能力	績效提升效果	說明
卡方選取關鍵字	無影響	持平	提升	使用整體的關鍵字選取策略，門檻值設為 30%
多字詞影響	增加	下降	下降	隨關鍵字選取方法不同而各有所長： TFIDF 偏好多字詞不保留單字詞 卡方偏好所有多字詞保留單字詞
擴詞影響	增加	下降	不穩定	TFIDF 部分有些許提升但不穩定 卡方加入擴詞後，績效均降低
分類器集成法	無影響	持平	提升或持平	在支援向量機部分對 TFIDF 搭配多字詞的策略有所提升，其他策略持平 貝式分類器則有明顯提升

5. 結論

本研究提出擷取多字詞以及領域知識擴詞，使用卡方檢定進行關鍵字選取並與 TF-IDF 比較，最後分類部分使用了支援向量機與貝式分類器分別搭配調適性多模激發法。實驗中本研究採用 50 個疾病分類碼的資料集，從實驗中觀察多字詞、擴詞、卡方檢定與分類器集成法所帶來的影響與績效提升情況。根據實驗結果，可提出以下幾點結論：

1. 卡方檢定所選取的關鍵字對於意義較不相同的類別，具有辨別能力，但面對相似類別時，仍無法選出區分相似類別的關鍵字，但整體而言具有比 TF-IDF 更好的績效。
2. 多字詞的擷取讓文件向量變得更加完整，但相似類別也同樣具有相似的關鍵字，反而造成績效的降低。

3. 領域知識擴詞的情況也與多字詞相同，雖然考慮了廣義詞、狹義詞以及縮寫詞的影響，但是擴充文件向量時，相似類別的關鍵字所擴字詞並不具備辨別相似類別的能力，造成分類績效的下降。
4. 從實驗結果中發現支援向量機的績效比貝式分類器都來的好，支援向量機搭配卡方最高可達 87.86%，較原本使用 TF-IDF 的最佳績效 87.47%提升了 0.39%。貝式分類器最佳能達到 80.93%，但仍不及支援向量機各種實驗組合的績效結果。

針對實驗結果，未來可朝下列方向持續進行，看是否能在進一步改善疾病碼預測績效。

1. 由於此資料集具有類別量不均衡之問題，未來可針對此問題進行處理，例如在多模激發法調整訓練資料抽樣權重時，將資料量較少類別的抽樣權重放大，讓這些類別的資料有更多的機會獲得訓練 (Sun et al., 2007)。
2. 本資料集包含若干相似類別，分類時很難將這些相似類別區分，而卡方檢定雖然可挑選出具有類別鑑別度的關鍵字，但面臨相似類別問題時，卻無法選取出具有分辨能力的字詞，因此未來選取關鍵字時，可將含有此關鍵字的類別數做為選取的依據，或者將該關鍵字分佈的類別數量納入關鍵字權重值計算之考量。
3. 可針對本研究前處理階段所使用的步驟做進一步的檢視，確認是否必要與其合理性。例如目前所使用的停用字集來自於 Bag-Of-Words Library (McCallum, 1996)，將來可尋找醫學領域專用的停用字集加以使用，於多字詞擷取與領域擴詞或能有更大的幫助。
4. 本研究使用支援向量機與貝式分類器做為分類模型，將來也可使用決策樹等其他類型的分類模型，並觀察其預測類別的能力。
5. 國際疾病分類碼屬於階層式架構，若能結合其階層式架構，於不同階層採取適合該階層的處理方式，例如於第一階段將整體資料分類至疾病碼架構中較上層的類別，接著進行第二階段，處理下層相似類別的分類，如此做法或許也能解決相似類別的問題。
6. 目前所處理的疾病碼分類問題均只有預測單一類別，但一份病歷摘要可能擁有多個疾病碼，在本研究中所使用的資料集皆為篩選過後，只擁有單一疾病碼的病歷摘要，其餘還有許多擁有兩碼以上的病歷摘要尚未處理，因此將來也可嘗試解決這些多類別的分類問題。

致謝

本研究承蒙國科會研究計畫編號 NSC 98-2410-H-224-010-MY2 輔助，特此致謝。

參考文獻

[1] 余金燕、潘德樑，2003，「**疾病分類實務**」，地點：合計圖書出版社。

[2] 林偉彥、黃一平、楊家韋、林昕彥，2006，「**由病歷摘要判斷疾病碼及自動化線上病摘拼字修正之研究**」，國立雲林科技大學資訊管理系大學部實務專題報告。

[3] 陳紹佩，2008，「**以領域用詞擴充病摘協助國際疾病碼判斷**」，國立雲林科技大學資訊管理系碩士論文。

[4] 廖艾貞，2008，「**兩階段處理類別數量不均問題—以判斷國際疾病分類碼為例**」，國立雲林科技大學資訊管理系碩士論文。

[5] 謝玉珊，2007，「**由病歷摘要判斷國際疾病分類碼**」，國立雲林科技大學資訊管理系碩士論文。

[6] 韓歆儀，2004，「**應用兩階段分類法提升 SVM 法之分類準確率**」，國立成功大學工業與資訊管

理研究所碩士論文。

[7] A. D. March, E. J. M. Laur ía, and J. Lantos. 2004. Automated ICD9-CM coding employing Bayesian machine learning: a preliminary exploration. in Proceedings of SIS2004 (Simposio de Informática y Salud - SADIO) 33rd Conference on Computer Science & Operational Research.

[8] A. McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/mccallum/bow.

[9] B. Ribeiro-Neto, A. H. F. Laender and L. R. S. de Lima. 2001. An experimental study in automatically categorizing medical documents. Journal of the American Sociaty for Information Science and Technology, 52(5), pp. 391-401.

[10] C. D. Manning, P. Raghavan, and H. Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press.

[11] G. Forman. 2003. An extensive empirical study of feature selection metrics for text classification. ACM JMLR, 3, pp. 1289-1305.

[12] G. Salton, and C. Buckley. 1988. Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management: an International Journal, 24(5), pp. 513-523.

[13] L. S. Larkey, and W. B. Croft. 1996. Combining Classifiers in Text Categorization. in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 289-297.

[14] Medical Subject Headings. Retrieved on November, 20, 2008 from http://en.wikipedia.org/wiki/MeSH.

[15] P. Franz, A. Zaiss, S. Schulz, U. Hahn, and R. Klar. 2000. Automated coding of diagnoses: Three methods compared. in Proceedings of Artificial Intelligence in Medicine Fall Symposium, pp. 250-254.

[16] S. Bedingfield and K. Smith-Miles. 2006. Two stage partial classification for inconsistent and imbalanced classes. IEEE International Conference on Information and Automation, pp. 167-171.

[17] T. M. Tu, C. H. Chen, J. L. Wu, and C. I. Chang. 1998. A Fast Two-Stage Classification Method for High-Dimensional Remote Sensing Data. IEEE Transactions On Geoscience And Remote Sensing, 36(1), January.

[18] V. N. Vapnik. (1995). The Nature of Statistical Learning Theory, Springer-Verlag, NY, USA.

[19] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. 2007. Knowledge-intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature. in Proceedings of Annual ACM SIGIR, pp. 655-662.

[20] Y. Freund, and R. E. Schapire. 1996. Experiments with a New Boosting Algorithm. Machine Learning: in Proceedings of the Thirteenth International Conference, pp. 148-156.

[21] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition 40, pp. 3358–3378.

[22] Y. Yang, and J. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. in Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412-420.

[23] Z. Zheng, X. Wu, and S. Rohini. 2004. Feature Selection for Text Categorization on Imbalanced Data. ACM SIGKDD, 6(1), pp. 80-89.