基於階層式類神經網路之自動新聞文件分類方法

陳彥呈　蔣榮先
國立成功大學資訊工程系
chenyc@ismp.csie.ncku.edu.tw
jchiang@mail.ncku.edu.tw

摘要

文件分類是一項決定一篇文件是否屬於一個或多個已事先定義好的類別之工作，而自動化分類則可以有效地幫助分類的處理。在本篇論文中，我們提出了一個以階層混合式的專家模組(hierarchical mixture of experts model)為基礎的文件分類方法。這個模組使用了分割－克服原理(divide-and-conquer principle)，在一個事先定義好的階層架構下定義較小的分類問題，而最後的分類器則是使用類神經網路中的倒傳遞網路來完成分類機制。另外，在特徵選取(feature selection)上，我們也做了一些有別於傳統方法的改變。最後，我們以部份路透社(Reuters-21578)的新聞性文件做為測試資料，實驗結果顯示我們所提出的方法能有效地改善文件分類的正確率。

1. 緒論

近幾年來，隨著網路技術不斷地進步，有用的資訊也相對地大量成長中。雖然網路上舉手可得的資訊方便人們對資訊的取得與傳遞，但是當網路資訊量愈來愈大時，如何有效、且快速地取得有用的資訊，便成為非常重要的事情。此時，文件分類(text categorization)技術，即透過演算法分析一電子文件後，將其分配(assign)給一或多個類別(categories)，便扮演著其中重要的角色。

傳統的文件分類工作都是由某個領域的人類專家(human experts in domain)所完成。但是，隨著文件數量快速地成長，對於專家而言，這樣的工作就變得更困難了。在這種情況下，文件的自動分類就顯得更加重要了。

很多在做文件分類的方法中，例如使用規則庫(rule-based)、知識庫(knowledge-based)、或樣本庫(instance-based)．．．等，都是依賴大量的樣本來決定和文件有關的規則或知識。一般而言，這些樣本集合必須由那些對應用領域有深入認識的專家來訂定與建立，也因此，這些方法常常因為相關樣本建立得不足或不完全，使得規則或知識也就相對地不齊全，因此，就無法對文件做全盤性的樣本比對，以致於造成了分類上的困難。

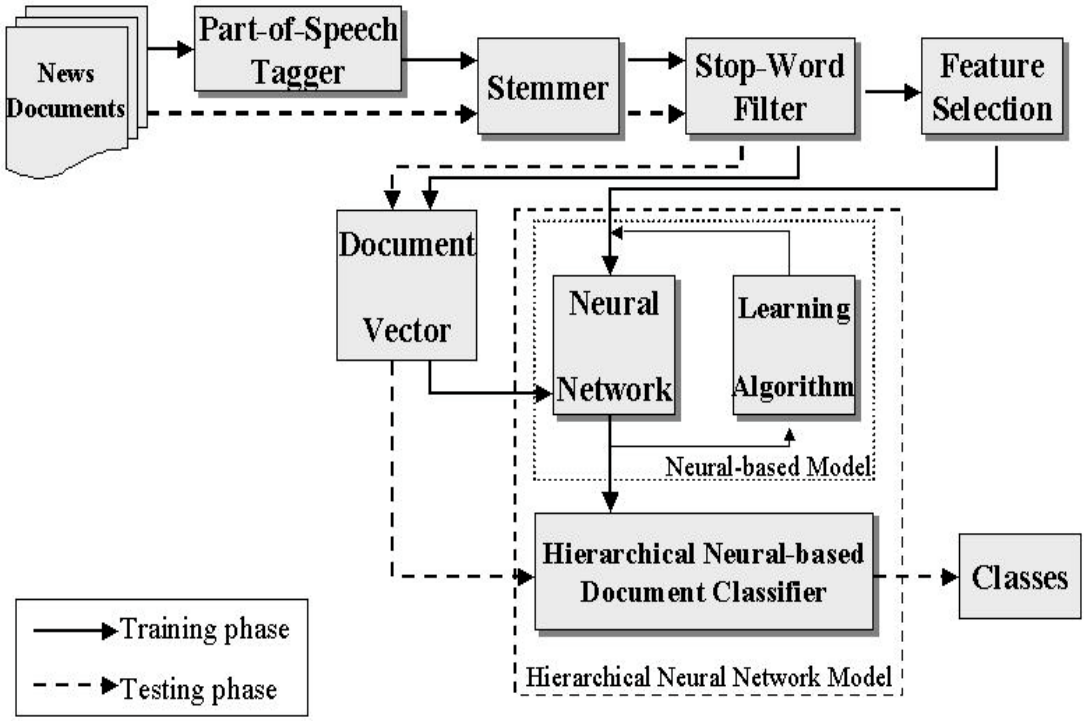
在本篇論文中，主要的動機在於改善目前文件分類的方法，我們不以關鍵字的存在否來決定一篇文件應屬於那一個或多個類別。進一步的，我們採用以類神經網路為基礎的階層式架構的機器學習的方法來決定文件的歸屬。而且，經由這樣學習的方法，可以使文件分類系統更容易地應用到其他的領域。

本篇論文除了緒論外，第二節將介紹我們所提的階層式模組，第三節將介紹特徵及訓練樣本集的選取，第四節則針對我們所使用的路透社新聞性資料集所做的一些自動化文件分類實驗的結果與分析。最後，我們為本篇論文提出總結。

2. 階層式模組

圖一所示，是我們所提出的自動化文件分類的完整模組。一個文件分類系統(text categorization system)的主要工作流程，是先用一組訓練樣本集來訓練系統中的文件分類器；然後再藉由已訓練好的分類器對測試樣本中的新文件做自動化分類的動作。在圖一的實線箭頭部份是整個文件分類的詳細訓練過程，首先決定一組已由專家分類好的樣本集，從此樣本集中，經過一連串的前處理程序後，選擇一組最能代表及識別(identification)此類別的特徵集(feature set)。並以向量方式表示之，如此就可得到一個以特徵向量表示的樣本組，而在階層式類神經網路模組中，主要是希望能透過每一個樣本組來訓練其所屬的分類器，使其能很正確地將每一個樣本分到正確的類別去。經過一連串的反覆學習後，我們得到一組已訓練好、具有相當辨識程度的分類器，以供測試階段時使用。

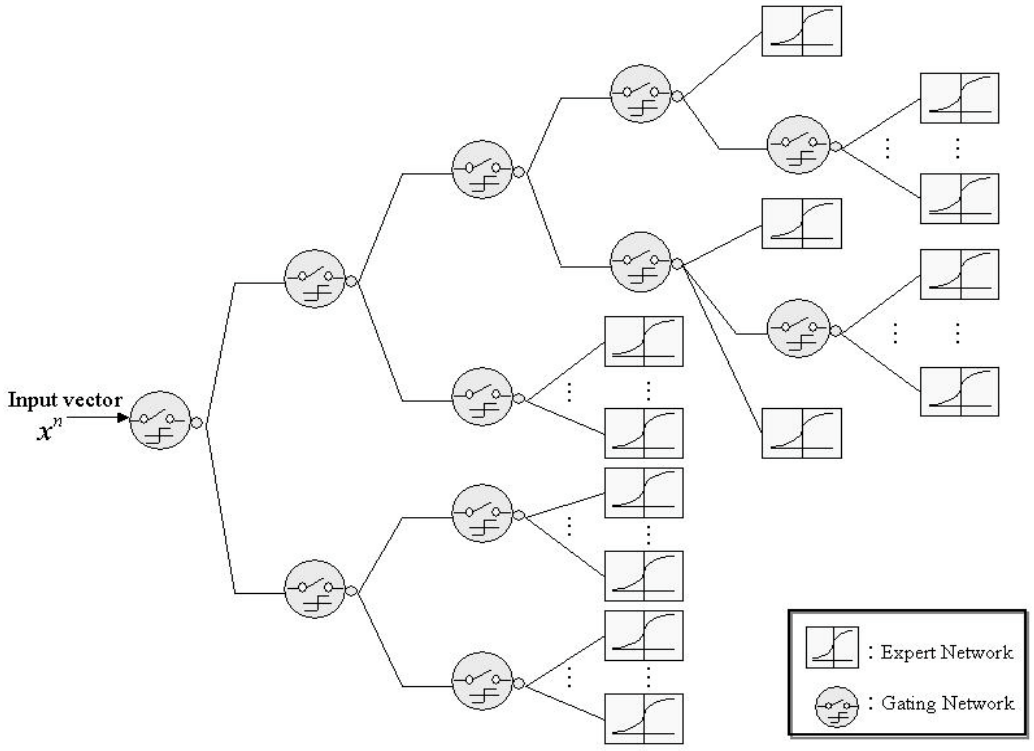
圖一中的虛線箭頭部份則是整個測試流程，起初也是將一新文件經過一連串的前序處理後，再依特徵集轉換成向量形式，最後透過階層式類神經網路模組，以決定新文件所屬的類別。



圖一　本論文所提出之自動化文件分類模組

在圖一用虛線方塊所圍成的，就是我們所提出的階層式類神經網路模組，其詳細的架構如圖二，此模組的主要的靈感是來自於 Jordan 和 Jacobs[1993]所提出的階層式混合的專家模型(hierarchical mixture of experts , HME model)。HME 模式是以分割-克服原理(divide-and-conquer principle)為基礎，其主要的想法是將一個大問題分割成若干個容易解決的小問題，然後再結合這些小問題的解答，以得到一般化的解答。而在分類一個減少範圍上(reduced domain)，HME 模型是經由將輸入空間(input space)劃分成一巢狀、順序的區域，然後訓練特定的較小分類器，以此求得一個分類問題的答案。HME 模型包含兩個基本的元件：閘門網路(gating networks)和專家網路(expert networks)。這些元件的結構類似於樹狀結構

(tree structure)，其內部節點是閘門、樹葉節點是專家。圖二就是我們提出的一個五層的階層式模組架構圖。



圖二　本論文所提出之階層架構圖

在我們的模型中，每個閘門所表示的是一份文件的一般概念，假如文件中包含著所表示的概念，則網路的輸出是 1，否則為 0。而專家所表示的是特定的類別[Ruiz, 1999]。所有的文件都以向量表示之。整個分類工作是由根節點(root node)開始，假如閘門的輸出值為真，則第二層的節點都會被啟動，如此的程序持續至它到達一個樹葉節點。

對於閘門和專家網路，由於類神經網路中的倒傳遞網路(back-propagation，BP Network)具有學習正確率高、理論簡明[Zurada, 1992]。因此，我們決定使用三層的倒傳遞類神經網路，其輸入層包含了 N 個特徵，隱藏層包含了(2N/3)個節點，而輸出層為單一個節點。而在神經元的架構中，我們使用 S 形函數(sigmoid function)作為轉換函數。此函數具有微分容易的優點，可配合降梯度法則來調整

神經元間的權重，此函數當自變數趨向正負無限大時，函數值趨近於常數，其函數值域在[0,1]之間。

3. 特徵選取和訓練資料集選取

一般而言，文件大部份都是人們以自然語言所書寫而成的，這些文件中的文字所要表達的，則是人們的想法與意見。我們相信在這些想法與意見中，主要是由一些重要的觀念所組成的，而我們認為文字中的名詞字詞最能表達一個觀念的形成。因此，在特徵選取過程中，我們首先使用了由 Eric Brill [1993]所提出的詞性分析器(part-of-speech tagger)為每個英文字標示其詞性資訊，然後選擇名詞集合的關鍵字詞。接下來則必須使用 stop word 過濾器模組，將上述所選取標示名詞的關建字詞中，過濾一些不足以代表文件本身特性字詞，以避免在接下來的處理過程中，引入太多不必要的雜訊(noise)。在做完 stop word 的處理後，其他剩下的名詞字集還不能算是最後想要的特徵集。因為根據人們的寫作習慣，對於那些出現頻率太過於頻繁或過於貧乏的字，通常都沒有太大的義意及重要性，對於符合這兩種情形的字集，我們可以經由字詞頻率－反文件頻率(term frequency and inverse document frequency , TFIDF)的分析而將其過濾掉，如此處理後所剩下的部份，我們稱之為特徵字詞(feature words)，這些字詞才是最重要的精華。

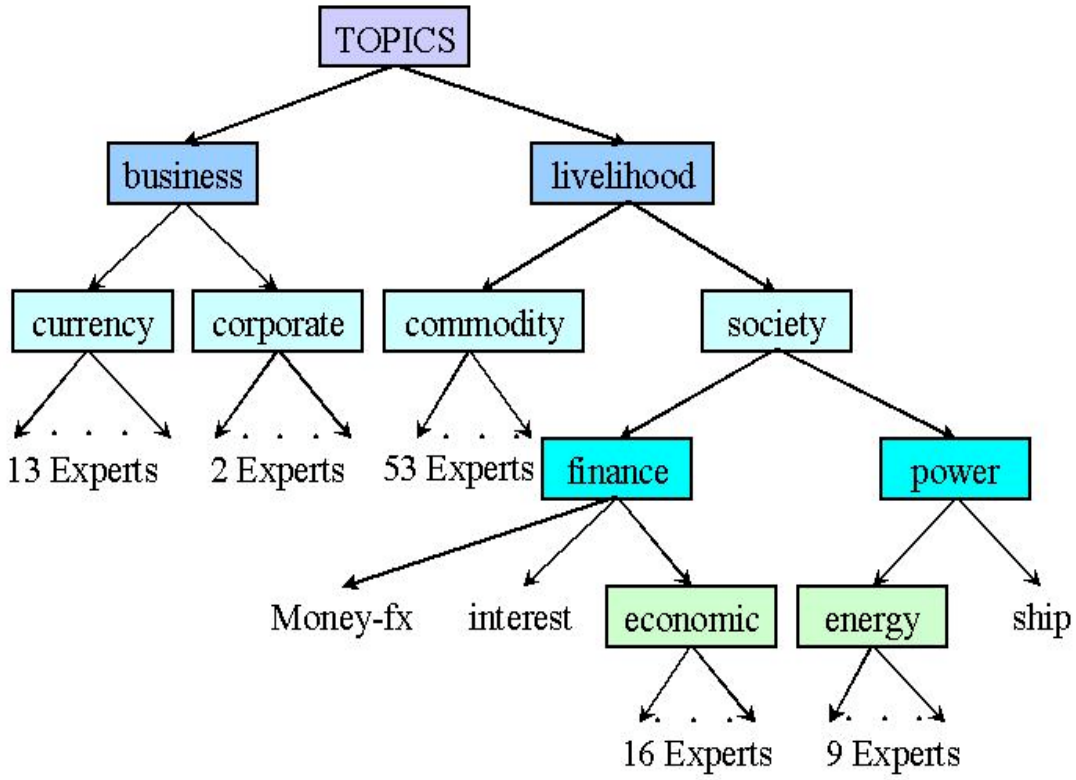
此外，在訓練分類器方面，對於同一類別的正負訓練樣本選取上，若兩者的選取差距過大，造成過度地不平均，很有可能會造成分類器在學習上的誤差，以致於造成最後分類上的錯誤。因此，對於訓練樣本的選取也是不可忽視的工作之一。在這一方面，我們採用了由 Ruiz [1999]所提出的"類別區(category zone)"的概念來選取訓練樣本集，其基本做法為選取屬於此類別的文件為正樣本，而選取最靠近此類別、卻不屬於此類別的文件做為負樣本。這樣的概念，最早是來自於 Singhal 等人[1997]為文件繞送(text routing)所提出來的想法。

4. 實驗結果與分析

4.1 資料集

本實驗所使用的測試資料集，是由 David D. Lewis [1996]和路透社人員所共同整理而成的路透社新聞性文件－Reuters-21578。在這個資料集中，總共包含了 21578 篇文件，分為五大類別(EXCHANGES, ORGS, PEOPLE, PLACES, TOPICS)，我們只拿五大類別中的 TOPICS 類別做為實驗之用。在這個類別中，包含了 135 個子類別，為了階層式模組的訓練及測試的需要，我們只選擇包含三篇文件以上的子類別做為測試類別。最後，我們使用了 96 個子類別、10555 篇文件作為實驗用的資料集。

對於 96 個子類別的階層架構，我們使用了 [陳彥呈, 2000] 所提出的架構圖，其架構如圖三。它基本的建構概念是依據文件在各類別之間的分佈來分析類別間的關連性所建立起來的。



圖三　在 TOPICS 中，96 個子類別的階層式架構圖

4.2 結果

在評估我們的模組效能之前，我們要先針對我們的模組提出兩個問題：1)在同樣使用類神經網路方法的情況下，有使用階層式架構和沒有使用階層式架構的效能差異。2)我們所提出的階層式架構和目前幾個有名的分類方法比較，其優劣為何？

在本實驗中所使用的評估方法，為在資訊擷取中最常被大家使用的正確率(precision)、召回率(recall)和 F_1 評估方法。

表格一所示，是我們所提出的階層式方法和沒有使用階層架構的方法的比較[Manevitz, 2000]，由表格中，我們可以很清楚地看出來，我們所提出的階層式方法，大大地提昇了分類的正確性。

表格一　使用階層式架構 V.S. 沒有使用階層式架構的平均效能比較

Class	NN (Hadamard)			NN (Frequency)			Proposed approach		
	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1
Earn	0.800	0.763	0.781	0.805	0.282	0.418	***0.837***	***0.851***	***0.844***
Acq	0.598	0.483	0.534	0.363	0.332	0.347	***0.850***	***0.844***	***0.847***
Money-fx	0.641	0.470	0.542	0.420	0.546	0.475	***0.826***	***0.841***	***0.833***
Grain	0.394	0.439	0.415	0.355	0.408	0.379	***0.831***	***0.862***	***0.846***
Crude	0.505	0.573	0.537	0.410	0.566	0.476	***0.853***	***0.867***	***0.860***
Trade	0.600	0.547	0.573	0.513	0.561	0.536	***0.842***	***0.847***	***0.845***
Interest	0.416	0.616	0.496	0.405	0.583	0.478	***0.836***	***0.899***	***0.866***
Ship	0.328	0.492	0.393	0.400	0.376	0.388	***0.827***	***0.865***	***0.846***
Wheat	0.446	0.588	0.507	0.430	0.400	0.414	***0.839***	***0.886***	***0.862***
Corn	0.451	0.236	0.310	0.434	0.247	0.315	***0.830***	***0.892***	***0.860***
Average (top 10)	0.517	0.520	0.508	0.453	0.430	0.422	***0.837***	***0.866***	***0.851***
Average (all)							0.867	0.892	0.879

表格二所示，則是我們所提出的方法和兩個著名的分類方法的比較－決策樹(decision tree) [Weiss, 1999]和 k-NN 方法[Aas, 1999]。由表格中，我們可以知道，我們所提出的模組在某些類別上，其效能比其他兩種方法好。而在正確率及召回率上的成長，也比其他兩種方法要來得穩定。

表格二　我們所提出的階層式分類模組和決策樹及 k-NN 之比較

Class	k-NN (k=30)			Decision Tree			Proposed approach		
	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1
Earn	0.950	0.920	0.935	***0.953***	***0.966***	***0.978***	0.837	0.851	0.844
Acq	***1.000***	0.910	0.953	0.961	***0.953***	***0.957***	0.850	0.844	0.847
Money-fx	***0.920***	0.650	0.762	0.771	0.758	0.764	0.826	***0.841***	***0.833***
Grain	***0.960***	0.700	0.810	0.953	***0.916***	***0.934***	0.831	0.862	0.846
Crude	0.820	0.750	0.783	***0.926***	0.850	***0.886***	0.853	***0.867***	0.860
Trade	***0.890***	0.660	0.758	0.812	0.704	0.754	0.842	***0.847***	***0.845***
Interest	0.800	0.710	0.752	0.649	***0.933***	0.766	***0.836***	0.899	***0.866***
Ship	***0.850***	0.770	0.808	0.769	0.861	0.812	0.827	***0.865***	***0.846***
Wheat	0.690	0.730	0.709	***0.972***	0.831	***0.894***	0.839	***0.886***	0.862
Corn	0.350	0.760	0.479	***0.982***	0.821	***0.894***	0.830	***0.892***	0.860
Average (top 10)	0.823	0.756	0.788	***0.879***	***0.879***	***0.879***	0.837	0.866	0.851
Average (all)	0.792	0.818	0.805	***0.878***	0.878	0.878	0.867	***0.892***	***0.879***

5. 結論

本論文主要是在文件分類上，提出一個結合機器學習方法的階層式模組，並且使用了詞性分析器，以擷取出真正有意義的特徵字詞。最後，我們將我們的方法和其他方法做比較。從實驗的結果我們得知，我們所提出的階層式模組確實能提高正確率和召回率。

本論文的未來研究方向主要有特徵的選取，在使用類神經網路做為分類模組時，特徵選取的好壞會直接影響到分類的正確性。此外，我們也希望在類別區上尋求其他的方法，以期能求得更合適的訓練樣本集。

參考文獻

Aas, K., and Eikvil, L., "Text categorization: A Survey", *Report No. 941, Norwegian Computing Center*, June, 1999. ISBN 82-539-0425-8

Eric Brill, ″A Corpus-based Approach to Language Learning″, *phD Dissertation, University of Pennsylvania*, 1993.

Jordan, M. I., and Jacobs, R. A., ″Hierarchical Mixtures of Experts and the EM algorithm″, *Technical Reports A. I. Memo No. 1440, Massachusetts Institute of Technology*, 1993

Koller, D., and Sahami, M., ″Hierarchical Classifying Documents Using very few Words″, in *ICML-1997: Proceedings of the 14th International Conference on Machine Learning*, 1997, pages 170-178.

Lewis, D. D., ″Reuters-21578 Text Categorization Test Collection Distribution″, in *AT&T Labs – Research*, 1996.

Manevitz, L. M., and Yousef, M., ″Document classification on neural networks using only positive examples″, *ACM SIGIR*, 2000, pages 304-306.

Ng, H. T., Goh, W. B., and Low, K. L., ″Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization″, in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, pages 67-73.

Ruiz, M. E., and Srinivasan, P., ″Combining Machine Learning and Hierarchical Indexing Structure for Text Categorization″, in *Proceedings of the 10th ASIS/SIGCR Workshop on Classification Research*, 1999.

Ruiz, M. E. and Srinivasan, P., ″Hierarchical Neural Networks for Text Categorization″, in *Proceedings of the 22nd ACM SIGIR International Conference on Information Retrieval*, 1999, pages 281-282.

Singhal, A., Mitra, M., and Buckley, C., ″Learning Routing Queries in a Query Zone″, in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, pages 25-32.

Weiss, S.M., Apte, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T., Hampp, T., "Maximizing text-mining performance", *IEEE Intelligent Systems*, Volume: 14 Issue: 4, July-Aug, 1999, pages 63-69.

Zurada, Jacek M., "Introduction to Artificial Neural Systems", *West Publishing Company, USA,* 1992.

陳彥呈, 蔣榮先, ″基於階層式類神經網路之自動文件分類模式″, 第八屆模糊理論及其應用會議, 2000.