The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 194-203

以多重表示選擇文章分類的樣本
Using Multiple Representations to Select Instances for Text Classification

陳耀輝 Yao-Hui Chen
國立嘉義大學資訊工程學系
ychen@mail.ncyu.edu.tw

王志偉 Jhih-Wei Wang
國立嘉義大學資訊工程學系
s0983042@mail.ncyu.edu.tw

摘要

以機器學習方式產生分類模型是一種常見將文章自動分類的技術，但若有標記不精確的訓練資料，就可能得到錯誤的分類結果，而樣本選擇能藉由減少訓練資料集的大小來改善這個問題。本論文提出以不同的文章表示法建立多個分類模型，並據以分析訓練文章是否標記錯誤，然後刪除標記錯誤的文章，以提升訓練資料的品質。我們以電影評論的文字資料集作為實驗語料庫，透過詞彙頻率、主題模型、及詞彙向量三種表示法分別建立文章分類模型，並依據分類結果選擇樣本。實驗結果顯示，本論文所提出來的樣本選擇方法可以提升文章分類的準確度。

關鍵詞：文章分類、樣本選擇、支援向量機。

一、緒論

將文章自動分類到事先標記好的類別裡是很多工作的基礎，所以改善文章分類方法的準確度便成為一個重要的研究題目。機器學習是一種從已知的訓練資料中自動分析出所蘊含的規律，並利用這些規律對未知資料進行預測的技術 [1] [14]，而使用機器學習來訓練文章分類器，是一種常見的文章分類方法 [10] [22]。

樣本選擇 (instance selection) 也被稱為資料縮減 (data reduction)，其主要的功能是從已被標記類別的資料集中，利用過濾雜訊或刪除不相關的資料來減少資料集的大小，使得文章分類演算法可以有效率地執行。刪去不相關的資料可以讓資料的類別更為集中，以滿足針對特定領域資料的文章分類演算法需求。過濾雜訊與刪除不相關的資料也可以提高訓練資料集的品質，而有了高品質的訓練資料，就能提高文章分類器的分類準確率

[12] [25]。

樣本選擇演算法在挑選資料時，需要考慮文章與文章之間的關係，如果不事先建立文章與文章的關係對照表，便很難得知文章與文章之間的關係。向量空間模型 (Vector Space Model) 是一種常被用於資料檢索的模型，它將語料庫所有的詞彙作為空間的維度，並將文章以其詞彙的頻率表達成空間中的向量，讓我們可以利用向量之間的關係計算文章跟文章之間的相似程度，或研究文章與詞彙之間的各種關係 [21]。使用向量空間模型將文章轉換成為向量後，我們可以得到文章與文章之間的關係，並利用這些關係建立文章分類器。

支援向量機 (Support Vector Machine, SVM) 是一種產生分類模型的演算法，它的目的是要找到一個分割不同類別資料的超平面。由於分類效果不錯，近年來 SVM 常被使用在各個領域 [17]。我們首先使用 SVM 分析訓練資料集以產生分類器，再使用該分類器將訓練資料集重新分類。原則上這個分類器應該能正確判定訓練資料的類別，但由於分類演算法的限制，在判定為同一類別的資料群中可能找到一些不屬於該類別的錯誤資料，而刪除這些錯誤資料，將可以提高訓練資料的品質 [5]。

本研究希望利用原始向量空間的詞彙頻率、由潛在狄氏配置 (Latent Dirichlet Allocation, LDA) 產生的主題模型 [2]、及由 Word2vec 產生的詞彙向量 [13]，將文章轉換成三種不同的表示法，再利用 SVM 在不同的表示法中將文章分類，最後找出並刪除可能標記錯誤的文章，以提高訓練資料的品質，達到改善文章分類準確度的目的。

本論文其餘部分的架構如下：第二節敘述樣本選擇與樣本選擇的相關研究，包含早期的樣本選擇方法及和我們方法相關的研究；第三節介紹所使用的工具；第四節定義問題以及介紹我們的方法流程；第五節說明實驗過程與結果；第六節提出結論及未來改善的方向。

二、相關研究

特徵選擇 [11] 以及樣本選擇是兩種常見改善分類演算法的技術，而本論文主要是針對樣本選擇的方法進行研究。在相關文獻中顯示，樣本選擇演算法將會對於機器學習

的分類模型有顯著的影響 [20]。在本節中我們會先介紹樣本選擇的概念，然後說明早期的樣本選擇方法，再比較一些相關的樣本選擇方法，最後介紹動態學習的概念。

(一) 樣本選擇

為了分類新的文章，需要給予機器學習演算法一些預先被標籤的資料集以產生分類器，這些資料集便被稱為訓練資料集 T；而要被進行分類的新資料集，則被稱為測試資料集 U。在現實情況下，T 往往會被混入一些對分類無用的資料 (例如：多餘資料或者錯誤資料)，導致產生不正確的分類器。一個子集 $S \subset T$ 是經過樣本選擇的演算法所選取出來不包含 T 中對分類無用資料的集合。為了比較不同樣本選擇演算法之間的好壞，會使用 U 來對 S 所產生出來的分類器進行測試 [18] [25]。

(二) 早期的樣本選擇方法

從 1968 年起樣本選擇方法被提出 [26]；有學者分別依照評估標準 (evaluation-type) [9] [18] 和應用類型 (application-type) [25] [26] 將樣本選擇方法進行分類。

1. 評估標準

評估選擇樣本的標準主要分成兩種，一種是包裝 (wrapper) 法，另一種是過濾 (filter) 法。包裝法的優點找到的樣本子集對於機器學習有很好的分類效果，而缺點因為過程中需要反覆驗證，時間複雜度很高。過濾法的優點是能快速地找到粗略的子集，而缺點是產生的子集還是包含許多無用的資訊，所以對於機器學習法的幫助較小。

2. 應用類型

樣本選擇的應用類型可分為三類，第一類為雜訊過濾法 (noise filters)，第二類為壓縮演算法 (condensation algorithms)，第三類為雛型搜尋演算法 (prototype searching algorithms)。雜訊過濾法的優點是訓練出來的分類模型對於測試資料可以有更佳的分辨率，缺點則是需要反覆對資料進行比較，需要花費較多的時間，而且保留許多同一類別的資料，導致儲存空間的浪費。壓縮演算法的優點是去除了大量資料，在訓練時可以用較少的資源進行訓練，缺點則是決策邊界較為模糊，對於測試資料分辨率的提升較無幫助。雛型搜尋演算法的優點是可以有效地減少資料集的大小且可以提升機器學習的分類效果，而缺點則是要比較資料之間是否類似，有較高的時間複雜度。

(三) 相關的樣本選擇方法

前面提到的方法大多是從 K-NN 的模型延伸及改良所產生出來的方法，以下將要介紹兩種利用 SVM 的概念所延伸出來的方法。SVOIS [25] 方法可以有效地刪去一些混雜在一起的資料，提高機器學習的分類效果，但是針對不同的資料集都必須重新挑選 IS_Range 跟 Margin 這兩個參數。這個方法也沒有考慮錯誤資料的處理，使得回歸模型可能會受到錯誤資料的影響，以致未能顯著地提升分類的正確率。而 SVM-IS [6] 方法在類別資料混合較為複雜的資料集中，需要使用核心函數將訓練資料投影到更複雜的空間上才有辦法進行切割，而越複雜的核心函數會需要更多的支援向量進行處理，導致他們所減少的樣本數量較少，而且使用更複雜的核心函數可能產生過度擬合 (overfitting) 的問題，所以無法有效提高分類的正確率。

(四) 動態學習 (Active Learning)

動態學習 (又被稱為 query learning 或 optimal experimental design) 是一種被應用於人工智慧的半監督式機器學習方法，其想法是在學習一些已被標記的訓練資料後可以自行標記新的資料，然後用這些新標記的資料來擴充訓練資料集 [15]。Query By Committee (QBC) 方法則是將標記的訓練資料分別用不同的學習演算法產生分類器，再將未標記的資料分別於不同的分類器中進行分類，然後根據分類的結果進行投票。若大多數的分類器都給予相同的類別，則將這個類別給予這個未標記的資料，然後將被標記後的資料納入訓練資料中重新進行訓練分類器，並對未標記的資料進行標記 [7] [23]。前面提到的 SVOIS 與 SVM-IS 兩種方法並沒有處理錯誤資料，而我們提出的方法則會結合 QBC 的概念產生多個分類模型，先進行錯誤資料的辨識並藉由去除錯誤資料來優化訓練資料的品質，讓之後的分類器可以有較好的分類正確率。

三、背景知識

本節介紹我們會使用到的相關知識與工具。

(一) 支援向量機 (SVM)

支援向量機 [5] 是一種被廣泛應用於分類議題上的監督式演算法，主要的概念是

建構一個超平面使得不同類別的資料可以被區隔，以達成分類資料的目的，主要是用來分類兩個不同類別的資料。但在許多狀況無法以簡單的一條直線來分割資料，所以用核心函數 (kernel functions) 將原始資料投影到另外一個空間再進行資料分類。

(二) 潛在狄式配置 (LDA)

潛在狄式配置 [2] 是一個以 bag of words 假設為前提的非監督式演算法，其目的是要建立主題模型。它分析並統計文章內的詞彙，並根據這些統計的資訊來判斷該文章含有哪些主題，以及主題在文章中的比例為何。其主要的想法是文章可以用隨機的潛在主題所代表，而這些潛在主題則是由不同詞彙的分佈機率所組成 [24]。在 LDA 的模型裡，所有文章都可以表示成主題機率分布，而透過分析這些主題資訊，便能讓文章與文章之間產生新的關聯性。

(三) 詞彙向量

Word2vec 是一種用來捕捉詞彙或片語意思並同時壓縮資料大小的技術，並可以分成 Continuous Bag Of Words (CBOW) 和 Skip-gram [13] 兩種不同的方法。CBOW 是希望藉由上下文來預測當前詞彙的機率，而 Skip-gram 則相反，是藉由當前詞彙來預測上下文的機率，且這兩種方法都是以類神經網路作為它們的分類演算法。開始時每個詞彙都是由一個隨機 N 維的向量構成，經過 CBOW 或 Skip-gram 的計算後，便能獲得每個詞彙最佳的向量結果。有了這些向量結果後就能以上下文的訊息發現詞彙跟詞彙之間的關係，而這些詞彙也可以代替 bag of words 來預測未知詞彙的情況。但是 Word2vec 的維度只能代表詞彙，所以需要結合文章中的所有詞彙才能代表文章。有研究是利用平均所有詞彙的向量作為文章的向量，然後就能比較文章跟文章之間向量的相似性，判斷文章的相似程度 [3]。

四、方法

本節一開始會先對研究的問題做一個完整描述，接著說明研究方法及流程。

(一) 問題定義

由於混入錯誤文章的訓練資料集將會降低分類器的分類正確率，而先前以 SVM

為核心的方法，並沒有去處理錯誤文章，使得他們所產生的分類器效果較差。因此我們利用詞彙頻率、LDA、Word2vec 分別得到文章與文章之間關係，並生成不同的模型，然後利用 QBC 的概念從不同的角度找出錯誤文章並予以刪除，希望可以藉此提高訓練資料集的品質。

(二) 方法流程

本研究的方法可分成資料前處理、訓練、及投票與測試三個階段。

1. 資料前處理

將所有文章去除 html 標記、數字、縮寫字等雜訊，然後計算每個詞彙出現的次數，去除出現次數最高的前 200 個詞彙以及只出現於一篇文章中的詞彙，收集剩餘的詞彙建立詞彙表，藉助詞彙表統一三種模型所輸入的文章。詞彙頻率模型的輸入格式是以詞彙表的詞彙當作特徵、詞彙的頻率當作特徵值；LDA 模型的輸入格式是第一行為總文章數量，第二行開始為去除沒出現在詞彙表中詞彙的原始文章；Word2vec 模型的輸入格式跟 LDA 表示法類似只是不用輸入總文章數量。

2. 訓練

LDA 訓練完成後會將每一篇文章以相關主題的機率代表，每一行代表一篇文章，數字則代表該文章屬於主題幾的機率，將這些主題當作特徵、主題的機率值作為特徵值，產生 LDA 的訓練結果。Word2vec 訓練完成後會產生詞彙的代表向量，將文章中各個詞彙向量取平均作為文章的代表向量，將文章的向量當作特徵、向量值作為特徵值，產生 Word2vec 的訓練結果。我們利用 SVM 分別對詞彙頻率模型、LDA 模型、Word2vec 模型所產生的訓練結果進行訓練，建立三個分類模型並以訓練資料進行測試，測試後會出現分類錯誤的文章，這些分類錯誤的文章便可能是訓練資料集中被標記錯誤的文章。

3. 投票與測試

統計三種表示法的錯誤資料結果，在越多表示法中被分類為錯誤的文章就越有可能是被標記錯誤的文章。將這些可能標記錯誤的文章從訓練資料集中刪除，將有助於提升訓練資料的品質。最後測試階段則是將被刪除過錯誤文章的訓練資料集利用 SVM 進行訓練產生分類器，再以測試資料集進行測試。

五、實驗

(一) 資料集

本論文所使用的資料集為 IMDB 這個電影評論的文字資料集 [16]。由於評論類的文章的評分大多都是自由心證，評分的情況可能會有一些跟評論不相關的文章，所以我們使用評論類的文章來進行實驗。這個資料集是由 25000 篇訓練文章及 25000 篇測試文章所組成，文章是從各個電影評論中隨機挑出最多 30 篇該電影的評論及對電影是否好看的評分，分數是採取 10 星等，越接近 1 星表示評論者認為該電影越難看，越接近 10 星表示評論者認為該電影越好看。我們將文章的評分結果當作正負面文章的判定，1~4 星為負面資料、6~10 星為正面資料。去除在經過前處理後完全一模一樣的文章後，剩餘訓練文章 24900 篇、測試文章 24797 篇，其中正面的訓練文章數量為 12470 篇、負面的訓練文章數量為 12430 篇，正面的測試文章為 12439 篇、負面的測試文章為 12338 篇。我們將訓練資料集分成 10 等份，在每次實驗取其中的 9 份當作訓練資料集，產生 10 個 folds，其中正面資料 11223 篇、負面資料 11187 篇。

(二) 分類器

對於文章分類所用的資料具有高維度、大規模的特性，核心函數的影響不大。而 Liblinear [8] 不計算核心函數，在訓練時比 LibSVM [4] 的計算複雜度低，所需時間也較少，所以我們利用 Liblinear 作為分類器。測試資料依照分類器的分類結果分成正確與錯誤兩種情形，統計正確資料後計算 Accuracy，其計算公式如下所示。

$$\text{Accuracy} = \frac{\text{分類正確的文章數量}}{\text{所有文章數量}}$$

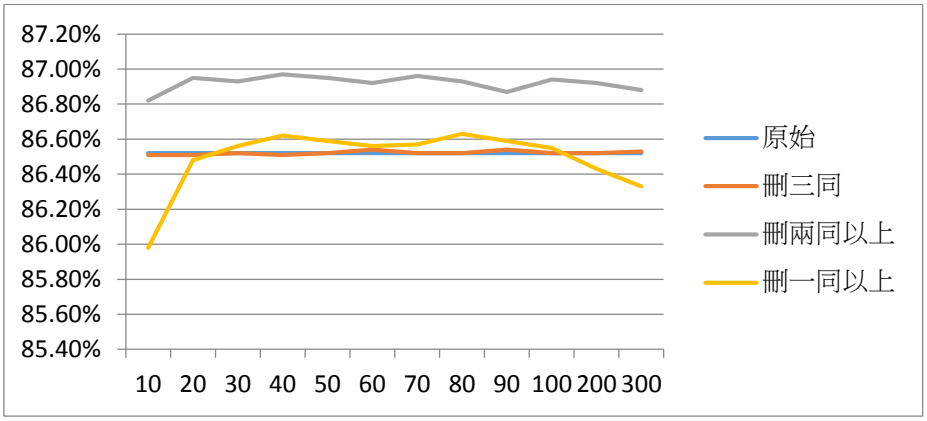
(三) 實驗結果

LDA [2] 的實作是使用 GibbsLDA++ [19]，參數設定是以 α 為 0.5、β 為 0.1 (α 是決定生成主題機率分布的參數，β 是決定主題詞彙機率分布的參數)、主題數量為 100、疊代 1000 次。Word2vec [13] 的實作是使用 T. Mikolov 在 Google 帶領的開發團隊所開發的軟體[1]，參數設定是以 window size 為 5 (window size 就是考慮一個詞彙的前 N 個和後

[1] https://code.google.com/p/words2vec/

N 個詞彙出現情形)、維度為 200 (詞彙要以多少維度的向量表示)。在實驗中我們以不同的投票結果作為刪除的條件，其中投票數代表該篇文章在幾個分類模型中被標示錯誤。我們將刪除文章後的訓練資料集分別進行測試，並與不刪除文章的訓練資料集進行比較，10 個 folds 平均的正確率，不刪除文章的正確率 86.52%；刪除投票數為 3 的文章時，正確率 86.52%；刪除投票數大於 2 的文章時，正確率 86.94%；刪除投票數大於 1 的文章時，正確率 86.55%。在刪除投票數大於 2 的文章後，分類正確率都優於其他的分類正確率，而在刪除投票數大於 1 或 3 的文章後，分類的正確率跟原始訓練資料集的分類正確率相比時好時壞。猜測在投票數為 3 的條件下，刪除的文章數量沒有很多，資料集中依然參雜許多的錯誤資料影響了分類的結果；投票數為 1 的條件下，因為刪除的資料數量太多，使得許多重要的資訊被刪除，所以導致分類的正確率下降。

在先前的實驗中，LDA 的主題數固定為 100，我們進一步比較不同主題數的平均正確率。實驗結果如下圖，顯示出不同主題數並不會影響實驗的結果。



圖一、不同主題數實驗的正確率結果

六、結論與未來工作

在本論文中我們使用詞彙頻率、主題模型、及詞彙向量三種文章表示法來進行樣本選擇，以提升文章分類的正確率。實驗結果顯示，當我們刪除投票數為兩票以上的錯誤資料時，其分類的正確率都比原始的訓練資料的分類正確率高，由此可見我們提出的樣本選擇方法可以提升文章分類的正確率。

未來我們也會繼續以不同的資料集進行實驗，以證明我們的方法可以推廣到其他不

同的資料集中。另外，本研究刪除錯誤資料的方法是採取滿足投票數的門檻後便刪除，這樣可能會去除掉一些有幫助的資料，未來我們會針對刪除資料的方法進行改進。我們未來也會研究如何結合各種表示法來產生新的模型，以更進一步提升文章分類的正確性。

參考文獻

[1] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, 2014.

[2] D.M. Blei, A.Y. Ng, M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[3] M. Campr and K. Ježek, "Comparing Semantic Models for Evaluating Automatic Document Summarization," *Proceedings of the 18th International Conference on Text, Speech, and Dialogue*, pp. 252-260, 2015.

[4] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, Article 27, 2011.

[5] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, No. 3, pp. 273-297, 1995.

[6] G.P. Figueredo, D.A. Augusto, H.J.C. Barbosa, and N.F.F. Ebecken, "A Support Vector Machine-based Technique for Instance Selection," *Proceedings of the XXXIV Iberian Latin-American Congress on Computational Methods in Engineering*, 2013.

[7] Y. Freund, H.S. Seung, E. Shamir, and N. Tishby, "Selective Sampling Using the Query by Committee Algorithm," *Machine Learning*, vol. 28, no. 2-3, pp. 133-168, 1997.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research,* vol. 9, pp. 1871-1874, 2008.

[9] S. Garcia, J. Luengo, and F. Herrera, "Instance Selection," *Data Preprocessing in Data Mining*, Springer International Publishing, pp.195-243, 2015.

[10] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques," *WSEAS Transactions on Computers*, vol. 4, is. 8, pp. 966-974, 2005.

[11] H. Kim, P. Howland, and H. Park, "Dimension Reduction in Text Classification with Support Vector Machines," *Journal of Machine Learning Research*, vol. 6, pp. 37-53, 2005.

[12] H. Liu and H. Motoda, "On Issue of Instance Selection," *Data Mining and Knowledge*

Discovery, vol. 6, no. 2, pp. 115-130, 2002.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," *Advances in Neural Information Processing Systems*, pp. 3111-3119, 2013.

[14] T.M. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[15] C. Mesterharm and M.J. Pazzani, "Active Learning using On-line Algorithms," *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 850-858, 2011.

[16] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* vol. 1, pp. 142-150, 2011.

[17] A.M. Nichat, and S.A. Ladhake, "Brain Tumor Segmentation and Classification Using Modified FCM and SVM Classifier," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 5, no. 4, pp. 73-76, 2016.

[18] J.A. Olvera-López, J.A. Carrasco-Ochoa, J.F. Martinez-Trinidad, and J. Kittler, "A Review of Instance Selection Methods," *Artificial Intelligence Review*, vol. 34, no. 2, pp. 133-143, 2010.

[19] X.-H. Phan and C.-T. Nguyen. "GibbsLDA++: A C/C++ Implementation of Latent Dirichlet Allocation (LDA)," http://gibbslda.sourceforge.net/, 2007.

[20] T. Reinartz, "A Unifying View on Instance Selection," *Data Mining and Knowledge Discovery,* vol. 6, no. 2, pp. 191-210, 2002.

[21] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol.18, no.11, pp. 613-620, 1975.

[22] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

[23] H.S. Seung, M. Opper, and H. Sompolinsky, "Query by Committee," *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pp. 287-294, 1992.

[24] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," *Handbook of Latent Semantic Analysis*, vol. 427, no. 7, pp.424-440, 2007.

[25] C.-F. Tsai and C.-W. Chang, "SVOIS: Support Vector Oriented Instance Selection for Text Classification," *Information Systems*, vol. 38, no. 8, pp. 1070-1083, 2013.

[26] D.R. Wilson and T.R. Martinez, "Reduction Techniques for Instance-Based Learning Algorithms," *Machine Learning*, vol. 38, no. 3, pp. 257-286, 2000.