股市新聞預測股價漲跌

程守雄
建國科技大學資訊管理系
shcheng@ctu.edu.tw

摘要

本研究主要運用文字探勘技術，建立一個股價漲跌分類預測模型。從股市新聞發佈的時間對應股價一小時內漲幅紀錄情形，透過支持向量機分類模型找出有用的樣式，預測股價漲跌趨勢。本研究藉由中文新聞文件斷詞後，利用新聞詞彙特徵選取法，萃取出關鍵詞彙再由關鍵詞彙組合、標記，將非結構化新聞文件資料轉換為結構化資料，再與股價做配對加入漲跌標記，並透過支援向量機建立分類預測模型。實證結果顯示本研究所提出的分類預測模型可以精準的預測股價漲跌趨勢。

關鍵字：文字探勘、股市新聞、支持向量機

一、 緒論

股市是眾多投資方法中最能獲取高額報酬的方法，但同時它也是眾多投資方法中最具風險的一種。一般在選擇股票投資標的時，最常採用的分析方式主要有基本面分析及技術面分析兩種。基本面分析主要考量上市公司的營運及財務狀況，藉以預測未來可能之盈虧，以作為選擇股票投資的依據;而技術面分析則著重過去歷史股價的變動，並從中找出股價趨勢間的特徵，藉以預測未來股價可能的漲跌趨勢，作為股票買賣的依據。然而，不論基本面分析或技術面分析都忽略了與股票上市櫃公司相關新聞消息對短期股價的衝擊。

文字探勘，一般而言，指的是從半結構化(semi-structured)或非結構化(un-structured)格式儲存的文件當中，發掘出文件中隱含的、有意義且重要的資訊，透過分析文件、特徵擷取的過程，從中粹取出隱性資訊，進而處理儲存成為可被再用的知識。依據Dau Sullivan (2001)的定義，文字探勘為「一種編輯、組織及分析大量文件的過程，主要提供分析人員或決策者等特定使用者對特定資訊(如摘要、關鍵字)，發現資訊特徵及其間的關聯性」。Gidófalvi(2001)曾提出window of influence的概念，指出新聞中所包含的資訊在一定的時間間隔，會對股市造成相當程度的影響。Mittermayer (2004)則按新聞發佈前後三分鐘之股價變動量，將影響個股漲跌的新聞分類成「Good News」、「No Movers」和「Bad News」三種類別，透過新聞分類模型進行股票買賣之建議與預測。然而「Good News」和「Bad News」所包含的關鍵字具有許多重複性，因此造成分類之

正確性較低。

鍾任明(2007)於研究中建構中文新聞與台灣股市之預測模型，透過啟發式詞性組合配合門檻值設定來萃取新聞關鍵詞。研究發現股價漲跌反應與詞性組合規則對正確率有顯著影響。

如何掌握消息面以便對股票買賣做出正確的決策，對於短期投資者而言，是相當重要的課題。本研究整合中文斷詞處理與關鍵詞彙萃取策略，再配合文字探勘相關技術的運用，透過支持向量機，預測個股價格漲跌的趨勢。本研究所建立預測模型，可以作為短期投資者的參考。

二、 研究方法

（一） 新聞文件前處理

1. 移除 HTML 標籤：
   由於有些標籤是與資料分析無關的內容，所以必須事先加以移除。
2. 中文斷詞:
   中文語系不如印歐語系，字與字間可利用空白符號區隔。因此，本研究利用中研院斷詞系統CKIP將新聞文件加以斷詞。
3. 詞性選取：
   中研院斷詞系統CKIP 也提供詞性標記，一般來說，中文句子中最重要詞性為動詞與名詞Tsai[8]。

（二） 新聞詞彙特徵選取法

由於中文詞集是一個開放集合，並不存在任何一個詞典或方法，可以盡數羅列所有的中文詞彙。因此關鍵詞彙的自動抽取成為分詞的先期準備步驟。一般而言，文件中高頻詞彙與文件主題有較高的關聯性，可以取之為特徵詞彙來代表整份文件。本研究利用中研院提供之中文詞頻統計，分類統計出該詞性在新聞文章出現的次數。再利用新聞詞彙權重的概念選取新聞文件中具鑑別度之關鍵詞。

（三） 新聞詞彙權重

語詞出現於各類別文件中的頻率(Term Frequency, TF)越高，代表該詞在文件中越重要。但若該詞在一篇文件中出現頻率很高，且在其他所有文件中出現頻率也很高，則代表此詞彙太普通不具有代表性，所以詞彙TF值高，不一定代表該詞較重要。故為了改善這樣的缺點，加入考量反向文件頻率(Inverse Document Frequency, IDF)，所以在某一特定文件內的高詞語頻率，以及該詞語在整個文件集合中的低文件頻率，可以產生出高權重的TFIDF。另外，基於每份文件的詞彙量並不相同，例如：長篇新聞中詞彙可能多達近200 個，但短篇新聞中詞彙可能不到50 個，當一重要詞彙出現在長篇新聞中，其權重相對較出現在短篇新聞中高，因此，本研究將詞彙TFIDF加以標準化。

（四） 詞性組合

由於中文斷詞後的候選詞太多，且單

一的字詞較無意義，透過啟發式的詞類序列，透過合併詞彙的方式降低候選詞的維度，並形成有意義的各類片語。

（五） 建立片語-文件矩陣

當某片語出現時，則標示為1，反之則為0，透過二元標記的方式表達該文件之特徵片語。如圖1所示。

	片語1	片語2…片語n
文件1	1	0 … 1
:	:	: :
文件m	1	1… 0

圖 1 詞彙-文件矩陣

（六） 標記股價漲跌類別

新聞文件透過前述處理步驟轉換為結構化資料，再與設定的反應時間內之歷史股價做配對加入漲跌標記，上漲標記為1，下跌標記為0，探討不同新聞文件中各類片語對股價漲跌的影響。

（七） 建立支援向量機分類模型

支援向量機(Support Vector Machine, SVM)擁有強大的推廣能力並使用統計學習理論為其理論基礎，是目前被廣泛應用在分類問題的一個方法，主要是希望將資料分佈在空間中，可能是二維、三維、甚至多維座標中，然後利用一個多項式或是三角函數組成的方程式將資料分割成兩邊，也就是它能夠原有的訓練資料所在的空間透過核心運算子(kernel operator)轉換成另一個更高維的空間 F。它的目標是自 F 中找出一個 最佳的分割超平面，這個超平面能夠達到將兩類點分的最開，利用找出來的核心函式(kernel function)將資料的座標輸入，即可知道資料是不是屬於這個類別，以此將資料分類。

本研究模型中整合了新聞文件與股價的量值資訊，利用所建立的詞彙-文件矩陣代表的新聞文件所內含的資訊，接著依據每篇新聞的發布時間，將對應的量值資訊作配對後，作為支援向量機的輸入資料。在新聞發佈的特定時間內股價的上漲與下跌作為輸出變數。

三、 資料來源

本研究預測模型將整合中文新聞與股價分時資料建立個股股價漲跌趨勢預測模型。中文新聞資料來自 Money DJ 理財網的個股相關新聞，股價分時資料包含了每一分鐘股價的變化，取自於兆豐證券。實驗資料收集的時間為 2007 年 1 月 1 日至 2009 年 12 月 31 日。另外，為了探討新聞文件對股價所產生的短期效應，取樣上需配合下列的限制：

1、新聞的發布時間限定於台股交易時間內。

2、新聞文件中必須包含所選定實驗個股之股票代號或是公司名稱。

3、新聞文件中不能出現兩個或超過兩個的股票代號或是公司名稱。

四、 實證分析

本實驗標的為台灣股票上市公司的鴻海精密工業股份有限公司（股票代號：2317），主要探究個股新聞發佈後一小時內對個股股價之影響，所以本實驗設定股價對新聞的反應時間為十五分鐘、三十分鐘、四十五分鐘與六十分鐘。實驗資料中，2007年1月1日至2008年12月31日共有138筆資料為訓練資料集，2009年1月1日至2009年12月31日共有55筆資料為驗證資料集。探討個股新聞發佈後一小時內個股股價的漲跌情形，實驗結果如下：

表1 反應時間十五分鐘之實證結果

預測 / 實際	上漲	下跌	正確率
上漲	26	0	100 %
下跌	0	29	100 %

表2 反應時間三十分鐘之實證結果

預測 / 實際	上漲	下跌	正確率
上漲	25	0	100 %
下跌	0	30	100 %

表3 反應時間四十五分鐘之實證結果

預測 / 實際	上漲	下跌	正確率
上漲	15	0	100 %
下跌	0	40	100 %

表 4 反應時間六十分鐘之實證結果

預測 / 實際	上漲	下跌	正確率
上漲	13	0	100 %
下跌	0	42	100 %

由實驗結果可以看出，本研究所建構的預測模型預測股價漲跌情形與實際股價漲跌情形正確率可到達100 %。

五、 結論

本研究提出一個股市新聞的文字探勘模型來預測一小時內股價漲跌情形，藉由中文新聞文件斷詞後，利用新聞詞彙特徵選取法，萃取出關鍵詞彙再由關鍵詞彙組合、標記，將非結構化新聞文件資料轉換為結構化資料，再與股價做配對加入漲跌標記，並透過支援向量機建立分類預測模型。由實例的驗證，驗證本方法是快速且有效率的。

致謝

本研究感謝國科會計畫補助，計畫編號：NSC99-2815-C-270-010-H 及建國科技大學校內專題計畫補助，計畫編號：CTU-99-RP-IM-006-038。

參考文獻

[1] 鍾任明，李維平，吳澤民，「運用文字探勘於日內股價漲跌趨勢預測之研究」，中華管理評論國際學報，10(1)，1-30 頁，2007 年。

[2] 陳俊達，王台平，劉昭麟，「以文件分類技術預測股價趨勢」，第十九屆自然語言與語音處理研討會論文集，347-361 頁，國立台灣大學，台北市，

台灣，2007 年。
[3] 陳振南，吳毓傑，「特徵選取與權重分配於中文新聞分類之比較」，第十三屆國際資訊管理學術研討會，721-728 頁，淡江大學，台北縣，台灣，2002 年。
[4] Sullivan, D. Document warehousing and text mining. Canada: WileyComputer Publishing. (2001).
[5] Gidófalvi, G. Using news articles to predict stock price movements.http://www.cs.ucsd.edu/users/gyozo/studies/cse254_AI/stock_price_prediction.pdf, 2004-06-15(2001).
[6] Mittermayer, M. A. Forecasting intraday stock price trends with textmining techniques. Proceedings of the 37th Hawaii international conference onsystem sciences, 64-73. (2004)
[7] R. P. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction Using Financial News Articles," *Proceedings of the 12th Americas Conference on Information Systems*, paper 185, Acapulco, Guerrero, Mexico, (2006).
[8] W.-Y. Ma and K.-J. Chen, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff," *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, vol. 17, pp. 168-171, Sapporo, Hokkaido, Japan, (2003).
[9] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412-420, Nashville, TN, USA, (1997).
[10] K. Aas and L. Eikvil, "Text Categorisation: A Survey," *Technical Report,Norwegian Computing Center*, (1999).