大數據以音訊來源之應用

各股市分析師與電視名嘴交叉分析尋找投資標的

楊志中 講師　　謝莉醇 學士

台灣科技大學　　中國文化大學

資訊管理系　　資訊管理系

ccyang.phd@gmail.com　　zxc88633@gmail.com

摘要

投資股票是現今相當普遍的一種金融活動,且有著迷人的高報酬率但相對的也伴隨著高度的風險，投資人買賣股票皆希望能夠提高報酬，所以如何準確地預測未來股價的變動一直是投資者所關注的目標。

一般投資人取得資訊於網路、報紙或電視,但以往的所研究的論文都是取得網路文字數據加以探勘，並未放入語音類型的資料，本研究將抓取各股市分析師與電視名嘴的語音訊息為資料來源，，並使用資料探勘方法中的傳遞類神經網路技術對籌碼分析與技術分析中的三大法人動向、主力買賣超、布林通道、KD值、MACD、成交量進行預測模型的建立，並找出最佳投資組標的，以降低投資風險。

關鍵字： 股票、基本分析、技術分析、籌碼分析、倒傳遞類神經網路。

1. 緒論

1-1 研究動機

近年來，台灣物價持續上漲但薪水卻不漲，此讓民眾的荷包大大縮水生活越來越辛苦，在這樣的情況下，我們應該懂得如何理財或是增加被動收入，而最廣為人知的理財方法就是購買股票。初入股票市場最難的一件事情就是如何選股，許多投資者進行股票投資多年，之所以獲利持平或是虧損，就是因為沒有良好的選股投資策略，每個人都想找到有效的投資標的，以降低風險並提高報酬，這是投資人最關心的事情。

而台灣屬於淺碟型市場經濟，常常因為一個新聞或是內線消息曝光就能使股市改變風向，雖然可以深入了解基本面、技術面、籌碼面，以及一些網路資料，例如:收盤價、主力買賣超、融資融券……等，卻很難從中去找到有用並且能提供決策的資訊。再者，以往的許多研究，都是以抓取網路上現有的新聞資料庫、台灣證券交易所的每日股市資料做為資料來源。但網路世界進步的速度有如光速，資料擷取不單單僅限於文字更包含了語音檔中夾帶的訊息，這些語音資料屬於非結構化資料，且情緒中隱含的意義有可能是重要指標。

因此，本研究不僅會延續前人已有的作法，並將加入更生動的元素—**電視名嘴的分析**，再與現行已有的資料做交叉分析，找出最具利多及分析師推薦的投資標的。相信多了這些語音資料，判斷其影響的幅度後，便能使投資人在決策上，更加提升準確率。

1-2 研究目的

在整個總體經濟大環境中，我們認為景氣有漲有跌；景氣好時，大部分公司的股價都會漲；股價跌的時候，由於法人機構急於套現，因此好公司的股票也會跟著跌。所以單純靠財務報表來預測股價可能失之偏頗。反之，財經節目的電視名嘴容易掌握國際情勢做每日股市分析，若能使用電視名嘴之分析當作資料來源想必能更據說服力。

因此，本研究最主要的研究目的是使用**倒傳遞類神經網路**，結合技術分析、籌碼分析與電視名嘴分析，來預測台股的未來股價走勢，並有下列三項研究目的：

1. 分析上市上櫃股票之投資績效表現與股價漲跌幅的因素。
2. 分析電視名嘴的語音檔使用類神經網路與情緒分析之投資標的。
3. 分析台灣證券交易所的每日收盤價資料、周線加上主力買賣超以及新聞資料庫的新聞事件。

1-3 研究架構

本論文分為六個章節，第壹章為緒論：主要闡述研究背景動機與目的，第貳章為研究架構：說明本研究之整體研究框架，第參章為參考文獻：主題為股市技術面分析與資料探勘相關技術文獻探討，第肆章為研究流程：敘述研究流程與做法，第伍章為實驗結果與分析：針對實驗結果探討與分析，並且與其他不同分類器方法進行比較，第陸章為結論與未來發展：說明本研究之發現，並提出未來進入研究所的發展。

本研究架構共分成六個章節，如下圖 1：

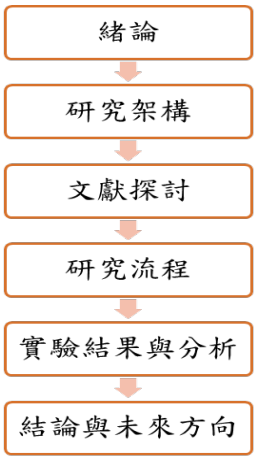

圖 1 研究流程圖

2. 相關文獻與技術探討

本研究之目的旨在探討語音大數據結合資料探勘技術應用於股市分析，並且找出最佳投資標的。為瞭解研究主題的應用，並建立研究架構，於是本章針對本研究之相關文獻進行探討。

2-1 類神經網路

類神經網路是利用模擬人類神經元的概念，進行模仿人類神經中的記憶能力與學習能力，找出資料中特定的規則，藉以預測股市的發展，由於此特性與股票投資需參考過去數據的特性非常相近，因此，可以應用類神經網路來進行對於股票買賣交易上的預測，但有學習速度比較慢與模型建構風險大的缺點。而倒傳遞網路是監督式學習，平行處理輸入、輸出變數間的非線性關係，具有學習的能力並且可解決互斥問題，達到準確分類的效果，是目前類神經網路學習模式中，最具代表性且應用最普遍的學習模式。以下是以相關文獻探討：

學者	年代	結論
陳昌捷	2015	此研究利用倒傳遞神經網路，預測台灣股市**次日收盤價**，以提供投資人除基本分析、技術分析及籌碼分析外的另一預測股市方法。技術指標先以皮爾森相關係數進行篩選，81 種指標中相關係數 r 值 0.7 以上共 17 種指標，分別以 r 值 0.7、0.8、0.9 三組進行預測，以 r 值 0.8 網路架構為 15-28-1 最佳，預測結果平均誤差百分比值(MAPE) 0.6315%，證明以倒傳遞類神經網路建構之預測模式台灣股市加權指數，具高度準確性。在預測趨勢方面，在7/22-7/30其間已預測出下跌趨勢，雖然實際指數有2天小幅 上漲，但此論文已成功預測下跌趨勢已來臨。
陳世芳	2007	此研究運用倒傳遞類神經網路，研究在交易的過程中的個股籌碼流動向量與股價波動關係，所使用資料是台灣證券交易所個股交易日資料，其中2003/10/01~2004/08/31 的資料為訓練用，2005/07/20~2005/10/28 的資料為驗證期 所使用，結果發現投資人進行個股短線投資時，對股價漲跌具有正向影響的融資交易人和外資交易人的籌碼變動向量，更值得投資人在買賣時機決策時作為判斷參考。
吳聲昌	2006	此研究指出沒有完美無缺的技術指標，技術分析不可能絕對正確。投資人除研判技術指標外，對於基本分析及籌碼分析方面，更要深入探討，才能在投入股市時，降低風險、提高報酬。所使用的資料是精業世紀贏家股票資料庫民國 91年1月2日至95年4月30日止。研究對象以摩根台指(MSCI)成分股與上市上櫃各產業龍頭股為主，而且以摩根台指權重的前20檔個股與上市上櫃各產業龍頭股 20檔做為分析比較的對象。類神經網路訓練樣本取自91年1月2日起至93年12月31日止，測試樣本取自94年1月2日起至95年4月30日止。運用了資料探勘技術於台灣股票市場尋找低風險投資組合，透過倒傳遞類神經網路模式協助投資人快速且有效地解決此一困擾，並且加入技術分析與籌碼分析的變數，幫助投資人在投資的過程能夠獲得最佳的報酬。

2-2 資料探勘

資料探勘是從原文的 Data Mining，其主要的意涵是 Mining From Data，從資料中挖掘金礦。而資料探勘就是從結構化或非結構化的文件當中，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。(鍾任明，2007)

不論現在還是過去，新聞媒體、FB、PPT充斥著許多以文字形式儲存的網路資料，以奇摩新聞來說，不論是經濟面、政策面或是技術報告等都有大量的新聞發佈，然而這些資料雖然隱含相當有價值的資訊，卻無法透過一般的方法直接分析取得。(Data Mining: Concepts and Techniques 2011）文中指出資料探勘就是從大型資料庫中抽取具有意義之資訊或模式的過程。主要為提供特定使用者（如：決策者、分析師）特定的資訊（如：摘要、關鍵字），以及發現某些特徵及其關聯。

以股市新聞來說，若使用情緒分析，則可利用作者的文章進行語意的解析來判定這則新聞為利空還是利多。如下圖5-2：建立語意資料庫後再利用資料探勘方式，電腦可判讀新聞中的樂觀詞語悲觀詞，再依據兩者的數量差距給予 SR 評分。此篇大立光新聞 SR ＝ 42.8571。(維京人酒吧 Viking Bar，預見雜誌)

圖5-2

在資料探勘中，其資料價值高低的評估，恰恰與**股票市場相似**，以收集資料、並

且取得其可能影響因素，以此因素來推測，最後結果驗證時，就能夠尋找到以往所未能發現到的資訊。

2-3 效率市場假說(Efficient Market Hypothesis)

根據 Fama(1970)的效率市場假說，若證券市場之價格可以正確且充分反應全部相關的資訊，如經濟、金融等相關的消息面，則稱該市場是有效率的。換言之，當影響市場股票價格的訊息出在市場時，該訊息會迅速且正確的反應在股價上，因此投資者無法在市場上賺取超額的報酬，其主要的假設有下列三項：

(1)投資者皆為理性。

(2)資訊即時公開，且獲得資訊無 須負擔額外的資訊成本。

(3)市場上沒有任何的投資者具有單獨影響股價的能力。

- 按照市場上資訊集中的不同類型，Fama 將市場效率劃分為以下三種型態：

(1)弱勢效率市場(Weak Form Efficient Market)

證券價格充分反應過去證券價格變化所提供的所有資訊，因此證券價格的未來走向與其歷史價格變化為互相獨立，並服從隨機漫步理論。投資人無法藉由運用技術分析來獲取超額報酬。

(2)半強勢效率市場(Semi-strong Form Efficient Market)

證券價格不但充分反應過去所有的歷史資訊，並且也完全反應了所有公開資訊，因此投資人無法藉由基本分析(Fundamental Analysis)來獲取超額報酬。

(3)強效率市場(Strong Form Efficient Market)

證券價格以充分反應『過去』及『現在』的所有公開與未公開資訊，因此投資人無法藉由公開或是未公開消息來獲取超額報酬。

杜金龍(2001)認為國內市場因為漲跌幅的限制，呈現出**比弱勢效率市場還弱**的資本市場，再加上投資管道不足，國內股市充斥著內線交易與短線交易、政府干預護盤、市場炒作、與非理性投資人等…，使得國內股市不具弱勢效率市場的條件。**當股市不具弱勢效率市場的條件時，投資者可藉由基本分析、技術分析與籌碼分析甚至是內線消息來獲取超額報酬。**

基於上述理由，在國內基本分析、技術分析與籌碼分析廣為一般投資者、法人投資機構擬定投資策略的重要參考指標。

2-3-1 選股策略分析

由於台灣股票市場呈現比弱勢效率市場還弱，因此基本分析、技術分析與籌碼分析

有其價值，其介紹如下：

基本分析

利用經濟指標找出公司股票的合理價值，分析各產業的興衰。並且更進一步的利用財報觀測各企業之營收的現況與未來前景，從中對企業作出客觀的評價，並且盡可能預測其未來的變化，作為投資的依據。基本分析方可利用以下三方面評估公司未來的成長與獲利的能力：

(1) 總體經濟面

通常包括國際和國內各項經濟指標，如：GDP 成長率、通貨膨脹、利率、匯率、生產力、失業率等，目的是分析出經濟景氣的好壞，從而可以提前應對。

(2) 產業面

包括整體產業的銷售量、指標產品價格變動、產品效能的進步、進入產業的公司數…等，最終目的是分析出**產業的供需狀況**，從而可以提前應對。

(3) 公司基本面分析

依各公司的實質營運狀況來分析研究。藉由公開發佈之財務報表來了解並分析期公司體質與期營運狀況，藉此擬定投資策略。

技術分析

技術分析主要為利用過去的資料如價格、成交量、波動度…等，藉由 統計方法來研判未來股價的走勢，其基本假設如下：

(1) 所有的資訊，都反映在股價上

股價不會被基本面影響，但**基本面的資訊都充分反映在股價上**。

(2) 股價的走勢是有規律的，所以可以預測未來走勢

過去的價格變動會形成一個走勢，而**未來的股價會有規律地跟隨這個走勢。**

(3) 歷史會不斷重演，所以趨勢會不斷循環

技術分析，**相信投資人的行為會一直重複**，也就是說，人們永遠不會記取教訓。

籌碼分析

籌砝碼分析主要為研究在市場上對股票的價格特別有影響力的人，我們簡稱為「大戶」，而大戶又分成三種

圖 5-3(來源：Cmoney)

(1) 大戶買進：

價格上漲，跟著買進！

(2) 大戶賣出：

價格下降，跟著賣出！因為大戶們一買一賣都是幾百張甚至幾千在賣，同一支股票他們一買進此檔股票需求大幅上升，價格自然上漲，但一旦賣出也會使股價下跌。

(3) 大戶的買進和賣出是連續的

當大戶開始買進，就會連續好幾天買，股價就會跟著上漲而相反的，只要開始賣就會連續賣好幾天，該公司股價就會跟著他們波動。

3. 研究流程

本研究步驟分為六個階段：

1. 深入了解目前國內外大數據股票分析的特性及方法，並蒐集股票相關文獻探討，並實際買下晶電、華航、旺宏、華通作為預測前後的樣本。
2. 假設資料來源加入電視名嘴語音檔且使用倒傳遞類神經網路分析，將能提高投資準確率。
3. 取得目前現有的語音辨識軟體 Virtual Audio Cable，**將**電視名嘴的語音**影音檔轉成文本。**
4. 利用網路爬蟲蒐集台灣證券交易所當日股票交易資料及相關新聞進行資料分析，並將電視名嘴文檔使用倒傳遞類神經進行分析。
5. 整合研究結果，觀察預測值是否符合市場變化，最後撰寫結論及研究發現以完成本研究。

4. 研究流程

本研究步驟分為六個階段：

1. 深入了解目前國內外大數據股票分析的特性及方法，並蒐集股票相關文獻探討，並實際買下晶電、華航、旺宏、華通作為預測前後的樣本。
2. 假設資料來源加入電視名嘴語音檔且使用**倒傳遞類神經網路分析**，將能提高投資準確率。
3. 取得目前現有的語音辨識軟體 Virtual Audio Cable，**將**電視名嘴的語音**影音檔轉成文本。**
4. 利用網路爬蟲蒐集台灣證券交易所當日股票交易資料及相關新聞進行資料分析，並將電視名嘴文檔使用**倒傳遞類神經**進行分析。

5. 整合研究結果，觀察預測值是否符合市場變化，最後撰寫結論及研究發現以完成本研究。

本研究流程如圖 5-4 所示：

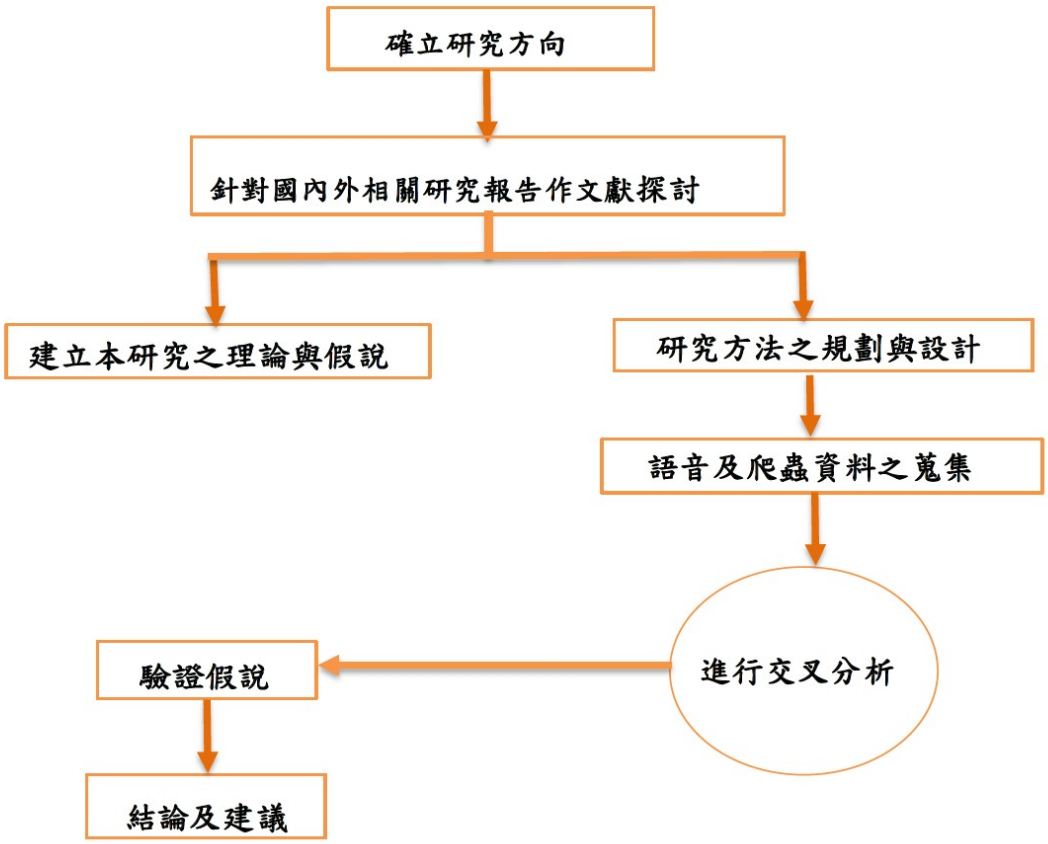

圖 5-4 研究流程圖

5. 實驗與預期成果

本研究預期將採用電視名嘴分析師語音資料與股市收盤價、周線、主力買賣超、融資融券等指標進行預測，資料皆採用網路公開資料。

台灣股票市場呈現比弱勢效率市場還弱，因此基本分析、技術分析與籌碼分析，再加上電視網路普及度高足以讓民嘴提到的股票形成強勢股，此對帶動風向確實有其價值。目前已完成前三個步驟，並抓取 2017/9/15 股市現場節目當作語音資料來源，將電視名嘴的語音**檔轉成文本，並擷取字串找出關鍵股票文字雲，如下圖 2：**

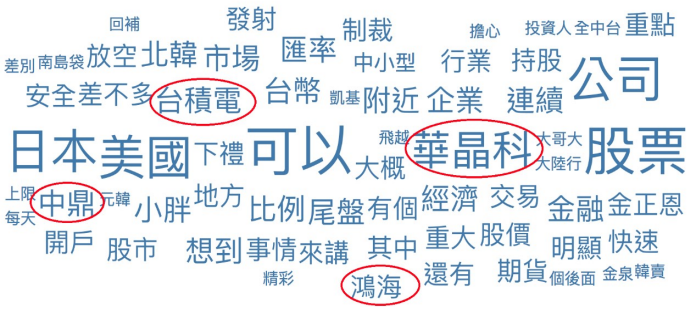

圖 2

未來本研究將使用**倒傳遞類神經網路分析**電視民嘴分析師提到的投資組合，在與奇摩股市新聞及網路股市資料進行比對後，進行預測準確率比較。

6. 預期貢獻

本研究預期貢獻為利用過去較少運用的電視名嘴影片聲音當作數據來源，希望透過本研究來證實語音是一個揣摩人心的重要指標，未來將使用**倒傳遞類神經網路分析**進行投資組合預測，最後幫助預測股市走向，提高投資人選股準確率。

7. 參考文獻

[1] 將旻瑋，以改良式支持向量機建立股票股價漲跌預測模型，碩士論文，2014。

[2] 鄭健毅，應用SVR支援向量迴歸模式來進行電子產業股價預測，碩士論文，2010。

[3] 陳昌捷，以倒傳遞類神經網路預測股市指數，碩士論文，2015

[4] 杜金龍，技術分析在台灣股市應用的訣竅，金錢文化出版公司，2002。

[5] 吳聲昌，以資料探勘技術於台灣股票市場尋找低風險投資組合之研究，世新大學資訊管理學研究所，碩士論文，2006。

[6] 陳世芳，運用類神經網路於個股籌碼流動向量股價預測系統之建立，南華大學財務管理研究所，碩士論文。，2007。

[7] 施正宏，「結合總體經濟指標及個股財報資料以預測個股漲跌-以台灣電子類股為例」，中原大學資訊管理學系，碩士論文，2004年。

[8] 鄒杰夫，台灣上市類股股價預測模型之研究-倒傳遞類神經網路模型之應 用，玄奘大學財務金融學系，碩士論文，2008。

[9] 李文智，以總體經濟變數與存託憑證探討對標的股股價預測模式影響之研 究-以台積電為例，大葉大學國際企業管理學系，碩士論文，2006。

[10] 喻欣凱，運用支援向量機與文字探勘於股價漲跌趨勢之預測，輔仁大學資 訊管理學系，碩士論文，2008。

[11] 蔡承益，使用SOM-SVR混合型系統搭配屬性篩選模式應用於臺灣股票指數期貨預測，國立高雄第一科技大學資訊管理所，碩士論文，2007。

[12] 劉翔瑜，倒傳遞類神經網路、支援向量迴歸於日經225現貨指數之預測及交 易策略之研究，輔仁大學金融研究所，碩士論文，2005。

[13] http://www.cmoney.tw/learn/course/michelle/topic/1245 ，股票籌碼面分析

[14] http://www.cmoney.tw/learn/course/0520/topic/678，股票基本面分析

[15] http://www.cmoney.tw/learn/course/technicals/topic/475，股市技術面分析

[16]https://zh.wikipedia.org/wiki/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98，維基百科-資料探勘