結合支援向量機與粒子群最佳化探索台灣股市預測模式

陳博文
國立台北科技大學
資訊與運籌管理研究所
powellchen@gmail.com

翁頌舜
國立台北科技大學
資訊與運籌管理研究所
wengss@ntut.edu.tw

摘要

本研究在於探討如何應用支援向量機(Support Vector Machine)在台灣股市中建立個股投資預測模型，並且輔以粒子群最佳化(Particle Swarm Optimization)進行參數最佳化與變數篩選。一直以來台灣股市被認為極難分析且預測，所以希望在本研究中，能夠有效地整合財務基本面與技術分析面的資訊，以建立穩健的投資預測模式。

本研究分析模式為根據公司各季所公布的財務報表與自行計算出的技術指標進行預測。在採用台灣股市歷史資料進行投資模擬後，實證本研究的方法可以產生不錯的獲利曲線。同時，本研究所得的模擬結果經過分析，也證明優於相同時期中的類神經網路預測模式或是買進及持有的投資策略。因此，本研究認為在同時整合財務基本面與技術分析面的資訊下，使用支援向量機和粒子群最佳化建立個股投資預測模型，不失為一種可以有效獲利而且相當穩健的投資策略。

關鍵詞：支援向量機、投資預測、粒子群最佳化

1.研究動機

股票的波動和趨勢，往往受到各種因素影響，從人為干擾和流言渲染，到公司的競爭生態變遷，以至於總體經濟或全球化的趨勢變化，產生台灣淺碟市場中所獨有的各種因素交互影響且隨時間變化結構的動態系統。所以採用傳統的統計方法或是線性模型，勢必無法產生令人滿意的預測結果，也因而發展出許多的研究分析模型來嘗試解決此一問題，譬如採用類神經網路 Huang et al. (2007b)等非線性預測模型。通常而言，類神經網路所建之預測模型是一個相當不錯的預測模型，但是也有研究指出類神經網路會有實際操作績效與訓練時期正確率相差太大的情況發生(Hann and Steurer, 1996)，以及經常伴隨著收斂至局部最佳解(local optima)和過度配適(over-fitting)的兩大缺點(Yu et al., 2009)。

為了克服這些問題和缺點，本篇論文嘗試採用 Vapnik (2000) 提出的 Support Vector Machine (SVM)，並且整合 Kennedy and Eberhart (1995) 提出的 Particle Swarm Optimization (PSO) 進行 SVM 的參數最佳化和分析模型的特徵篩選。並且將分析結果與類神經網路分析模型進行比較，以確認此種分析模式是否的確具有優勢。

1.1 研究目的

本研究以有限但是動態的分析變數範圍，藉由 SVM 與 PSO 分析模型的整合，嘗試在台灣股市的普通股交易行為中，根據每季公布的公司財務數據和一般技術指標，找出具有獲利可行性的投資預測模型，以便在每季財報公布後制定選股策略。

因為這些投資標的中的公司股價變化，受到太多因素干擾，而這些因素的交互影響關係又會隨著時間而產生變化。所以，此種問題很明顯地無法以單純的線性模型求得解答，其可能存在的預測模型也不可能是永恆不變的，而所有可能影響變數的強度和範圍也可能隨著時間而動態改變。

本研究之研究目的彙整如下：

1. 在各種雜音(noise)以及個人情緒干擾下，建立一種客觀且經過驗證的投資準則。讓一般投資人在各種明牌和謠言中，找到客觀的選股策略。
2. 除了建立一般傳統的靜態財務分析模型，同時也考慮動態的時間變化，進行相對應的變數調整。
3. 採用 SVM 技術希望降低過度學習的風險。以及整合 PSO 的最佳化能力，以期找到參數與分析變數的最適組合。
4. 整合上述所有特徵，建立一種有效且穩健的投資預測選股模型。

1.2 研究流程

本研究嘗試結合支援向量機(Support Vector Machine, SVM)的分類能力與粒子群最佳化(Particle Swarm Optimization, PSO)的搜尋能力，尋求在台灣股市中，建立可以獲利的選股策略。同時為了避免資料探勘偏差(data mining bias)，所以採用近十年的股市交易資料作為分析基礎。而主要的分析角度將從基本面分析出發，著重於公司的歷史財務指標表現，發掘公司價值與未來股價變化之間的關係，並且從不同的資料探勘模型(data mining model)，進行績效比較，希望找出最具有獲利能力的模型，以便作為建立投資組合的參考依據，同時也希望找出不同時期中最具有影響力的分析變數，期望以最少的變數，建立最佳的模型，達到最高的獲利率。並且與同時期的買進及持有策略(Buy & Hold)獲利率進行比較，計算其超額(excess)獲利率，作為客觀的投資獲利評估基準。

2. 文獻回顧

在股市預測上，目前已經存在相當多的文獻和研究。因此，在文獻整理上，需要進一步加以收斂整理的範圍。所以，本研究首先從應用資料探勘(data mining)技術於股市預測模型上的論文作為一般性方法論的切入點。然後就本研究所採用的支援向量機(SVM)與粒子群最佳化(PSO)分別進行探討。

2.1 一般股市相關文獻整理

通常新興市場股價較已開發國家的股價更容易受到某些因素的影響(Cao et al., 2005)，亦即新興市場比較容易預測。同時，以上海股市為例，可以發現若採用類神經網路模型預測模式，則比使用傳統統計的線性模式具有一定的優勢。

通常股市預測模型可以分成分類(classification) 模型與迴歸 (regression) 模型，Olson and Mossman (2003) 特別指出，在使用 6 年移動窗口資料以預測下一年股價變化的實驗研究中，比較 point estimation (BPNN vs. OLS)和 classification (BPNN vs. logit)兩種模式下，使用 61 種財務比率，並且將訓練資料切分為 2、4、8、16 等不同的類別數目，以比較在不同分類數目下，是否對獲利率產生影響。並且同時採用逐步法 (stepwise) 進行變數篩選，其研究發現：1. 分類模型的表現較佳。2. 不需要 61 個變數一起進行分析。3. 類神經網路的表現與羅吉斯迴歸 (logistic regression) 的表現近似。

在基本面分析與技術面分析的取捨上，Quah (2008) 特別指出技術面分析適用短期操作，而基本面分析則較適合長期操作，用於檢視投資組合是否健全。而於軟性計算上，其也認為 Soft computing 追求的是在可容許的不精準、不確定、部分正確、近似解的可能情況下，達到可追蹤性、強韌性、以及低解答成本的目的。並且也認為系統存在 the curse-of-dimensionality issue，並非是要把所有可得的變數都納入分析模型之中，重點在於是要挑選最適合的變數。

但是最適合的變數也是會隨著時間變化的。Gilli and Roko (2008) 採用決策樹 (decision tree) 分析不同時間點，產生的第一個節點 (node) 規則所代表的決定性變數(最重要的分析變數)。實證發現變數組合是會隨著時間而變化，同時也同時納入基

本面與技術面分析數據，希望克服以往傳統基本面分析主要不便之處，無法預測股價所可能發生的急速趨勢反轉。

例如 Eakins and Stansell (2003) 採用類神經網路作為分析模型。因為像類神經網路此種非線性模型，相對於傳統統計的方法，在訓練 (learning) 階段，類神經網路具有比迴歸模型 (regression model) 更穩定的權重係數(處理 noise 的能力)。同時，在處理時間序列 (time-series) 資料上，類神經網路也優於傳統的線性多項式。類神經網路的優勢為：1. 發掘隱藏於資料中的複雜資訊組成。2. 本身具有非線性處理能力。3. 分割樣本空間的能力。4. 針對不同解空間，產生不同函數對應。

至於基本面分析所著重的價值衡量標準，也是會隨著時間而變化的。如 Ren et al. (2006) 所指出，過往經濟學家所提出的選股模型已經不再具有高獲利率，以往的重心在於如何選取合適的投資組合，至於參考的變數則依不同經濟學派而不同。不同的投資大師皆有其規則，但是每一個規則皆有其生成的時代背景以及適用的政經情勢，因此這些所謂的大師規則大部分已經不能再產生令人滿意的結果。

2.2 支援向量機相關文獻整理

Lin et al. (2008) 採用 PSO 對 SVM 的參數做最佳化與變數篩選，同時編碼方式與本研究相同。其與本研究之差異點則在於其研究數據來自於 UCI 公開資料集，而本研究則在於如何產生具有獲利的投資組合。Lin et al. (2008) 認為 SVM 的參數設定在訓練過程中，對於模型的正確度有顯著影響。所以在不損失正確性的前提下，其研究整合 PSO 成為 PSO+SVM，並且採用公開的資料集作為比較基準，然後與 Grid Search 與 GA+SVM 進行比較，發現 PSO+SVM 優於 Grid Search，並且與 GA+SVM 有相類似的表現 (PSO 優於基因演算法(GA)，但是沒有具有絕對優勢)。同時，本研究亦參考 Lin et al. (2008) 所得到的實證數據，進行程式結果驗證，以及實施單元測試。

Su and Yang (2008) 認為在特徵選取問題上，一般可分為 Wrappers，Filters 與 Embedded 三種不同的類別。Wrappers 與 Filters 會選取特徵變數的部份集合，Wrappers 方式是根據對於應變量的貢獻進行自變量評估，而 Filters 方式則通常發生於前處理步驟，與應變量的內容是互相獨立的。至於 Embedded 方式則是為了避免繁重的計算成本，所發展出來的特徵排序方法。

至於 SVM 與類神經網路在分類問題上是否具有優勢？Chen and Shih (2006) 給了初步答案，其研究應用 SVM 至信用評等分類系統上，並且使用三種變數：證券市場資訊, 政府財務支援, 以及大戶財務支援，以增加分類的效度，而其研究範圍則涵蓋三年之久的資料。其實驗結果在與類神經網路 (BPNN) 模型比較之後，產生結論為：1. 變數越多正確率不一定越高，涵蓋分析時間越長不一定越正確。2. SVM 對類神經網路與線性迴歸有非常顯著的優勢。

在 Tay and Cao (2001) 的研究中，也得出類似結論，其使用 SVM 進行時間序列資料預測，並且根據 NMSE, MAE, DS, WDS 做為評估基準，與類神經網路 (BPNN) 模型進行比較之後，得到結論是 SVM 優於類神經網路 (BPNN)。

Kim (2003) 亦佐證此一論點。在預測財務時間序列方法中，SVM 是一種獲得肯定的方法，因為其將實證誤差包含在風險函數內，並且以最小結構風險原則求解。此研究應用 SVM 於股市價格指數，並且與 BPN 與案例推理 (Case-Based Reasoning, CBR) 進行比較。結論為：SVM 優於類神經網路 (BPNN)以及案例推理(CBR)這兩種模型的表現。

在分類問題中，Huang et al. (2008) 發現若是採取 Wrapper approach 在 23 種技術指標中進行 Feature selection，並採用 Voting scheme 合併多種分類器以預測韓國與台灣股市的趨勢，實證顯示 Wrapper approach 較 Filter approach 較好，此外 Voting scheme 也較單一分類器表現較佳。

2.3 粒子群最佳化相關文獻整理

Banks et al. (2007) 與 Banks et al. (2008) 詳細彙整 PSO 的相關發展：雖然 PSO 相較於其他 natural computing paradigms 稍嫌年輕，但是 PSO 已經在這麼短的時間內吸引眾多研究學者的注意，並且取得不錯的表現。上述兩篇論文對 PSO 進行編年史的回顧並且檢視 PSO 遭遇到的挑戰與機會。Banks et al. (2007) 從 natural computing 的觀點檢視 PSO 的定位，回顧理論的發展，以及為了避免 swarm stagnation 與 tackle dynamic environment 所做出的改善措施。Banks et al. (2008) 則檢視 PSO 目前在 Hybridisation, combinatorial problems, multi-criteria and constrained optimization 等研究領域的發展。

通常財務問題都採用分類問題模式解決 (Marinakis et al., 2009)，而在分類問題中，如何挑選變數便為一重要課題。因此 Marinakis et al. (2009) 使用蟻群法 (Ant Colony) 與 PSO 進行 feature selection。其研究嘗試使用三種分類器，分別為：1-nearest，k-nearest，weighted k-nearest。而變數篩選則包含：ACO，PSO，GA，與 Tabu 等方法。所以總共有 3 x 4 = 12 種分析模型。結論為：PSO 結合 k-nearest 較具有優勢。

在混合型 (hybrid) 演算法中，Zhang et al. (2007) 則嘗試合併 PSO 與類神經網路 (BPNN)，其認為 PSO 的優點在於 global search，而 BPNN 的優點在於 global optimum 和 local search，所以建議使用 PSO 進行一開始的類神經網路權重決定，並且在不同應用情況中採用三種粒子編碼方式，以便適用於三種不同的問題，同時，也顯示出此種新的演算法在正確度和速度上，要優於原有的 Adaptive Particle swarm optimization algorithm (APSOA) 和 BPNN。

Kuo (2009) 嘗試合併 PSO 與模糊 (Fuzzy) 理論，因為以往使用基因演算法 (GA) 和模擬退火法 (SA) 結合 Fuzzy Time Series 的預測性無法令人滿意。因此，在此研究中採用 PSO 與 Fuzzy Time Series 結合，其成果經證實要優於任何已知的現行 model。其所採用方法為使用 PSO 組合測試 k-階時間序列因素，以找出最佳組合(最低 MSE)。

就混合型 (hybrid) 演算法整體而言，Grosan and Abraham (2007) 進行了一些初步的整理。其首先強調 hybrid evolutionary algorithm 的重要性，接下來則列舉出各種不同組合的可能性。然後呈現最近幾十年所演化產生幾種不錯的混合型 (hybrid) 演算法作為參考。

至於在處理投資組合最佳化問題中，PSO 究竟是否具有優勢呢？Cura (2009) 嘗試給了一個初步解答。其認為投資組合最佳化的問題是混合了 Quadratic 和 Integer Programming 的問題，因此不存在有效解答。雖然之前有人嘗試過 Heuristic 的方式，但是沒人試過 PSO。所以此研究採用 cardinality constrained mean-variance model，並且將 PSO 的結果與 GA，SA 和 Tabu 進行比較。其結論認為 PSO 要較其他方法更具有優勢。

Atsalakis and Valavanis (2009) 也整理了超過 100 篇正式出版的學術論文，其彙整主題則是圍繞於軟性計算 (soft computing) 在股市及財務投資上的各種應用模式，因此可以得知此一研究領域相當具有挑戰性，並且匯集相當多研究者的學術成果。

3.研究方法

一般的股票投資策略，大致可以分為基本面分析與技術面分析兩大類。基本面分析著重於股票市場中公司的基本財務資訊作為未來投資的主要參考，其投資策略背後的理念認為目前公司的財務表現(已經發生的歷史)，會影響目前或是未來公司的價值(市場上的股價表現)。而在另外一方面，技術面分析則認為影響未來股價波動的因素太多，不如從過去歷史的價格或是成交量的變化形態得出下一個轉折的蛛絲馬跡。因為已經公布的財務資訊可能已經反應在價格上，同時未來的財務展望又無法得到完全的確認，但是這些因素會以不

同方式影響股價與量的趨勢變化，所以可以藉由判斷股票未來的獲利可能性。

本研究主要著重於基本面分析的數據變數，亦即採用每季公布的公司財務報表中所得到的相關財務指標，作為模型分析所需要的自變量主要組成。同時，因為分析的基本時間單位為以季為單位，所以無法採用一般市場上慣用的技術指標(如 RSI, KD 等)，但是會採用少數的簡單技術指標作為分析用的自變量。

本研究分析的對象為在台灣證券交易市場中的上市上櫃股票 (不包含金融相關類股)，分析時間從 2000/01/01 開始，到 2009/12/31 為止，以每一季做為分析的基本時間單位。模型的預測能力評估方式為計算選股策略的獲利率，亦即藉由本研究產生的預測訓練模型，以及上一季所公布的財務指標，挑選下一季的股票投資組合，計算這些投資標的在下一季所產生的獲利率，且與買進並持有策略(Buy & Hold)所產生的獲利率進行比較，以避免可能的趨勢偏差(bias)。

3.1 分析模型及研究架構

本研究中，訓練模型藉由 2009 年第二季的財務指標資料以及技術指標資料作為自變數，應變數為第四季的股價變化 =(期末股價 - 期初股價)/期初股價，接著藉由最佳適應函數值找出最佳訓練模型，輸入第三季的自變數，產生 2010 年第一季的選股範圍。最後，以此選股範圍進行計算預測報酬率= (期末股價 - 期初股價)/期初股價。

本研究希望以最少的變數訓練出最準確的模型。因此將採用 Wrapper 的方式結合 PSO 與 SVM(以及 SVR) 進行變數篩選與參數最佳化，找出表現最好的組合。此選擇變數於每一個訓練時期皆會進行一次，以模擬時間變遷所可能造成的影響。在第二階段產生投資組合部分，則是找出表現最佳的訓練模型作為接下來測試期所採用的分析參數與模型，得出投資組合並且計算報酬率。

本研究所考慮的分析模型包括：PSO+SVM、PSO+SVR、Vote 模型 (使用投票機制決定訊號結果，投票者包括 PSO+SVM、PSO+SVR，各模型權重相同)。同時，也納入倒傳遞類神經網路 (BPNN) 模型作為 benchmark 比較之用。如圖一所示，此圖大致描繪出本研究架構為從「研究動機目的」開始，然後進行「資料蒐集及分析」，之後將相同的資料進行不同的模型分析(包含 PSO+SVM，PSO+SVR，Voting，BPNN)，最後將實驗數據進行「模型分析與績效比較」和得出「實驗結論」。

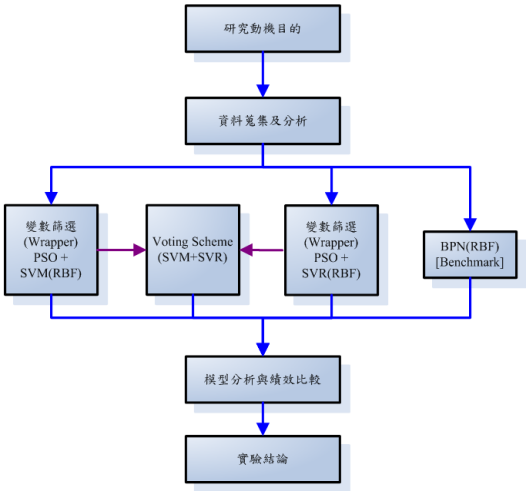

圖 1 分析模型架構

3.2 自變量之基本面財務指標

自變量的基本面財務指標部分以公司所公佈季報資料為計算基礎。季報實際公布時間，將根據證券交易法第 36 條的條文內容：

『已依本法發行有價證券之公司，應於每營業年度終了後四個月內公告並向主管機關申報，經會計師查核簽證、董事會通過及監察人承認之年度財務報告。其除經主管機關核准者外，並依左列規定辦理：

一、於每半營業年度終了後二個月內，公告並申報經會計師查核簽證、董事會通過及監察人承認之財務報告。

二、於每營業年度第一季及第三季終了後一個月內，公告並申報經會計師核閱之財務報告。

三、於每月十日以前，公告並申報上月

份營運情形。』

至於季報公布的內容,其資料來源來自於 TEJ 的財務資料庫:「Unconsolidated Finance DB」的「財務(單季),一般產業 IV」和「財務(累計),一般產業 IV」。同時,個股範圍並不包含「金控公司」,「金融業」,「壽險業」,「產險業」,和「證券業」。另外一方面,本研究之所以不採用 TEJ 新版財務資料庫的原因為,企業因應財務報表會計準則第 34 及 36 公報科目重分類自 2006 年起實施。所以其資料範圍無法涵蓋本研究所需。

3.3 自變量之技術面指標

自變量的技術面指標則以市場上的股價變化以及成交量資料作為計算基礎,但因為考慮到除權息效應,所以採用經過除權息調整過後的股價資料,其資料來源來自於 TEJ 資料庫:「TEJ Equity」的「調整股價(月)-除權息調整」資料庫。然後將月資料彙整成為季資料。本研究因為時間單位採取季為單位,所以採用較為簡單的技術指標作為基礎。

3.4 應變量

應變量以當季市場上的股價變化為計算基礎,但因為考慮到除權息效應,所以採用經過除權息調整過後的股價資料。其資料來源來自於 TEJ 資料庫:「TEJ Equity」的「調整股價(月)-除權息調整」資料庫,然後將月資料彙整成為季資料。

$$Return = (Ct - Ot) / Ot \qquad (公式 1)$$

公式 1 的符號定義為:Ct 代表計算當季的收盤價,Ot 代表計算當季中的開盤價。

應變量在進行 SVM 分類模型分析時,將進一步轉換為+1(漲)與-1(跌)兩種數值。而在預測階段,則將大於 0 的輸出值歸類為+1,小於 0 則歸類為-1,作為計算後續分析步驟的基礎,亦即使用 0 作為 cutoff point。至於 SVR 迴歸模型分析,則無需做轉換。

3.5 分析資料前處理

分析資料前處理為過濾股票代碼在四碼以上的股票。如第一次無擔保轉換公司債或是特別股。本研究的分析時間為 2000 年第三季至 2009 年第四季,分析時間長度為 38 季。計算相關變數的參考資料範圍則為 1999 年第一季至 2009 年第四季,共有 11 年的財務股價資料供作計算研究變數之用。

財務資料欄位中若有缺漏值 (missing value) 則處理邏輯如下:

1. 若是該欄位有超過 1/4 皆為遺漏值,則刪除此欄位。其餘欄位則予以補充。補充方式為檢視此欄位性質,若是此欄位代表數值越大越好時,則以 0 取代此遺漏值。反之,以當期最大值取代。
2. 對於被刪除之欄位,則盡量根據財務公式所參考的季財報數據重新計算,以補齊所遺漏的資料。
3. 若是匯整之後還有遺漏值,則予以剔除其代表公司的所有資料樣本。

在本研究中,股價資料的過濾原則為:

1. 若是此筆資料的股價為缺漏值 (missing value),則將此筆資料予以刪除。
2. 若是此筆資料股價開盤價低於 10 塊以下,則將此筆資料予以刪除。

本研究中映射 (mapping) 資料的時間平移方式為:

1. 因為股利資料是相對於實際獲利的全年度資料,但實際發放股利是在隔年第三季左右。所以先將全年度資料拆成四個季,再將季資料平移六個季後,如 2005 年第一季股利資料將平移至 2006 年第三季。
2. 單季普通股股本與單季財務明細資料,則依照實際季報公布時間,平移至兩個季之後,以大致模擬取得資料的時間,如 2005 年第一季財務資料將平移至 2005 年第三季。
3. 股價資料則由月股價資料,彙整成季資料後,視應變量所在季別為基礎參考時間點,另行額外分別製作

平移一(t)、二(t-1)、三(t-2)、五(t-4)個季後的資料作為自變量計算之用。

3.6 實作程式

本研究分析用程式，採用 Microsoft Visual Studio C#作為實作程式。在 SVM 方面，採用 Lin (2001) 所公開的 LibSVM 程式，並且採用改寫過後的C#版本。至於PSO部分，則參考 PSO 公式，自行開發相關程式。從實驗數據萃取、轉換、匯入，到資料清理、對照、關連，以致到後續的整合PSO(變數篩選與參數最佳化)和 SVM 於建立預測模型上，和最後的投資績效報酬率計算，皆由本研究完成相關程式的實作，且進行各部分程式的單元測試，以確認相關資料以及功能運作正確無誤。

3.7 投資組合定義

本研究根據前一期(季)資料所訓練的模型，並根據本期(季)的自變量(財務指標 F1~F52 與技術指標 T1~T13)，進行預測本期樣本中最有上漲潛力的前十名股票作為本期投資組合，且以相對應的上漲幅度作為投資績效。同時計算本期樣本中所有股票的績效作為買進及持有策略的投資績效，將目標績效扣除此績效後便為超額利潤。

3.8 PSO 編碼方式及設定參數

PSO 將同時進行變數篩選與 SVM 參數最佳化的作業。因為 SVM 參數有兩項：C 和 γ，而研究變量有 65 個(財務指標 F1~F52 與技術指標 T1~T13)，所以將 PSO 編碼為 2+65=67 維度的向量。其中 C 的搜尋範圍為 2^{-15}~2^{15} 之間，γ 的搜尋範圍為 2^{-15}~2^{15} 之間，其餘則介於 0 與 1 之間，若大於 0.5 則表示選取此變量。

PSO 在進行搜尋近似最佳解時，相關參數設定為：回合數為 10，迭代數為 100，粒子群數目為 50。在搜尋迴圈中，若是適應函數值達到停止門檻值(設定為 0.9)，則停止搜尋。

3.9 適應函數

PSO 在進行 SVM 分類時，因為本研究重點在於多方操作，所以改寫原有 LibSVM 程式，將分類的適應函數值(Fitness value)由正確率 (accuracy) 改為計算 F1 Score。若是為 SVR (regression) 時，則採用 mean squared error (MSE) 的概念計算適應函數值：1.0 – 10.0 * MSE，其原因在於調整最佳適應值的方向(越接近 1 越好)，以及放大調整 MSE 的敏感度。

3.10 Vote 模型

Vote 模型同時整合了 SVM 和 SVR 的預測值。傳統的作法是一種模型一票，但是難免會產生同票的情形。所以本研究將各模型產生的輸出值轉換成 0~1 之間的實數後進行加總，挑出排序在前面的十名個股作為當季的投資組合。

3.11 類神經網路模型

本研究另外採用類神經網路模型中的倒傳遞網路 (Back-propagation) 作為比較基準(benchmark)。輸入變數為所有 65 個研究自變量，輸出變數則為一個應變量，應用分類資料集，並且將應變量調整為 1(漲)與 0(跌)。網路架構為輸入層有 65 個節點，隱藏層有 100 個節點，輸出層有 1 個節點。迭代次數為 10,000 次，學習速率為 0.3，採用 Sigmoid activation function。

最後計算分類正確率時，若是類神經網路輸出值大於等於 0.5，則視為 1，小於 0.5 則視為 0，亦即使用 0.5 作為 cutoff point。因為此模型並未進行參數最佳化或是變數篩選，所以並不需要設定適應函數作為評估依據。為了可以與其他分析模型有相同的比較基礎，所以設定傳統的正確率作為計算適應函數值的公式，並且統計訓練期與預測期的數據作為日後比較之用。此模型產生投資組合方式一樣是以輸出值降冪排序的前 10 名股票作為當季的投資標的，

試算其相對報酬率。

4.實驗結果及數據分析

本研究的分析資料從 2000 年第 3 季到 2009 第 4 季為止，計有 38 個檔案。但是，因為實驗需求將其分成前後期為訓練/測試成對組合，因此產生 37 組實驗組數據，如第一個實驗組為 2000/09 (訓練組)與 2000/12 (測試組)，第二個實驗組為 2000/12 (訓練組)與 2001/03 (測試組)，其餘依此類推。

4.1 買進及持有績效

本研究計算各個測試組的全部樣本投資績效作為買進及持有策略 (Buy & Hold Strategy) 的投資績效，其測試組資料從 2000 年第 4 季到 2009 第 4 季為止，計有 37 個檔案。其投資績效計算方式為統計所有樣本資料的當季投資績效的平均值。每季平均投資績效則為所有投資績效加總後除以季別總數(37)。

4.2 訓練階段的適應函數值

各訓練階段適應函數值所對應的時段從 2000 年第三季 (2000/09) 到 2009 年第三季 (2009/09)。因為所有模型的適應函數值定義並不相同，所以本研究將一一檢視各模型的實際適應函數值所代表的意義。

若以 PSO+SVM 為分類模型，其適應函數的定義為計算 F1 score，而非單純計算正確率。以訓練階段所得到的實際值分布來看，最差為 0.92，最好為 1.00。所以可知其訓練用模型已經高度收斂至訓練樣本的分布特性。

至於 PSO+SVR 為迴歸模型，其適應函數的定義為計算公式：1.0 – 10.0 * MSE。從其適應函數值的實際結果分析，其值介於 0.92 與 0.98 之間。若以最差的 0.92 換算回 MSE，則 MSE = (1-0.92) / 10 = 0.008。因此也可以瞭解到此迴歸模型訓練階段的平均誤差不致於過大。

最後檢視類神經網路模型 (BPNN) 在經過 10,000 世代的訓練模型結果。因為類神經網路為分類模型，其適應函數採用傳統的分類正確率公式 = 分類正確數目 / 總樣本數目。其實際訓練值的分布情況為：0.90~0.99 有 16 組，0.80~0.89 有 16 組，0.70~0.79 有 5 組。所以 86%以上的分類正確率在 80%以上，而剩餘的 14%的分類正確率也在 71%以上。所以，觀察可知此類神經網路模型應該也可以視為收斂至訓練樣本的表現行為上。

4.3 預測階段的適應函數值

在此列出各預測階段的適應函數值，所對應的時段從 2000 年第四季 (2000/12) 到 2009 年第四季 (2009/12)。因為所有模型的適應函數值定義並不相同，所以將一一檢視各模型實際適應函數值在測試階段的不同表現。

若是將 PSO+SVM 的適應函數值與實際每季獲利率進行對照，則可以發現發生虧損的幾個時期，其對應的適應函數值大部分皆在 0.5 以下。但是，卻並非所有適應函數值在 0.5 以下的時間，皆會發生虧損。此現象的可能解釋為模型的適應函數公式為計算 F1 score，所以可能即使在 0.5 以下，卻仍然可以避開 Type-I error。

而 PSO+SVR 的適應函數值介於-0.05 與 0.97 之間。若以最差的-0.05 換算回 MSE，則 MSE = (1- (-0.05)) / 10 = 0.105。其對應負報酬率的適應函數值範圍為 0.64~0.96，所以並非適應函數值越低越會發生虧損。

至於類神經網路模型 (BPNN) 的實際適應函數值的分布情況為：0.80~0.99 有 6 組，0.79~0.30 有 25 組，0.00~0.29 有 6 組。所以 68%以上的分類正確率在 30%~79%左右，換句話說，大部分的預測率落在 50%的上下 20%範圍內。這部份的現象也符合對於類神經網路的預期表現，在訓練階段過度收斂至訓練樣本，但是在測試階段則無法有效將樣本進行分類。若是將實際每季報酬率與適應函數值進行對照，則可以發現發生虧損的幾個時間點，其適應函數

值則介於 0.29 到 0.9 之間，並無一定的對應關係。

4.4 平均投資績效比較

本研究發現若是單純將各季的報酬率進行平均，將所得到的季平均報酬率進行比較，各模型的優劣順序排名則如表 1 所示，從最佳者開始依次為 Vote 模型、PSO+SVM 模型、PSO+SVR 模型、買進及持有策略以及表現最差的 BPNN 模型。

但是，此種平均計算方式，並不能代表實際的獲利能力。因為此種計算方式只是單純將各季的報酬率加總後再除以總季數，並未考慮投入本金後，經過各期的報酬率和虧損幅度對於資產的實際變化影響，所以只能作為一種較為單純的模型比較基礎與初始的績效概略認知。

在此認知基礎上，本研究可以發現 Vote 模型與 PSO+SVM 模型的投資表現上差異不大，而買進及持有策略與 BPNN 模型的投資表現上雖有差異，但也不算太大。而在 5 個投資模型中，表現中等的為 PSO+SVR 模型，但是其投資績效明顯接近 Vote 模型與 PSO+SVM 模型，而非買進及持有策略與 BPNN 模型。所以，可以初步認知到，本研究所提出的投資分析模型明顯優於傳統分析模型 (BPNN) 或是大盤趨勢 (買進及持有策略)。而其最差的投資績效也至少有 2 倍以上 (6.81/3.01 = 2.26)，因此可以發現此種投資模型具有穩健性 (robustness) 的優勢。

表 1 各模型平均投資績效比較

投資策略	季平均投資績效
買進及持有策略	3.01%
PSO+SVM 模型	8.67%
PSO+SVR 模型	6.81%
Vote 模型	8.75%
BPNN 模型	2.28%

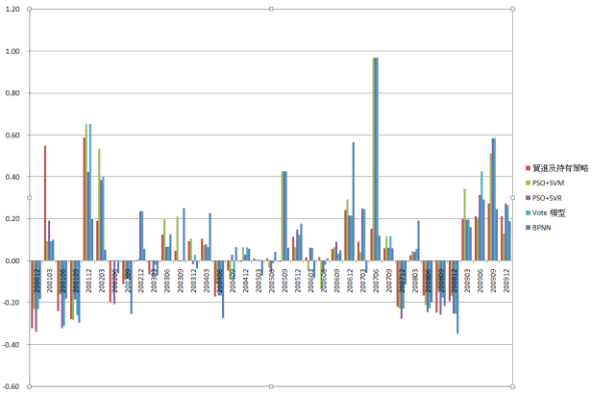

圖 2 各模型季投資績效的時間分布

本研究由圖 2 可以發現，各個模型在不同期間的績效表現，通常呈現一致的方向(正向或是負向)，但是幅度上卻有相當大的差異，因此可以得知此為造成各個模型表現差異的最大原因。

4.5 投資組合模擬

在投資組合模擬中，假設於期初投資一塊錢，不計入交易成本，經過各個時期的投資組合變化，累計其投資盈虧對於總資產的影響程度，最後得以比較各模型的累積投資報酬，其累計資產的變化趨勢如圖 3 所示。

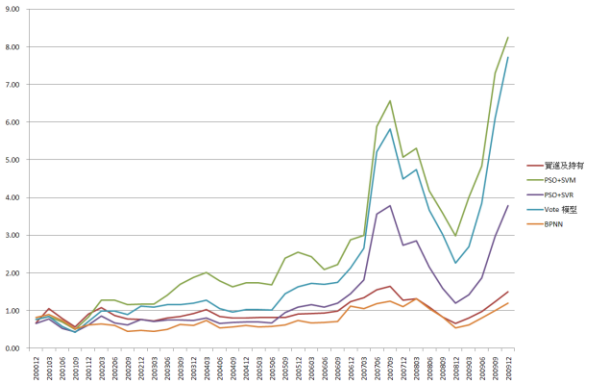

圖 3 累計投資報酬

從圖 3 可以得知，表現最好的模型為 PSO+SVM，其次依次為 Vote 模型，PSO+SVR，買進及持有策略，最後則為 BPNN 模型。若是以表現最好的 PSO+SVM 為例，其期末累積資產可以到達 8.25 元，亦即 9.25 年的累積報酬率可以達到 (8.25-1)/1 = 725%。所以，若是將此報酬率平均分攤到每一年，則平均年報酬率為 725%/9.25=78.38%。若是改採複利計算方

式，則可以換算成年複利率為 26%。($1.26^{9.25}$ = 8.48)

以期末資產來檢視各模型的優勢，則為 PSO+SVM 優於 Vote 模型，Vote 模型優於 PSO+SVR，PSO+SVR 優於買進及持有策略、買進及持有策略優於 BPNN 模型。其中可以發現 BPNN 與買進及持有策略的資產相差一直不是太大，不過買進及持有策略一直優於 BPNN 的資產，雖然在 2008 年兩者的資產幾乎重疊在一起。

雖然 PSO+SVR 模型在 2005 年上半年前皆弱於買進及持有策略，但是之後便逐漸拉開差距。至於 Vote 模型因為同時參考 SVR 與 SVM 的評分，所以其資產也一直維持於兩者之間，但是在 2007 年之後則明顯接近 SVM 的獲利表現。至於 PSO+SVM 則在 2002 年之後，逐漸拉開與其他分析模型的投資資產，此後便一直維持優勢直到期末。

所有分析模型在 2007 年第四季到 2008 年第四季皆承受重大損失。此一現象則與同時期所發生的金融海嘯有關，同時也由於本研究只限制在多方操作，所以在空頭時期，無法有效避開下跌風險。不過，本研究也發現一個有趣現象，在 2009 年第四季時，SVR、BPNN 與買進及持有策略皆已回復金融海嘯前的資產總值，但是，SVM 與 Vote 模型的投資組合卻可以在發生重大虧損之後，以優良的報酬率打敗其他投資模型，超越原先的資產總額。

本研究希望能夠計算投資模型的超額利潤 (excess return)，亦即希望能夠剔除因為整體趨勢的漲跌所帶來的利潤，而計算純粹因為投資模型本身的預測能力所得到的獲利報酬。以此基礎來計算 PSO+SVM 的超額獲利為 8.25-1.5=6.75，所以 9.25 年的整體超額報酬率為 (6.75-1)/1 = 575%，平均而言則年超額報酬率為 575/9.25 = 62.16%。若是改採複利計算方式，則可以換算成年複利率為 23%。($1.23^{9.25}$= 6.79)

本研究採取類神經網路模型 (BPNN) 作為比較基準 (benchmark)，經過投資組合模擬之後，可以清楚發現其投資表現跟買進及持有策略相似，也符合一般效率市場假說的認知，亦即投資模型不會產生超額報酬。同時，因為類神經網路是採用所有 65 個自變量，所以也可以發現即使把所有變數一起加入分析模型中，也不會增加投資預測上的優勢。

4.6 最佳化參數

本研究的訓練資料從 2000 年第 3 季到 2009 年第 3 季為止，計有 37 個檔案。根據這 37 組訓練資料，SVM 與 SVR 分別利用 PSO 搜尋最佳的參數組合。可以得知不論是在 SVM 的參數組合(C，γ)與選取的分析特徵上，並無一致性的規律特性，亦即沒有哪一個變數會跨越不同分析時段同時出現。此一現象可以被解釋為在不同時期中，可能會著重於不同性質的財務數據或是技術分析指標。譬如在某些時期，負債比例是判斷一家公司是否值得投資的重要指標，而在另一個時間點，卻可能是營收數據佔有分析上的重要比例。因為這些指標具有隨時間變化的特性，所以每一季重新搜尋具有代表意義的各種分析變數便為一件重要的工作。並且，這也與原先本研究對於股市為一動態系統的認知相符，因為若是有一分析變數維持相同的重要特性，在一段時間之後，應該會被市場分析大眾所得知，並且因而失去其內含的預測特性。

5. 結論

5.1 研究發現

本研究發現分類正確度並不能保證可獲利性，不論是用正確率或是 F1 score 去衡量，因為分類正確率是根據「分類」而產生的，而「分類」則是根據模型輸出值加上 cut-off point 判斷的產物，所以其操作方式是使用單純且不具彈性的方式進行分類切割。但是，在產生獲利的投資組合上，即使沒有產生目標分類的結果，一樣可以產生獲利結果，只要具有獲利潛力的股票可排序在前面即可。換句話說，獲利的投資組合並不要求分類界線一定非得位於 0

不可，只要是具有獲利潛力的股票可以排序在前面，即使所有樣本都並未超過 cut-off point，這也是為什麼 F1 score 雖然為 0，卻可以產生獲利組合的原因。

本研究同時也發現各時期具有影響力的變數組合並不相同，此項觀察也與 Gilli and Roko (2008)以及 Ren et al. (2006)的研究發現一致。可見股票市場所內含的系統模式是動態變化的，當某些變數被發現是具有預測影響力之後，在短時間之內就會被傳播出去而被過度濫用，以致於消失預測能力。同時也可以引申為若是採用長時間(譬如一年以上)的分析區間，則極有可能無法有效擷取在短時間內即產生變化的影響力變數。

本研究各分析模型的漲跌呈現幾乎一致波動，雖然會造成蒙受相似的虧損，但是在獲利幅度上卻有天壤之別。因此各模型內在的區辨能力仍存有本質上極大的差異，至於虧損的相似性，則應該是受到只做多方操作的研究所限制。

5.2 結論建議

本研究所提出的模型 (不論是 PSO+SVM 或是 PSO+SVR) 經過實證驗證，皆證實的確可以產生穩健且有效獲利模型，且證實優於傳統的類神經網路模型 (BPNN) 與同時期大盤表現 (買進及持有策略)。

本研究與 Huang et al. (2008) 不同之處，在於發現 Voting scheme 不見得一定是最佳模型，但是卻也可以佐證至少其具有相當優良的獲利表現。在此，本研究認為應該是由於 Voting scheme 是合併各參與投票模型的建議，因此其績效表現介於 PSO+SVM 與 PSO+SVR 之間。

在變數篩選與參數最佳化的部份，也證實在不同的變數組合下，其搭配的 SVM 參數也會隨之不同。因此在使用 PSO 進行變數篩選與參數最佳化這方面，實有其必要性。

最後根據本研究實驗結果證實 SVM 在財務最佳組合的分類問題上，的確較類神經網路模型 (BPNN)有較優良的表現。

5.3 研究限制以及研究貢獻

綜合上述，本研究的限制為：

1. 因為台灣上市公司的季財務報表公布時間與每季結算時間間隔不一致，同時公布時間也有彼此重疊之處，因此需要做相對的分析時段調整。
2. 並未考慮公司自行公布結算財報的效應與其他總體經濟影響因素等。
3. 僅限定為「作多」方向操作。

本研究的貢獻為：

1. 採用 PSO 篩選變數以及進行模型參數最佳化。
2. 動態調整模型參數以適應市場結構變遷。
3. 比較 SVM 與 SVR 在選股上的效率，並且引進 Vote 模型綜合評估。
4. 同時整合基本面財務比例與技術指標於建立投資組合上。
5. 以季為單位並且長時間 (超過 9 年) 驗證模型之有效性。

5.4 未來研究方向

受限於本研究主題著重於在找出具有高獲利特徵的投資組合，所以後續研究可以進一步加以分析各時段的重要影響變數，所代表的潛在意義和分析構面。譬如俗諺云：『漲時重勢，跌時重質』便指出在不同時間點，各種不同的分析變數所代表的重要性也會隨之而有不同的表現和更迭。

至於在另外一方面，本研究並未對不平衡 (imbalanced) 資料做進一步的處理，而只是單純地將資料進行模型分析與驗證，以確認 PSO +SVM 與 PSO+SVR 辨識出可獲利投資組合的能力。雖然在某些時期可能產生一些不必要的虧損，或許在對這些不平衡資料做進一步的處理之後，可以更好的提昇投資組合獲利率以及預測正確度。

參考文獻

[1] Atsalakis, G. and Valavanis, K. Surveying stock market forecasting techniques-Part II: Soft computing methods. Expert Systems with Applications, 36, 3, 2009, 5932-5941.

[2] Banks, A., Vincent, J. and Anyakoha, C. A review of particle swarm optimization. Part I: background and development. Natural Computing, 6, 4 2007, 467-484.

[3] Banks, A., Vincent, J. and Anyakoha, C. A review of particle swarm optimization. Part II: hybridisation, combinatorial, multicriteria and constrained optimization, and indicative applications. Natural Computing, 7, 1 2008, 109-124.

[4] Burges, C. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2, 2 1998, 121-167.

[5] Cao, Q., Leggio, K. and Schniederjans, M. A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. Computers & Operations Research, 32, 10 2005, 2499-2512.

[6] Chen, W. and Shih, J. A study of Taiwan's issuer credit rating systems using support vector machines. Expert Systems with Applications, 30, 3 2006, 427-435.

[7] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[8] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, A Practical Guide to Support Vector Classification, 2003.

[9] Cura, T. Particle swarm optimization approach to portfolio optimization. Nonlinear Analysis: Real World Applications, 10, 4 2009, 2396-2406.

[10] Eakins, S. and Stansell, S. Can value-based stock selection criteria yield superior risk-adjusted returns: an application of neural networks, International review of financial analysis, 12, 1 2003, 83-97.

[11] Gilli, M. and Roko, I. Using Economic and Financial Information for Stock Selection. Computational Management Science2008, 317?V335.

[12] Grosan, C. and Abraham, A. Hybrid evolutionary algorithms: Methodologies, architectures, and reviews. Hybrid Evolutionary Algorithms2007, 1-17.

[13] Hann, T. and Steurer, E. Much ado about nothing Exchange rate forecasting: Neural networks vs. linear models using monthly and weekly data* 1. Neurocomputing, 10, 4 1996, 323-339.

[14] Huang, C., Chen, M. and Wang, C. Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 33, 4 2007, 847-856.

[15] Huang, C., Yang, D. and Chuang, Y. Application of wrapper approach and composite classifier to the stock trend prediction. Expert Systems with Applications, 34, 4 2008, 2870-2878.

[16] Huang, W., Lai, K., Nakamori, Y., Wang, S. and Yu, L. Neural networks in finance and economics forecasting. International Journal of Information Technology and Decision Making, 6, 1 2007, 113-140.

[17] Kennedy, J. and Eberhart, R. Particle swarm optimization. Perth, Australia, City, 1995.

[18] Kim, K. Financial time series forecasting using support vector machines. Neurocomputing, 55, 1-2 2003, 307-319.

[19] Kuo, I. An improved method for forecasting enrollments based on fuzzy time series and particle swarm optimization. Expert Systems with Applications, 36, 3 2009, 6108-6117.

[20] Lin, S., Ying, K., Chen, S. and Lee, Z. Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Systems with Applications, 35, 4 2008, 1817-1824.

[21] Marinakis, Y., Marinaki, M., Doumpos, M. and Zopounidis, C. Ant colony and particle swarm optimization for financial

classification problems. Expert Systems with Applications, 36, 7 2009, 10604-10611.

[22] Olson, D. and Mossman, C. Neural network forecasts of Canadian stock returns using accounting ratios. International Journal of Forecasting, 19, 3 2003, 453-465.

[23] Quah, T. DJIA stock selection assisted by neural network. Expert Systems with Applications, 35, 1-2 2008, 50-58.

[24] Ren, N., Zargham, M. and Rahimi, S. A decision tree-based classification approach to rule extraction for security analysis. International Journal of Information Technology & Decision Making, 5, 1 2006, 227-240.

[25] Scholkopf, B. and Smola, A. Learning with kernels. Citeseer, 2002.

[26] Su, C. and Yang, C. Feature selection for the SVM: An application to hypertension diagnosis. Expert Systems with Applications, 34, 1 2008, 754-763.

[27] Tay, F. and Cao, L. Application of support vector machines in financial time series forecasting. Omega, 29, 4 2001, 309-317.

[28] Vapnik, V. The nature of statistical learning theory. Springer Verlag, 2000.

[29] Yu, L., Chen, H., Wang, S. and Lai, K. Evolving least squares support vector machines for stock market trend mining. Evolutionary Computation, IEEE Transactions on, 13, 1 2009, 87-102.

[30] Zhang, J., Zhang, J., Lok, T. and Lyu, M. A hybrid particle swarm optimization back propagation algorithm for feedforward neural network training. Applied Mathematics and Computation, 185, 2 2007, 1026-1037.