支撐向量迴歸分析在股票價格上之預測與軟體開發

柯坤志
國立東華大學
企業管理所
學生
e-mail：
carno75116@yahoo.com.tw

吳方儒
國立虎尾科技大學
資訊工程系
學生
e-mail：
william30101@hotmail.com

李威成
國立虎尾科技大學
資訊工程系
學生
e-mail：
ichiro4407@yahoo.com.tw

鄭錦聰
國立虎尾科技大學
資訊工程系
教授
e-mail：
tsong@nfu.edu.tw

摘要

在本論文中我們使用支撐向量機中的支撐向量迴歸(Support Vector Regression)對台灣的股票價格作曲線近似的模擬，再用從資料學習中所產生的 model 檔對未來的股價走勢做預測分析，並繪出趨勢圖，讓投資者了解該趨勢所在。我們提出在訓練檔案中所用的格子點搜尋演算法找出最佳化的參數組合。所謂的格子點搜尋法就是將我們所輸入的參數分出上下限，並依此上下限，對參數做切割，而從其中去找出，符合我們所期望的最佳 model 檔。最後我們採用 C#去開發成一軟體系統，及進行軟工分析。

關鍵詞：支撐向量機、股票價格、格子點搜尋

1. 前言

本論文的目的是將使用者想預測的股票從網路上下載下來，產生符合的檔案格式，再從這個檔案讓使用者去訓練出較佳的model檔，最後利用model 檔預測出未來可能會出現的趨勢。目前我們使用的輸入變數只有開盤價，但期望在未來能加入多個經濟指標及公式下，得出更完整及準確的學習及預測圖形，以便經濟學者亦或是一般使用者能運用此軟體得到未來的趨勢或者走向，幫助學者及使用者能達到預期的目標進而改善社會經濟。

1.1 投資概論

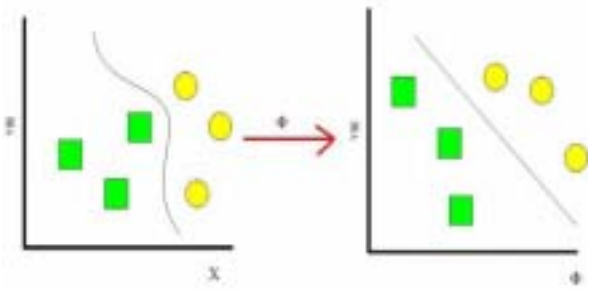
投資一詞經常在日常生活中出現，到底什麼是投資，應當每個人都具備了廣義與狹義的概念，可斟酌當時情境作適當的解釋。由於投資是著眼於未來且具有不確定性，可能賺很多錢，也可能賠錢，因此選擇投資工具亦是一門學問，我們使用SVM以及C#撰寫了股票價格預測程式，能讓有此需求之投資者，運用此程式，判斷及預測出股票日後之價格趨勢。

1.2 支撐向量機

支撐向量機(SVM，Support Vector Machines)[1~2]在近年來被廣泛應用於生物科技[3]、影像辨識[4]、文字分類[5]等領域以解決各個領域分類、預測等相關的議題。支撐向量機中包含了一個學習演算法和一個輸入空間，輸入空間內含一個訓練集和一個測試集。藉由訓練集的輸入，學習演算法可以找出一個辨識器來處理分類問題，或是找到一個迴歸式子來估算其預測值，並經由測試集我們可以知道此辨識器或是迴歸式子的正確度。

1.3 支撐向量回歸

支撐向量機(SVM)[6]，是近年來被提出用於預測的方法之一，主要是根據統計學習理論(Statistical Learning Theory)為基礎[7]，並用結構化風險最小誤差法(Structural Risk Minimization principle,SRM)之原理的學習演算法[8]。支撐向量迴歸的主要思想是針對二元分類問題，在高維度空間中尋找一個超平面作為二類的分割，以保證最小的分類錯誤率，而且SVM 一個重要的優點就是能處理線性不可分的情況。支撐向量迴歸是將自變術與應變數間的對應關係(透過一個函數 Φ)從原本較低維度提高到高維度的特徵空間中，透過這個方法尋找一新的對應函數，使投射出的預測效果最佳，如圖一所示。



圖一 SVM 概念圖

1.4 RSV 指標

威廉R指標(%R，Williams Overbought / Oversold Index)。其公式如下：

*%R=(Hn Cn)/(Hn Cn)*100*

Hn：n日內最高價。
Cn：當日收盤價。
Ln：n日內收盤價。

RSV(未成熟隨機值)就是威廉指標補數值。其公式如下：

*RSV=100-(Hn-Cn)/(Hn-Cn)*100*

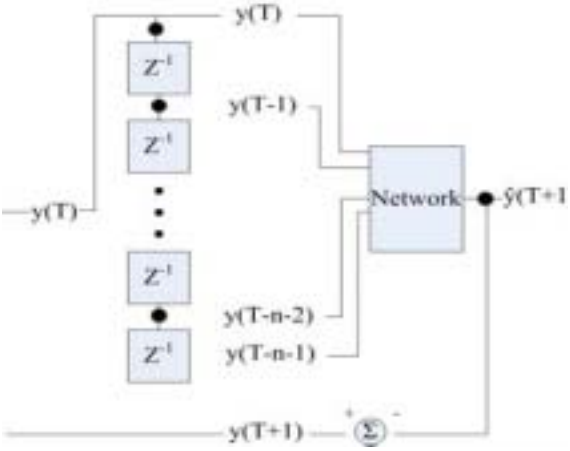
RSV是以衡量買方力道來表現，而%R是以衡量賣方力道來表現，但二者是一體二面，這二個值都是界於0~100，且RSV+%R=100。

其應用正好與威廉指標相反，亦就是RSV指標的數值越大代表賣方力道越小，數值越小代表買方力道越小，故當100-%R小於20 時，表示賣超，%R大於80時代表買超。在本論文中，採用的是RSV指標做為買賣盤訊號偵測。

2.系統架構與分析

2.1 SVR Model 之架構

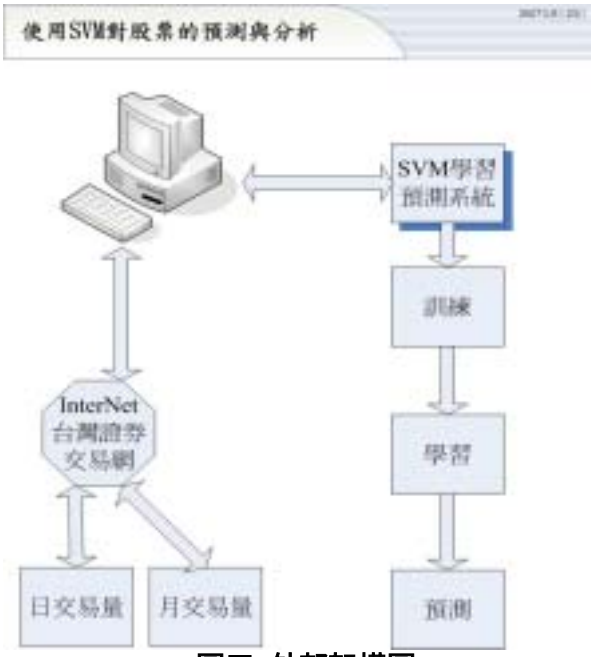
根據圖二中可看出，若我們對於原本之輸入檔予以多點輸入對一點輸出之下，學習出來之圖形會較原本準確，並且使用較少的支撐向量點即可擁有不錯的學習效率及效果，我們也用此架構寫出以舊有之參數點當做學習之經驗輸入，進而得出未來可能之預測點。在訓練檔案時，日價格預測一共分為：單點與十對一點兩種格式。其中單點為一輸入對一輸出，我們所採用的是輸入是天數，輸出是股價開盤價；而在十對一為十點輸入對一點輸出，我們所採用的輸入是其股票的前十天開盤價，對應隔天的開盤價，主要作此分別是想知道我們實際輸入股價來做預測是不是真的會比單純用天數來的準確。在月均價與趨勢預測則是一律使用十對一格式。



圖二 SVR Model 架構圖

2.2 系統外部架構

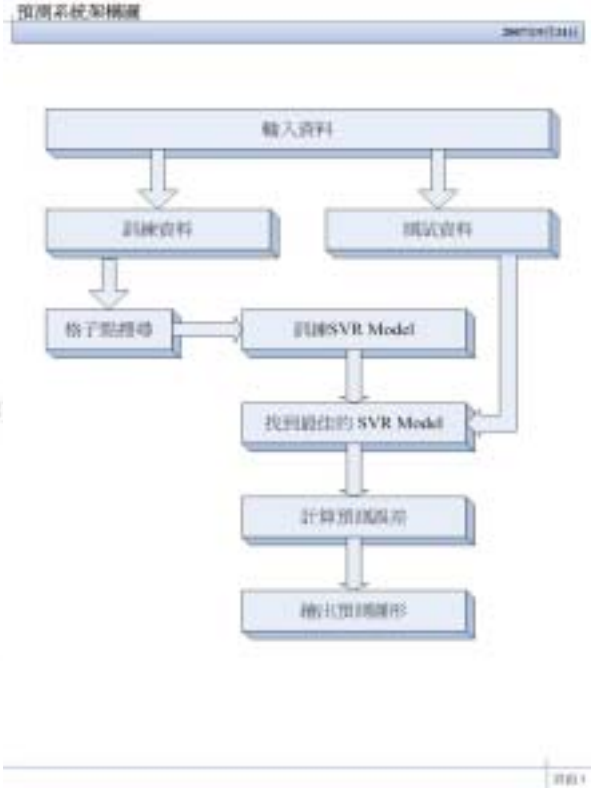
在程式環境方面，我們運用C#編寫並加入支撐向量機作為主體，針對股價做學習及預測，使用者能在我們的程式中下載過去到現在的日或月的股價，並透過系統繪出學習圖及預測圖，以便分析與判斷用，請參考圖三。



圖三 外部架構圖

2.3 系統內部架構

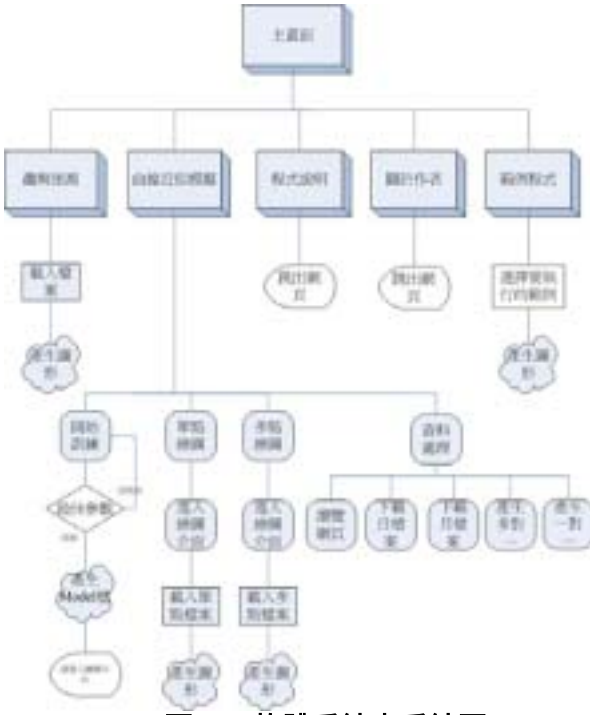
此系統主體概念在與使用者輸入資料後，可經過訓練資料後，透過格子點搜尋法找出最佳之Model檔以便未來預測之用，再透過測試資料及計算誤差後，可得出想要之預測圖形，如圖四所示。



圖四 預測系統架構圖

2.4 主介面介紹

本軟體系統之系統圖如圖五所示，



圖五　軟體系統之系統圖

使用者安裝軟體後，進入主畫面會看到五個按鈕可點擊，在個別點擊後會有對應視窗彈出，分別有曲線近似模擬、趨勢預測、程式說明、關於作者、範例程式五方面：

(1) 曲線近似模擬：

提供讓使用者在程式中即可下載需要的檔案，並且提供使用者選擇股票代碼及時間，讓使用者能多選擇是此程式較人性化部份，下載完檔案後能進入學習區塊模擬及學習曲線。

(2) 趨勢預測：

使用者若學習出參數好的Model檔後，能運用此區塊判斷接下來幾天或幾月的趨勢，當然，在此的判斷只是經過SVM學習後的假設曲線，並不足以真正影響使用者對股票漲跌的判斷，僅供參考。

(3) 程式說明：

給予初次使用此程式的使用者類似Help檔的區塊，若有操作或是流程方面不懂的部份，能到此查詢。

(4) 關於作者：

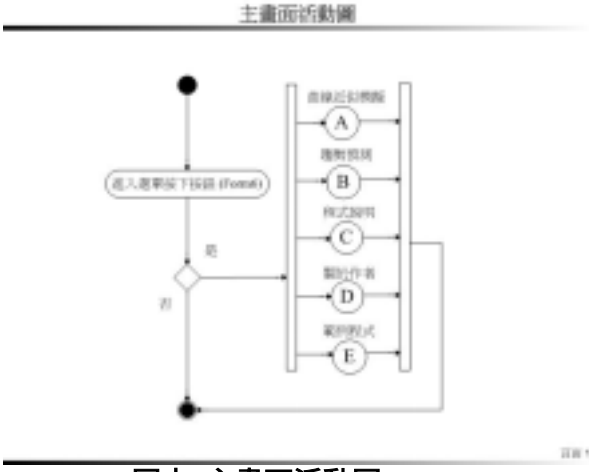
此區塊放置我們自己的檔案及專題老師的資訊，給予對此有興趣的同好能互相切磋及交流的方法及訊息。

(5) 範例程式：

對於較需範例參考方可了解此程式的使用者設計的區塊，畢竟文字只能概述無法真正傳達精隨於使用者，為了解決此問題，特別想到若先做好範例，給使用者直接看了後，應能有概略了解加上簡單實際操作後，及可熟悉此程式。

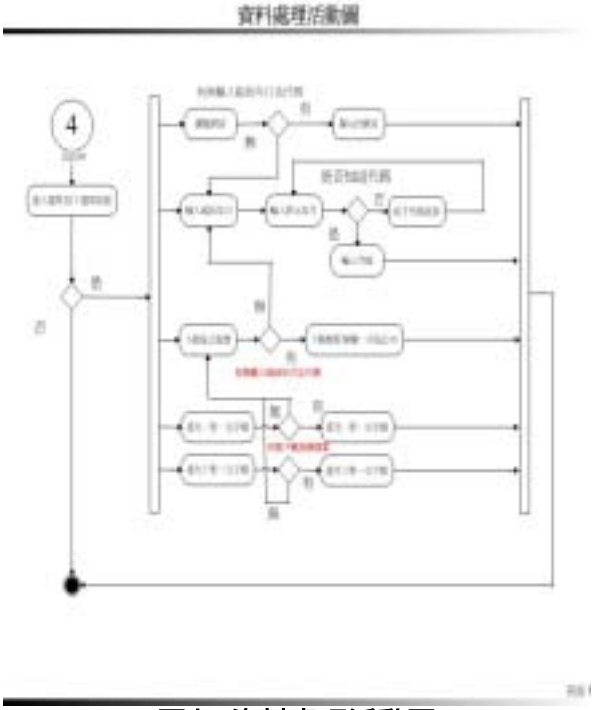
2.5 系統之活動圖

經由圖六所示之活動圖，可清楚看出我們程式的主體由主畫面開始延伸建構，再經由五個子選項可得使用者所欲之資訊，分別試曲線近似模擬、趨勢預測、程式說明、關於作者、範例程式，以下先以進入曲線近似模擬為例說明。



圖六 主畫面活動圖

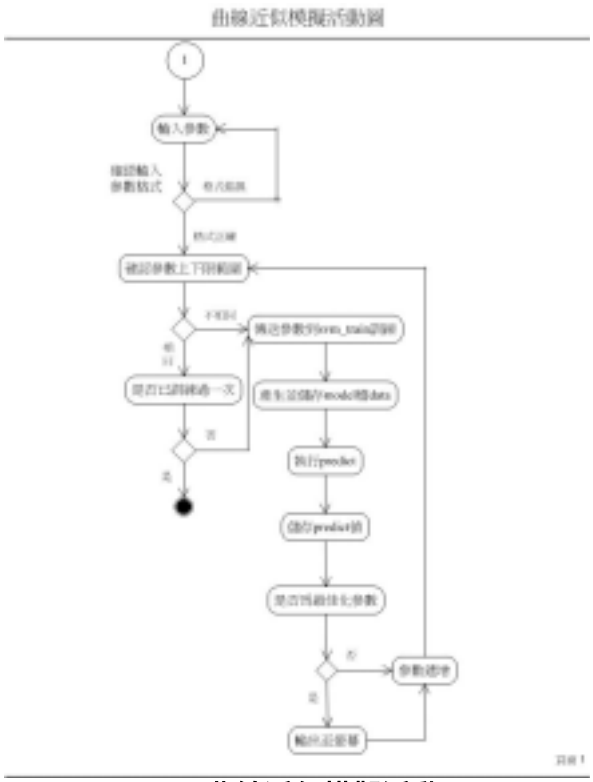
再透過主畫面進入曲線近似模擬後，能進而進入資料處理如圖七所示，可下載股價檔案。



圖七 資料處理活動圖

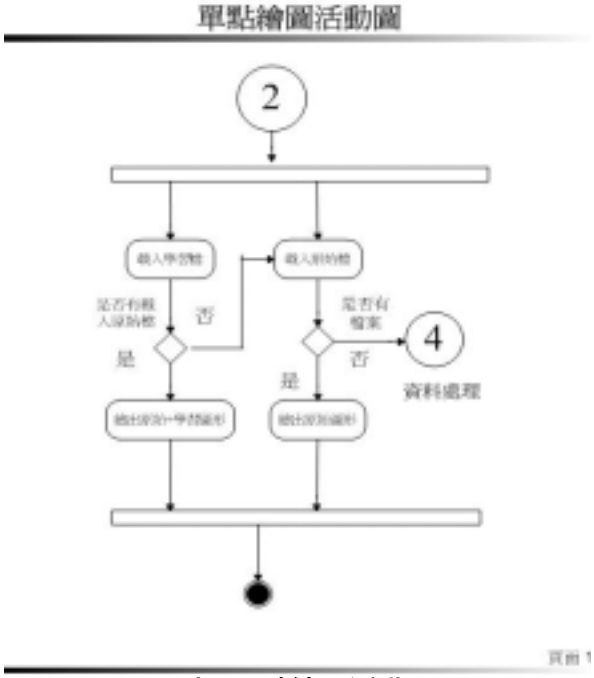
並且在此載好所需檔案後，能夠轉換成程式中所需之必要格式，而在轉換之中，RSV 指標亦會跟著必要格式檔轉換而出，在往後趨勢

預測時能夠使用此檔判斷目前的股價及趨勢。由資料處理後轉成必須之檔案，緊接著回曲線近似模擬視窗，可進行股價檔案之訓練及學習，從圖八看出此動作之流程架構。

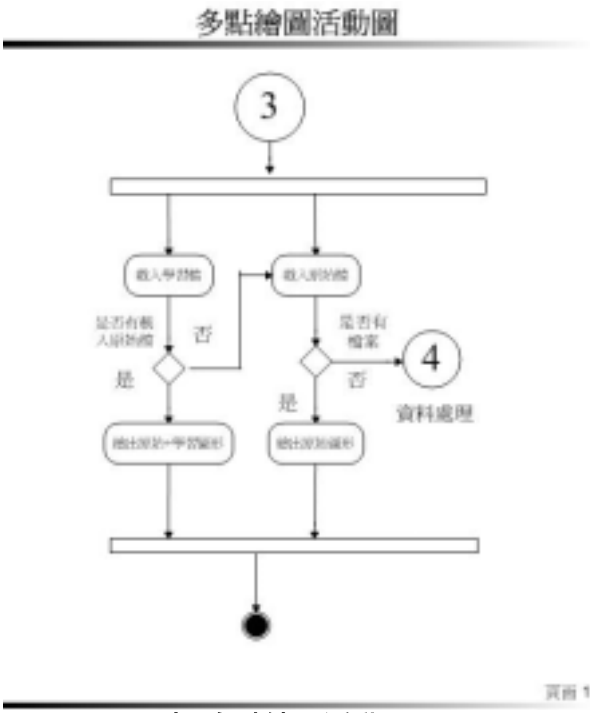


圖八 曲線近似模擬活動圖

在訓練出Model檔後，能經由單點繪圖或者多點繪圖與原始檔進行比較，進而觀察學習出檔案之相似度，如圖九及圖十所示，可看出進行繪圖之動作流程。

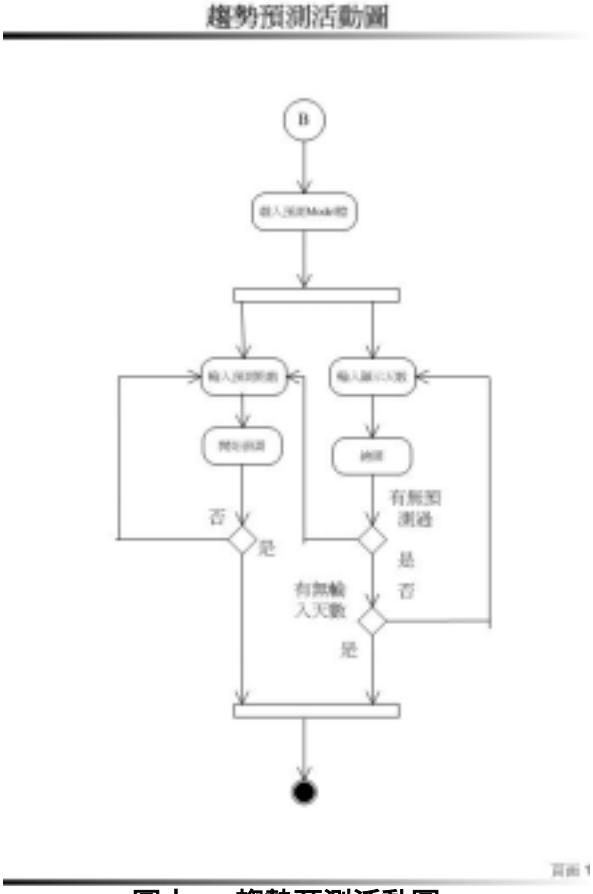


圖九 單點繪圖活動圖



圖十 多點繪圖活動圖

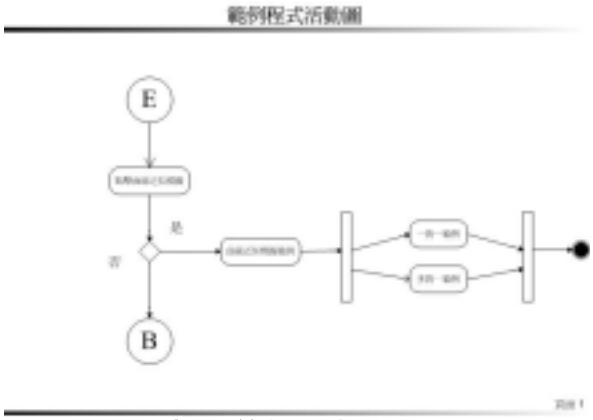
完成上述步驟後，即可退回主畫面進行趨勢預測步驟，如圖十一所示，能讀入訓練出之Model檔進行預測模擬，並透過RSV指標的輔助，得到適切的資訊。



圖十一 趨勢預測活動圖

另外在主畫面亦有範例程式可供使用者

參考，如圖十二所示。



圖十二 範例程式活動圖

3.結果與討論

對於支撐向量機迴歸模型之預測績效，以MSE(mean square error)及總支撐點向量(Total_SV)等統計量來評估，統計量的定義如表一所示。當MSE愈小則預測出之值愈接近真實值，而總支撐點向量愈小則代表使用較少點數可學到一樣準確的數值，能減少程式負擔且有較大延展性於爾後更多點之學習與預測之用。

表一

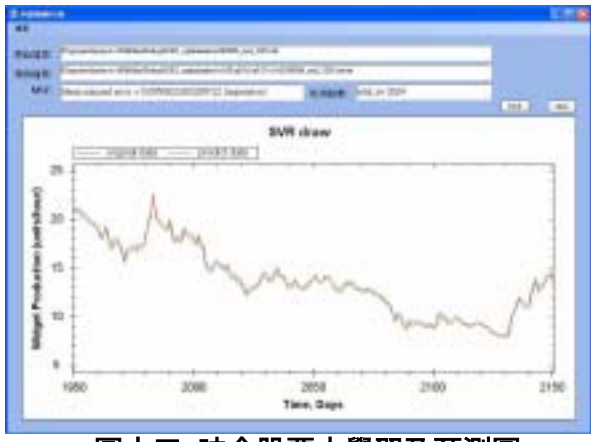
統計量	定義
MSE	$MSE = \Sigma(a_i - p_i)^2 / n$
Total_SV	使用總支撐向量數

a i：真實值　　p i：預測值

本論文使用C.J.Lin所開發之LIBSVM (aLibrary for Support Vector Machines)工具及C#編寫之程式進行開發及驗證。

在本實驗之驗證階段，使用格子點演算法技術於訓練集上找出最佳化之參數組合，再利用此參數組合以訓練資料集建實際的支撐向量機迴歸預測模型，最後再以測試資料集來評量支撐向量機迴歸預測模型對於新進資料之預測能力如何。而在此我們實際跑了兩個範例程式以供觀看與測試，分別為聲寶以及味全兩家公司。

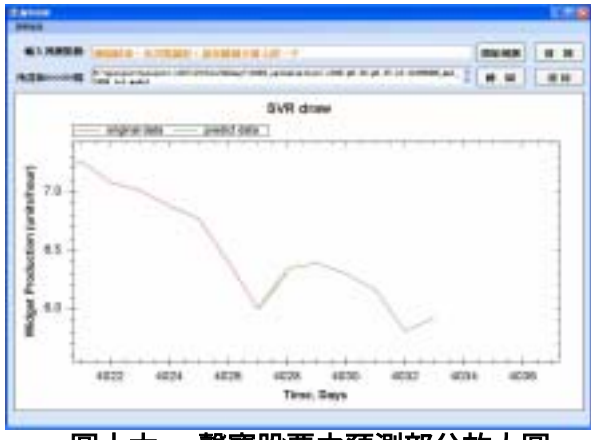
請參考圖十三及圖十四為範例一，此兩圖為味全公司從民國81年至96年個股日成交資訊之圖形，大約有4000多點以供學習，由圖可看出，學習及預測出之圖形準確度還蠻高的，證明我們的程式是有準確性在的，而在爾後證明MSE及TSV的過程差異性不大，鑑於此，我們只貼出學習時之圖形以及最後預測出的圖形，而證明MSE以及TSV我們以下一個範例作詳解。圖十五及圖十六為範例二。



圖十三 味全股票之學習及預測圖

圖十四　味全股票之預測部分放大圖

圖十五　聲寶股票之學習及預測圖



圖十六　聲寶股票之預測部分放大圖

另外我們將聲寶從民國81年至96年個股日成交資訊做為範例程式，大約有4000多個參數點供學習，以驗證預測出之迴歸模型是否接近真實值，以此我們整理出表二(a)及表二(b)所示。

表二(a)

	C = 1	C = 10
G=0.04	1831(TSV) 1.615(MSE)	1701(TSV) 0.1996(MSE)
G=0.08	2072(TSV) 3.135(MSE)	1929(TSV) 0.1771(MSE)
G=0.16	2412(TSV) 5.8052(MSE)	2192(TSV) 0.2731(MSE)
G=0.32	2829(TSV) 11.197(MSE)	2601(TSV) 0.5514(MSE)
G=0.64	3242(TSV) 21.241(MSE)	2969(TSV) 0.8911(MSE)
G=1.28	3621(TSV) 32.194(MSE)	3336(TSV) 1.4564(MSE)

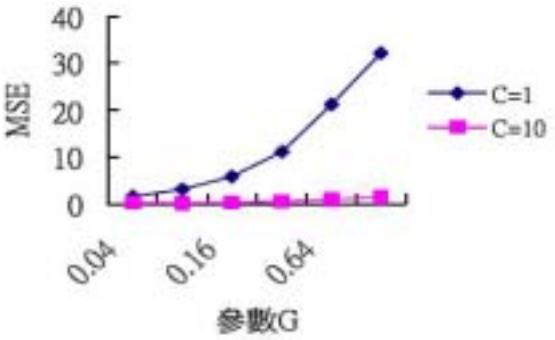
表二(b)

	C = 100	C = 1000
G=0.04	1761(TSV) 0.1285(MSE)	2005(TSV) 0.0983(MSE)
G=0.08	2004(TSV) 0.0981(MSE)	2224(TSV) 0.0856(MSE)
G=0.16	2257(TSV) 0.0872(MSE)	2319(TSV) 0.0855(MSE)
G=0.32	2542(TSV) 0.0913(MSE)	2537(TSV) 0.0909(MSE)
G=0.64	2923(TSV) 0.0992(MSE)	2921(TSV) 0.0992(MSE)
G=1.28	3286(TSV) 0.1072(MSE)	3286(TSV) 0.1072(MSE)

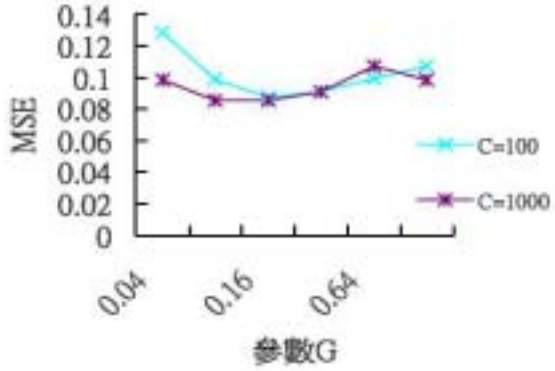
相關參數調整比較之結果可參圖十七(a)、(b)、圖十八、圖十九及圖二十這四個圖形。

從MSE觀點，由圖十七與圖十八看出，在固定參數P之下，參數C愈大，MSE跟著變小，但是MSE不一定會隨著調整參數G的比例而增加或減少，以此証明出，若調整參數C對MSE的影響變化會較好。

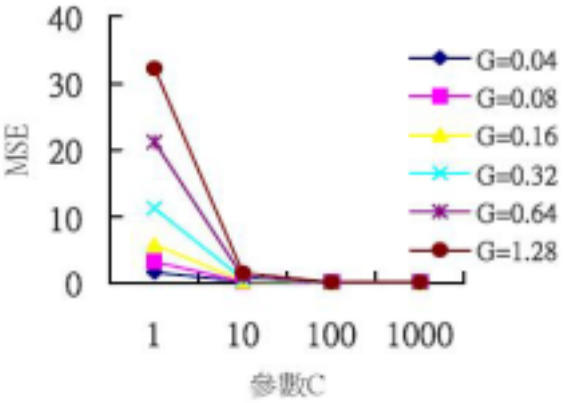
從TSV觀點，由圖十九可看出，若調整參數G愈大，則使用之TSV愈多，再反觀圖二十，儘管參數C愈大，但TSV不一定隨著增加或減少，由此可知，TSV最主要參數應為G，而TSV大小會帶來的影響即是，當使用較多TSV時，Model檔亦會學習到較多之極端值，而模擬出來之Model對於預測未知檔案時的容許誤差度會較小，造成圖形之不準確。



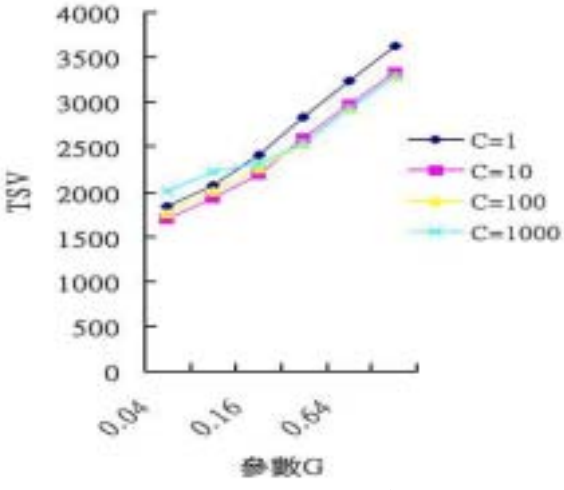
圖十七(a) 參數G對MSE影響



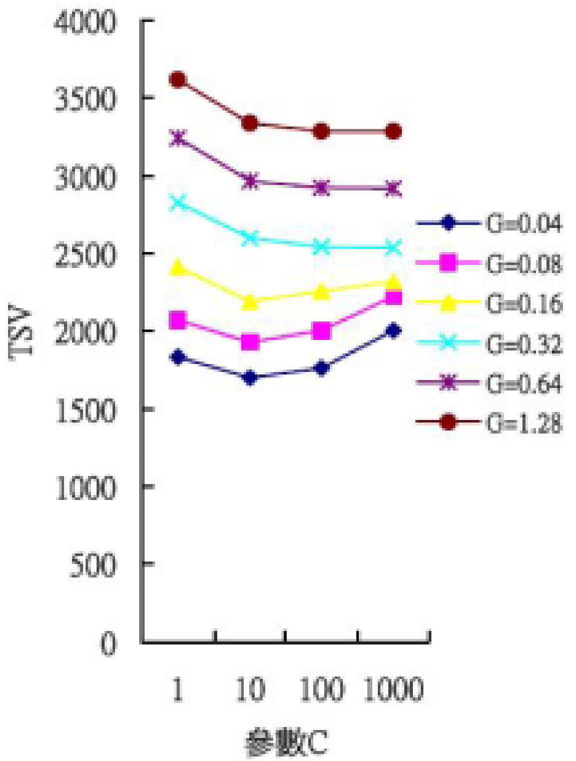
圖十七(b) 參數G對MSE影響



圖十八 參數C對MSE影響



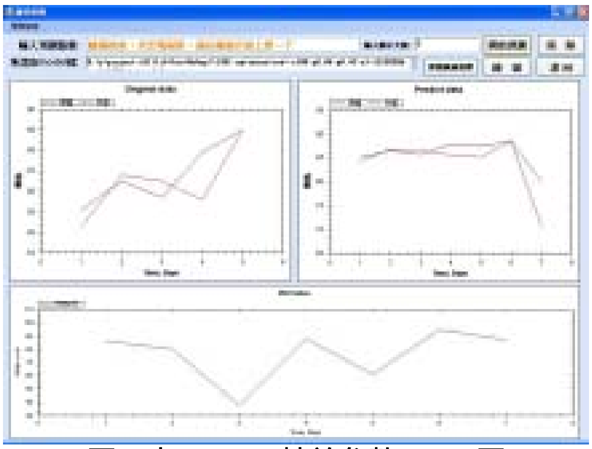
圖十九 G對TSV影響



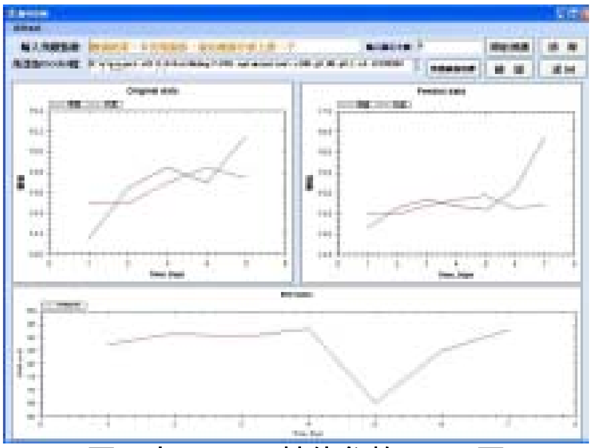
圖二十 參數C對TSV影響

目前對於TSV知多少對於未知檔案學習的影響，僅知介於原始檔案的一半左右為最佳，過多則會陷入區域最佳化即過度學習，過少則需要判斷MSE之大小，MSE小而TSV 使用少則學習出來之檔案為較好的，反之則為可忽略之學習不好的檔。

在知道如何找到較佳之Model後，亦可預測出較準確之趨勢及圖形，圖二十一所示之，此圖形為較差之Model預測之趨勢，與圖二十二運用較好之Model預測得出之趨勢可清楚看見，較好之Model預測出之趨勢較為準確許多，以此可知，參數的選擇正確性是非常重要的。



圖二十一 較差參數Model圖



圖二十二 較佳參數Model圖

本系統分別台塑一對一、十對一，聲寶一對一、十對一，味全一對一、十對一和聲寶，單點與多點分別找出的最佳化參數，以及他所使用的支撐向量機點數與MSE之結果請參表三。

表三 最佳化參數、使用的支撐向量機點數及MSE之結果

檔名	資料點數	最佳化參數	TSV	MSE
台塑一對一輸入(天)	4030	C100、G0.16、P0.25	3984	0.184127980088845
台塑十對一輸入(天)	4029	C100、G0.16、P0.35	2678	0.240231234258847
聲寶一對一輸入(天)	4030	C100、G0.16、P0.2	2651	0.0437820008832077
聲寶十對一輸入(天)	4029	C100、G0.32、P0.35	2542	0.0913632398366051
味全一對一輸入(天)	4038	C100、G0.16、P0.35	2569	0.0887141605485018
味全十對一輸入(天)	4040	C100、G0.04、P0.45	2634	0.0398823383200712
聲寶十對一輸入(月)	162	C100、G0.02、P0.3	140	0.0820383478818383

4. 結論與討論

本論文之軟體剛開始只對下載來之檔案中的開盤價作為訓練預測之指標，而尚未加入其他之經濟指標或可影響判斷Model準確性之數值，不過後來我們將這方面的功能改善，加入了開盤、收盤、最高、最低價的數據，並加入RSV指標作為另一判斷之依據，比先前只能透過開盤價來預測整體股價體制有明顯的改善。在經過這段時間以SVR參數調整對於股票價格的預測，發現參數C對於MSE的影響較大，而參數G則是TSV的影響較大。我們就試著跑一些範例，讓我們這個理論能夠得到證實，果然在P變動不大的情況下，只要藉著調整參數C就能夠控制MSE的大小，而在調整參數G的時候又發現TSV的變化。此時發現一個重要的問題，要兼顧MSE及TSV的情況之下，調整出來的參數往往在預測圖形上會不夠準確，所以我們以MSE為主，TSV為輔的情況下去找較佳的參數，以此組合跑範例後發現，預測出來比先前準確許多了。

誌謝

The authors wish to thank that this work was supported by National Science Council Under Grant NSC 96-2221-E-150-070-MY3.

參考文獻

[1] A. Marti, "Support Vector Machine," IEEE Intelligent Systems , pp.18~28 , 1997.

[2] V. Cherkassky, F. Mulier, "Learning from Data:Concepts , Theory , and Methods , New York:Wiley-Interscience," 1998.

[3] W. Brock, L.J. Josef, and B. Lebaron, "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns," Journal of Finance, vol. XLVII No. 5, 1731-1764,1992.

[4] M. Pontil and A. Verri, "Object recognition with support vector machines," IEEE Trans. On PAMI, 20, pp. 637-646, 1998.

[5] T. Joachims, "Text categorization with support vector machines, " In Proceedings Of European Conference on Machine Learning(ECML), 1998.

[6] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola and V. Vapnik, "Support vector regression. machines," In. M.C. Mozer, M.I. Jordan, and T. petsche, editors, Advances in Neural Information Processing Systems, Vol. 9, MIT press, Cambridge, MA, pp. 155-161, 1997.

[7] V.N. Vapnik, "The Nature of Statistical Learning Theory," Springer, New York, USA , 1995.

[8] B. Schkopf, C. Burges and V. Vapnik, "Extracting support data for a given task," In First International Conference on Knowledge Discovery & Data Mining, Menlo Park, pp. 252-257, 1995.

[9] A.K Keng and Q. Chai, "Stock Trading Using RSPOP:A Novel Rough Set-Based Neuro-Fuzzy Approach," IEEE Transactions On Neural Networks, Vol. 17, No.5, pp. 1301-1315, September 2006.

[10] 祁亨年 "支持向量機及其應用綜述"， 計算機工程期刊， 2004 年 5 月第 30 卷第 10 期。

[11] 黃承龍、蔡承益 "支援向量迴歸結合自我組織特徵映射圖用於預測台灣股票指數期貨"， 中華民國第十四屆模糊理論及其應用會議， 2006 年 12 月。

[12] 宋曉峰、陳德釗 胡上序 "支持向量迴歸估計性能分析"， 計算機與應用化學期刊， 2005 年 7 月第 22 卷第七期。

[13] 齊志泉、田英杰、徐志浩 "支持向量機的何參數選擇問題"， 控制工程期刊，2005 年 7 月第 12 卷第 4 期。

[14] 林長青 支撐向量機應用於科學探索，國立雲林科技大學電子與資訊工程研究所碩士論文， 2003 年 6 月。

[15] 謝宗霖 資料探勘於半導體測試工程之應用， 國立高雄第一科技大學資訊管理系碩士論文， 2007 年 5 月。

[16] 黃承龍、陳穆臻、王界人 "支援向量機於信用評等之應用"， 計量管理期刊，2004 年第一卷第二期第 155~172 頁。