科技部補助專題研究計畫成果報告

（□期中進度報告 / □期末報告）

**計畫名稱：程式交易之深度學習架構與基因演算法探討**

計畫類別：□個別型計畫 □整合型計畫

計畫編號：MOST 109 - 2813 - C - 004 - 013 - E

執行期間：2020年07月01日 至 2021年02月28日

執行機構及系所：國立政治大學資訊科學系

計畫主持人：彭彥璁

共同主持人：無

計畫參與人員：林顥倫

本計畫除繳交成果報告外，另含下列出國報告，共 \_0\_ 份：

□執行國際合作與移地研究心得報告

□出席國際學術會議心得報告

□出國參訪及考察心得報告

中華民國 110 年 03 月 05 日

科技部補助專題研究計畫成果自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值(簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性)、是否適合在學術期刊發表或申請專利、主要發現(簡要敘述成果是否具有政策應用參考 價值及具影響公共利益之重大發現)或其他有關價值等，作一綜合評估。

|  |
| --- |
| 1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估  □ 達成目標  □ 未達成目標（請說明，以 100 字為限）  □ 實驗失敗  □ 因故實驗中斷  □ 其他原因  說明： |
| 2. 研究成果在學術期刊發表或申請專利等情形（請於其他欄註明專利及技轉之 證號、合約、申請及洽談等詳細資訊）  論文：□已發表 □未發表之文稿 □ 撰寫中 □ 無  專利：□已獲得 □申請中 □ 無  技轉：□已技轉 □洽談中 □ 無  其他：（以 200 字為限） |
| 3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性，以 500 字為限）。  我們開發出LSTM-Based Soft Stock Trend Model(LSSTM)此ㄧ以深度學習為基礎的股價漲跌趨勢模型，具備相當的能力能夠預測出股價的漲跌趨勢，達到88%的精準度。 另外，再搭配我們自定義的進出場策略，能夠達到47.13%的投資報酬率，勝過大盤的 32.09%。未來可望將股票的特性進行分類，依照不同類股做模型訓練，已達到更好的效果以及報酬率。 |
| 4. 主要發現  本研究具有政策應用參考價值：□否 □是，建議提供機關\_\_\_\_\_\_\_ (勾選「是」者，請列舉建議可提供施政參考之業務主管機關) 本研究具影響公共利益之重大發現： □否 □是  說明：（以 150 字為限） |

（壹）摘要

財富管理一直以來都是非常重要的議題，尤其是近幾年，因為ETF 等商品出現，使得許多人的投資意識提高，而在資產配置方面，又多了一種選擇。過去，多數人經常保守地利用定存，當作唯一的投資手段，不過，隨著銀行利率逐年調低，在考慮過通貨膨漲後，該定存資產將逐年縮水。意識到這件事之後，許多人開始逐步研究各種基金、股票、期貨、選擇權，甚至是國外債券等。除此之外，在金融界也升起一股程式交易的浪潮，像是最著名的橋水基金(Bridgewater Associates)，或者是美國最大的金融機構摩根大通集團(JPMorgan Chase)，都開始研發人工智慧選股、智能投資，進一步幫助投資人作更好的投資策略建議。我們開發出LSTM-Based Soft Stock Trend Model(LSSTM)此ㄧ以深度學習為基礎的股價漲跌趨勢模型，利用多層的LSTM以及資料的前處理，我們具備相當的能力能夠預測出股價的漲跌趨勢，達到88%的精準度。另外，再搭配我們自定義的進出場策略，能夠達到47.13%的投資報酬率，勝過大盤的32.09%。

（貳）關鍵字

Program Trading、Stock Market Prediction、程式交易、股市預測

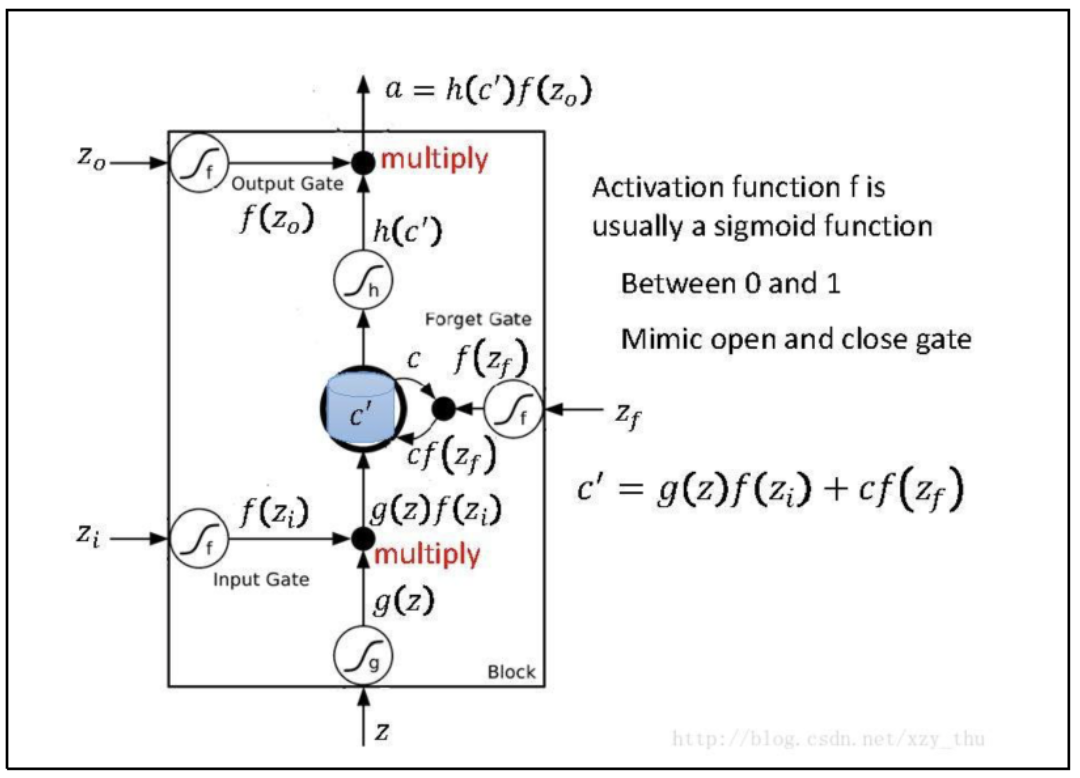
（參）研究目的

在投資的路上，除了擁有好的投資的標地外，最重要的就是「資產配置」。一般而言，資產配置與投資標的選擇的同等重要，好的資產配置不但可以使我們的風險降低，也可以優化我們投資效益，就算市場出現波動也不會造成投資人的資產崩盤。因此我們也會在這項研究中，經由我們所設計的基因深度學習模型，應用在股票投資上，預測股票未來的漲跌，並將這些預測，結合進出場策略，觀測是否能擊敗大盤指數及其他類似標的的基金，以幫助投資人增加獲利績效。另外，我們也會藉由所設計的系統，試著找出最佳的資產配置。

由過往的文獻及資料中，我們不難看出不管是長短期記憶模型Long Short-Term Memory (LSTM)或是基因演算法對於股票市場的預測，只要進行過回測、調整好參數皆有能力打敗大盤。因此在這次研究項目之中我們會先個別做出LSTM及基因演算法之模型，再進行LSTM與基因演算法之組合模型。本次研究著重的重點在於LSTM、基因演算法、LSTM 與基因演算法之組合模型，我們希望能透過兩者的結合找出最大獲利。另外，在LSTM與基因演算法的架構中，由於「記憶性」、「選擇性」等等兩個演算法所著重的地方不盡相同，因此在設計的過程中我們也會考慮到這兩者結合時所產生可能不合的問題以及觀測其表現。

最後，在此研究架構的基礎上，由於股市每日都會有新的變化，我們希望這套系統除了能夠回測過去的紀錄之外，也同時希望在未來能實際幫助到投資人，因此取得每日最新的股價是必要的條件，而由於新股價的進入對於LSTM記憶性的影響，對於「買進」、「賣出」訊號的看法也會因為時間性而不同。

（肆）文獻探討



圖一、LSTM架構（取用自[2]）

如圖一所示，LSTM[1]是基於遞迴神經網路（Recursive Neural Networks, RNN）所改良而產生的架構，RNN 限制於於短期記憶的問題，因為RNN只參考到附近資料的數值，當預測的資料讀取到後面時，越前面的數據對當前數據的影響力慢慢消失，容易完全捨棄，因而，如果嘗試處理一段時序較長的資料，在開始時RNN 極可能就會遺漏重要資訊。為能解決上述RNN 短期記憶的問題，LSTM 其架構具有三個Gate：Input、Forget 和Output gate，用來強化記憶力，避免梯度消失。

LSTM 有四個計算步驟[3]：Forget Gate 能決定哪些數據應該被記得或遺忘，來自當前輸入值𝑥𝑡和先前隱藏狀態ℎ𝑡−1同時輸入到Sigmoid 函式，輸出值𝑓𝑡位於0和1之間，越接近0表示越應該被遺忘，越接近1表示越應該被留下，如式（1）。

𝑓𝑡=𝑠𝑖𝑔𝑚𝑜𝑖𝑑(𝑥𝑡+ℎ𝑡−1) (1)

Input Gate用來對當前單元狀態𝑐𝑡做更新，將當前輸入值𝑥𝑖和先前隱藏狀態ℎ𝑡−1輸入到Sigmoid函式，調整輸出的數值到0跟1之間，得到𝑖𝑡，來決定哪些資訊該被更新，如式（2）。

𝑖𝑡=𝑠𝑖𝑔𝑚𝑜𝑖𝑑(𝑥𝑖+ℎ𝑡−1) (2)

另一方面，當前輸入值𝑥𝑖和先前隱藏狀態ℎ𝑡−1傳輸給tanh 函式，並將數值壓縮在-1

和1之間以對網路數值進行調節，得到𝑐̅𝑡，如式（3）。

𝑐̅𝑡 = 𝑡𝑎𝑛ℎ( 𝑥𝑡 + ℎ𝑡−1 ) (3)

New Cell State負責計算新的單元狀態𝑐𝑡。將上述計算所得之𝑖𝑡和𝑐̅𝑡相乘，Sigmoid輸出值𝑖𝑡將決定在tanh 輸出值𝑐̅𝑡中哪些資訊需要被保留且是重要的，然後將Forget Gate的輸出值𝑓𝑡和上一個的單元狀態𝑐𝑡−1相乘，如果它乘上接近零的數值，則表示在新的單元狀態中可能將會丟棄這個值，最後相加得到一個新單元狀態𝑐𝑡，如式（4）。

𝑐𝑡 = 𝑖𝑡 \* 𝑐̅𝑡 + 𝑓𝑡 \* 𝑐𝑡−1 (4)

Output Gate用來計算下一個隱藏狀態的值ℎ𝑡，隱藏狀態中包含了過去輸入的相關資訊。將先前隱藏狀態ℎ𝑡−1和當前輸入數值𝑥𝑖輸入給Sigmoid 函式，得到Output Gate 輸出值𝑜𝑡，如式（5）。

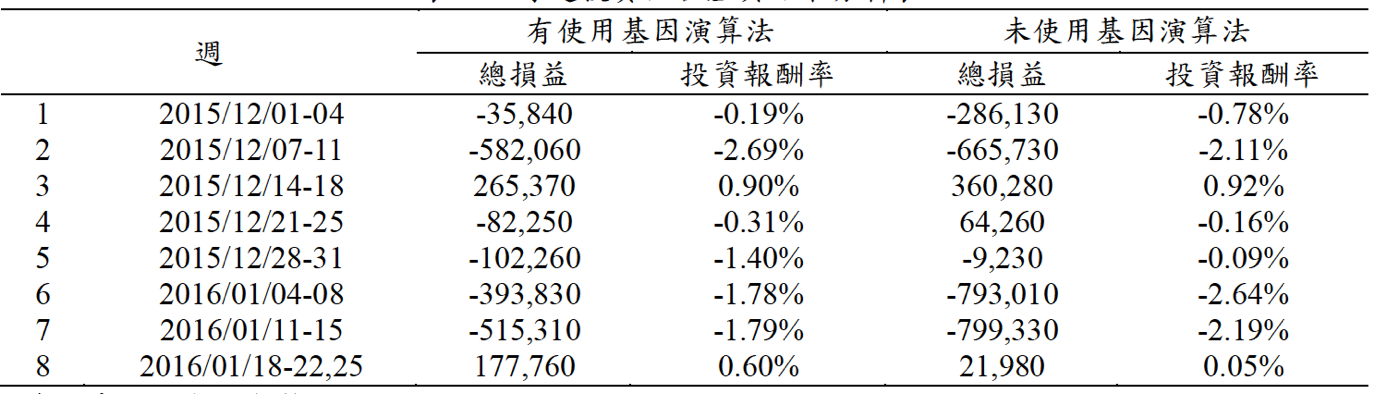
𝑜𝑡 = 𝑠𝑖𝑔𝑚𝑜𝑖𝑑( 𝑥𝑡 + ℎ𝑡−1 ) (5)

接著將計算所得之新的單元狀態𝑐𝑡傳輸給tanh 函式；最後將Output Gate 輸出值𝑜𝑡和tanh 輸出值相乘，以計算出新的隱藏狀態ℎ𝑡應擁有的資訊，如式（6）。

ℎ𝑡 =𝑜𝑡 \* 𝑡𝑎𝑛ℎ(𝑐𝑡) (6)

最後把新的單元狀態𝑐𝑡和新的隱藏狀態ℎ𝑡傳輸給下個時間序列的單元，新的隱藏狀態ℎ𝑡則作為目前單元輸出。在輸入特徵方面，除了基本歷史數據的輸入外，加入技術指標亦能有效預測未來價格的變動趨勢，此外，由Chen, K. et al.和黃華山 et al.所提出的論文中將相關股票的數據資料一併作為特徵輸入[5][6]，其實驗研究出相關股票之趨勢也可以是對預測目標造成影響的因素。劉昭雨、顏士淨的實驗結果中則證實特徵越多不一定提升準確率[4]，還是需要擁有足夠大量的數據，參照上述這些論文，我們的資料搜集將會使用台灣全部股票的歷年分時滑動時窗，並將股票分門別類，找出相關股票的正相關性，而特徵則會簡化，目標是找出可信度最高指標。John Holland在1992年提出的論文“Genetic Algorithms”中引用達爾文進化論中所提到的「物競天擇，適者生存」，首次提到他所設計的基因演算法(Genetic Algorithms)，將變數模擬為生物的染色體(Chromosome)，並設計一個適應函數(Fitness Function)對於環境之適應程度標準量化以比較其優劣，再透過隨機的交配(Crossover)以及交配過程中機率性產生的突變(Mutation)演化出新的子代，重複操作使子代之適應性優於父代，並在競爭生存的過程中將不符合標準者逐漸淘汰，最終求得代保留之最優秀的品種即為最佳適應性的染色體，即為其演算所得出的結果。基因演算法在過去幾年已被廣泛應用與研究在金融模型上，適合處理最佳化問題，如葉美均和周宗南論文中應用股票五大面（公司基本面、財務面、技術面、市場面及籌碼面）共18種選股法挑選出排名較前的公司使用基因演算法找出最佳投資組合[8]，證實基因演算法所選出的投資組合報酬率能擊敗大盤。如表一所示。

表一、每週投資組合差異結果分析表（引用自[8]）



基因演算法被視為是有效率搜尋方法的原因之一，在龐大的搜尋空間中有系統的多點搜尋可行解的特性。應用基因演算法求解組合最佳化問題的基礎原理是：將欲搜尋問題的參數編碼成0與1的二進位碼型式，根據問題的條件或目標函數來客製化一套適應函數。適應值程度較佳的個體就有較高的機率被挑選進入交配池來進行突變與交配，交配即是在搜尋空間中做較大的跳躍式搜尋，突變則是在附近區域進行局部搜尋，設計基因演算法求解問題，大致可分成下列七個步驟：

（1）編碼設計方式

（2）決定群體規模

（3）設計適應函數

（4）決定挑選與複製的方式

（5）定義交配與交配率

（6）定義突變與突變率

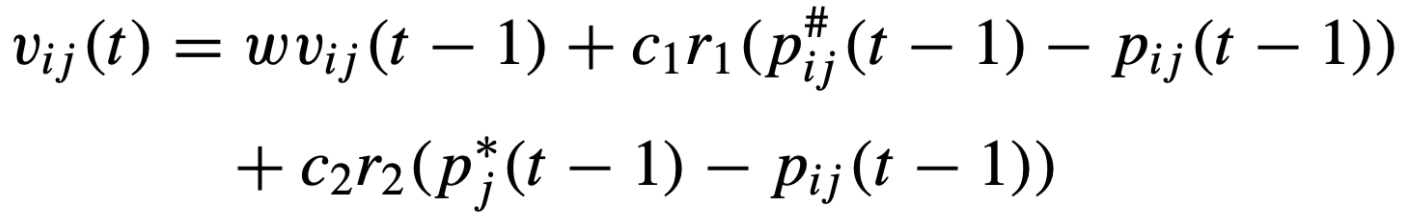
（7）決定終止條件

吳忠敏 et al.的研究中則是以移動平均線Moving Average(MA)為基礎來計算趨勢值[9]，並將股價分為三種趨勢：漲勢、盤整、跌勢，接著使用基因演算法的特色：最佳化解性質，成功建構出貼合各趨勢使用的最適投資策略，將分類結果套上趨勢投資策略系統來進行模擬交易。而實驗結果也表示，本研究提出之趨勢分類方法結合趨勢投資決策系統，其投資績效相較葛蘭碧法則(J. Granville Rules)策略有更高的報酬率、也比KD指標策略與買入持有策略更好。Chang et al.以基因演算法為基礎，建立一個關聯分類法(Associative Classification)[10]，分類股價資料與技術指標，並從中找出交易規則與最佳進出時點。結果表示出以基因演算法做為基礎技術所發展的分類方法能夠做出有效且具有較高的預測率。Nunez-Letamendia 則利用多種技術指標結合基因演算法來驗證馬德里證券交易所指數成分股之交易績效[11]。結果表示透過基因演算法所建立出的一套交易策略其操作效果比買入持有策略更好，作者並認為技術指標具有與股價未來習習相關互相影響的能力，而這些指標微變化是影響股價未來趨勢的重要關鍵。

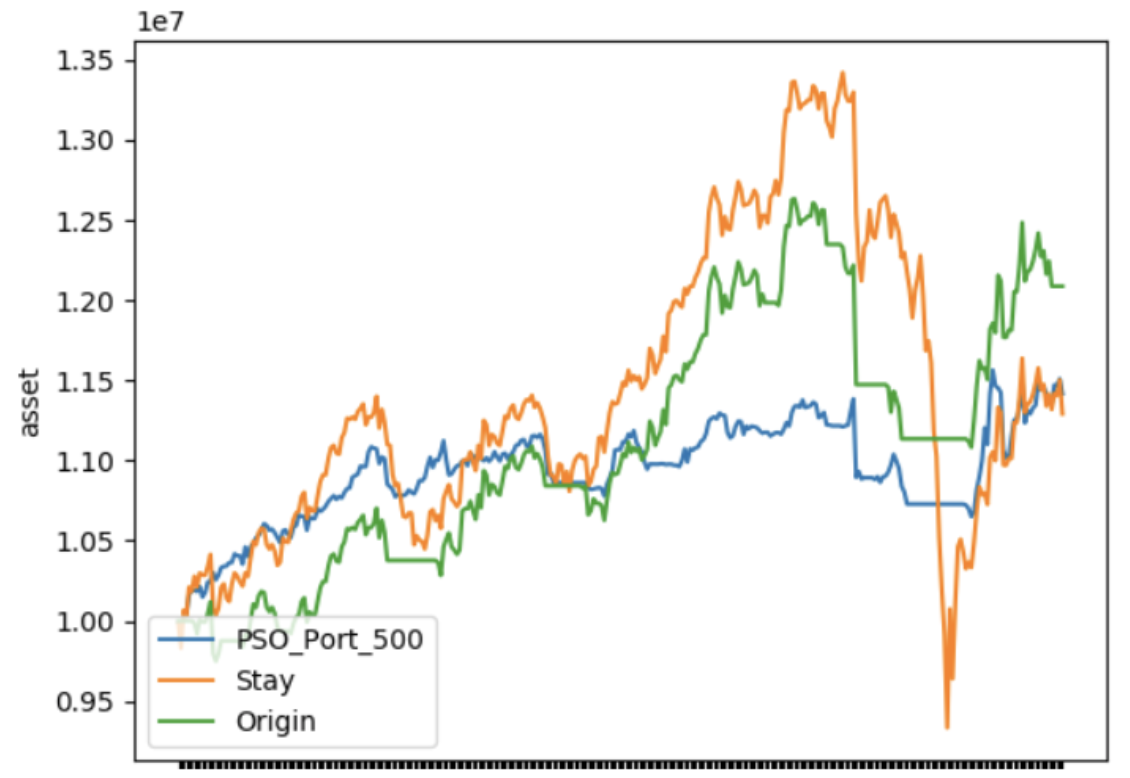
Wei-Lun Yeoh et al.則在基因演算法研究中為了讓交易更加彈性[12]，加入了放空，並抓下臺灣50指數(0050)資料及當時市值上前三大的股票(台積電 2330、富士康、2317、中華電信 2412)這四種藍籌股為投資標的，並比較月對月，季對季與年對年，最後得出結論：採用年對年滑動視窗(Year to Year sliding window)來解決股票景氣循環問題，且績效會優於傳統滑動視窗(sliding window)。

（伍）研究方法

在研究初期，我們先使用PSO演算法作為起始嘗試，假設每一顆粒子為一檔股票，給予每一顆粒子一個初始值，藉由每一次的迭代中，記錄此粒子是否為本身的最佳，以及紀錄整體的最佳將每個粒子進行速度與位置更新如下式（7），以達到找出區間最佳值的需求。

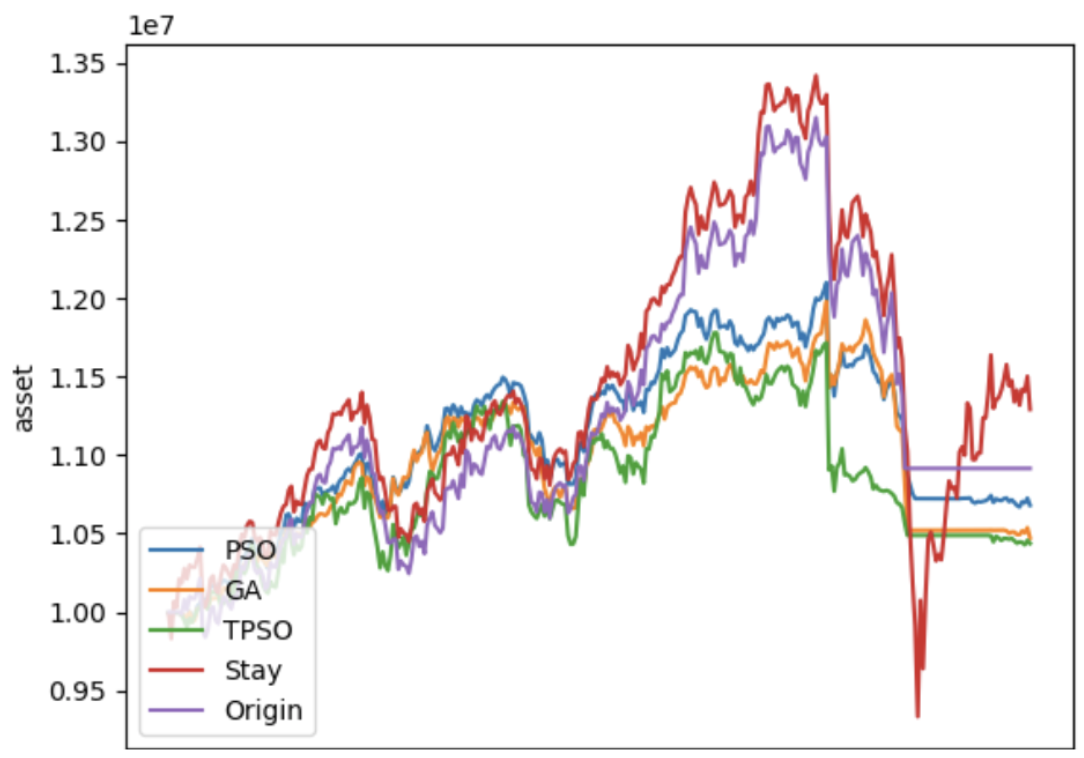
 (7)

在搜索的過程中，Pij會越來越接近目前搜索紀錄中最佳的點，速度更新式的第二項與第三項會越來越接近0， 造成Vij會以w的速度進行衰減。但我們在實際將此演算法加入我們對初始資產的配置後，我們卻發現所得到的實驗數據不如想象中滿意，如圖二。圖二中的X軸為時間，Y軸為資產變化，其中stay為在回測時間第一天買進並在最後一天賣出，期間的資產變化。origin則為對0050進行最佳化Moving Average(MA)進出場後的回測資產變化圖， 圖二為我們所做出的PSO演算法訓練500回合後的成果進行資產配置與0050持有和MA策略操作的比較，可以看出PSO演算法並沒辦法呈現出我們期待的結果。



圖二、 回測歷史數據 (2019/1/1~2020/5/31)

之後我們也嘗試加入了其他演算法如Particle Swarm Optimization(PSO)[13]、Turbulent PSO(TPSO)[14]、Genetic Algorithm(GA)[15]，如圖三之比較可看出仍無法有一個較為顯著的改善，所以我們將方向轉為現在大眾較為熟知的機器學習訓練。

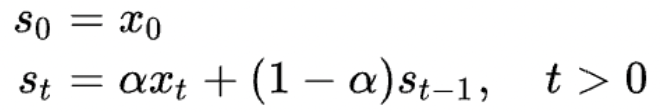


圖三、各演算法回測歷史數據（2019/1/1~2020/5/31）

經過不斷的嘗試後，我們開發出一套深具效果的LSSTM模型。在資料的部分，我們抓取基本的開高低收、成交量、法人買賣以及融資融券表做為我們的特徵值。針對開、高、低、收，以及成交量，這是股市交易裡面最基本的資訊特徵，熟稱的K棒即是由此組成。此外，我們也加入法人買賣表作為特徵，會使用法人買賣表示因為，外資、投信、自營商，三大法人都是一群資金雄厚、經驗豐富的投資人，對市場的影響力非常大，他們持股超過全部股票的65%，尤其外資占比約40%；通常三大法人如果每天一直大幅賣出，那未來股價可能下跌機率大。但相反的三大法人如果每天一直偷偷買，未來股票往上漲的的機會就比較大。而最後，我們納入融資融券表作為特徵值，因為當人們預期股價上漲到一定程度時，才會使人們有足夠的誘因融資來買股票；或是當人們預期股價下跌到一定程度時，才會使人們有足夠的誘因融券來賣股票，這部分的量價交易資訊存在相當大的重要性。

經過整理合併後，我們共有23個特徵值，分別為：開盤價、收盤價、最高價、最低價、成交量、調整後的價位、保證金購買量、保證金購買現金還款、保證金購買限制、保證金賣出限制、本日保證金購買餘額、昨日保證金購買餘額、抵銷貸款與空頭、賣空交易、賣空限制、賣空量、本日賣空餘額、昨日賣空餘額、法人買入總額、法人賣出總額、法人融資融券比、整體市場融資融券比、整體市場買入量、整體市場賣出量。

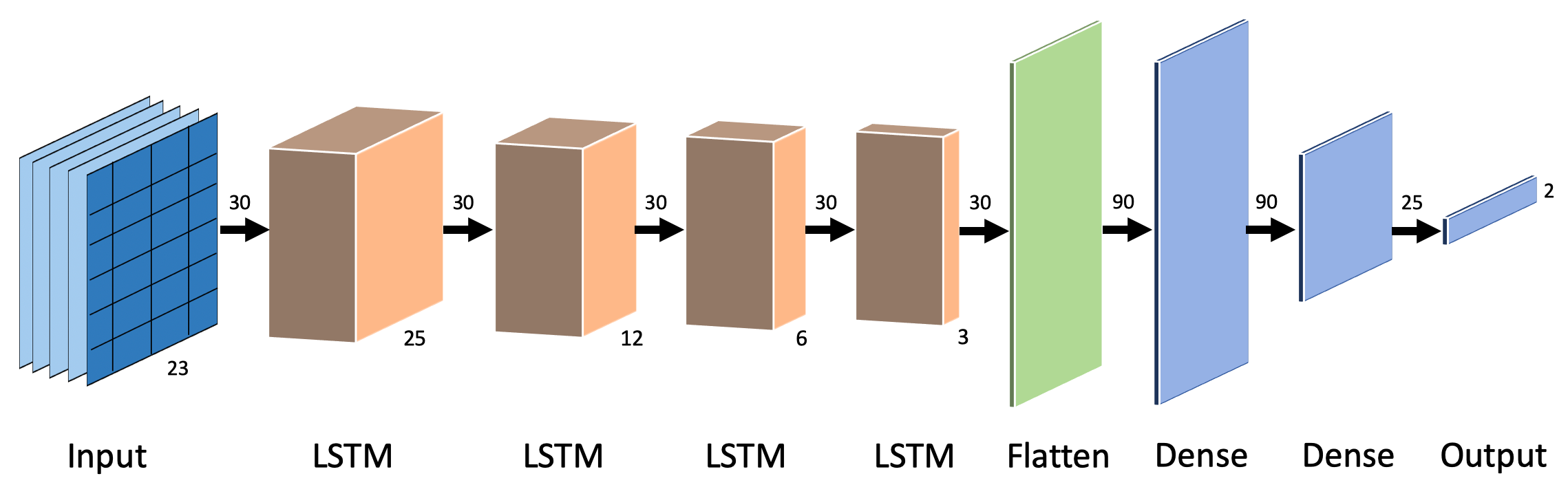
接著我們會對這些特徵進行指數平滑（alpha=0.2，如式8），且取5天資料平滑，會這麼做是因為，我們認為資料本身有其時序性上重要性的差異，資料越靠近未來，其權重是越高，反之，當資料距今越久，權重佔比應越低。原始資料與經過指數平滑後的資料如下圖四，藍色的線為指數平滑後的收盤價格。此外，為了避免梯度消失的問題出現，我們統一將各個特徵標準化，映射到同一個資料分佈。特別留意的是，我們使用的標準化方法不是以整體的資料進行，而是以Mini-Batch來進行計算其平均值與標準差來進行標準化，原因是因為這樣才能顯示在三十天內股價的波動。否則，若未來的股價上升較大時，會造成你現在的區間內，經過標準化的股價都很相近，無法顯現明顯差異。

 (8)



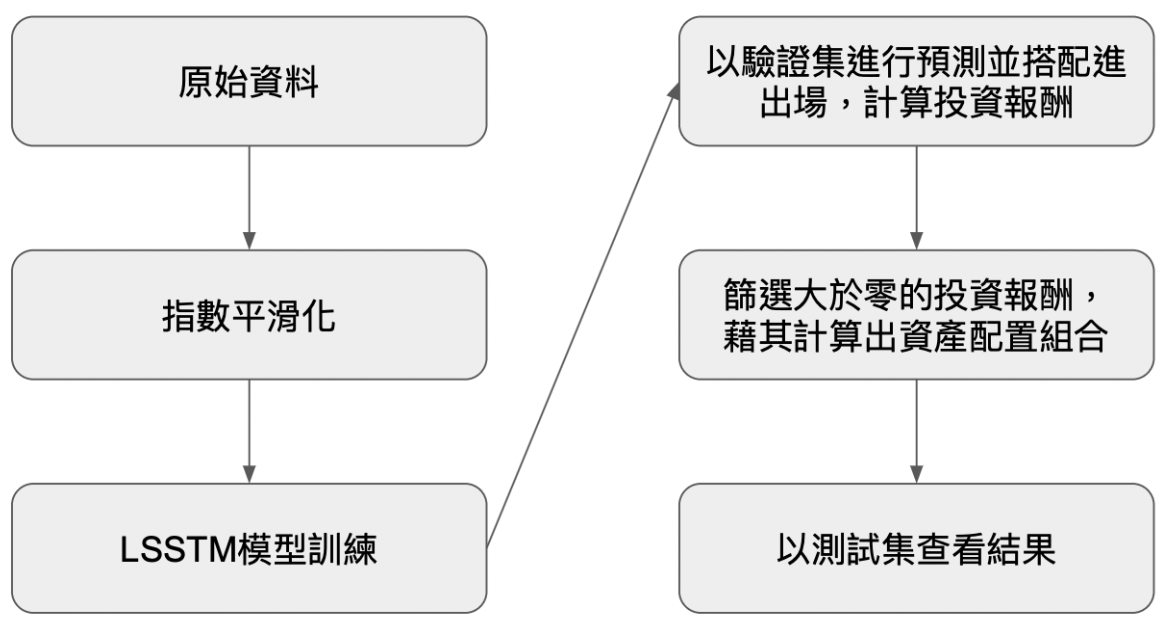
圖四、圖形化表示資料

在模型的架構部分，我們嘗試了多類別與二元的分類，由於類別造成的資料不平衡，多分類的操作勢將股市分為漲、 跌、與盤整，其中我們需要去對盤整作出定義，因交易需扣除其中的手續費與交易稅，所以我們假設正負2％以內為盤整，此時會造成資料訓練不均，所以我們還是採取平滑後的二分類，改為預測上漲趨勢或下跌趨勢。而我們提出的LSSTM，是以LSTM為主體，但我們的SSTM是Soft Stock Trend Model的縮寫，Soft表示我們對股價作指數平滑，減少其變異程度。模型架構如下圖五。

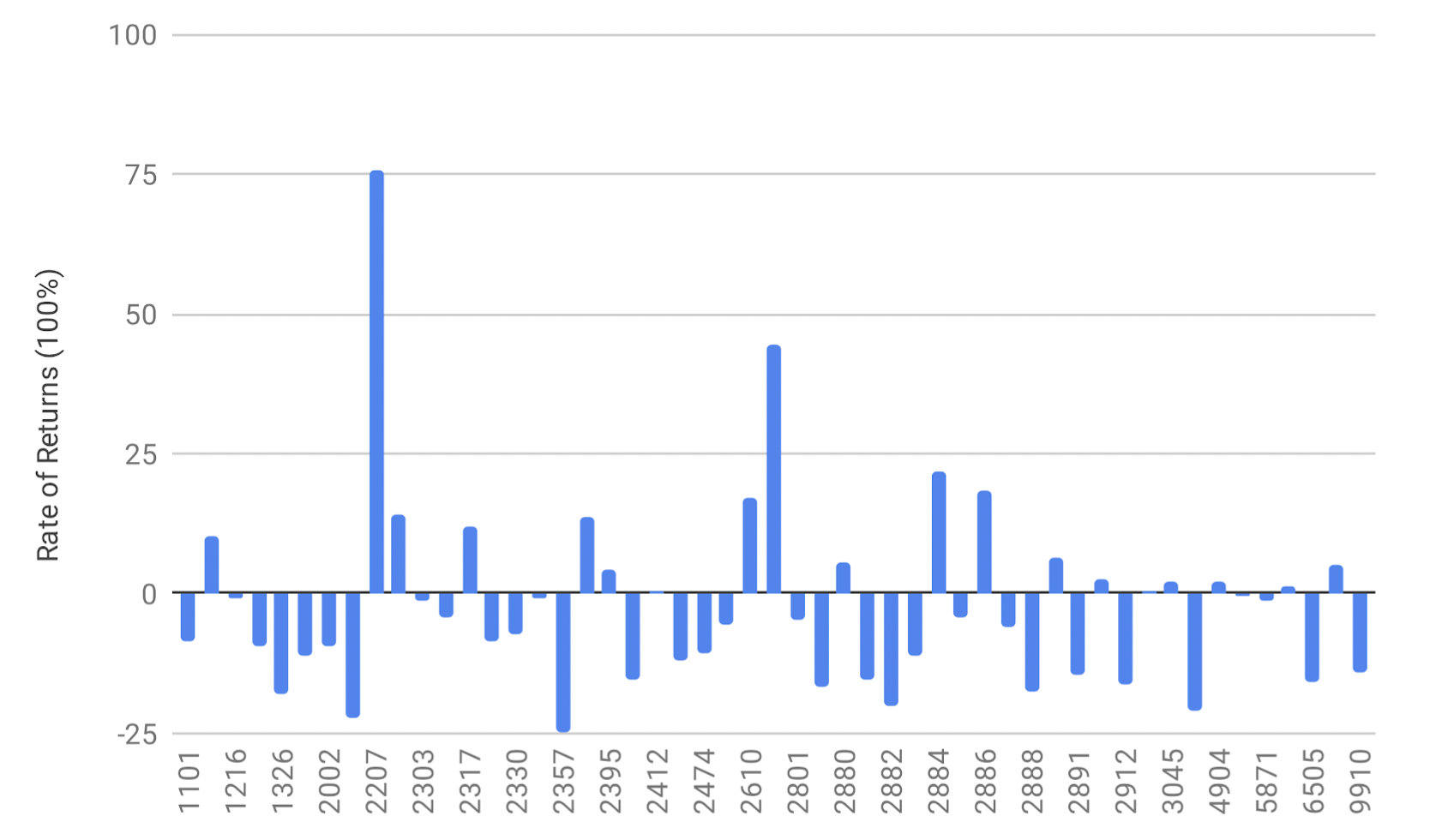


圖五、模型架構圖

在演算法流程（如圖六）中，我們擷取含當日前三十天交易日的資料，及當日五天後的價格作為標籤，將這些資料成捆傳入LSSTM進行訓練。模型訓練完成後，我們將驗證資料集實際藉由模型預測漲跌，並搭配不同的進出場策略，計算出各支股票的投資報酬率。接著我們會藉由這個驗證集所算出的投資報酬率（如圖七），捨棄那些負的數值，只參考正的數值來進行加權平均，以計算測試時的投資組合權重。得到權重後，我們將其運用測試資料來進行計算最終報酬的回測。



圖六、演算法流程



圖七、由驗證集所計算出的投資報酬率

（陸）結果與討論

從下表可以得知，針對目前回測區間的表現，相對於元大台灣卓越50基金(0050)，若於測試起始日期入場，於結束日期出場，投資報酬率為32.09%，而我們提出的LSSTM搭配進出場策略的演算法最高可以達到47.13%的投資報酬率，表現有顯著提升。

為確認資金配置獲得的利益，我們也嘗試了其他許多不同的進場方式，舉例來說：若以知我們的模型有著極高的準確度，那若我們模型產生的訊號為漲，則在第五天時必然會比現今的價位高，那我們就有兩種選擇，其一是第四天再進場，賺取價差，另一個則是產生訊號的當天進場，如表二中的前兩列所呈現。

另一個有趣的想法是如果我們對所擁有的總資產做切割的話又會如何呢，打個比方，若我們有1000萬資產，我們將其切割成五天，所第一天所產生的訊號是上漲，則我們使用200萬進場，第二天若產生下跌訊號，則我們不進場，目前進場200萬，剩餘資金800萬，若第三天產生上漲訊號，則我們對資金做三份切割（因剩餘天數為三天），進場266萬，以此類推。由第三列可看出實驗結果。

市場上有一種較為合乎人性的操作是設立止損點，若我們進場，後來價格跌到我們所設立的固定比例，則我們認定此次模型預測失誤，那我們就認賠撤出，以此想法下去做實驗，結果如表二的第四列。

最後，我們在這樣的想法上加入一些市場上常見的操作：逢低加碼，前提是模型在這一次的預測是準的，我們對每一部分資金分別去做追蹤，若我們進場的隔天市場價格下跌，則我們加碼總剩餘資金的固定比例，若模型無誤則後面幾天必然會回漲，以此想法下作延伸，則結果如我們表二的第五列。



表二、回測結果

未來期許能將模型的訓練提升一個層次，我們現在是針對0050進行一個偏向整體大盤的訓練，之後希望能各細分成各種模型專門去訓練每個特定類別的股票，例如：金融股、電子股、科技股，以利進一步的資產預測。此外，計劃將此模型的應用拓展到其他金融領域，不論是期貨，抑或是現在逐漸成為主流的虛擬貨幣，希望能將這份研究不斷鑽研，以幫助一般散戶投資客獲取更多利潤及報酬。

（柒）參考文獻

[1] Elman, Jeffrey L.“Finding structure in time.” Cognitive science 14.2 (1990): 179-211.

[2] 李宏毅機器學習課程筆記9 URL https://blog.csdn.net/xzy\_thu/article/details/74930482

[3] Nguyen, Michael. “Illustrated Guide to LSTM’s and GRU’s: A step by step explanation.” URL https://towardsdatascience. com/illustrated-guide-to-lstms-and-grusa-step-by-step- explanation-44e9eb85bf21 (2018).

[4] 劉昭雨, and 顏士淨. “卷積神經網路在金融技術指標之應用”(2017).

[5] Chen, Kai, Yi Zhou, and Fangyan Dai. “A LSTM-based method for stock returns prediction: A case study of China stock market.” 2015 IEEE international conference on big data (big data). IEEE, 2015.

[6] 黃華山, and 邱一薰. “類神經網路預測臺灣50股價指數之研究.”(2005): 19-42.

[7] 沈沛瑄,and魏廉臻. “以 LSTM-RNN預測ETF 50股價趨勢並結合交易策略以獲取最大獲利率”(2019).

[8] 葉美均, and 周宗南. "應用基因演算法整合五大構面選股策略." (2016): 1-81.

[9] 應用基因演算法建構以趨勢為基礎之股票投資決策系統. 資訊科學應用期刊第8卷第 1期, 2013.

[10] Chien, Ya-Wen Chang, and Yen-Liang Chen. "Mining associative classification rules with stock trading data–A GA-based method." Knowledge-Based Systems 23.6 (2010): 605-614.

[11] Núñez-Letamendia, Laura. "Trading systems designed by genetic algorithms." Managerial Finance (2002).

[12] Yeoh, Wei-Lun, *et al.* "Automatic Stock Trading System Combined with Short Selling Us- ing Moving Average and GQTS Algorithm." 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2018.

[13] https://en.wikipedia.org/wiki/Particle\_swarm\_optimization

[14] https://www.softcomputing.net/fatpso.pdf

[15] https://en.wikipedia.org/wiki/Genetic\_algorithm