

明新科技大學 校內專題研究計畫成果報告

具有股票交易擇時之投資組合最佳化投資程序
A portfolio optimization investment procedure with timing of
trading stocks based on the analytic hierarchy process, support
vector regression and genetic algorithms

計畫類別：任務型計畫 整合型計畫 個人計畫

計畫編號：MUST-108 企管-1

執行期間：108 年 01 月 01 日至 108 年 09 月 30 日

計畫主持人：徐志明

共同主持人：

計畫參與人員：

處理方式：公開於校網頁

執行單位：企業管理系

中 華 民 國 108 年 10 月 1 日

明新科技大學校內專題成果報告

公開授權書

(提供本校辦理紙本與電子全文授權管理用)

本授權書為明新科技大學校內專題研究計畫成果報告授權人：徐志明

在明新科技大學 管理 學院 企業管理 系 108 年度校內專題研究計畫。

研究計畫編號：MUST-108 企管-1

研究計畫名稱：具有股票交易擇時之投資組合最佳化投資程序

計畫類型：個人計畫

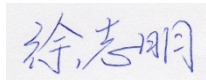
執行期限：108年01月01日至108年09月30日

茲同意將授權人擁有研究之上列成果報告：紙本授權全文公開陳列於本校圖書館，為學術研究之目的以各種方法重製，或為上述目的再授權他人以各種方法重製，不限地域與時間，惟每人以一份為限；成果報告之電子檔（含摘要），本校圖書館保留以供文獻典藏使用，但可依使用權限授權於網路公開，提供讀者非營利性質之免費線上檢索、閱覽、下載或列印。

成果報告之電子檔案使用權限授權，請勾選下列一項：

- 校內外立即公開全文(含摘要)
- 校內外立即公開摘要，校內立即公開全文，一年後校外公開全文
- 校內外僅於公開摘要，校內立即公開全文，校外永不公開全文

授權人：



E-Mail：cmhsu@must.edu.tw

中 華 民 國 108 年 10 月 1 日

摘要

在金融投資領域，投資股票相對是比較容易的，因為透過在其低廉的價格購買股票。然而，並在其相應的高價賣出股票以賺取利潤相，對於其他投資商品是較為簡單，選擇具有潛在獲利能力的股票，或是決定這些選取股票的財務投資比例，是一件很困難而深具挑戰性的問題。選擇股票以進行投資的目的是藉由選擇財務表現突出公司發行的股票，以期待避免股票價格的不利表現，而確定這些所選股票的最佳資本配置，目的是為了能最大幅度地減少投資組合風險，從而保證預期的利潤可以實現。而股票交易擇時則可以幫助投資者決定最佳買入/賣出股票的時機和數量，以產生更好的利潤或預防更多的損失。此計畫利用層級分析法、支援向量迴歸和遺傳演算法，以設計一個三階段投資組合最佳化程序，用以循序性地解決投資組合選擇，投資組合最佳化和股票交易時間問題。

關鍵詞：股票投資、投資組合選擇、投資組合最佳化、交易擇時、層級分析法、支援向量迴歸、遺傳演算法

Abstract

In the fields of financial investments, investing stocks is relatively easy while comparing to the other investment commodities since making a profit through buying a stock at its low price and selling the stock at its corresponding high price is intuitive. However, it is really a challenge work for an investor to choose stocks which might be profitable or determine the capital allocations for these selected stocks or even timing the transactions for stocks. The purpose of selecting some stocks for investing is to choose the stocks issued by the corporations with outstanding financial performances thus expecting to avoid unfavorable performances of stock prices; the purpose of determining the optimal capital allocations for these selected stocks is to minimize the portfolio risk thus ensuring the expected profit can be reached; the transactions' timings for stocks can help an investor to decide the optimal moment and share amount for buying/selling a stock thus yield a better profit or preventing a more loss. In this study, the analytic hierarchy process (AHP), support vector regression (SVR), and genetic algorithm (GA) are employed to design a three-stage portfolio optimization approach for sequentially solve the portfolio selection, portfolio optimization, and transaction timing.

Keywords: Stock investment, Portfolio selection, Portfolio optimization, Transaction timing, Analytic hierarchy process, Support vector regression, Genetic algorithm

目錄

摘要.....	I
Abstract	II
目錄.....	III
表目錄.....	V
圖目錄.....	VI
1. 緒論.....	1
2. 研究目的.....	1
3. 文獻探討.....	1
4. 研究方法.....	6
4.1. 層級分析法.....	6
4.2. 支援向量迴歸.....	8
4.3. 遺傳演算法.....	10
5. 股票投資組合最佳化程序架構.....	11
5.1. 投資組合選擇.....	11
5.2. 投資組合最佳化.....	12
6. 實例探討.....	13
6.1. 建構投資組合.....	13
6.2. 投資組合最佳化.....	14
6.3. 交易股票.....	16
6.4. 評估投資績效.....	17
7. 結論.....	17
參考文獻.....	18
研究計畫執行成果自評表.....	22
研究計畫運用於教學成果記錄表.....	24
計畫發表之相關論文(1).....	26
計畫發表之相關論文(2).....	34
計畫發表之相關論文(3).....	46
計畫發表之相關論文(4).....	53

表目錄

表 1. 某些決策日投資組合中股票的最佳資本配置.....	16
表 2. 股票投資績效整理.....	17

圖目錄

圖 1. 層級分析法的例子.....	8
圖 2. 股票投資組合最佳化程序架構.....	11
圖 3. 以層級分析選擇股票的決策問題.....	13
圖 4. Expert Choice 11 所選擇的 10 檔股票	14

1. 緒論

在金融投資領域中，相對於其他投資商品而言，投資股票是相對較為容易的，因為透過以相對較低的價格買入股票，而以相對較高的價格賣出股票，對投資人而言是非常直覺而容易的。然而，對於投資者而言，選擇可能有利可圖的股票，以及確定這些選定投資的股票之資本配置，的確是一項具有挑戰性的工作。選擇股票投資的目的，是基於相信選擇具有優秀財務業績公司發行的股票，可以避免股價表現不佳的風險。而另一方面，確定這些選定投資股票的最佳資本配置的目的，則是讓投資組合的風險能最小化，從而確保投資者的預期利潤能夠實現。

2. 研究目的

根據上述研究背景之描述，本計畫之研究目的列述如下：

- (1) 應用層級分析法(analytic hierarchy process, AHP)、支援向量迴歸(support vector regression, SVR)和遺傳演算法(genetic algorithm, GA)，建立一個三階段的投資組合最佳化方法，以同時解決”投資組合選擇”、”投資組合最佳化”和”股票交易”問題
- (2) 依據候選股票的公司財務報告，利用層級分析法選擇有利可圖的股票，以形成投資組合中的成份元素。
- (3) 利用支援向量迴歸為投資組合中所選擇的股票建構股票預測模型，以預測未來的股票收盤價格，並以遺傳演算法最佳化投資組合問題
- (4) 將在每個投資決策日所獲得的最佳投資組合與在前一個投資決策日所獲取的最佳投資組合進行比較，從而進行股票交易。
- (5) 以台灣股票市場的半導體和鋼鐵類股為例，驗證本計畫所提出的投資組合最佳化程序，從而驗證其有效性和有效性。
- (6) 彙整研究計畫結論並提出未來研究方向之建議。

3. 文獻探討

(一)本研究與本校之中程校務發展計畫關聯性

本計畫結合層級分析法、支援向量迴歸與遺傳演算法，建構一個具有股票交易擇時之投資組合最佳化投資程序。此演算程序足以做為國內外學者在探討股票交易擇時與投資組合最佳化問題時的重要參考文獻；同時，本研究所發展之整合性股票投資程序，可以確實協助投資者充分掌握股票市場的上漲、下跌或盤整等的變化趨勢，以期投資者可以選擇最佳的買賣時間點與最佳投資組合，獲得比市場大盤或銀行定存利率更佳之投資報酬率。因此，本計畫之研究成果應足以對實務界產生一定之實用性。此外，此計畫將利用 Visual C++程式，將股票價格預測程序發展為可真實應用之軟體，對學生未來在財務經融公司或軟體公司就業皆具有顯著幫助，與本校中程校務發展之「精進教學品質，確保學習成效」與「強化產學鏈結，縮短學用落差」兩個面向，具有高度之相關性。

(二)研究背景

在金融投資領域中，相對於其他投資商品而言，投資股票是相對較為容易的，

因為透過以相對較低的價格買入股票，而以相對較高的價格賣出股票，對投資人而言是非常直覺而容易的。然而，對於投資者而言，選擇可能有利可圖的股票，以及確定這些選定投資的股票之資本配置，的確是一項具有挑戰性的工作。選擇股票投資的目的，是基於相信選擇具有優秀財務業績公司發行的股票，可以避免股價表現不佳的風險。而另一方面，確定這些選定投資股票的最佳資本配置的目的，則是讓投資組合的風險能最小化，從而確保投資者的預期利潤能夠實現。

(三)研究目的

根據上述研究背景之描述，本計畫之研究目的列述如下：

- (7) 應用層級分析法(analytic hierarchy process, AHP)、支援向量迴歸(support vector regression, SVR)和遺傳演算法(genetic algorithm, GA)，建立一個三階段的投資組合最佳化方法，以同時解決”投資組合選擇”、”投資組合最佳化”和”股票交易”問題
- (8) 依據候選股票的公司財務報告，利用層級分析法選擇有利可圖的股票，以形成投資組合中的成份元素。
- (9) 利用支援向量迴歸為投資組合中所選擇的股票建構股票預測模型，以預測未來的股票收盤價格，並以遺傳演算法最佳化投資組合問題
- (10) 將在每個投資決策日所獲得的最佳投資組合與在前一個投資決策日所獲取的最佳投資組合進行比較，從而進行股票交易。
- (11) 以台灣股票市場的半導體和鋼鐵類股為例，驗證本計畫所提出的投資組合最佳化程序，從而驗證其有效性和有效性。
- (12) 彙整研究計畫結論並提出未來研究方向之建議。

(四)重要性

要從”股海”(亦即金融市場上所發行的眾多上市股票)當中，選擇潛在有利可圖的股票，對投資者是一個難題。在過去的研究中可以發現，多準則決策(MCDM)技術是對企業經營績效進行評估和排序的常用工具，從而挑選出具有高績效表現的所謂”好”企業。同時，與資料包絡分析、TOPSIS 和 VIKOR 等相比，層級分析法較不常使用。另外，軟計算(soft computing)技術，例如遺傳演算法、粒子群演算法、人工蜂群、和諧搜尋和細菌覓食最佳化演算法及變異等，廣泛地應用在於獲得投資組合最佳化的近最佳解。然而，投資者過去研究並無考慮無法獲得估計利潤的機率(即風險)，而只關注可以從投資組合中獲得的估計利潤。此外，”投資組合選擇”和”投資組合最佳化”，甚至”股票交易”，被視為三個獨立的問題，然而，實際上這些問題是相互依賴的。

(五)重要參考文獻與評述

Wang *et al.* (2017)提出了一個系統化的方法，以幫助組織確定最佳可持續發展目

標的合作夥伴。作者使用了超級鬆馳模型(super slacks-based model, SBM)對公司進行排名，並依據 Malmquist 生產力指數(Malmquist productivity index, MPI)衡量 2012 至 2015 年間的效率、技術和生產力的變化，並以 GM (1,1)模型預測其未來在 2016 至 2019 年的績效表現。最後，結合了資料包絡分析(data envelope analysis, DEA)和灰方法(grey methods)，以確定在 2012 至 2019 年間的最佳綠色物流合作夥伴關係及其競爭力水平。Chen and Chen (2011)運用資料包絡分析和 Malmquist 生產力指數，針對 2004 至 2007 年間，台灣晶圓製造企業的經營業績進行調查，其中，輸入變數包括了總資產、運營成本以及銷售和行政支出，而產出變數則是淨銷售額。作者並應用 GM (1,1)模型以預測 2008 至 2010 年台灣晶圓製造業的增長趨勢，然後利用 GM (1,N)模型探討輸入變數對輸出變數的影響。依據實驗結果，銷售和行政支出對於產出變數的影響最大。Tsaur *et al.* (2017)利用其所提出的方法，調查 2009 至 2012 年間，台灣六家薄膜晶體液晶顯示器(thin film transistor liquid crystal display, TFT-LCD)公司的運營表現。其所提出的方法包含四個階段，首先，透過資料包絡分析以評估企業每年的效率，然後，第二階段以 Malmquist 指數探索各個公司的效率變化，而在第三階段，則使用修正的灰色關聯分析(grey relation analysis, GRA)和熵係數(entropy coefficients)以排序這些 TFT-LCD 公司，並根據分析結果提出結論和建議。從實證結果來看，作者所提出的方法論，可以指出哪些公司可以透過提高其規模收益率(variable returns to scale, VRS)或規模效率(scale efficiency)，以提高其運營效率。此外，具有熵加權的灰色關聯分析方法，可以評估企業當前的表現，並且可以預測其未來的表現。Hsu (2015)將資料包絡分析和改進的灰色關聯分析(improved grey relational analysis, IGRA)相結合，設計了一個決策模型，以衡量相對效率。作者以評估 2010 年在台灣上市的半導體公司之效率和運營績效為例，首先，將半導體公司分成兩組，即高效和低效，然後，透過節合了多標準決策 VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR)、IGRA 和熵加權方法，分別評估高效和低效組的運行績效。另外，作者建議投資者應該選擇效率更高，經營業績更好的公司以作為投資標的，而不能只考慮過去候選投資目標的業績排名。Wang *et al.* (2015)考慮台灣電子公司經營業績的變化，以探討資產減值是否會為盈餘管理提供機會。因此，作者首先運用動態資料包絡分析(dynamic data envelopment analysis, dynamic DEA)模型，針對 2004 至 2013 年間企業的經營績效進行評價，然後，以統計方法檢驗對企業完成資產減值後的平均效率變動進行評價。Song and Guan (2015)根據三個一級指標，採用超級鬆馳模型，對安徽省 16 個城市環保部門的電子政務績效，包括公眾參與程度、網站服務質量和公眾滿意度進行評估。在他們的研究中，由於決策制訂單位(Decision Making Unit, DMU)的效能可以成功地進行評估，而邊際效率研究則可與實際商業世界更相關。Hsu *et al.* (2015)的研究在於為公司建立一個可持續性的績效評估標準和模式。因此，作者將企業財務、信用風險、環境和社會責任的衡量標準，納入可持續業務績效的評估標準。然後，以灰色關聯分析和改進的 TOPSIS 方法，構建企業可持續性的績效評估模型。作者並運用其提出的方法，探索 2011 年台灣高科技公司的可持續性的經營業績和排名情況。其實證結果可以為投資者或銀行信用審計工作提供重要的參考依據。Gul *et al.* (2016)對 VIKOR 的應用研究進行了最新的文獻回顧。其研究發現，VIKOR 方法被廣泛地應用於績效評估領域和標竿評價領域，特別是應用在模糊決策的環境中。Yalcin *et al.* (2012)提出了一種應用模糊層級分析法(fuzzy analytic hierarchy process, FAHP)以確定準則權重的方法，並透過 TOPSIS 和 VIKOR 方法對公司各部門進行排序。其方法透

過對土耳其製造業上市公司各部門的業績進行排名以展示，而實證結果顯示，對於每家公司而言，各部門所獲得的排名幾乎相同。Ghadikolaie *et al.* (2014)提出了一個在德黑蘭證券交易所(Tehran stock exchange, TSE)市場上，評估汽車公司財務業績的混合方法。他們首先利用模糊層級分析法(fuzzy analytic hierarchy process, FAHP)以找出最佳的準則權重。同時，應用 VIKOR、模糊綜合評價法(fuzzy additive ratio assessment, ARAS-F)和模糊綜合比例評價法(fuzzy complex proportional assessment, Fuzzy COPRAS)進行排序。從分析結果中發現，經濟價值對於衡量公司績效的重要性高於會計指標。Shaverdi *et al.* (2011)通過構建一個使用包括 TOPSIS、VIKOR 和 ELECTRE 在內的多重標準決策工具，以評估三家非政府伊朗銀行的業績，其中，選定了 BSC 概念的 21 個指數以進行評估。此外，其應用模糊層次分析法以計算每個選擇指標的相對權重，從而得以容忍信息的含糊性和模糊性。

一旦投資者從“股海”中挑選出投資目標，亦即形成投資組合最佳化問題，接著，投資者就必須確定每個選定投資目標的資本配置，以同時達到最大化整體預期利潤，並最小化整體投資風險的目標。然而，投資組合優化是一個 NP-Hard 的問題。因此，對於具有決策時間限制的投資組合最佳化問題，尤其是在投資組合最佳化模型中增加了一些額外限制式時，要獲得一個“硬”的(亦即全域)最佳解，是一項艱鉅的任務。因此，軟計算(soft computing)技術最近廣泛地被應用於在可接受的時間範圍內，以獲得“軟”(亦即近乎全域)的最佳解。Kumar and Mishra (2016)提出了一種混合共變異數和人工蜂群(artificial bee colony, ABC)的方法，以解決具有多個衝突目標的投資組合最佳化問題，從而更加精確地加速收斂速度。透過執行 OR 資料庫中的幾個組合最佳化問題，以驗證其提出方法的效能和效率。實驗結果顯示，其所提出的方法可以透過同時處理現實的限制以找到各種最佳的折衷解決方案。Ni *et al.* (2017)利用動態隨機族群的拓撲結構，以利抽象成無向連通圖可以依據預定義的規則和程度隨機生成抽象的無向連通圖，用來改進傳統的粒子群最佳化(particle swarm optimization, PSO)演算法。並利用修改後的 PSO 透過調整規則和程度以增強演化過程中演化的通信機制和 PSO 的解決效能。透過基於動態隨機族群的拓撲結構的幾種改進 PSO 演算法與傳統 PSO 變種的比較，以求解廣義證券組合選擇模型，實驗數據包括了恆生指數、DAX 100 指數、FTSE 100 指數、S&P 100 指數和日經 225 指數。實現結果顯示，其所提出的動態隨機族群的拓撲結構確實可以顯著地改善傳統 PSO 方法的計算效能。Metawa *et al.* (2016)利用遺傳算法(genetic algorithm, GA)提出了一種動態自我組織方法以組織銀行貸款決策。其所提出的模型同時考慮了銀行利潤的最大化和銀行違約概率的最小化，以構建一個最優的貸款組合。具體而言，透過貸款特徵和債權人評級的幾個因素與 GA 染色體的編碼相結合，從而獲得最有效的貸款決策。Seyedhosseini *et al.* (2016)將和諧搜尋法(harmony search, HS)和人工蜂群(artificial bee colony, ABC)進行混合，以解決由 Markowitz 均值-半變異數模型(mean-semi variance model)形成的投資組合最佳化問題。透過比較由作者提出的方法與 HS 和 GA，以獲得有效邊界來評估效率和準確度。計算結果顯示，其所提出的方法比 HS 和 GA 更成功而可以找到最佳解。Lwin *et al.* (2014)考慮了現實世界中存在的基數、數量、預配置和其他很多限制式，擴展了 Markowitz 均值-半變異數模型組合最佳化模型。其混合了學習導向的解決方案生成策略，以提出一種多目標進化演算法，以透過進化搜索引導到搜索空間有前途的區域，解決擴展的組合優化問題，從而促進有效的收斂。作者將包含 1318 個資產的 7 個市場指標公共 OR 資料庫用於比較其所提出的演算法和四個多目標進化演算法，包括非支配排序

遺傳演算法(non-dominated sorting genetic algorithm, NSGA-II)、強度 Pareto 進化算法(strength Pareto evolutionary algorithm, SPEA-2)、基於 Pareto 包絡選擇算法(Pareto envelope-based selection algorithm, PESA-II)和 Pareto 存檔演化策略(Pareto archived evolution strategy, PAES)。執行結果顯示，其所提出的方法可以顯著地優於其他四個多目標進化演算法，以提供了一個更高質量的有效前沿。Mishra *et al.* (2014)應用多目標細菌覓食最佳化(multi-objective bacteria foraging optimization, MOBFO)演算法以解決考慮多種實際限制(包括最小購買閾值、最大限制和基數)的資產組合選擇(portfolio asset selection, PAS)問題。利用五個基準數據集將其提出的方法，基於六個性能度量、Pareto 前沿和計算時間，與一組競爭性多目標進化演算法進行比較。依據實驗結果，其提出的 MOBFO 演算法，可以在保持可接受的多樣性，同時識別出一個好的 Pareto 解，並為不同的基數限制提供適當的解決方案。Hsu (2014)提出了利用資料包絡分析、人工蜂群演算法和基因規劃(genetic programming, GP)以解決投資組合最佳化問題的系統化程序。在其提出的方法中，首先，基於公司的歷史財務數據，透過資料包絡分析選擇一些有望可以獲利的股票。接下來，使用人工蜂群演算法以優化投資組合最佳化問題，而該投資組合最佳化問題是透過 Markowitz 平均數-變異數(mean-variance, M-V)模型來描述，並使用低於目標的半變異數(semi-variance)以測量關於股票的投資風險。最後，通過基因規劃技術以構建股票價格預測模型，並設計股票交易規則，以決定股票買賣的最佳時機。Gupta *et al.* (2014)提出了一種基於支持向量機(support vector machine, SVM)和遺傳演算法的多準則方法，利用投資者的偏好以選擇特定類別的資產以滿足給定的投資者類型。首先，其使用支持向量機以分類候選資產，然後通過求解考慮四個財務標準(包括短期收益、長期收益、風險和流動性)的投資組合選擇問題，應用遺傳演算法以獲得最佳投資組合選擇。Hajinezhad *et al.* (2017)開發了一種稱為混合禁忌機(mixed Tabu machine, MTM)的人工類神經網絡(artificial neural network, ANN)，在離散搜尋空間和連續搜尋空間中，透過 Tabu 搜索機制調節狀態轉換機制，以協助搜尋過程中從局部最小狀態逃逸的能量，從而找到全域最優解。其利用 MTM 解決投資組合優化問題，並使用香港恆生、德國 DAX 100、英國 FTSE 100、美國 S&P100 和日本日經 225 等 5 個資本市場指數數據集以驗證 MTM 的效率。實驗結果顯示，其提出的 MTM 方法可以在非常短的 CPU 時間內獲得理想而出色的結果。Strumberger *et al.* (2016)混合了蝙蝠演算法(bat algorithm, BA)和人工蜂群演算法(artificial bee colony, ABC)，簡稱為 BA-ABC 演算法，以解決典型的均值-變異數(mean-variance)投資組合選擇問題。其使用 Sefiane and Benbouziane (2012)提出的歷史數據集以比較 BA-ABC、其他群體智能演算法以及遺傳演算法的三種變體，以驗證其提出的方法的穩健性。基於實驗結果，我們可以認為 BA-ABC 演算法具有處理有許多限制式投資組合問題的巨大潛力。Chen (2015)引入了一種稱為 MABC 的改進人工蜂群演算法(modified artificial bee colony)，該演算法利用混沌初始化方法(chaotic initialization approach)，並將 ABC 和粒子群最佳化(particle swarm optimization, PSO)混合以解決考量了一些實際限制(如交易成本、基數和數量限制)的投資組合最佳化問題。作者透過 Chen *et al.* (2006)的實例驗證以展現 MABC 演算法的有效性。Gunasekarana and Ramaswami (2014)將自適應神經模糊推理系統(adaptive neuro-fuzzy inference system, ANFIS)和資本資產定價模型(capital asset pricing model, CAPM)，簡稱為 ANFIS-CAPM 模型，以解決股票投資組合最佳化問題。作者根據孟買證券交易所(Bombay Stock Exchange, BSE)和 SENSEX (SENSitive IndEX)的歷史數據以及一些知名技術指標，應

用 ANFIS 預測股價。然後將 CAPM 嵌入到投資組合最佳化模型中，以找到能夠同時最大化預期回報率和最小化投資組合風險的股票組合。透過一個包含 2009 年至 2010 年 BSE SENSEX 指數中 30 支股票的案例研究以測試其所提出的 ANFIS-CAPM 方法的效率。計算和比較結果顯示，其所提出的 ANFIS-CAPM 方法表現優於 BSE、SENSEX、動量投資組合和買入持有策略。Elhachloufi *et al.* (2012) 提出一個基於人工智能的兩階段程序，以優化投資組合中股票的選擇。第一階段選擇具有低風險和高回報率的股票，透過迴歸神經網路(regression neural networks)形成初始投資組合。在第二階段，透過應用遺傳演算法對投資組合中每個元素的投資比例進行最佳化，其中，投資組合風險是以一個半變異數(semi-variance)模型來描述。Chen *et al.* (2013) 提出了一種改進的人工蜂群算法，稱為 IABC (improved artificial bee colony)，以獲得投資組合最佳化問題的效率邊界，其中，包括利用整數和實數變數的混合編碼以滿足投資組合問題的特徵，例如基數和邊界限制。透過 OR 資料庫的全球股票市場指數進行 IABC 驗證，並與模擬退火、禁忌搜索、變數鄰域搜索(variable neighborhood search)和 ABC 進行比較。執行結果顯示，若同時考慮到多樣性、收斂性和有效性，IABC 具有最好的表現。Liu *et al.* (2010) 運用風險偏好係數，以考慮投資房地產的半變異數投資組合模型，並應用人工蜂群演算法以處理所建構的模型。利用中國北京的一個真實案以證明其所提的方法的績效表現，並獲取了滿意的結果。

從上述文獻研究可以發現，多準則決策(MCDM)技術是對企業經營績效進行評估和排序的常用工具，從而挑選出具有高績效表現的所謂”好”企業。同時，與資料包絡分析、TOPSIS 和 VIKOR 等相比，層級分析法較不常使用。另外，軟計算(soft computing)技術，例如遺傳演算法、粒子群演算法、人工蜂群、和諧搜尋和細菌覓食最佳化演算法及變異等，廣泛地應用在於獲得投資組合最佳化的近最佳解。然而，投資者過去研究並無考慮無法獲得估計利潤的機率(即風險)，而只關注可以從投資組合中獲得的估計利潤。此外，”投資組合選擇”和”投資組合最佳化”，甚至”股票交易”，被視為三個獨立的問題，然而，實際上這些問題是相互依賴的。因此，本計畫應用層級分析法(analytic hierarchy process, AHP)、支援向量迴歸(support vector regression, SVR)和遺傳演算法(genetic algorithm, GA)，建立一個三階段的投資組合最佳化方法，以同時解決上述三個問題。首先，依據這些候選股票的公司財務報告，利用層級分析法以選擇被認為是有利可圖的股票，以形成投資組合中的成份元素。然後，利用支援向量迴歸為投資組合中所選擇的股票建構股票預測模型，以預測未來的股票收盤價格，從而計算每支股票的估計”平均”利潤。接下來，在每個投資決策日期，例如，在一周的最後一個交易日中，針對投資組合的”平均”利潤和”變異”(風險)進行評估，並以遺傳演算法最佳化投資組合問題，亦即確定投資組合中每支股票的最佳資本配置，並同時最大限度地降低估計利潤無法獲得的可能性(機率)，和保持投資者預期的利潤水平。最後，將在每個投資決策日所獲得的最佳投資組合與在前一個投資決策日所獲取的最佳投資組合進行比較，從而確定每支股票的最佳持有量，即確定每支股票應當買入或賣出的量，藉此以進行股票交易，而投資期末則可以對股票投資的績效進行評估。

4. 研究方法

4.1. 層級分析法

層級分析法(analytic hierarchy process, AHP)是由 Satty (1980)所提出一個可以組織與分析複雜決策問題的結構化工具。在層級分析法中，決策問題必須被分解成子問題

(決策標準)的層次結構，其中，每個子問題必須可以獨立分析，且比原問題更容易理解。假設一個具有 m 個可分解決策標準的決策問題，而決策者則擁有 n 個選擇，如圖 1 所示。層級分析法有如下的三個步驟：

步驟 1. 計算決策標準向量

首先，層級分析法先透過比較決策標準的相對重要性以創建決策標準比較矩陣 \mathbf{A} ($m \times m$)。矩陣 \mathbf{A} 的元素 a_{ij} 代表第 i 個決策標準相對於第 j 個決策標準的重要性。如果 $a_{ij} > 0$ 則代表第 i 個決策標準比第 j 個決策標準重要，而 $a_{ij} < 0$ 則代表第 j 個決策標準比第 i 個決策標準重要，至於第 i 個決策標準跟第 j 個決策標準一樣重要，則 a_{ij} 等於 1，且對於所有 $i = j$ ，其則 $a_{ij} = 1$ 。接著，以下列公式計算正規化決策標準比較矩陣 \mathbf{A}_{norm} ($m \times m$)：

$$\bar{a}_{ij} = \frac{a_{ij}}{\sum_{k=1}^m a_{ik}}, i=1, \dots, m, j=1, \dots, m. \quad (1)$$

因此，可以透過每一列元素的平均值而決策標準向量 \mathbf{w} (m 維度的行向量)，公式如下：

$$w_i = \frac{\sum_{k=1}^m \bar{a}_{ik}}{m}, i=1, \dots, m. \quad (2)$$

步驟 2. 計算分數矩陣

首先，每個選擇 k 與選擇 l 比較其對於每個決策標準 i 的績效表現，以形成評估矩陣 $\mathbf{B}^{(i)}$ ($n \times n$ 的矩陣) ($i=1, 2, \dots, m$) 的元素 b_{kl}^i 。 $b_{kl}^i > 1$ 代表，對於決策標準 i ，第 k 個選擇比第 l 個選擇來得好， $b_{kl}^i < 1$ 則代表，對於決策標準 i ，第 l 個選擇比第 k 個選擇來得好，而 $b_{kl}^i = 1$ 顯然是代表，對於決策標準 i ，第 k 個選擇與第 l 個選擇具有同樣的績效表現。同時， b_{kl}^i 與 b_{lk}^i 必須滿足：

$$b_{kl}^i \times b_{lk}^i = 1, k=1, \dots, n, l=1, \dots, n \quad (3)$$

與

$$b_{kk}^i = 1, k=1, \dots, n. \quad (4)$$

接著，每個評估矩陣 $\mathbf{B}^{(i)}$ 的每個元素除以該元素同行元素的總和，然後，每一列元素再求平均值以獲得一個 $n \times 1$ 的分數向量 $\mathbf{s}^{(i)}$ ($i=1, \dots, m$)。因此，分數矩陣 \mathbf{S} ($n \times m$) 可以透過下列公式獲得：

$$\mathbf{S} = [\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \dots, \mathbf{s}^{(m)}]. \quad (5)$$

步驟 2. 選擇排序

計算整體分數向量 \mathbf{v} ($n \times 1$) 如下：

$$\mathbf{v} = \mathbf{S} \cdot \mathbf{w}. \quad (6)$$

其中，向量 \mathbf{v} 裡面的第 k 個元素代表層級分析法給予第 k 個選擇的整體分數，亦即優先順序。

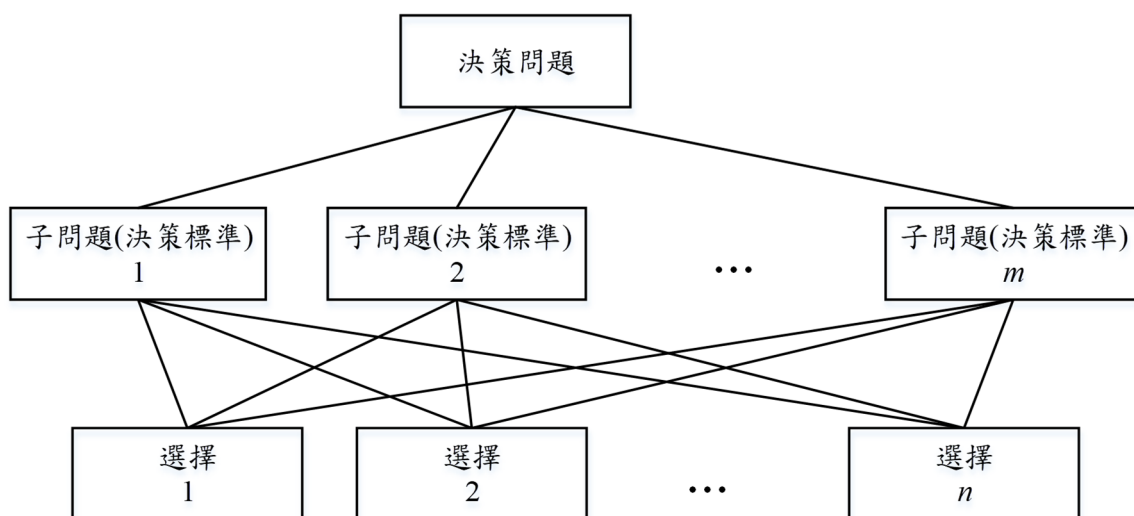


圖 1. 層級分析法的例子

層級分析法被廣泛用於解決各種應用中的群體決策問題，例如 Rhew *et al.* (2017)、Khademalhosseiny *et al.* (2017)、Tapia *et al.* (2017)與 Krmac and Djordjevic (2017)。

4.2. 支援向量迴歸

在現實世界中，研究者經常需要模擬一些輸入和輸出變數之間的函數關係。尤其是當輸入對輸出有非線性影響時，函數建模任務變得更加複雜和困難。為了解決非線性迴歸問題，Vapnik *et al.* (1997, 1997)引進了一種稱為支援向量迴歸(support vector regression, SVR)的迴歸技術。支援向量迴歸使用映射 Φ 將輸入映射到高維度的特徵空間以構建線性迴歸模型。假設我們要構建一個迴歸模型以描述 Q 對的 n 維中的一個輸入向量 $X_k = (x_{k1}, x_{k2}, \dots, x_{kn}) \in \mathbb{R}^n$ 與輸出 $y_k \in \mathbb{R}$ 之間的關係。首先，將原始輸入 X_k ($k=1, 2, \dots, Q$)轉換到高維度空間(m 維)中的 $\phi_i(X_k)$ 。接著，研究者想要獲得轉換輸入 $\phi_i(X_k)$ 的近似權重，以構建獲取預測輸出 y'_k 的線性迴歸模型如下：

$$y'_k = f(X_k, W) = \sum_i^m w_i \phi_i(X_k) + w_0 = W^T \Phi(X_k) + w_0, k=1, 2, \dots, Q. \quad (7)$$

其中， W 是由 w_i 構成的權重向量； $\Phi(X_k)$ 是由 $\phi_i(X_k)$ 所組成的特徵向量； w_0 是偏誤。此外，Vapnik (1998)認為，如果預測的輸出接近實際輸出在可接受的誤差 ε 內，那麼損失可以被認定為零，因此提出一個 ε -insensitive 損失函數以評估預測誤差如下：

$$L_\varepsilon(y_k, y'_k) = \begin{cases} 0 & \text{if } |y_k - y'_k| \leq \varepsilon \\ |y_k - y'_k| - \varepsilon & \text{otherwise} \end{cases}, k=1, 2, \dots, Q. \quad (8)$$

換句話說，損失可以重新表達如下：

$$y_k - W^T \Phi(X_k) - w_0 - \varepsilon \leq \xi_k, k=1, 2, \dots, Q \quad (9)$$

$$W^T \Phi(X_k) + w_0 - y_k - \varepsilon \leq \xi'_k, k=1, 2, \dots, Q \quad (10)$$

$$\xi_k \geq 0, k=1, 2, \dots, Q \quad (11)$$

$$\xi'_k \geq 0, k=1, 2, \dots, Q. \quad (12)$$

值得注意的是，實際值 y_k 高於和低於預測值 y'_k 的誤差分別由非負數鬆弛變數 ξ_k 和 ξ'_k 來評估。接著，Vapnik (1995, 1998) 將經驗風險最小化問題定義如下：

$$\frac{1}{2} \|W\|^2 + C \left(\sum_{k=1}^Q \xi_k + \sum_{k=1}^Q \xi'_k \right). \quad (13)$$

並滿足方程(9)-(12)中描述的限制條件，而參數 C 是由使用者預先指定以平衡複雜性和損失的參數。令 $\Xi = (\xi_1, \dots, \xi_Q)^T$ 與 $\Xi' = (\xi'_1, \dots, \xi'_Q)^T$ 代表鬆弛變數向量，而對應於公式(9)、(10)、(11)與(12)的 Lagrangian 乘數向量分別記為 $\Lambda = (\lambda_1, \dots, \lambda_Q)^T$ 、 $\Lambda' = (\lambda'_1, \dots, \lambda'_Q)^T$ 、 $\Gamma = (\gamma_1, \dots, \gamma_Q)^T$ 與 $\Gamma' = (\gamma'_1, \dots, \gamma'_Q)^T$ 。因此，上述的公式(13)的最佳化問題，可以透過求解以 Lagrangian 主變數(primal variables)建構的方程如下：

$$\begin{aligned} L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma') \\ = \frac{1}{2} W^T W + C \left(\sum_{k=1}^Q \xi_k + \sum_{k=1}^Q \xi'_k \right) - \sum_{k=1}^Q \lambda_k (W^T \Phi(X_k) + w_0 - y_k + \varepsilon + \xi_k) - \end{aligned} \quad (14)$$

$$\sum_{k=1}^Q \lambda'_k (y_k - W^T \Phi(X_k) - w_0 + \varepsilon + \xi'_k) - \sum_{k=1}^Q (\gamma_k \xi_k + \gamma'_k \xi'_k)$$

接著，透過將 L_P 對主變數的偏導數取其鞍點以獲得最佳解性：

$$\frac{\partial L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma')}{\partial W} = 0 \Rightarrow W = \sum_{k=1}^Q (\lambda_k - \lambda'_k) \Phi(X_k). \quad (15)$$

$$\frac{\partial L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma')}{\partial w_0} = 0 \Rightarrow \sum_{k=1}^Q (\lambda_k - \lambda'_k) = 0. \quad (16)$$

$$\frac{\partial L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma')}{\partial \xi_k} = 0 \Rightarrow \gamma_k = C - \lambda_k. \quad (17)$$

$$\frac{\partial L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma')}{\partial \xi'_k} = 0 \Rightarrow \gamma'_k = C - \lambda'_k. \quad (18)$$

定亦 $K(X_k, X_l) \equiv \Phi(X_k) \cdot \Phi(X_l)$ 為核函數，並以方程式(15)、(17)及(18)帶入方程式(14)以產生簡化的對偶型式 L_D 如下：

maximize

$$L_D(\Lambda, \Lambda') = \sum_{k=1}^Q d_k (\lambda_k - \lambda'_k) - \varepsilon \sum_{k=1}^Q (\lambda_k + \lambda'_k) - \frac{1}{2} \sum_{k=1}^Q \sum_{l=1}^Q (\lambda_k - \lambda'_k) (\lambda_l - \lambda'_l) K(X_k, X_l). \quad (19)$$

with the constraints

$$\sum_{k=1}^Q (\lambda_k - \lambda'_k) = 0 \quad (20)$$

$$0 \leq \lambda_k \leq C, k = 1, 2, \dots, Q \quad (21)$$

$$0 \leq \lambda'_k \leq C, k = 1, 2, \dots, Q. \quad (22)$$

一些常用的核函數包括線性、多項式(齊次)(homogeneous polynomial)、多項式(非齊次)(inhomogeneous polynomial)、徑向基函數(radial basis function)和雙曲線正切函數(hyperbolic tangent)等。此外，”支持”向量是其相對應的 λ_i 或 λ'_i 不為零的成對資料

(X_k, y_k) 。最後，透過最佳化 Lagrangian 函數以獲取權重向量 W 的最佳近似值 \hat{W} 如下：

$$\hat{W} = \sum_{k=1}^{n_s} (\hat{\lambda}_k - \hat{\lambda}_k) \Phi(X_k). \quad (23)$$

值得注意的是，索引 k 只針對總數為 n_s 的支持向量執行。至於偏差 w_0 ，則是透過 Karush–Kuhn–Tucker (KKT)(Karush, 1939, Kuhn and Tucker, 1951)條件，以獲得最佳近似偏差估計值 \hat{w}_0 ：

$$\hat{w}_0 = \frac{1}{n_{us}} \sum_{k=1}^{n_{us}} \left(y_k - \sum_{l=1}^{n_s} \beta_l K(X_l, X_k) - \varepsilon \text{sign}(\beta_k) \right). \quad (24)$$

其中，無界限的且使 Lagrangian 乘數滿足 $0 < \lambda_k < C$ 與 $\beta_k = \hat{\lambda}_k - \hat{\lambda}_k$ 的支持向量總數為 n_{us} 。因此，基於公式(23)與(24)，方程式(7)的線性迴歸估計模型可以獲得如下：

$$f(X, \hat{\lambda}_k, \hat{\lambda}_k) = \sum_{k=1}^Q (\hat{\lambda}_k - \hat{\lambda}_k) K(X_k, X) + \hat{w}_0. \quad (25)$$

其中， X 是由 X_k (for $k = 1, 2, \dots, Q$) 所構成的輸入矩陣。

支援向量迴歸中的參數設置會顯著地影響所建構迴歸模型的準確性。因此，使用必須事先確認使用哪種方法以確定支援向量迴歸中參數的最佳設置。搜索支援向量迴歸中最佳參數設置的方法有很多種，例如，梯度下降演算法 (gradient descent algorithm)(Chapelle *et al.* 2002)、演繹演算法 (evolutionary algorithm)(Chen 2006) 及格狀搜尋法 (grid-search approach)(Hsu *et al.* 2008) 等。由於支援向量迴歸具有非常良好建構非線性函數關係的能力，支援向量迴歸已經被廣泛地應用於解決在許多領域出現的現實世界問題。例如，Cheng and Lu (2018)、Gould *et al.* (2017) 與 Hua *et al.* (2017) 利用支援向量迴歸建構了合適的分析預測模型，並取得很好的預測結果。

4.3. 遺傳演算法

達爾文(Charles Robert Darwin)是一個著名的自然學家、地質學家兼生物學家，其提出了一個著名的重要理論-自然選擇，用以描述世界上有機體的演化現象。基於這個演化機制，Holland (1975) 提出了廣泛使用的啟發式算法，以獲得一個最佳化問題的近似最佳解，稱為遺傳演算法 (genetic algorithm, GA)。為了應用遺傳演算法，第一步就是透過編碼方案以產生的個體，稱為染色體，以表示問題的解。對於一個最佳化問題而言，會有一組的染色體同時形成，稱為群體。接著，我們必須定義一個目標函數以評估染色體對最佳化問題的適合度 (fitness)。適合度越高，染色體的質量越好。然後，設計一個選擇機制以選擇所謂較好的染色體，亦即具有更高適合度值的染色體，以組成交配池，稱為父母。此外，設計一個交配機制以用於創建後代染色體，稱為兒童。透過匹配父母染色體，並在父母，之間交換基因信息，以預期後代染色體，亦即兒童，將會更優於父母染色體。為了模擬自然界中的基因突變，亦設計一個突變機制以打亂後代染色體。一般性遺傳演算法的主要和執行程序說明如下：

步驟 1：為最佳化問題設計一個用以產生染色體的遺傳表示方法。

- 步驟 2：基於目標函數，定義適合度函數。
- 步驟 3：隨機地或按照某種預先設定的方法產生一群染色體。
- 步驟 4：透過對群體中的每個染色體進行解碼以獲得解，從而為每個染色體產生相對應的適合值。
- 步驟 5：基於群體中每個染色體的適合度值，透過應用一定的規則，例如俄羅斯輪組，以選擇一些染色體形成交配池。
- 步驟 6：隨機配對交配池中的兩條染色體，以創建配對父母。
- 步驟 7：每對配對父母，利用交配機制以產生它們的後代染色體。
- 步驟 8：對這些後代染色體，實施突變機制。
- 步驟 9：評估這些後代染色體的適合值，並隨機從原始染色體，亦即步驟 4 中所述群體中的染色體，和後代染色體中，選擇生存的染色體，以形成新的群體，亦即步驟 4 中的群體已經更新。
- 步驟 10：如果滿足終止標準，則停止並產生最終最佳解(染色體)，即具有最佳適合度的染色體;否則，回到步驟 4 繼續執行。

由於遺傳演算法對於求解不同種類最佳化問題的高度彈性，遺傳演算法及其混合演算法已經成為在短時間內獲得近似最佳解的一種流行工具。例如，Wodecki *et al.* (2018)、Wang and Mak (2018)、de Faria *et al.* (2017)與 Nakata *et al.* (2017)已經利用遺傳演算法，為不同領域的問題獲得可接受的最佳解。

5. 股票投資組合最佳化程序架構

本計畫運用層級分析法、支援向量迴歸和遺傳演算法，發展一個三階段投資組合最佳化程序，以同時解決”投資組合選擇”、”投資組合最佳化”和”交易”問題。所提出程序的概念簡要地如圖 2 所描述，並以下面的幾個小節詳細解釋本計畫所提出的演算程序。

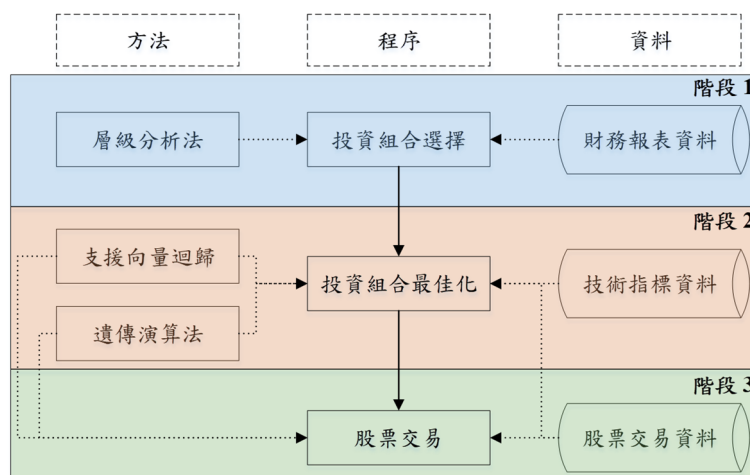


圖 2. 股票投資組合最佳化程序架構

5.1. 投資組合選擇

在第一階段，運用層級分析法對候選公司的財務報表進行排序。值得注意的是，圖 1 中的”決策問題”(目標)是被認為是”選擇股票”(選擇有利可圖的股票)，而與圖 1

中”決策問題”(目標)所對應的”子問題”則是在本計畫中所考慮的財務指標。換句話說，子問題的總數即是投資者所研究的財務指標總數。而可以透過同時最大化每個子問題的績效表現以優化目標(即決策問題)。此外，依據投資者的偏好，將財務指標 A (子問題 A)相對於財務指標 B 的重要性設定為 IM_A/IM_B ，而 IM_A 與 IM_B 分別為財務指數 A 和財務指標 B 的重要性。然而， IM_A/IM_B 必須滿足限制式 $(IM_A/IM_B) \times (IM_B/IM_A) = 1$ 與 $(IM_A/IM_B) \times (IM_B/IM_C) = (IM_A/IM_C)$ ，其中， IM_C 為財務指數 C 的重要性，而 IM_A/IM_C 為財務指標 A 相對於財務指標 C 的重要性。公司 A 對於公司 B ，在財務指標 C 的相對表現則設定為 FI_A^C/FI_B^C ，其中， FI_A^C 與 FI_B^C 分別為公司 A 和公司 B 的財務指標 C 。而如果財務指數 C 是一個負向指數(即越小越好)，則公司 A 對公司 B 的相對表現應該設為 FI_B^C/FI_A^C 。層級分析法中所考慮的候選公司可以透過綜合(synthesizing)這些公司財務指標對於決策目標(決策問題)的表現以進行排序，亦即選擇股票。因此，投資者可以選擇具有較高財務績效表現的公司，以期這些所選取的公司所發行的股票在未來可以表現良好，從而降低投資風險。本階段即是確定投資組合中所包含的股票。值得注意的是，那些具有負面財務指標的公司，而此指數則越大越好，例如負的 EPS，此公司則必須排除在後續層級分析法的進一步分析。

5.2. 投資組合最佳化

首先，蒐集在第一階段層級分析法所選定公司所發行股票的技術指標和交易資料。對於每個選定的股票和每個交易日，以技術指標作為輸入變數(預測變數)，並以落後於目前交易日為決策期長度，例如五個交易日，的收盤價格作為輸出變數(應變數)，以形成列數據。其中，為了避免技術指標之間的巨大差異可能會干擾建模的過程和準確性，每列中的數據應予以進行正規化。另外，根據交易日，將整理完成的正規化數據分為兩部分。投資期內整理完成的正規化數據構成驗證(validation)資料，而其餘整理完成的正規化數據則構成第一部分數據。而第一部分的數據則進一步按比例隨機分為訓練(training)資料和測試(test)資料。接著，利用支援向量迴歸工具以構建預測模型，並採用交叉驗證(cross validation)技術評估預測效能。透過最佳化支援向量迴歸中的參數，我們可以建立最佳的支援向量迴歸預測模型。

依據包含在第一階段所建構的投資組合，在驗證資料期間，即投資期間，股票 j 在目前決策制定(交易)日 i 的技術指標，透過訓練好的支援向量迴歸模型應用以預測股票在下一個決策制定(交易)日的收盤價格。定義無法達到預期利潤的風險如下：

$$RS = NOP\left(\sum_{j=1}^n {}_c\hat{P}_i^j w_i^j, \sum_{j=1}^n {}_cP_{i-1}^j w_i^j, \sum_{j=1}^n \sum_{k=1}^n w_i^j w_i^k Cov_{j,k}\right)$$

其中， ${}_cP_{i-1}^j$ 是股票 j 在決策制定(交易)日 $(i-1)$ 的收盤價，而 ${}_c\hat{P}_i^j$ 與 w_i^j 分別是股票在決策制定(交易)日 i 的預測收盤價和資金分配比例。此外， $Cov_{j,k}$ 是股票 j 與股票 k 的共變異數； n 是投資組合中股票的總數；函數 $NOP(a,b,c)$ 是標準常態變數 Z 比 $(a-b)/\sqrt{c}$ 來得小的累積機率。故投資組合利潤低於預期的懲罰可以定為：

$$PY = \text{Max}\left(TP - \frac{\left(\sum_{j=1}^n {}_c\hat{P}_i^j w_i^j - \sum_{j=1}^n {}_cP_{i-1}^j w_i^j\right)}{\sum_{j=1}^n {}_cP_{i-1}^j w_i^j}, 0\right) \times M \quad (27)$$

其中， TP 是投資者想要的目標利潤，而 M 是一個非常大的數字。因此，投資最佳化組合可以表達為

$$\text{Minimize } RS + PY \quad (28)$$

subject to

$$\sum_{j=1}^n w_i^j = 1 \text{ for } i=1,2,\dots,v \quad (29)$$

其中， v 是投資期間的決策時間點，例如，每周的最後一個交易日，的總天數。

然後，將遺傳演算法應用於求解方程式(28)和(29)中所述的投資組合最佳化問題，從而決定在投資期間每個決策制定日($i=1,2,\dots,v$)的最佳資金分配。

7.3 股票交易

透過比較上決策日 i 與上一個決策日 $i-1$ 的最佳資本分配，投資者可以決定投資組合中的每一檔股票 j 在每個決策日 i 時，應該買入或賣出的股數數量如下：

$$SH_i^j = \left(\frac{w_i^j}{c_i^j} - \frac{w_{i-1}^j}{c_{i-1}^j} \right) \times CAP \text{ for } i=2,\dots,v \text{ and } j=1,\dots,n \quad (30)$$

其中， CAP 為總投資資本，而如果當 SH_i^j 為正數或負數時，投資者應該在決策日 i 買入或賣出股數數量為 SH_i^j 的股票 j 。買賣股票在決策日後的下一個交易日執行。此外，假設所有股票的持有量是可以任意分配，且可以在決策日後的下一個交易日，以股票開盤價買入或賣出。最後，每個獲得每個決策日的交易利潤並產生投資期間的最終投資利潤。

6. 實例探討

6.1. 建構投資組合

本研究考慮了台灣股票市場的半導體和鋼鐵類股的股票。研究期間為 2012 年 1 月 1 日至 2017 年 6 月 13 日。首先，自 TEJ(台灣經濟日報)數據庫所公布之 2017 年第一季季末的財務報告，收集了 222 家股票上市公司的資料。接著，透過七個重要的財務指標，包括每股收益(EPS)，資產報酬率(ROA)，股東權益報酬率(ROE)，毛利率(GPM)，營業利潤率(OPM)，債務比率(DR)，和本益比(P/E)當作層級分析的評估指標。因此，透過層級分析選擇投資組合中的有利可圖的股票，可以被認為是由七個子問題所組成的決策問題，如圖 3 所示。

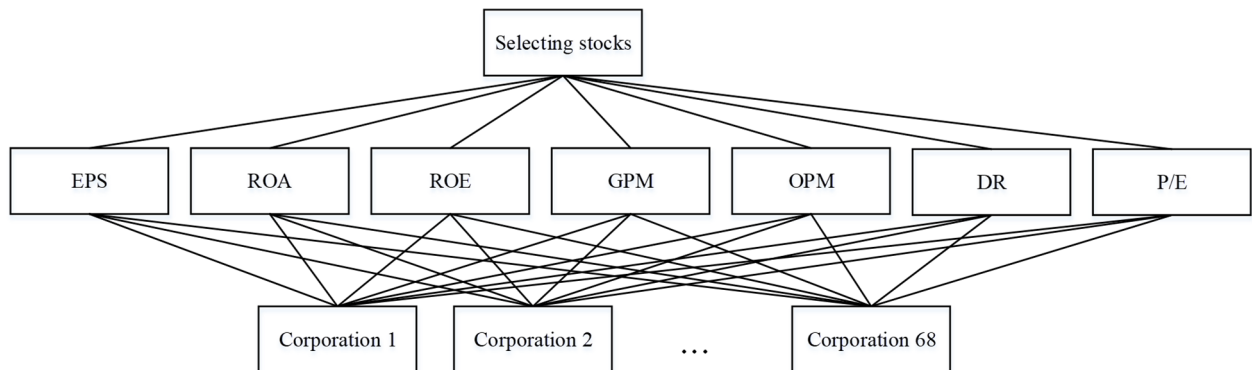


圖 3. 以層級分析選擇股票的決策問題

值得注意的是，在排除了具有負值的“越大越好”財務指數的公司之後，進一步分析的候選公司總數為 68 家。且每個準則(財務指標)與其他標準(財務指標)是一樣重要的。因此，表示第 i 個標準相對於第 j 個標準的重要性，以 a_{ij} 表示，設定為 1，以形成一個 7×7 比較矩陣 A 。因此，根據公式(1)計算正規化標準比較矩陣 A_{norm} 。同時，利用公式(2)獲得標準向量 w (七維列向量)。接著，透過能夠滿足等式(3)和(4)的元素 b_{kl}^i ，將每個備選(公司) k 對於第 i 個標準(財務指標)，與備選(公司) l 進行比較。因此，透過對每列上的元素求平均以產生 68×1 分數向量 $s^{(i)}(i=1, \dots, 7)$ ，並且可以透過等式(5)以獲取分數矩陣 $S(68 \times 7)$ 。在 AHP 結束時，計算 AHP 分配給每個備選(公司)的總得分，進而透過等式(6)以構建總得分數向量 $v(68 \times 1)$ 。因此，可以依據透過合成總目標獲得的分數，以選擇股票出被認為具有相對操作效率高的公司。

本研究考慮了 10 家公司，Expert Choice 11 軟體的執行結果如圖 4 所示。共有 0 檔股票被選取，包括代碼 3529、6510、6462、2408、5274、8150、5269、2330、6568 和 3532 的股票。

Synthesis with respect to: Goal: Selecting Stocks

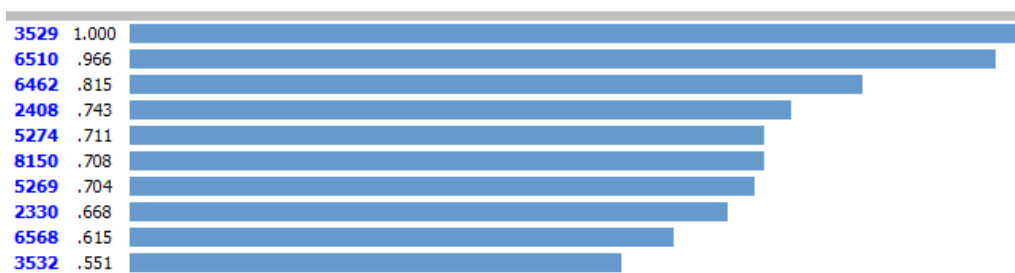


圖 4. Expert Choice 11 所選擇的 10 檔股票

6.2. 投資組合最佳化

首先從 C-Money 和 TEJ 資料庫收集在圖 4 中所選定的 10 檔股票在 2012 年 1 月 1 日至 2017 年 5 月 25 日期間的技術指標和交易數據。本研究考慮了 16 個技術指標，(1)10 日均線、(2)20 乖離率、(3)指數平滑異同移動平均線、(4)9 日隨機指標 K、(5)9 日隨機指標 D、(6)9 日威廉指標、(7)10 日變動率、(8)5 日相對強弱指標、(9)24 順勢指標、(10)26 日成交量變化率、(11)13 日心理線、(12)14 天正向趨向指標、(13)14 天負向趨向指標、(14)26 天買入/賣出動量指標、(15)26 天買/賣意願指標和(16) 10 天動量，這些技術指標是根據 Kim and Han (2000)、Kim and Lee (2004)、Tsang *et al.* (2007)、Chang and Liu (2008)、Ince and Trafalis (2008)、Huang and Tsai (2009)、Lai *et al.* (2009) 和 Hsu (2013) 的研究所選擇出的。

對於每個交易日，將 16 個技術指標以及與距離目前交易日具有一定決策期間的交易日股票收盤價格排列成一列。這些技術指標和收盤價分別作為輸入變數和輸出變

數。然後，根據相對應列的最大值和最小值，將每列中的數據正規化到 0 到 1 之間的範圍。接著，將 2012 年 1 月 1 日至 2017 年 5 月 25 日期間(即模型構建期間)排列的正規化數據，包含 86411 筆資料，以形成資料集(I)。此外，資料集(II)則由 2017 年 5 月 26 日至 2017 年 8 月 25 日期間所排列的正規化資料(即投資期間)所組成，以形成驗證數據。根據 3:1 的比例，隨機分割資料集(II)以產生訓練和測試數據。因此，訓練、測試和驗證數據分別包含 6481、2160 和 650 筆資料。

接著，以 LIBSVM 支援向量迴歸工具應用於訓練和測試數據以構建預測模型，其中，本研究利用交叉驗證技術以評估預測效能，並透過網格搜尋法(Kuhn and A. Tucker, 1951)以最佳化支援向量迴歸中的參數。因此，支援向量迴歸中的重要參數，包含 C 、Gamma 和 Epsilon 分別最佳地設定為 32、0.707106781187 和 0.00390625。此外，使用徑向基函數(radial basis function)內核。獲得的支援向量迴歸模可以為訓練數據獲得 0.994363 和 0.00005428 的 R^2 和 MSE，同時，為測試數據提供 0.991832 和 0.00006052 的 R^2 和 MSE。這些值反映了支援向量迴歸能夠以足夠的準確度對輸入(技術指標)和輸出(股票收盤價格)之間的非線性函數關係進行建模，並可用於預測未來的股票價格。

根據 2017 年 5 月 26 日至 2017 年 8 月 25 日投資期間當前決策(交易)日的技術指標，預測決策期間下一個決策(交易)日的收盤價格，透過構建良好的支援向量迴歸模型，以遺傳算法軟體 Evolver 7 軟體以解決如公式(28)和(29)所示的投資組合最佳化問題，以確定下一個投資期間決策中投資組合裡的每檔股票最佳資本配置。

公式(27)中的目標利潤 TP 和非常大的數值 M ，分別設定為 0.02048% 和 1,000,000。值得注意的是，台灣銀行的一年期定期存款利率為 1.065%，因此，目標利潤設定為 $1.065\% / 52 = 0.02048\%$ ，因為投資組合中的資本設定與分配是每週進行最佳化一次。遺傳演算法中的母體大小、交配率和突變率分別設定為 50、0.5 和 0.15。遺傳演算法程序在每個決策日實施 10 次，表 1 表示了在某些決策日投資組合中股票的最佳資本配置。

表 1. 某些決策日投資組合中股票的最佳資本配置

決策日	股票代碼	資本配置
2017.05.26	2330	0
	2408	0
	3529	0.004100
	3532	0
	5269	0
	5274	0
	6462	0.114866
	6510	0
	6568	0.881033
	8150	0
2017.06.02	2330	0
	2408	0
	3529	0.010048
	3532	0.989952
	5269	0
	5274	0
	6462	0
	6510	0
	6568	0
	8150	0
2017.06.09	2330	0
	2408	0.024577
	3529	0
	3532	0.893456
	5269	0
	5274	0.081967
	6462	0
	6510	0
	6568	0
	8150	0

6.3. 交易股票

在投資期間，將每個決策日的最佳資本配置與前一個決策日獲得的最佳資本配置進行比較，並根據公式(30)計算要買入或賣出的投資組合中的股票數量。此外，總投資資本 *CAP* 設定為 100 萬新台幣。

以 2017 年 6 月 9 日的決策日為例，應購買或出售代碼 3532 的股票數量計算如下：

$$\left(\frac{0.893456}{84.78} - \frac{0.989952}{88.75}\right) \times 1,000,000 = -615.865 \quad (31)$$

其中 84.78 和 88.75 分別是 2017 年 6 月 2 日和 2017 年 6 月 9 日的股票收盤價。因此，投資者於 2017 年 6 月 12 日(即 2017 年 6 月 9 日決策日後的下一個交易日)以 83.39 的開盤價出售代號為 3532 的股票，股票數量為 615.865 股。投資者將獲利新台幣 $615.865 \times 83.39 = 51356.98$ 元。

透過對每個決策日的投資組合中的每檔股票，採用類似的程序，可以獲得最終的投資利潤。

6.4. 評估投資績效

有 64 家鋼鐵公司在台灣股市發行股票。在同一投資期間，第 6.1 至 6.3 節中說明的投資程序也適用於台灣股票市場鋼鐵類股的股票。支援向量迴歸中的參數 C 、 Γ 和 ϵ 設置為 8、1 和 0.00390625，它們是透過網格搜尋法獲得的，並採用徑向基函數(radial basis function)內核。遺傳演算法的參數，包括母體大小、交配率和突變率分別設定為 50、0.5 和 0.15。表 2 整理了台灣股票市場半導體和鋼鐵類股股票的投資表現。

根據表 2，半導體和鋼鐵類股股票的年投資報酬率分別為 15.36%和 6.15%。台灣一年期定存的利率約為 1%。因此，根據實驗結果，本研究所提出的方法是現實世界股票市場中的實用投資工具。

表 2. 股票投資績效整理

類股	半導體	鋼鐵
投資期間	2017.05.26~2017.08.25	2017.05.26~2017.08.25
期初資本	1,000,000	1,000,000
期末資本	1,038,395	1,015,446
三個月投資報酬率	3.8395%	1.5446%
年投資報酬率	15.36%	6.18%

7. 結論

過去的研究中已經廣泛研究了投資組合最佳化問題。然而，最佳化模型通常只考慮投資者是否可以獲得他/她的估計利潤。然而，並沒有考慮投資風險的問題，亦即投資者無法獲得他/她所期望的估計利潤的可能性。其次，大多數研究都獨立地處理了“投資組合選擇”，“投資組合最佳化”和“股票交易”。但是，上述三個問題實際上是相互影響的。因此，本研究利用層次分析法、支援向量迴歸和遺傳演算法以設計一個三階段的投資組合最佳化程序，以便系統性的同時解決這些問題。

投資於台灣股票市場半導體和鋼鐵類股的股票的案例研究，證明了本研究所提出方法的有效性與可行性。根據實驗結果，半導體和鋼鐵類股的股票年投資報酬率分別達到 15.36%和 6.15%，遠優於台灣的一年期定存利率(約 1%)。因此，本研究提出的方法可以被認為是在現實世界股票市場中，投資股票的實用且有用的工具。

總結來說，本研究所提的方法具有以下貢獻：1、投資者可以同時考慮幾個財務指標，並根據自己的喜好隨意地改變其相對重要性；2、透過使用明確的技術指標隱性地了解投資者的心態，以預測未來股票價格的表現；3、個人可以考慮投資利潤和風險，從而建立他/她的最佳投資組合；4、可以清楚地確定交易股票的最佳時機。

未來可能的研究可以有幾個方向，例如：(1)研究股票表現與公司運營效率之間的

關係；(2)考慮股票交易的實際考慮因素，如股票分割和交易成本的分割，以及(3)最佳化支援向量迴歸和遺傳演算法中參數的設置。

參考文獻

- Andrade, C. E.; Ahmed, S., Nemhauser, G. L. and Shao, Y. F. (2017), A hybrid primal heuristic for finding feasible solutions to mixed integer programs, *European Journal of Operational Research*, **263**(1), 62–71.
- Cortes, C. and Vapnik, V. (1995), Support-vector networks, *Machine Learning*, **20**(3), 273–297.
- Chang, P.-C., Liu, C.-H. (2008), A TSK type fuzzy rule based system for stock price prediction, *Expert Systems with Applications*, **34**(1), 135–144.
- Dorigo, M., and Maniezzo, V. (1996), Colomni A. The ant system: optimization by a colony of cooperating agents, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **26**(1), 29–41.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. (1997), Support vector regression machines, in M. C. Mozer, M. I. Jordan, and T. Petsche (Eds.), *Advances in Neural Information Processing Systems*, **9**, 151–161, Cambridge, MA: MIT Press.
- Eberhart, R. C., and Kennedy, J. (1995), A new optimizer using particle swarm theory, in *Proceedings of the sixth international symposium on micro machine and human science*, Nagoya, Japan.
- Gandomi, A. H., Yang X.-S., and Alavi, A. H. (2013), Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems, *Engineering with Computers*, **29**(1), 17-35.
- Geem, Z. W., Kim, J.-H., and Loganathan, G. V. (2001), *A New Heuristic Optimization Algorithm: Harmony Search*, Institutional Knowledge at Singapore Management University, Singapore.
- Goh, C. Y., Jiang, F., Tu, J. and Zhou, G. (2011), Forecasting bond risk premia using technical analysis, *Simulation*, **76**(2), 60–68.
- Guo, L. H., Wang, G. G., Gandomi, A. H., Alavi, A. H., Duan, H. (2014), A new improved krill herd algorithm for global numerical optimization, *Neurocomputing*, **138**, 392–402.
- Hofmann, E. E., Haskell, A. G. E., Klinck, J. M., and Lascara, C. M. (2004), Lagrangian modelling studies of Antarctic krill (*Euphasia superba*) swarm formation, *ICES Journal of Marine Science*, **61**(4), 617–631.
- Hossain, A., and Nasser, M. (2011), Comparison of finite mixture of ARMA-GARCH, back propagation neural networks and support-vector machines in forecasting financial returns, *Journal of Applied Statistics*, **38**(3), 533–551.
- Hsieh, L. F., Hsieh, S. C., and Tai, P. H. (2011), Enhanced stock price variation prediction via DOE and BPNN-based optimization, *Expert Systems with Applications*, **38**(11), 14178–14184.
- Hsu, C. W., Chang, C. C. and Lin, C. J. (2008), A practical guide to support vector classification, available at: <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
- Hsu, C.-M. (2013), A hybrid procedure with feature selection for resolving stock/futures price forecasting problems, *Neural Computing & Applications*, **22**(3–4), 651–671.
- Huang, C.-L. and Tsai, C.-Y. (2009), A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting, *Expert Systems with Applications*, **36**(2), 1529–1539.
- Ince, H. and Trafalis, T. B. (2008), Short term forecasting with support vector machines and

- application to stock price prediction, *International Journal of General Systems*, **37**(6), 677–687.
- Kaveh, A., and Talatahari, S. (2010), A novel heuristic optimization method: charged system search, *Acta Mechanica*, **213**, 267–289.
- Kim, K.-J., and Han, I. (2000), Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, *Expert Systems with Applications*, **19**(2), 125–132.
- Kim, K.-J. and Lee, W. B. (2004), Stock market prediction using artificial neural networks with optimal feature transformation, *Neural Computing & Applications*, **13**(3), 255–260.
- Lahmiri, S. (2017), Modeling and predicting historical volatility in exchange rate markets, *Physica A-Statistical Mechanics and Its Applications*, **471**, 387–395.
- Lai, R. K., Fan, C.-Y., Huang, W.-H. and Chang, P.-C. (2009), Evolving and clustering fuzzy decision tree for financial time series data forecasting, *Expert Systems with Applications*, **36**(2), 3761–3773.
- Lu, C. J., J., Ramjugernath, D. (2013), Hybridizing nonlinear independent component analysis and support vector regression with particle swarm optimization for stock index forecasting, *Neural Computing & Applications*, **23**(7–8), 2417–2427.
- Moodley, K., Rarey, J., Ramjugernath, D. (2015), Application of the bio-inspired Krill Herd optimization technique to phase equilibrium calculations, *Computers & Chemical Engineering*, **74**, 75–88.
- Moodley, K., Rarey, J., Ramjugernath, D. (2015), Application of the bio-inspired Krill Herd optimization technique to phase equilibrium calculations, *Computers & Chemical Engineering*, **74**, 75–88.
- Morin, A., Okubo, A., and Kawasaki, K. (1988), Acoustic data analysis and models of krill spatial distribution, *Scientific Committee for the Conservation of Antarctic Marine Living Resources, Selected Scientific Papers, Part I*, 311–329.
- Mukherjee, A., and Mukherjee, V. (2016), Chaos embedded krill herd algorithm for optimal VAR dispatch problem of power system, *International Journal of Electrical Power & Energy Systems*, **82**, 37–48.
- Ortiz-Garcia, E. G., Salcedo-Sanz, S., Perez-Bellido, A. M. and Portilla-Figueras, J. A. (2009), Improving the training time of support vector regression algorithms through novel hyper-parameters search space reductions, *Neurocomputing*, **72**(16–18), 3683–3691.
- Oztekin, A., Kizilaslan, R., Freund, S., and Iseri, A. (2016), A data analytic approach to forecasting daily stock returns in an emerging market, *European Journal of Operational Research*, **253**(3), 697–710.
- Pradeepkumar, D. and Ravi, D. (2017), Forecasting financial time series volatility using particle swarm optimization trained quantile regression neural network, *Applied Soft Computing*, **58**(September), 35–52.
- Price H. J. (1989), Swimming behavior of krill in response to algal patches: a mesocosm study, *Limnol Oceanogr*, **34**(4), 649–659.
- Ren, Y., Chen, Suganthan, P. N. and Srikanth, N. (2016), A novel empirical mode decomposition with support vector regression for wind speed forecasting, *IEEE Transactions on Neural Networks and Learning Systems*, **27**(8), 1793–1798.
- Sermpinis, G., Stasinakis, C., Theofilatos, K., Karathanasopoulos, A. (2015), Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization, *Analytica Chimica Acta*, **544**(1–2), 292–305.

- Skouras, S. (2001), Financial returns and efficiency as seen by an artificial technical analyst, *Journal of Economic Dynamics & Control*, **25**, 223–244.
- Tang, W. J., and Wu, Q. H. (2009), Biologically inspired optimization: a review, *Transactions of the Institute of Measurement and Control*, **31**(6), 495–515.
- Trafalis, T. B., and Ince, H. (2000), *Support Vector Machine for Regression and Applications to Financial Forecasting*, School of Industrial Engineering University of Oklahoma, OK.
- Tsang, P. M., Kwok, P., Choy, S. O., Kwan, R., Ng, S. C., Mak, J., Tsang, J., Koong, K., and Wong, T.-L. (2007), Design and implementation of NN5 for Hong Kong stock price forecasting. *Engineering Applications of Artificial Intelligence*, **20**(4), 453–461.
- Ustun, B., Melssen, W. J., Oudenhuijzen, M. and Buydens, L. M. C. (2005), Modeling, forecasting and trading the EUR exchange rates with hybrid rolling genetic algorithms-Support vector regression forecast combinations, *European Journal of Operational Research*, **247**(3), 831–846.
- Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, New York: Springer-Verlag.
- Vapnik, V. N. (1998), *Statistical Learning Theory*, New York: Wiley.
- Vapnik, V., Golowich, S., and Smola, A. (1997), Support vector method for function approximation, regression estimation, and signal processing, in Mozer, M., Jordan, M. and Petsche, T. (Eds.), *Advances in Neural Information Processing Systems*, **9**, 281–287, Cambridge, MA: MIT Press.
- Versace, M., Bhatt, R., Hinds, O. and Shiffer, M. (2004), Predicting the exchange traded fund DIA with a combination of genetic algorithms and neural networks, *Expert Systems with Applications*, **27**(3), 417–425.
- Wang, J., and Wang, J. (2015), Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks, *Neurocomputing*, **156**, 68–78.
- Wei, L.-Y., Cheng, C.-H., Wu, H.-H. (2014), A hybrid ANFIS based on n-period moving average model to forecast TAIEX stock, *Applied Soft Computing*, **19**, 86–92.
- Yang, X.-S. (2008), *Nature-Inspired Metaheuristic Algorithms*, Cambridge: Luniver Press.
- Zang, H., Zhang, S., and Hapeshi, K. (2010), A review of nature-inspired algorithms, *Journal of Bionic Engineering*, **7**(S), 232–237.
- Zhang, X. L., Chen, X. F. and He, Z. J. (2010), An ACO-based algorithm for parameter optimization of support vector machines, *Expert Systems with Applications*, **37**(9), 6618–6628.
- Zhong, X. and Enke, D. (2017), Forecasting daily stock market return using dimensionality reduction, *Expert Systems with Applications*, **67**, 126–139.
- 王翎聿(2015)，應用倒傳遞類神經網路與支援向量機預測加權股價指數。國防大學管理財務管理學系碩士班，碩士論文。
- 石濟豪(2014)，技術分析指標複合的搭配及預測:以台灣加權股價指數為例。暨南大學國際企業學系碩士班，碩士論文。
- 林婉茹(2004)，類神經網路於台灣50指數ETF價格預測與交易策略之應用。輔仁大學金融研究所，碩士論文。
- 黃敏菁(2005)，支援向量機在財務時間序列預測之應用。輔仁大學金融研究所，碩士論文。
- 陳昌捷(2015)，以倒傳遞類神經網路預測股市指數。國立宜蘭大學多媒體網路通訊數位學習碩士在職專班，碩士論文。
- 陳淑玲(2010)，臺灣股票市場技術指標之研究-不同頻率資料之分析。東海大學財務金

- 融學系碩士在職專班，碩士論文。
- 陳鄺貞(2011)，以財務指標及技術指標建構股價預測模型—類神經網路模型之應用。國立臺北大學國際財務金融碩士在職專班，碩士論文。
- 張瑞杰(2009)，變幅波動與GARCH模型之波動預測績效比較—台灣加權股價指數之實證。淡江大學財務金融學系碩士在職專班，碩士論文。
- 鄒紹輝(2006)，隱含波動率之模型及預測：以台灣市場為例。國立中央大學統計研究所，碩士論文。
- 蔡承益(2007)，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。
- 鄭健毅(2010)，應用SVR支援向量迴歸模式來進行電子產業股價預測。明新科技大學工業工程與管理研究所，碩士論文。
- 劉翔瑜(2006)，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。
- 戴淑瑩(2007)，臺灣50指數ETF整合型預測之研究。國立成功大學統計學系碩博士班，碩士論文。

明新科技大學 108 年度 研究計畫執行成果自評表

計畫類別： <input type="checkbox"/> 任務導向計畫 <input type="checkbox"/> 整合型計畫 <input checked="" type="checkbox"/> 個人計畫 所屬院(部)： <input type="checkbox"/> 工學院 <input checked="" type="checkbox"/> 管理學院 <input type="checkbox"/> 服務學院 <input type="checkbox"/> 人文社會科學院 執行系別：企業管理系 計畫主持人：徐志明 職稱：教授 計畫名稱：具有股票交易擇時之投資組合最佳化投資程序 計畫編號：MUST-108 企管-1 計畫執行時間：108 年 01 月 01 日 至 108 年 09 月 30 日	
計畫執行成效	教學方面 1.對於改進教學成果方面之具體成效： <u>透過本專題研究之執行，可讓計畫主持人對「股票投資組合最佳化」問題有更深入之瞭解，這將有助於計畫主持人未來在「財務管理」之教學活動上更加得心應手。另透過層級分析法、支援向量迴歸、遺傳演算法之實際應用，將有助於計畫主持人未來在「柔性演算法」及「最佳化」等課程之教學活動。</u> 2.對於提昇學生論文/專題研究能力之具體成效： <u>此專題研究已經協助訓練專題學生在尋找研究主題、蒐集文獻、發展研究架構，以及使用層級分析法、支援向量迴歸、遺傳演算法等工具之實際經驗。</u> 3.其他方面之具體成效： <u>此專題研究順利訓練專題學生具有相當程度的研究能力，並且可做為學生在未來擬定畢業專題或研究所論文之重要參考資料及研究方向。</u>
	學術研究方面 1.該計畫是否有衍生出其他計畫案 <input type="checkbox"/> 是 <input checked="" type="checkbox"/> 否 計畫名稱：_____ 2.該計畫是否有產生論文並發表 <input checked="" type="checkbox"/> 已發表 <input type="checkbox"/> 預定投稿/審查中 <input type="checkbox"/> 否 發表期刊名稱： <u>International Journal of Engineering Research and Management (兩篇)</u> 發表期刊日期： <u>2019 年 2 月 1 日與 2019 年 7 月 1 日</u> 發表研討會名稱： <u>2019 管理與服務創新國際學術研討會(兩篇)</u> 發表研討會日期： <u>2019 年 5 月 17 日</u> 3.該計畫是否有要衍生產學合作案、專利、技術移轉 <input type="checkbox"/> 是 <input checked="" type="checkbox"/> 否 請說明衍生項目：_____

<p>成 果 自 評</p>	<p>計畫預期目標：</p> <ol style="list-style-type: none"> (1)完成投資組合選擇與最佳化問題的文獻回顧及評述。 (2)依據投資組合選擇與最佳化問題文獻評述，擬定投資組合最佳化程序之初步想法。 (3)完成層級分析法、支援向量迴歸以及遺傳演算法等研究方法的文獻探討與整理。 (4)完成利用研究方法發展解決投資組合最佳化問題的系統化程序。 (5)完成使用台灣經濟新報及 CMoney 財經資料蒐集軟體、層級分析法、支援向量迴歸以及遺傳演算法軟體。 (6)以臺灣股市的兩個類股為例，驗證本計畫提出之系統化投資組合最佳化程序之可行性與有效性。 (7)將本計畫所提出之股票投資組合最佳化程序程序，與股票大盤和定存績效進行比較。 (8)檢討整體研究計畫之成果、彙整結論並提出未來研究方向之建議。 (9)整理研究計畫報告、投稿研討會和學術期刊。 <p>計畫執行結果：</p> <p>完整達成計畫預期目標。</p> <p>預期目標達成率：100 %</p>
	<p>其它具體成效：無 (若不敷使用請另加附頁繕寫)</p>

明新科技大學 108 年度校內專題研究計畫 運用於教學成果記錄表

計畫類型	<input checked="" type="checkbox"/> 個人型 <input type="checkbox"/> 整合型 <input type="checkbox"/> 任務導向型		計畫編號	MUST-108 企管-1	
計畫名稱	具有股票交易擇時之投資組合最佳化投資程序				
計畫主持人 資料	姓名	徐志明		職稱	教授
	學院	管理學院		系所	企業管理系
聘用助理	系科班級	學號	姓名	聘僱起訖時間	工作內容
融入課程	開課班級	課程名稱		修課 人數	課程內容概述
	四技企二甲 二專企二甲 夜四技企三甲	生產與作業管理		56 15 42	在課程中,教導學生如何利用生產與作業管預測與模擬之概念,進行預測模型建構。
	四技企二乙 四技企二甲	應用統計學 統計學		41 61	在課程中,教導學生如何利用統計之概念,進行投資組合最佳化模式建構與層級分析法之使用。
	四技企一甲 四技企四乙	資訊素養 商用套裝軟體		51 50	在課程中,教導學生如何利用資料分析、資料庫、程式語言和發展軟體之概念,做出一個可以為企業界實用之軟體。
指導專題或 碩士論文	指導班級	專題(論文)名稱		分組 人數	專題(論文)內容概述
指導學生參 與活動或競 賽	活動或競賽名稱			參與 人數	活動或競賽成果概述
製作教材與	教材與教具名稱			教材與教具概述	

<p style="text-align: center;">教具</p>		
<p style="text-align: center;">其他促進教學之成果說明</p>	<ol style="list-style-type: none"> 1. 此專題研究已經訓練學生在尋找研究主題、蒐集文獻、發展研究架構，以及使用層級分析法、支援向量迴歸以及遺傳演算法等工具之實際經驗。 2. 此專題研究已經順利讓學生具有相當程度的獨立研究能力，並且可做為畢業專題或論文之重要參考資料及研究方向。 	

An Empirical Study of Institutional Investors' Net Buy/Net Sell on Forecasting Stock Prices Using Genetic Programming in Taiwan

Chih-Ming Hsu

Abstract— Among various traditional investment tools, stock investment is one of the easily understood targets for ordinary people because the concept of trading stocks is relatively clear and simple. However, for investors, the essential task of accurately forecasting future trends or market prices of stocks is difficult, since the factors that can affect the performance of stocks in the trading market are diverse. Therefore, the problems involved in forecasting stock prices continuously attracts great interest from both researchers and practitioners. Investors making short-term stock investments, especially, prefer to forecast stock prices via technical indicators calculated based on stock-trading information. In addition, an investor pays very close attention to the institutional investors' net buy or net sell of each day, because its value can reflect future expectation about the overall performance of a corporation, as well as judgement regarding the future trends of stock prices. Therefore, this study proposes a stock price forecasting procedure based on genetic programming (GP) and cluster analysis, using technical indicators as predictors in an empirical study on the effects of institutional investors' net buy or net sell in forecasting stock prices in Taiwan. The feasibility and effectiveness of the proposed procedure are illustrated through examining five stocks with the highest trading volumes in the semiconductor section of the TAIEX (Taiwan Capitalization Weighted Stock Index). The implementation results show that the stock price modelling procedure applying the GP method is a robust and practical forecasting technique. Institutional investors' net buy/net sell of stocks can indeed improve forecasting performance. Furthermore, forecasting accuracy can be further refined through classifying the trends in institutional investors' net buy/net sell.

Index Terms— Cluster Analysis, Genetic Programming, Institutional Investors, Stock Price Forecasting.

I. INTRODUCTION

Investing in stocks is one of the traditional and easily understood investment tools for most people, since the concept of trading stocks is clear and simple compared to other investment targets, e.g. options, futures, or swaps. No matter what investing strategies an investor applies, forecasting future trends, even future market prices, is essential work for an investor. Some people believe that the financial condition of a corporation is the most critical to the market prices of stocks issued by the company, thus preferring to employ information in financial reports to project the achievements of a stock, called fundamental analysis. Some

investors, especially those making short-term stock investment, however, think that the status in the stock trading market and current events recently are key to future stock prices. This group of investors prefer to forecast stock prices via the indices, i.e. technical indicators, based on the stock trading information, called technical analysis. Whether employing fundamental or technical analysis, forecasting the trends or prices of a stock is always a tough task, since the factors that can affect its performance in a trading market are diverse; thus this topic continuously attracts high interest of researchers coming from both academic and practical worlds. For example, Behravesh [1] applied regression models to forecast the stock prices of the major Iranian petrochemical companies, where the predictors (independent variables) included (1) stock price in the last month, (2) capital of the company, (3) P/E (price-earnings ratio), (4) DPS (dividend per share), and (5) EPS (earnings per share). He used E-Views software to run the regression models, and discussed choosing the best decision among petroleum companies for beneficiary business markets. The author also discovered the application and effectiveness for making stock investment decisions, and assigned a limited budget to each petroleum company. Yeh, Huang and Lee [2] incorporated the sequential minimal optimization and gradient projection methods to develop a two-stage multiple-kernel learning algorithm to resolve stock market forecasting problems. Their proposed algorithm can take several combined advantages from different hyperparameter settings and improve the overall system performance. In addition, the hyperparameter settings need not be specified in advance, and the trial-and-error procedure for determining the optimal values for the hyperparameters can be avoided. The proposed algorithm is demonstrated and compared to single kernel support vector regression (SKSVR) [3], autoregressive integrated moving average (ARIMA)[4], and TSK-type fuzzy neural network (FNN) [5] by carrying out experiments on datasets taken from the TAIEX (Taiwan Capitalization Weighted Stock Index). The experimental results revealed that their approach can provide performance superior to other methods. Zuo and Kita [6] transformed the continuous P/E ratio to a set of digitized values via a clustering algorithm, and forecast the P/E ratio by applying a Bayesian network to the set of digitized values. They took the NIKKEI stock average (NIKKEI225) and Toyota Motor Corporation stock price as examples. The results showed that their approach can attain similar accuracy and a better correlation coefficient compared to time-series forecast algorithms. In addition, their algorithm, using the Ward method, can improve the computational accuracy by

Manuscript received February, 2019.

Chih-Ming Hsu, Department of Business Administration, Minghsin University of Science and Technology, Hsinchu, Taiwan (R.O.C.)

An Empirical Study of Institutional Investors' Net Buy/Net Sell on Forecasting Stock Prices Using Genetic Programming in Taiwan

15% and 20% for the NIKKEI stock average and Toyota Motor Corporation stock price, respectively, against the traditional AR (Auto Regressive), MA (Moving Average), ARMA (Auto Regressive Moving Average) and ARCH (AutoRegressive Conditional Heteroskedasticity) methods. Hsu [7] combined the backpropagation (BP) neural network, feature selection, and genetic programming (GP) techniques to develop a hybrid procedure for resolving stock and futures price forecasting problems with the technical indicators as predictors. He first used the BP neural network to construct a preliminary forecasting model, then utilized feature selection through simulation to probe the built neural network, thus selecting the critical technical indicators for forecasting stock and futures prices. Furthermore, the vital technical indicators were also automatically screened out by employing the GP method. Finally, the final forecasting model using the selected technical indicators was established using the BP neural network. The author used TAIEX futures of the spot month to demonstrate the feasibility and effectiveness of the proposed procedure. Based on the experimental results, the forecasting performance was significantly improved through selecting appropriate technical indicators by applying the feature selection method or solely based on the preliminary GP forecasting model. Liu and Hu [8] proposed a feature-weighted support vector machine regression algorithm by providing different weights for different features of the samples, thus improving the performance of traditional SVM. A case study on examining sample stock data sets selected from China was conducted to demonstrate their proposed method, and the result showed that using GCD (grey correlation degree) as the weight value had good generalization capability, and the prediction accuracy improved. Xiong, Bao and Hu [9] applied multi-output support vector regression (MSVR), whose parameters are determined by their proposed approach based on the firefly algorithm (FA), to forecast the interval-valued stock price index series over short and long horizons. Three globally traded broad market indices (the S&P 500 for the US, the FTSE 100 for the UK, and the Nikkei 225 for Japan) were used to illustrate their method. The experimental results showed that their proposed FA-MSVR method outperformed some well-established counterparts based on statistical criteria regarding the forecasting accuracy measure and the accuracy of competing forecasts. Xiao, Xiao, Lu and Wang [10] proposed a three-stage nonlinear ensemble model based on neural networks, improved particle swarm optimization (IPSO), and support vector machines (SVM). In their study, three different types of neural-network-based models, including the Elman network, generalized regression neural network (GRNN) and wavelet neural network (WNN), all further optimized by improved particle swarm optimization (IPSO), were constructed. Later, the support vector machines (SVM) neural network was utilized to generate a neural-network-based nonlinear meta-model. Their proposed approach was able to explore complex nonlinear relationships better, and the built forecasting model was validated by three daily stock indices' time series, including the Shanghai composite index, Shenzhen component index, and Shanghai-Shenzhen 300. The empirical results demonstrated that their proposed ensemble approach significantly

improved prediction performance over other individual models and linear combination models. Dan, Guo, Shi, Fang and Zhang [11] presented deterministic echo state network (ESN) models, which construct reservoirs randomly to simplify their structure and applications relative to the standard ESN, for stock price forecasting. They used two benchmark datasets, including the Shanghai Composite Index and S&P 500, to investigate the forecasting performance of their presented method. The experimental results showed that the deterministic ESN was able to outperform the standard ESN by about 20% and 52% in accuracy and stability, respectively, on average. Furthermore, about 23% improvement in efficiency, as well as insignificant improvement in forecasting accuracy, was found for the S&P 500 dataset. Singh and Borah [12] introduced a new type-2 fuzzy time series model, which was then enhanced by applying particle swarm optimization (PSO), in order to utilize more observations while forecasting. The authors' purpose was to tune up the lengths of intervals in the universe of discourse which are used while forecasting, but not to increase the number of intervals. The performance of their proposed model was evaluated by making a study on the daily stock index price data set of SBI (State Bank of India), as well as on the daily stock index price of Google. The experimental results showed that their proposed model is effective and robust compared to the existing fuzzy and conventional time series models. Rounaghi, Abbaszadeh and Arashi [13] utilized the multivariate adaptive regression splines (MARS) model and semiparametric splines technique to predict stock prices. In their study, the MARS model and semi-parametric smoothing splines technique serve as adaptive and nonparametric regression methods, respectively. They utilized 40 variables, including 30 accounting variables and 10 economic variables, to predict stock price by using the MARS model and semi-parametric splines technique. Four accounting variables: (1) book value per share, (2) predicted earnings per share, (3) P/E ratio and (4) risk, were selected by investigating the models to be the influencing variables on stock price forecasting via the MARS model. On the other hand, another combination of four accounting variables, which included dividends, net EPS, EPS forecast and P/E ratio, were chosen as effective variables while forecasting the stock prices. The performance regarding the multi-step ahead forecasting of their proposed approach was evaluated by comparing it to the traditional global linear model through simulation, and the results indicate the nonparametric model can yield superior forecasting performance compared to the global linear model. Furthermore, the intraday data of the Japanese stock price index and time series of heart rates are also analyzed and forecast in their study, and the experimental results revealed that the forecasting performance does not differ significantly in the Japanese stock price index, but the nonparametric model can provide significantly better performance while analyzing heart rates. Guo, Han, Shen and Li [14] applied the support vector machine regression (SVR) technique to tackle the characteristics, including discreteness, non-normality and high noise, which are significantly different in different periods for the same stock, or in the same period for different stocks in high-frequency data. In his study, an adaptive SVR with dynamic optimization of the

learning parameters through particle swarm optimization (PSO) is developed to resolve the stock data at three different time scales (daily data, 30-min data, and 5-min data). Compared to the traditional SVR and backpropagation neural networks, his proposed approach can yield better results based on the experimental results. Wang [15] proposed a method based on the big data framework with fuzzy time series to forecast stock prices. She applied fuzzy time series to historical stock big data to predict the fuzzy trend regarding the forecast data, then determined the amount of fluctuation about the forecast data by using an autoregressive model. By integrating trend prediction with fluctuation quantity together, the forecast stock prices are finally obtained. Her proposed forecasting framework was illustrated by forecasting the TAIEX, and produced superior forecasting accuracy compared to existing methods. Chou and Nguyen [16] proposed a sliding-window metaheuristic optimization model by hybridizing the FA (firefly algorithm) and LSSVR (least squares support vector regression), called MetaFA-LSSVR, to predict the stock prices of Taiwanese construction companies one step ahead. Their proposed system is a stand-alone application with a graphical user interface, which is greatly interesting to home brokers that have insufficient investment knowledge. In addition, their proposed model is a favorable predictive technique for dealing with the highly nonlinear time series that traditional models have difficulty capturing. Their experiments indicated that outstanding prediction performance and improved overall profit were attained by using their developed hybrid system. Cheng and Yang [17] utilized the rough set rule induction to develop a fuzzy time-series model to forecast stock indices. They employed rough sets to generate forecasting rules for replacing fuzzy logical relationship rules according to the lag period, and utilized an adaptive expectation model to improve forecasting performance. Furthermore, they also presented buy and sell rules to provide investment suggestions as references for investors based on three different scenarios. A dataset consisting of the TAIEX, Nikkei, and HSI stock prices from 1998 to 2012 was used to evaluate their proposed model, which was able to outperform existing models with comparing to the listing models under three error indices and profits criteria.

Thus, in the literature, some relatively new techniques, e.g. support vector regression (SVR), neural network (NN), genetic programming (GP), firefly algorithm (FA), and particle swarm optimization (PSO), have been broadly utilized to construct stock price forecasting models to improve the shortcomings or limitations of traditional statistical methods, thus obtaining fairly excellent experimental results. In addition, the trading information regarding institutional investors' net buy/net sell announced every trading day is important reference material when diagnosing the trends of stock prices for an investor. Therefore, this study intends to conduct an empirical study on the effects of institutional investors' net buy/net sell in forecasting stock prices in Taiwan based on genetic programming (GP). The information of institutional investors' net buy/net sell is used to recognize and classify the stock prices' trends, and GP is employed to construct the forecasting model corresponding to each classification

(group) of stock prices' trends with technical indicators as predictors. The feasibility and effectiveness of the proposed procedure are demonstrated by examining the five stocks with top trading volumes in the semiconductor section of the TAIEX. The rest of the paper is organized as follows. Section 2 briefly introduces the genetic programming (GP) method, as well as the technical indicators. The proposed stock price forecasting procedure is presented and illustrated by a case study in Sections 3 and 4, respectively. Finally, conclusions and possible research directions are given in Section 5.

II. METHODOLOGIES

A. Genetic Programming

Based on the principles of Darwinian natural selection and biologically inspired operations, Koza (1992) invented an evolutionary method for generating programs or functions automatically, called genetic programming (GP), to solve a user-defined problem through evolving a population of chromosomes. The GP utilizes a tree-based structure consisting of terminal and function sets, as shown in Figure 1, to represent an individual program (chromosome). The tree (program) can be interpreted from the left to the right and from the top to the bottom as $(5.2 - x/15) + (9 * \sin(y))$. The available terminal elements to the branch in an evolving chromosome such as the constants, variables or zero-argument functions, etc., are defined in the terminal set. The function set consists of the functions of the program, e.g. the square root, minus, logarithm, or sine, etc.

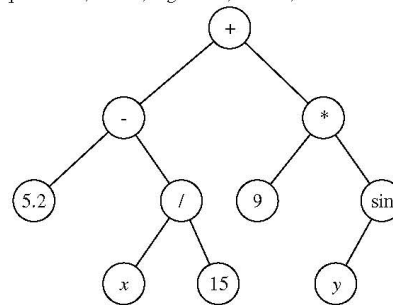


Figure 1. An example of tree structure expression in GP.

The fundamental steps of GP are briefly illustrated as follows [18–20]

Step 1: Create solutions of an initial population

First, create some solutions (chromosomes) of an initial population with a pre-specified population size, called generation 0, which are computer programs consisting of elements from the functional and terminals sets according to the characteristics of the problem.

Step 2: Evaluate the fitness of each chromosome

Execute each program decoded from the corresponding chromosome in the population and measure the degree of how well the program can deal with the problem at hand via a pre-defined fitness function, called fitness value.

Step 3: Select the elite chromosomes

Based on the fitness value of each chromosome, the

An Empirical Study of Institutional Investors' Net Buy/Net Sell on Forecasting Stock Prices Using Genetic Programming in Taiwan

probability corresponding to the chromosome is first obtained. Some chromosomes (programs) are then picked from the population by using the Russian roulette mechanism according to the obtained probability. These selected chromosomes form a matching pool.

Step 4: Apply genetic operators

To the selected programs in the matching pool, the genetic operators, including reproduction, crossover, mutation, and architecture-altering operator are randomly applied, thus creating an offspring population, called generation $g+1$, by replacing the elements in the population of the current generation g with the chromosomes in the offspring population based on a certain strategy.

Step 5: Examine the criteria for terminating the GP

If one of the termination criteria is satisfied, the best chromosome, i.e. the chromosome that can provide the best execution result (the highest fitness value) is designated as the final result of the GP run. Otherwise, repeat Steps 2 to 5 iteratively.

The GP widely extends the applications of genetic algorithms to the solution space consisting of computer programs. Hence, a lot of successful practical work using the GP in various application fields has been reported in the literature, e.g. [21-24]

B. Technical Indicators

Fundamental analysis and technical analysis are two major analytical tools for choosing appropriate investment stocks. The fundamental analysis is to analyze a business's financial statements (such as assets, liabilities, and earnings), health and competitors, and markets, as well as consider the overall state of the economy and other influential factors (such as the interest rates, production, earnings, employment, GDP, housing, manufacturing and management). With regard to the technical analysis, it is a method for forecasting stock price trends through studying past market data, primarily prices and volumes. The technical analysts widely use market indicators of many sorts, called technical indicators, which are mathematically calculated based on historic prices, volumes, and other inputs, thus aiming to forecast financial market direction. For example, the commodity channel index (CCI), average convergence/divergence (MACD), relative strength index (RSI), and stochastic oscillator (KD) are some common technical indicators.

III. PROPOSED FORECASTING PROCEDURE

This study intends to study the effects of institutional investors' net buy/net sell upon forecasting stock prices through genetic programming modelling. The research model is designed as depicted in Figure 2 and described as follows:

Step 1: Collect stock trading data

The trading data regarding the investment stocks are first collected.

Path I

Step 1-1: Calculate technical indicators

Some important technical indicators are calculated based on the collected stock trading data.

Step 1-2: Prepare training, testing and validation data

For each investment stock, the technical indicators obtained in Step 1-1 and stock trading data collected in Step 1 are first arranged day by day. For each trading day, the input variables, i.e. predictors, are the technical indicators, and the output variable, i.e. response, is the stock closing price in three days. The arranged data are then divided into two parts. The first part is further separated into the training and testing data based on an appropriate proportion, and the second part serves as the validation data.

Step 1-3: Build GP models

The GP technique is applied to construct the stock forecasting model based on the training and testing data prepared in Step 1-2. The GP will be executed several times, and the best GP model is designated by evaluating the weighted forecasting error associated with the training and testing data.

Path II

Step 2-1: Calculate technical indicators along with trends of net buy/net sell

Based on the collected stock trading data, several important technical indicators are calculated. Furthermore, the net buy/net sell by the institutional investors is also confirmed.

Step 2-2: Prepare training, testing and validation data

The obtained technical indicators, information about the trends of net buy/net sell by the institutional investors, and collected stock trading data are organized for each investment target and for each trading day. The data in each row include the predictors and response. The technical indicators serve as the predictors, and the stock closing price in three days is treated as the response. Then, the well-organized data are split into two groups. At the same time, the training and testing data are yielded according to a pre-specified ratio and the second part data form the validation data.

Step 2-3: Build GP models

The GP algorithm is implemented several times to establish the stock forecasting model by using the prepared training and testing data obtained in Step 2-2. The forecasting errors associated with the training and testing data are weighted to determine the best GP stock forecasting model.

Path III

Step 3-1: Calculate technical indicators along with trends of net buy/net sell

According to the stock trading data collected in Step 1, several important technical indicators are figured out. In addition, the trends of net buy/net sell by the institutional investors are also affirmed.

Step 3-2: Group data based on trends of net buy/net sell

First, the trends of net buy/net sell by the institutional investors are classified into three types. The first type's trend illustrates the rising situation, i.e. the

net buys of the institutional investors in the previous three trading days are all positive. If all of the net buys of the institutional investors are negative in the previous three days, the trend is determined to be falling and is categorized as the second type. Remaining trading days which do not belong to type one or two are grouped into the trend of the third type.

Step 3-3: Prepare training, testing and validation data for each group of data

For each group of data classified in Step 3-1, the calculated technical indicators, information about the trends of net buy/net sell by the institutional investors, and collected stock trading data are put in order according to the sequences of trading days for each investment stock. Each row of data involves the technical indicators, serving as the predictors, and the corresponding stock closing price in three days, treated as the response. Then two sets are acquired by separating the arranged data. A pre-specified ratio is utilized to split the first part of the data into the training and testing data, while the validation data are the second part of the data.

Step 3-4: Build GP models for each data group

For each data group obtained in Step 3-2, the GP algorithm is executed several times to construct the GP stock forecasting models, where the required data come from the training and testing data corresponding to each data group prepared in Step 3-3. The forecasting performance of each GP model is evaluated by weighting the corresponding training and testing errors, thus settling the optimal GP forecasting model.

Step 2: Compare forecasting performance

The forecasting performance of the GP models resulting from Path I-III are compared by evaluating their RMSE (root-mean-square error), R^2 (R-square), and MAPE (mean absolute percentage error). Then, the effects of institutional investors' net buy/net sell on forecasting stock are concluded.

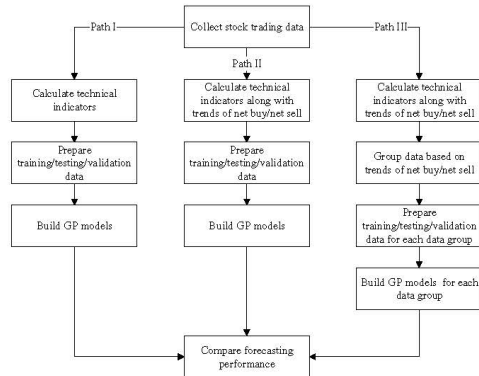


Figure 2. Proposed forecasting procedure.

IV. CASE STUDY

In this section, we utilize the GP algorithm to construct the stock price forecasting models and compare their forecasting performance for several stocks, thus empirically examining the effects of institutional investors' net buy/net sell on forecasting stock prices.

A. Collection of Stock Trading Data

In this study, the examined stocks include five stocks with the highest yearly trading volumes in the semi-conductor section based on the statistics provided by the TWSE (Taiwan Stock Exchange Corporation). The five investigated stocks include stocks of codes 2303, 2330, 2337, 2344, and 6182. Then, the daily trading data include the open prices, highest prices, lowest prices, closing prices, and trading volumes along with the trading volumes of institutional investors' net buy/net sell for the selected five stocks from 2010/01/03 (the first trading day in 2010) to 2018/09/28 (the last trading day in September 2018).

B. Calculation of Technical Indicators

For each studied stock, the daily trading data are used to calculate its technical indicators. There are sixteen technical indicators considered in this study in accordance with Kim and Han [25], Kim and Lee [26], Tsang *et al.* [27], Chang and Liu [28], Ince and Trafalis [29], Huang and Tsai [30], and Lai, Fan, Huang and Chang [31]. Notably, the technical indicators are normalized into the range (-1, +1) based on the corresponding technical indicator's maximum and minimum values to avoid the technical indicators with larger indices dictating the forecasting models.

C. Determination of trends regarding net buy/net sell

For each trading day of each examined stock, the trends regarding net buy/net sell are classified into three types. If the trading volumes of institutional investors' net buy for an examined stock are all positive in the previous successive three days (e.g. $t-2$, $t-1$, and t days), the trend on trading day t is determined as rising (type 1). On the other hand, the trend on trading day t is considered falling (type 2) if all of the trading volumes of institutional investors' net buy in the previous successive $t-2$, $t-1$, and t days are negative. The net buy/net sell trends on one trading day are categorized as type 3.

D. Preparation of training, testing and validation data

For each trading day, the corresponding data are arranged in a row in the form (\mathbf{X}, y) , where \mathbf{X} is the vector of input variables and y is the output variable. The output variable is the stock closing price three days later. However, there are three different situations for determining the input variables. In the first one, the technical indicators of each trading day are used for the input variables. In the second situation, the input variables are consisted by the technical indicators each trading day along with the trading volumes of institutional investors in the previous three days. Finally, in the last situation, the input variables are the same as those in the second situation, but the paired input/output data are further segmented into three types based on the trends regarding net buy/net sell determined previously. Next, for the first or

An Empirical Study of Institutional Investors' Net Buy/Net Sell on Forecasting Stock Prices Using Genetic Programming in Taiwan

second situation, the arranged data from 2011/01/03 to 2016/12/27 form the first group, and the arranged data from 2016/12/28 to 2018/09/28 make up the second group. The first data group are then divided into the training and testing data randomly according to a ratio 3:1, while the second data group becomes the validation data. As regarding each type's data in the third situation, the arranged data between 2011/01/03 to 2016/12/27 form the first group, and are split into the training and testing data with a proportion of 3:1 randomly, while the arranged data between 2016/12/28 to 2018/09/28 turn into the validation data.

E. Building forecasting models by GP

For the first and second situations as well as each type in the third situation, in the preparation of training, testing and validation data, the training data are used to construct the stock price forecasting models by applying the GP techniques. Here, Discipulus software is utilized in this study. The main parameters in GP are set as software's default values (population size = 500, mutation rate = 0.95, and crossover rate = 0.5). The fitness value of each chromosome is measured by the RMSE (root-mean-square error), for which smaller is better. In addition, the built models are also evaluated on their forecasting performance via their corresponding testing data to assess the flexibility of the established GP models when applied to unknown data. The implementation results for stock code 2303 are summarized in Table 1. Based on Table 1, the execution result in bold represents the optimal result evaluated according to the weighted training and testing R^2 among each of the ten implementations. All of the coefficients of variation (CV) regarding the training RMSE, testing RMSE, weighted training and testing RMSE, training R^2 , testing R^2 , and weighted training and testing R^2 are considered to be sufficiently small, as shown in Table 1, where the largest CV is just 0.064993. Therefore, the stock price modelling procedure via the GP technique can be thought of as a robust and useful tool.

Similarly, the same modelling method is also implemented for the other four investigated stocks. The selected best GP models are given in Table 2.

Table 1. The Implement Results of Stock Code 2303.

(A) Situation 1						
GP Run No.	Training RMSE	Testing RMSE	Weighted Training and Testing RMSE	Training R^2	Testing R^2	Weighted Training and Testing R^2
1	0.004459	0.004429	0.004444	0.90334	0.90371	0.90353
2	0.004608	0.004633	0.004651	0.90706	0.90747	0.90726
3	0.004937	0.004934	0.004936	0.89696	0.89597	0.89696
4	0.004469	0.004504	0.004496	0.90481	0.90355	0.90418
5	0.005186	0.005366	0.005276	0.89379	0.89843	0.89161
6	0.004679	0.004583	0.004616	0.90274	0.90235	0.90300
7	0.004859	0.004720	0.004790	0.90588	0.90623	0.90666
8	0.004904	0.005105	0.005004	0.90156	0.89636	0.89896
9	0.004457	0.004404	0.004431	0.90562	0.90392	0.90578
10	0.004660	0.004703	0.004682	0.90039	0.89888	0.89963
Mean	0.004758	0.004772	0.004764	0.90200	0.900797	0.901492
Standard deviation	0.000226	0.00031	0.000273	0.004205	0.00563	0.004873
Coefficient of variation	0.049562	0.064993	0.057341	0.004661	0.00625	0.005405

(B) Situation 2

GP Run No.	Training RMSE	Testing RMSE	Weighted Training and Testing RMSE	Training R^2	Testing R^2	Weighted Training and Testing R^2
1	0.004107	0.003846	0.003977	0.91105	0.91635	0.91370
2	0.004284	0.004272	0.004278	0.90950	0.90862	0.90906
3	0.004310	0.004166	0.004238	0.90744	0.91124	0.90934
4	0.004279	0.004189	0.004234	0.90763	0.90930	0.90822
5	0.004056	0.003851	0.003955	0.91560	0.91705	0.91533
6	0.004024	0.004000	0.004012	0.91273	0.91360	0.91317
7	0.004316	0.004012	0.004164	0.90787	0.91332	0.91059
8	0.004040	0.003830	0.003944	0.91211	0.91746	0.91479
9	0.003990	0.003991	0.003991	0.91338	0.91507	0.91422
10	0.004014	0.003825	0.003955	0.91257	0.91636	0.91467
Mean	0.004147	0.004013	0.004080	0.910803	0.913549	0.912177
Standard deviation	0.000137	0.000162	0.000141	0.002501	0.003234	0.002680
Coefficient of variation	0.032957	0.040258	0.034557	0.002746	0.003529	0.002938

(C) Situation 3

GP Run No.	Training RMSE	Testing RMSE	Weighted Training and Testing RMSE	Training R^2	Testing R^2	Weighted Training and Testing R^2
1	0.004031	0.003781	0.003906	0.91794	0.92233	0.92013
2	0.004011	0.003675	0.003843	0.91907	0.92536	0.92221
3	0.004130	0.003601	0.003809	0.91544	0.92667	0.92115
4	0.003858	0.003410	0.003634	0.92180	0.92979	0.92580
5	0.004168	0.003547	0.003858	0.91594	0.92682	0.92138
6	0.004065	0.003640	0.003822	0.91798	0.92530	0.92151
7	0.003921	0.003500	0.003710	0.91967	0.92829	0.92398
8	0.004099	0.003518	0.003808	0.91604	0.92763	0.92183
9	0.004178	0.003597	0.003897	0.91643	0.92659	0.92151
10	0.004106	0.003764	0.003925	0.91711	0.92278	0.91995
Mean	0.004060	0.003578	0.003819	0.917718	0.926603	0.922147
Standard deviation	0.000104	0.000118	0.000091	0.001992	0.002321	0.001754
Coefficient of variation	0.025627	0.032924	0.023877	0.002170	0.002504	0.001902
Type 2						
1	0.003188	0.00362	0.003404	0.91181	0.89236	0.90254
2	0.002987	0.003329	0.003158	0.91771	0.90164	0.90968
3	0.002981	0.003355	0.003168	0.91881	0.90026	0.90953
4	0.003137	0.003476	0.003307	0.91392	0.89651	0.90521
5	0.003142	0.003515	0.003329	0.91474	0.89536	0.90505
6	0.002985	0.003408	0.003197	0.91863	0.90065	0.90964
7	0.003048	0.00335	0.003209	0.91639	0.9031	0.90975
8	0.003023	0.003552	0.003287	0.91796	0.89812	0.90804
9	0.003213	0.003201	0.003207	0.91242	0.90623	0.90633
10	0.00334	0.003512	0.003426	0.91021	0.89712	0.90566
Mean	0.002997	0.003411	0.003254	0.915643	0.89988	0.907766
Standard deviation	0.000119	0.000126	0.000096	0.003092	0.003895	0.002830
Coefficient of variation	0.038495	0.036848	0.029580	0.003377	0.004329	0.003118
Type 3						
1	0.004151	0.003904	0.004028	0.90494	0.91121	0.90808
2	0.003786	0.003829	0.003808	0.91333	0.91281	0.91307
3	0.002887	0.003926	0.003811	0.91551	0.91108	0.91330
4	0.004134	0.003882	0.004008	0.90697	0.91239	0.90923
5	0.003708	0.003519	0.003644	0.91426	0.91904	0.91710
6	0.003658	0.003710	0.003684	0.91708	0.91629	0.91668
7	0.003726	0.003805	0.003765	0.91573	0.91482	0.91528
8	0.002877	0.003924	0.003901	0.91520	0.91118	0.91319
9	0.004039	0.003829	0.003958	0.90646	0.91336	0.91141
10	0.003952	0.003574	0.003763	0.91612	0.91957	0.91559
Mean	0.003847	0.003784	0.003816	0.912584	0.914604	0.913761
Standard deviation	0.000182	0.000148	0.000132	0.004421	0.003337	0.003047
Coefficient of variation	0.047430	0.039048	0.034646	0.004844	0.003648	0.003333

Table 2. The Selected Best GP Models.

Stock #	Situation	Type	Training RMSE	Testing RMSE	Weighted Training and Testing RMSE	Training R^2	Testing R^2	Weighted Training and Testing R^2
2303	1	N/A	0.004466	0.004435	0.004461	0.90706	0.90747	0.90726
	2	N/A	0.004049	0.003839	0.003944	0.91211	0.91746	0.91479
	3	1	0.003858	0.003410	0.003634	0.92180	0.92979	0.92580
2330	1	N/A	0.003213	0.003201	0.003207	0.91242	0.90623	0.90633
	2	2	0.003786	0.003519	0.003644	0.91426	0.91994	0.91710
	3	1	0.003448	0.003729	0.003684	0.90697	0.91239	0.90923
2337	1	N/A	0.000490	0.000466	0.000478	0.92517	0.92556	0.92537
	2	N/A	0.001400	0.001361	0.001391	0.90169	0.91150	0.91139
	3	1	0.000740	0.000544	0.000642	0.98776	0.91447	0.98662
2344	1	N/A	0.000642	0.000611	0.000627	0.99205	0.99211	0.99208
	2	2	0.000772	0.000657	0.000715	0.98731	0.98881	0.98816
	3	1	0.001288	0.000750	0.000744	0.99577	0.90944	0.99251
6182	1	N/A	0.001669	0.001660	0.001664	0.98299	0.98040	0.98135
	2	N/A	0.001060	0.001164	0.001112	0.98721	0.98583	0.98652
	3	1	0.000932	0.001073	0.001002	0.99069	0.98921	0.98995
	1	2	0.002421	0.002142	0.002282	0.98121	0.98170	0.98145
	2	3	0.001152	0.001163	0.001138	0.98561	0.98567	0.98564

F. Comparing the forecasting performance

To realize the institutional investors' net buy/net sell on forecasting stock prices, the forecasting performance for the

selected GP models in Table 3 is appraised via MAPE (mean absolute percentage error). Table 3 also reveals the forecasting performance by applying the selected GP models to the never-met validation data while constructing and choosing the optimal models in order to examine the GP models' generalizability. The numbers in bold represent the best result, i.e. the highest R^2 or lowest MAPE, among the three situations for a certain stock. Based on Table 3, we can find that the three GP forecasting models constructed in the third situation for each stock can jointly provide the optimal forecasting for the stock prices in general, except that the GP model built in the second situation yields the smallest testing MAPE. This implies that the information regarding the institutional investors' net buy/net sell is helpful to forecast stock prices. Furthermore, the classification for the trends of net buy/net sell by the institutional investors is also beneficial to improve the forecasting accuracy further. In conclusion, the gathering of institutional investors' net buy/net sell can really assist investors in forecasting future stock prices in addition to collecting common technical indicators.

Table 3. Forecasting Performance.

Stock #	Situation	Training R^2	Testing R^2	Validation R^2	Training MAPE	Testing MAPE	Validation MAPE
2203	1	0.90706	0.90747	0.92250	0.025657	0.025465	0.027656
	2	0.91211	0.91746	0.95873	0.023123	0.022880	0.031976
	3	0.92088	0.92150	0.95002	0.022122	0.02077	0.021480
2330	1	0.99560	0.99290	0.84177	0.020967	0.021897	0.029619
	2	0.99595	0.99511	0.92424	0.022525	0.020252	0.026072
	3	0.99407	0.99542	0.94611	0.019202	0.019157	0.019156
2337	1	0.98108	0.98150	0.87923	0.009902	0.009959	0.036586
	2	0.98752	0.98943	0.92075	0.007811	0.007509	0.025909
	3	0.98852	0.99025	0.94435	0.007660	0.007273	0.021687
2344	1	0.96957	0.96944	0.81570	0.010749	0.010821	0.046546
	2	0.97257	0.97135	0.87827	0.010264	0.010409	0.035349
	3	0.97675	0.97170	0.94157	0.009547	0.009342	0.023642
6182	1	0.98289	0.98040	0.90135	0.067000	0.068830	0.091938
	2	0.98721	0.98582	0.96717	0.065276	0.063301	0.091338
	3	0.98840	0.98804	0.97356	0.063385	0.065347	0.063076

V. CONCLUSIONS

Stock investing is easy to understand for an investor among various investment targets. The ordinary operation for investing in a stock is to buy a stock at a relatively low price and sell it at a high price later. In addition, first selling a stock at a high price and buying back the stock at a low price later, called selling short, is also a possible method. For either, it's an important and critical issue to forecast future stock prices, thus assisting in making proper investment decisions. However, forecasting is a complex and difficult task in itself, with the behavior of a stock's prices being intangible, since the factors that affect the market prices regarding a stock are multitudinous. Traditionally, an investor utilizes fundamental analysis of a business's financial conditions and other influential factors to map the future long-term financial performance of the issued stock. The technical analysis, applying the technical indicators, is more suitable for short-term forecasting. In addition, an investor pays much closer attention to the institutional investors' net buy/net sell of each day in Taiwan, as its value indicates the degree of optimism about the overall performance of a corporation, which assists in judging the future trend of stock prices. Therefore, the present empirical study in Taiwan of the effects of institutional investors' net buy/net sell on forecasting stock prices uses genetic programming (GP), where the technical indicators serve as predictors. An examination procedure is proposed and five stocks with the highest trading volumes in

the semiconductor section of the TAIEX are used to illustrate the proposed procedure. The implementation results show that the stock price modelling procedure using the GP method can be considered a steady and practical technique, with low coefficients of variation in the execution results. In addition, the forecasting performance can indeed be improved by gathering information on institutional investors' net buy/net sell of stock. Classification of the trends of net buy/net sell by the institutional investors can further help an investor to raise the forecasting accuracy. In conclusion, the trading volumes and classification of institutional investors' net buy/net sell in Taiwan have significant influence on the behavior and forecasting of stock prices according to this empirical study. Furthermore, genetic programming is shown to be a robust and effective tool for constructing forecasting models for an investor. Researchers can further apply clustering techniques to segment institutional investors' net buy/net sell trends, which may yield superior forecasting performance.

ACKNOWLEDGMENT

The author would like to thank the Minghsin University of Science and Technology, Taiwan, R.O.C. for supporting this research under Contract No. MUST-108BA-1.

REFERENCES

- [1] Behravesh, M., "Forecasting stock price of Iranian major petrochemical companies," *African Journal of Business Management*, Vol. 5, No. 1, 2011, pp. 7-12.
- [2] Yeh, C. Y., Huang, C. W., and Lee, S. J., "A multiple-kernel support vector regression approach for stock market price forecasting," *Expert Systems with Applications*, 2011, Vol. 38, No. 3, pp. 2177-2186.
- [3] Tay, F. E. H., and Cao, L., "Application of support vector machines in financial time series forecasting," *Omega: The International Journal of Management Science*, 2001, Vol. 29, No. 4, pp. 309-317.
- [4] Box, G. E. P., and Jenkins, G. M., *Time Series Analysis: Forecasting and Control*, 5th ed., New Jersey: John Wiley & Sons, 2016.
- [5] Chang, P.-C., and Liu, C.-H., "A TSK type fuzzy rule based system for stock price prediction," *Expert System with Applications*, 2008, Vol. 34, No. 1, pp. 135-144.
- [6] Zuo, Y., and Kita, E., "Stock price forecast using Bayesian network," *Expert Systems with Applications*, 2012, Vol. 39, No. 8, pp. 6279-6737.
- [7] Hsu, C.-M., "A hybrid procedure with feature selection for resolving stock/futures price forecasting problems prediction," *Expert Systems with Application*, 2013, Vol. 22, No. 3-4, pp. 651-671.
- [8] Liu, J.-N.-K., and Hu, Y.-X., "Application of feature-weighted Support Vector regression using grey correlation degree to stock price forecasting," *Neural Computing and Applications*, 2013, Vol. 22, pp. S143-S152.
- [9] Xiong, T., Bao, Y.-K., and Hu, Z.-Y., "Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting," *Knowledge-based Systems*, 2014, Vol. 55, pp. 87-100.
- [10] Xiao, Y., Xiao, J., Lu, F.-B., and Wang, S.-Y., "Ensemble ANNs-PSO-GA approach for day-ahead stock e-exchange prices forecasting," *International Journal of Computational Intelligence Systems*, 2014, Vol. 7, No. 2, pp. 272-290.
- [11] Dan, J.-P., Guo, W.-B., Shi, W.-R., Fang, B., and Zhang, T.-P., "Deterministic echo state networks based stock price forecasting," *Abstract and Applied Analysis*, 2014, Document No. 137148.
- [12] Singh, P., and Borah, B., "Forecasting stock index price based on M-factors fuzzy time series and particle swarm optimization," *International Journal of Approximating Reasoning*, 2014, Vol. 55, No. 3, pp. 812-833.
- [13] Rounghi, M. M., Abbaszadeh, M. R., and Arashi, M., "Stock price forecasting for companies listed on Tehran stock exchange using multivariate adaptive regression splines model and semi-parametric splines technique," *Physics A- Statistical Mechanics and its Applications*, 2015, Vol. 438, 625-633.

An Empirical Study of Institutional Investors' Net Buy/Net Sell on Forecasting Stock Prices Using Genetic Programming in Taiwan

- [14] Guo, Y.-H., Han, S.-M., Shen, C.-H., and Li, Y., "An adaptive SVR for high-frequency stock price forecasting," *IEEE Access*, 2017, Vol. 6, pp. 11397-11404.
- [15] Wang, W.-N., "A big data framework for stock price forecasting using fuzzy time series," *Multimedia Tools and Applications*, 2018, Vol. 77, No. 8, pp. 10123-10134.
- [16] Chou, J.-S., and Nguyen, T.-K., "Forward forecast of stock price using sliding-window metaheuristic-optimized machine-learning regression," *IEEE Transactions on Industrial Informatics*, 2018, Vol. 14, No. 7, pp. 3132-3142.
- [17] Cheng, C.-H., and Yang, J.-H., "Fuzzy time-series model based on rough set rule induction for forecasting stock price," *Neurocomputing*, 2018, Vol. 302, pp. 33-45.
- [18] Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., and Lanza, G., *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*, New York: Springer, 2005.
- [19] Ciglaric, I., and Kidric, A., "Computer-aided derivation of the optimal mathematical models to study gear-pair dynamic by using genetic programming," *Structural and Multidisciplinary Optimization*, 2006, Vol. 52, No. 2, pp. 153-160.
- [20] Koza, J. R., Streeter, M. J., and Keane, M. A., Routine high-return human-competitive automated problem-solving by means of genetic programming," *Information Sciences*, 2008, Vol. 178, No. 23, pp. 4434-4452.
- [21] Kazemnia, A., Kaedi, M., and Ganji, B., "Personality-based personalization of online store features using genetic programming: analysis and experiment," *Journal of Theoretical and Applied Electronic Commerce Research*, 2019, Vol. 14, No. 1, pp. 16-29.
- [22] Saghafi, H., and Arabloo, M., "Development of genetic programming (GP) models for gas condensate compressibility factor determination below dew point pressure," *Journal of Petroleum Science and Engineering*, 2018, Vol. 171, pp. 890-904.
- [23] Pathi, A., and Sadeghi, R., "A genetic programming method for feature mapping to improve prediction of HIV-1 protease cleavage site," *Applied Soft Computing*, 2018, Vol. 72, pp. 56-64.
- [24] Shen, J., and Jiménez, R., "Predicting the shear strength parameters of sandstone using genetic programming," *Bulletin of Engineering Geology and the Environment*, 2018, Vol. 77, No. 4, pp. 1647-1662.
- [25] Kim, K.-J. and Han, I., "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index," *Expert System with Applications*, 2000, Vol. 19, No. 2, pp. 125-132.
- [26] Kim, K.-J., and Lee, W. B., "Stock market prediction using artificial neural networks with optimal feature transformation," *Neural Computing and Applications*, 2004, Vol. 1, No. 3, pp. 255-260.
- [27] Tsang, P. M., Kwok, P., Choy, S. O., Kwan, R., Ng, S. C., Mak, J., Tsang, J., Koong, K., and Wong, T.-L., "Design and implementation of NN5 for Hong Kong stock price forecasting," *Engineering Applications of Artificial Intelligence*, 2007, Vol. 20, No. 4, pp. 453-461.
- [28] Chang, P.-C., and Liu, C.-H., "A TSK type fuzzy rule based system for stock price prediction," *Expert System with Applications*, 2008, Vol. 34, No. 1, pp. 135-144.
- [29] Ince, H. and Trafalis, T. B., "Short term forecasting with support vector machines and application to stock price prediction," *International Journal of General System*, 2008, Vol. 37, No. 6, pp. 677-687.
- [30] Huang, Z. W., Li, M. Z., Chousidis, C., Mousavi, A., and Jiang, C. J., "Schema theory-based data engineering in gene expression programming for big data analytics," *IEEE Transactions on Evolutionary Computation*, 2018, Vol. 2, No. 5, pp. 792-804.
- [31] Lai, R. K., Fan, C.-Y., Huang, W.-H., and Chang, P.-C., "Evolving and clustering fuzzy decision tree for financial time series data forecasting," *Expert System with Applications*, 2009, Vol. 36, No. 2, pp. 3761-3773.



Chih-Ming Hsu Chih-Ming Hsu is currently a Professor in the Department of Business Administration at Minghsin University of Science and Technology, Taiwan. He holds a PhD in Industrial Engineering and Management from National Chiao Tung University, Taiwan. His present research interests include quality engineering, optimization methods in industrial applications and data mining applications in CRM.

A Mixture of Expert-Based Prediction Approach by Using Genetic Programming and Clustering: A Case Study on Predicting the On-Time Percentages of Local TRA Trains in Taiwan

Chih-Ming Hsu

Abstract—The accurate prediction of on-time percentages of trains is an important issue since it can significantly affect the management of trains' operation in many fields, such as making an appropriate timetable, organizing trains' waiting, arranging trains' running tracks, settling the required manpower etc. However, the problem of predicting the on-time percentages of trains is very complex and difficult since the on-time percentages of trains can be influenced by numerous factors. To deal with such a problem, genetic programming (GP) and clustering techniques are used to develop a prediction procedure by means of a mixture of experts in this study. The GP method is utilized to construct prediction models, as well as to identify the feature variables, i.e. critical independent variables. The clustering approach is applied to partition the data into several clusters where data in each cluster are as similar as possible. To illustrate the usefulness and effectiveness of this approach, prediction of the on-time percentages of local trains operated by the Taiwan Railway Administration (TRA) in Taiwan are demonstrated as a case study. The execution results show that the GP can construct an adequate model for predicting the on-time percentages of local trains. Furthermore, a comparison also shows that the technique of mixture of experts can yield a superior GP prediction model by evaluating the performance through MSE, R^2 , and MAPE. Hence, we can consider our proposed prediction approach to be a useful and effect procedure for resolving a problem of prediction in the real world.

Index Terms—Mixture of Expert, Genetic Programming, Clustering, On-time Percentage, TRA.

I. INTRODUCTION

Prediction problems exist almost everywhere in our living world, e.g. prediction of stock prices, power consumption, rainfall, etc. For prediction problems involving trains, many methods that apply techniques from different fields have been proposed, and their effectiveness and usefulness have been demonstrated. For example, [1] developed a hybrid model called GAANN that utilizes genetic algorithms (GAs) to optimize the network architecture of artificial neural networks (ANNs) so as to forecast passenger volume in each month on Serbian railways. Hence, the number of neurons in the middle layer of ANNs can be determined by using the selected population in GAs. A time series of the total monthly number of passengers flows gathered from the SORS (the

Statistical Office of the Republic of Serbia) was used to assess the predicting performance of the GAANN. A comparison between the GAANN and the traditional SARIMA (Seasonal Autoregressive Integrated Moving Average) model revealed that their proposed approach can yield better forecasting results. [2] developed a model that uses the Holt-Winters model with consideration of the changes in TSF (train service frequency) for the OD (origin to destination) at different times during an operating day to forecast passenger flow in the short term on high-speed rail. The Holt-Winters model can take advantage of characteristics in the time series of passenger flow. In addition, the changes of TSF for the OD at different times in a day are also considered. The final hybrid model is generated through integration based on the minimum absolute value method. To verify the effectiveness, the operational data of the high-speed railway from Beijing to Shanghai railway during 2012 to 2016 were used to demonstrate their proposed model. Furthermore, their method can be further applied to forecast the effects of the TSF with a definite formation. [3] combined temporal forecasting based on a radial basis function neural network (RBF NN) and spatio forecasting based on spatial correlation degree to develop an approach for forecasting the passenger flow status in a high-speed railway transport hub (HRT). The temporal forecasting based on RBF NN is utilized to forecast passenger flow status in the bottleneck position, as well as combining a spatio forecasting approach based on spatial correlation degree to improve the forecasting precision. The computational experiments on actual passenger flow status of a specific bottleneck position and its correlation points in the Chinese HRT revealed the effectiveness of their proposed approach in forecasting the passenger flow status with high precision. [4] developed several railway level crossing (LX) accident prediction models that can highlight the main parameters' influences. Based on LX accidents, they utilized the ordinary least-squares (OLS) and nonlinear least-squares (NLS) methods to estimate the respective coefficients for variables in the prediction models. The dedicated accident database provided by SNCF (National Society of French Railway Networks) Réseau was used as a case study to validate the performance of their proposed model, and a comparison process was made through evaluation by statistical means for examining how well their models' estimations can fit the reality. The experimental and compared results prove that their proposed improved accident prediction model can produce statistic-based approbatory quality, while the

Manuscript received July 23, 2019
Chih-Ming Hsu, Department of Business Administration, Minghsin University of Science and Technology, Hsinchu, Republic of China

A Mixture of Expert-Based Prediction Approach by Using Genetic Programming and Clustering: A Case Study on Predicting the On-Time Percentages of Local TRA Trains in Taiwan

combination of the proposed model with the negative binomial distribution can yield relatively high accuracy of prediction for the probability of accident occurrence. [5] aimed to develop several primary delay recovery (PDR) predictor models to predict recoverable train delay accurately. They first utilized operation records from the Wuhan-Guangzhou high-speed railway (HSR) to identify the main variables, as well as individual sections' influence, that affect the delay of a train by developing a general framework that can be applied to any HSR line. Random forest regression (RFR), multiple linear regression (MLR), support vector machine (SVM), and artificial neural networks (ANN) are then applied to predict the PDR. The validation results on test datasets showed that the RFR prediction model can outperform the other three alternative models while measuring the prediction accuracy. In addition, the RFR model can achieve a prediction accuracy of more than 80%, while the prediction tolerance is less than one minute based on their evaluation results. [6] applied the neural network and origin-destination (OD) matrix estimation to propose a divide-and-conquer method for forecasting the short-term passenger flow in a high-speed railway system. They first collected the numbers of arriving and departing passengers at each station to form the OD matrices. The neural network was then used to comprehend the short-term forecasting for the arriving/departing passengers' flow. Finally, an OD matrix estimation method was utilized to obtain the OD matrices for a short-term timeframe. Verification through a case study on a high-speed railway with fifteen stations in China showed that the proposed divide-and-conquer method can indeed perform adequately in forecasting the short-term passenger flow on a high-speed railway. [7] applied fuzzy logic relationship recognition techniques to build a fuzzy-temporal-logic-based passenger flow forecast model (FTLPFFM) where the past sequences of passenger flow are considered by using fuzzy logic relationship recognition techniques in the searching process to predict the short-term passenger flow for a high-speed railway. The implementation results using real-world data indicated that the FTLPFFM model can significantly improve the forecast accuracy in terms of measuring with MAE, MAPE, and RMSE compared to the ARIMA (autoregressive integrated moving average model) and KNN (k-nearest neighbor) models. [8] utilized a timed event graph with dynamic arc weights to develop a microscopic model to predict train event times. Through using processed historical track occupation data to reflect all phenomena of railway traffic captured by train describer systems and preprocessing tools, the process times in the model can be dynamically obtained. Next, the characteristics regarding the graph structure of the model allow fast algorithms to be applied to estimate the event times even for large networks. Incorporation of predicted route conflicts on train running times due to braking and reacceleration can further increase the prediction accuracy. Furthermore, the expected prediction error is continuously minimized adaptively by detecting the train runs with process times that successively deviate from their estimates in a certain pattern, as well as the downstream process times. The train describer log files on the busy corridor of Leiden-The Hague-Rotterdam-Dordrecht in the Netherlands were used to test and validate their proposed tool, and adequate

experimental results were obtained. However, various errors in logging of event times can still occur even though detailed data with high quality have been used. [9] proposed an approach based on fuzzy rules and time series analysis for predicting online failure in railway transportation systems. They model the relationships among different variables while applying univariate time series analysis to describe the evolution of each variable. Two predicted values can be obtained for a dependent variable, where one variable is produced from the time series model, and the other one is computed from the fuzzy rules by implementing fuzzy inference. A failure in some time period ahead can be declared when the difference between the two values has exceeded a pre-specified threshold. The authors claim that their proposed method differs from the existing methods since it not only considers the evolutionary trend of each variable but also reflects the relationships among different variables. In addition, no prior knowledge of the system model or failure patterns are required. An ATP railway transportation system was used to illustrate their proposed method, and experimental results also proved that the proposed method can predict online system failures effectively. [10] presented a neural network model aiming at predicting the delay of passenger trains of Iranian Railways. They utilized three methods, including normalized real number, binary coding, and binary set encoding, to define inputs. In addition, three different strategies, called quick method, dynamic method, and multiple method, are investigated to find an appropriate architecture for a neural network to a specific task. The data regarding passenger train delays are also divided into three parts, training, validation, and testing sets, to prevent the occurrence of overfitting while modeling a neural network by making cross validation. The execution results by implementing three different data input methods and three different architectures are compared with each other, as well as with well-known prediction methods including the decision tree and multinomial logistic regression. A time-accuracy graph is plotted to make a fair comparison among all models, and the results indicate that their proposed model indeed can yield higher prediction accuracy. [11] exploited big data technologies, learning algorithms, and statistical tools to build a data-driven train delay prediction system (TDPS) for railway networks with large scale. Their intention was to exploit the recent in-memory large-scale data processing technologies for predicting train delays by developing a fast learning algorithm for shallow and deep extreme learning machines. It was demonstrated that their proposed method can perform up to twice as well as the current state-of-the-art methods through a case study and comparison on the train movement data provided by RFI (Rete Ferroviaria Italiana) in the real world. The authors also present methods for tuning the hyperparameters of the learning algorithms efficiently and effectively. In addition, robust models with high performance with respect to the actual train delay prediction system of RFI are derived by deeply exploiting the historical data on train delays. [12] proposed a method of calculating spatial correlation degree between a key area and a correlated surveillance area, and developed an algorithm to forecast passenger flow risk according to the spatial correlation. The effectiveness of their proposed approach was verified

through implementing computational experiments on a specific key area in a high-speed railway transport hub, and adequate results were obtained. [13] proposed a booking model by using a framework of case-based reasoning based on reservation data. There are four modules containing distinctive functions for similarity evaluation, instance selection, arrival projection, and parameter search included in his proposed method. The forecasting capability was validated by testing the proposed model on fourteen collected data series and comparing the out-of-sample accuracy based on the four traditional benchmarks. According to the empirical results, his proposed self-learning model could reduce at least 11% of mean square errors (MSE) on average, and the MSE can be significantly reduced through the learning scheme in his proposed approach in comparison with the other naive versions' prediction performance.

As the above literature review shows, problems regarding the prediction of the on-time percentages of trains has not been investigated previously. However, predicting the on-time percentages of trains is a critical issue for the operation of a railway corporation since the on-time percentages of trains have considerable effects on making an appropriate timetable, organizing the stations for waiting trains, organizing the running tracks of trains, arranging the sufficient manpower etc. However, various factors can influence the on-time percentages of trains, e.g. the total number of passengers, passengers' types, operating days of trains, the initiation and terminal stations or time, the running distances for trains etc., and it is not easy to identify and simultaneously consider all of them completely. In addition, it is also a difficult task to gather the operation data of trains. The prediction for the on-time percentages of trains is therefore considered a very complicated and thorny problem, and has been the subject of far fewer studies. Next, the mixture of experts is a practical technique for resolving a single problem by partitioning the original problem into several sub-problems with smaller sizes, and tackling each sub-problem to create an expert that is dedicated to solving its corresponding specialized sub-problem. Therefore, this study develops a prediction approach by using a technique of a mixture of experts based on genetic programming (GP) and clustering analysis. The remaining sections are organized as follows. The analyzing and modelling methodologies are briefly introduced in Section 2. Then, Section 3 presents our proposed prediction approach. Section 4 gives a real case study on predicting the on-time percentages of local trains operated by the Taiwan Railway Administration in Taiwan. Finally, conclusions are provided in Section 5.

II. ANALYZING AND MODELLING METHODOLOGIES

There are two main analyzing and modelling methodologies including the two-step clustering and genetic programming in our proposed prediction approach. This section briefly interprets these two research methods by introducing two-step clustering first.

A. Two-step Cluster Analysis

Cluster analysis is a method which groups a set of data such that the data in the same group, i.e. cluster, are more similar to each other than to other groups' data. Cluster analysis can be achieved by various approaches. TwoStep

cluster analysis is a famous one among these methods. The TwoStep cluster analysis procedure utilizes a likelihood distance measure where the variables in the cluster model are assumed to be independent. Furthermore, the distribution for each continuous variable is assumed to be normal, and a multinomial distribution for each categorical variable is assumed. The TwoStep cluster analysis differs from the traditional clustering techniques in that it has several desirable features including: (1) It can create clusters based either on categorical or on continuous variables; (2) It can select the optimal number of clusters automatically; (3) It can analyze large scales of data files efficiently. The procedure of TwoStep cluster analysis is summarized as follows:

Step 1. Construct a cluster features (CF) tree

The cluster features (CF) tree places the first case at the root of the tree in a leaf node where the variable information of that case is contained. Then, each successive case is added to an existing node or it can form a new node according to its similarity to the existing nodes, as well as based on the similarity criterion that uses the distance measure. Therefore, a node can summarize the variable information about the cases clustered in the node. The CF tree hence can give capsule summary information for the data file.

Step 2. Group the leaf nodes

An agglomerative clustering algorithm is applied to group the leaf nodes of the CF tree to produce a range of solutions. Then, Schwarz's Bayesian Criterion (BIC) [14] or the Akaike Information Criterion (AIC) [15] can be used as the clustering criterion to determine the optimal number of clusters.

B. Genetic Programming

Through observing the evolution progress of organisms in the natural world, C. R. Darwin put forward his famous theory of natural selection and evolution. [16] presented the well-known optimization method, called genetic algorithms (GAs), based upon inspiration by Darwin's evolution theory, to solve an optimization problem by imitating the evolutionary procedure of living beings. In GAs, a feasible solution (individual) for an optimization problem is represented by using a series of genes, called a chromosome, that mimics the chromosome of a living thing. All individuals form a population. The excellence of each feasible solution in the population is evaluated by the fitness, assessed according to the fitness function designed based on the objective function in an optimization problem. The mechanism of natural selection and matching is designed to simulate the marriage of individuals to form a matching pool. Then, the paired individuals, called parents, in the matching pool can hopefully produce superior new individuals, called offspring, by using well-designed crossover functions which closely relate to the fitness corresponding to a feasible solution in an optimization problem. In addition, a mutation function is also designed to represent the unusual situation of crossover, i.e. extraordinary genetic changes. Finally, the individuals of offspring are assessed by the fitness function, and the better individuals among the offspring replace the worse (weak)

A Mixture of Expert-Based Prediction Approach by Using Genetic Programming and Clustering: A Case Study on Predicting the On-Time Percentages of Local TRA Trains in Taiwan

individuals in the previous generation, i.e. their parents' generation, thus forming a new population in the next generation. Later, [17] developed genetic programming (GP), which extends the GAs into a field of computer programs by expressing a feasible solution (program) through a tree-based structure as shown in Figure 1. The tree in Figure 1 represents a computer program; we can decode the tree from left to right, as well as from bottom to top, as follows:

$$3 - \frac{x}{15} + 8 \times \sqrt{y}$$

The tree in GP is made up of elements from two parts including the terminal and function sets. The terminal set defines the elements that are available for each terminal branch of the GP program (chromosome). It can be the independent variables, zero-argument functions, random constants, etc. For example, the 3, x, 15, 8, and y are the elements in the terminal set. The function set is a set of primitive functions available to each branch of the tree-structure program, such as addition, square root, multiplication, sine and others. The +, -, ×, ÷, and √ are the elements from the function set. The fitness corresponding to the above equation can be evaluated by feeding variables x and y into Equation (1), as well as referring the objective function that it is intended to optimize. Next, the crossover and mutation operators in GAs can also be transformed to the styles that can fit the tree-based GP as illustrated in Figures 2 and 3, respectively. In Figure 2, the original paired solutions include:

$$3 - \frac{x}{15} + 8 \times \sqrt{y}$$

$$4 + \cos(x) - \frac{\log(y)}{6z}$$

The new paired solutions will be:

$$3 - \cos(x) + 8 \times \sqrt{y}$$

$$4 + \frac{x}{15} \cos(x) - \frac{\log(y)}{6z} \quad (1)$$

Similarly, the original tree, i.e. $3 - \frac{x}{15} + 8 \times \sqrt{y}$, in Figure 3 mutates into a new program, i.e. $3 - \frac{x}{15} + 8 \times (5 + y)$.

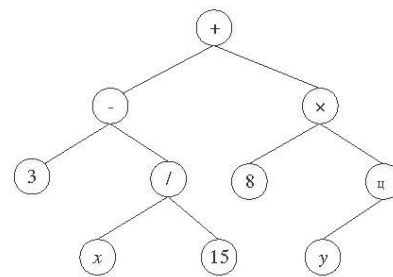


Figure 1. Tree-based genetic programming

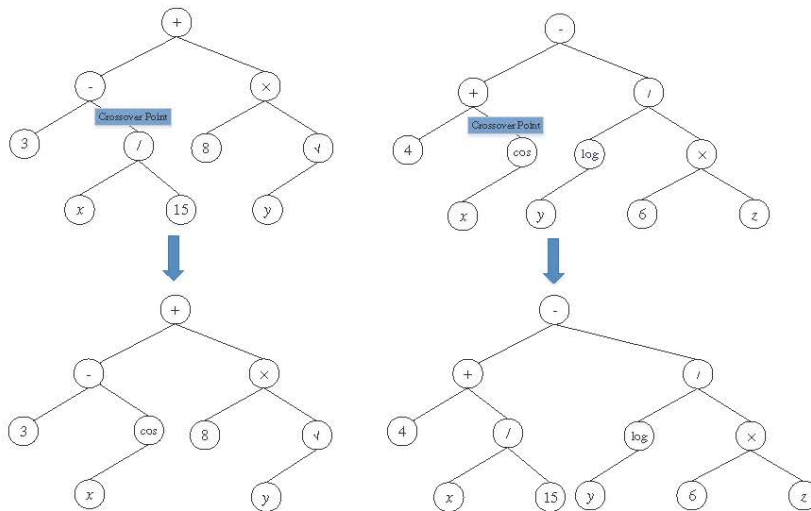


Figure 2. Crossover in the tree-based GP

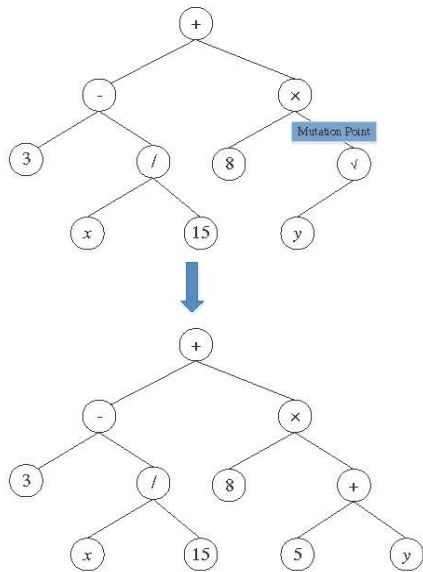


Figure 3. Mutation in the tree-based GP

On the basis of the above definitions, the simple GP can be depicted by Figure 4 and stated by the following procedure [18-20]:

1. Initialization

The control parameters in GP, e.g. population size, maximum size of programs, crossover rate, mutation rate etc. are first specified. Then, a population of initial solutions (programs or individuals) is generated based on some well-designed mechanism or randomly. In general, the programs should be generated with the condition of a pre-specified maximum size, and the individuals can have different sizes and shapes.

2. Evaluation

Each program in the population is executed and its fitness (adaptability) in the population is explicitly or implicitly measured in terms of how well it can solve the optimization problem through a pre-defined fitness function. The ways for evaluation can be various, e.g. the amount of error between its output and target, the total cost/time for bringing the system to a desired state, or the classification accuracy. The evaluated result is called the fitness.

3. Create the next generation

Individuals are selected from the population with a probability that is based on fitness. Then, the genetic operators are applied to the selected individuals (programs), including:

- (1) Reproduction: Copy the selected program to create a new individual.
- (2) Crossover: Randomly recombine chosen parts of paired selected programs (called parents) to form two new individuals (called children) in the offspring.
- (3) Mutation: Randomly mutate a randomly chosen part of the selected program to generate a new offspring individual.
- (4) Architecture-altering: Alter the architecture of the selected program to produce a new offspring individual.

The individuals in the current population (the now-old generation) are replaced by the individuals in the offspring population based on a certain strategy, e.g. elitist strategy, to create individuals in the new population (the next generation).

4. Check the termination criterion

The single best program ever encountered during the run (i.e. the best-so-far individual) is designated as the final result when the termination criterion is satisfied. Steps 2 to 4 will run iteratively if the termination criterion cannot be fulfilled.

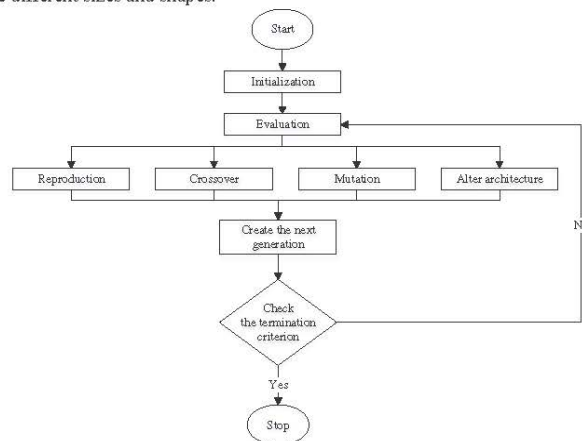


Figure 4. Simple GP flowchart

A Mixture of Expert-Based Prediction Approach by Using Genetic Programming and Clustering: A Case Study on Predicting the On-Time Percentages of Local TRA Trains in Taiwan

III. PROPOSED APPROACH

This study proposes an approach for predicting the on-time ratios of trains by using the technique of mixture of experts as shown in Figure 5, illustrated as follows:

Step 1: Data preparation

The required data are first collected. To avoid having a variable with a wide range dominate a variable of relatively narrow range, the data in each variable are identically normalized into the range $(-1, 1)$ according to the maximum and minimum values of the corresponding variables. Next, the normalized data are divided into training and test data groups based on a pre-determined ratio, e.g. 3:1. The training data are used to construct a GP prediction model, and the test data are applied to evaluate the generalizability of the well-trained GP model.

Step 2: Construct preliminary GP models

The GP technique is utilized to construct preliminary GP models for predicting the on-time percentages of trains. The prediction performance is measured by the root-mean-square error (RMSE), R squared (R^2), and mean absolute percentage error (MAPE). The GP tool must be implemented several times to obtain multiple prediction models.

Step 3: Select features from preliminary GP models

This step selects features, i.e. important variables that have great impacts on the on-time ratios of trains, according to the preliminary GP models built in the previous step. Each preliminary GP model constructed in Step 2 is analyzed to identify which variables appeared very frequently in its building progress, called appearing variables in this study. Therefore, the featured variables that significantly influence the on-time ratios of trains can be identified through synthesizing the information of appearing variables of each GP model built in Step 2.

Step 4: Divide data into sub-problems

The TwoStep cluster analysis is used where the featured variables identified in Step 3 act as the bases for clustering. The optimal number of clusters is determined according to Schwarz's Bayesian Criterion (BIC) (Burnham and Anderson, 2002) or the Akaike Information Criterion (AIC) (Akaike, 1969) criterion. Hence, the original data prepared in Step 1 can be segmented into several sub groups. The data of each sub-group form a new prediction problem.

Step 5: Construct GP models for each sub-problem

As in Step 1, the data in each sub-group clustered in Step 4 are partitioned into training and test data groups according to a pre-specified ratio. Then, the GP method is applied to the training data to establish several GP prediction models, and the model with the optimal prediction performance, measured by RMSE, R^2 , and MAPE, is designated to be the final GP prediction model. In addition, the generalizability of the final GP model to unknown data is evaluated by predicting the on-time ratios of trains in the corresponding test data group for each sub-problem.

Step 6: Mix GP models of each sub-problem into a final model

In Step 5, the final GP model of each sub-problem can be viewed as an expert for predicting on-time percentages of trains for each corresponding sub-problem. Hence, each final GP model selected in each sub-problem is then combined to make an integrated GP prediction model that can predict on-time percentages of trains for all sub-problems, i.e. all of the original data.

Step 7: Evaluate prediction performance

The original GP model built in Step 2 and the integrated GP model combined in Step 6 are compared to contrast their prediction performance, as well as to confirm the superiority of the integrated GP. The criteria for evaluating prediction performance can include MSE, R^2 , MAPE etc.

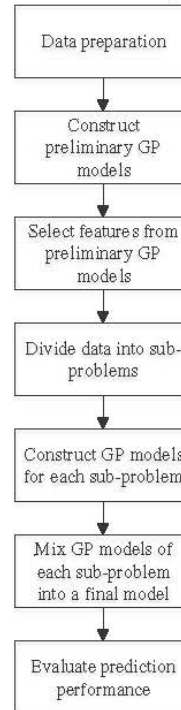


Figure 5. Proposed prediction approach

IV. CASE STUDY

In this section, a case study on prediction of on-time ratios of trains is presented to demonstrate the usage of our proposed approach.

A. Introduction to TRA

Taiwan Railway Administration (TRA) is a traditional railway transportation company owned and operated by the government of the Republic of China (ROC). The first railway was constructed in 1891 with 28.6km of track. Now, the total length of railways operated by TRA has reached

1065km, and contains twelve lines as shown in Table 1 and Figure 6. Since the population concentrates mainly in the western areas, especially in the northwest areas, the traffic of railways between Zhunan and Keelung is considered the busiest. In addition, there are several types of passenger trains, e.g. Tzu-Chiang Limited Express, Puyuma Express, Taroko Express, Chu-Kuang Express, Fu-Hsing Semi Express, Fast Local, and Local Trains operated simultaneously on the same tracks. Among these trains, the Local Trains must often wait for an Express or Fast Local Train to pass, thus the on-time percentages of Local Trains are influenced by many factors and are not easily predicted. Therefore, this study focuses on predicting the on-time percentages of Local Trains which operate covering the region between Zhunan and Keelung stations.

Table 1. TRA operation lines

No.	Lines
1	Chengzhui Line
2	Eastern Trunk Line
3	Jiji Line
4	Liujia Line
5	Neiwan Line
6	Pingxi Line
7	Shenao Line
8	Shalun Line
9	South Link Line
10	Suaoxin-Suao Line
11	Western Trunk Line
12	Western Trunk Line (Coast Line)

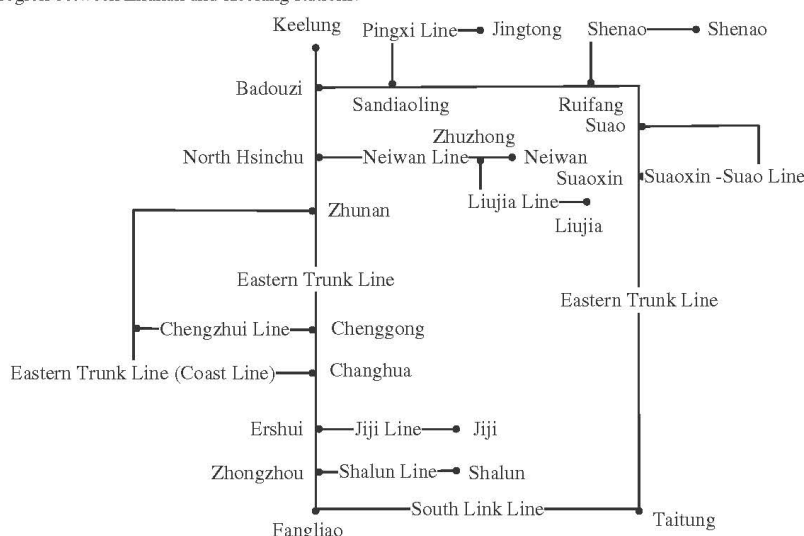


Figure 6. Map of TRA operation lines

B. Data Preparation

First, several variables that are thought to be possible factors, i.e. independent variables, that might affect the on-time percentages of Local Trains are first determined, as summarized in Table 2 (also refer to Figure 6). The response, i.e. dependent variable, is the on-time percentage of each Local Train. The on-time percentage of a Local Train is defined as the total number of trains that can arrive at the terminal station on time divided by the total number of trains that arrive at the terminal station during an operating period, usually one month. Furthermore, a train is deemed to be on time if the train can arrive at its terminal station within five minutes of its scheduled time. For each Local Train, the thirty-three prediction variables, i.e. independent variables, as shown in Table 2 along with its corresponding on-time percentage, i.e. dependent variable, are arranged into a row. Based on the operation data collected by TRA on March 2019, this study gathered the data including 278 rows. Among these independent variables, the ranges of some

predictors are large relative to certain other ones. For example, the minimum and maximum values for the “Initial time” (independent variable #5) are 100 and 1400, respectively. However, the “Train-kilometer” (independent variable #6) only ranges from 2.1 to 130.9. In order to avoid the effects of the large-range variables dominating the effects of variables with relatively small ranges on the dependent variable, all variables (including the independent and dependent variables) in the collected data with 278 rows are first normalized linearly to a range from -1 to 1 based on their corresponding maximum and minimum values. The normalized data are then randomly partitioned into the training and test data sets with 209 and 69 rows, respectively, based on a ratio of 3:1.

A Mixture of Expert-Based Prediction Approach by Using Genetic Programming and Clustering: A Case Study on Predicting the On-Time Percentages of Local TRA Trains in Taiwan

Table 2. Possible factors that might affect the on-time percentages of Local Trains

No.	Factors (Independent variables)	Explanation
1	Northbound/Southbound	The train is northbound/southbound.
2	Operation days	Train runs daily/on Sundays/daily except Saturdays/daily except Saturdays.
3	Operation lines	Train runs via Eastern Trunk line/Eastern Trunk Line (Coast Line)/Cross-Line (Mixed Lines).
4	Initial kilometer marker	Kilometer marker ("kmarker") of station from which train initiates. Zhunan station kmarker is set at 0. Kmarker of each station is calculated based on distance from Zhunan station.
5	Initial time	Time when train sets out.
6	Train-kilometers	Distance of train's trip.
7	Traveling time	Duration of train's trip.
8	Train-kilometers before train enters region	Northern point at which Western trunk and Western trunk line (coast line) merge is Zhunang station. Eastern trunk line starts at Badouzi station. Trains not setting out from points between Zhunang and Keelung stations have run for a distance before entering region between Zhunang and Badouzi stations. This distance is defined as "Train-kilometers before train enters region".
9	Travel time before train enters region	Similar to explanation in previous factor. Trains not initiating between Zhunang and Keelung stations have run for a duration before entering region between Zhunang and Badouzi stations. This duration is defined as "Travel time before train enters region".
10	Number of waiting times before train enters the region	Similar to situation explained in factor 8. Trains not initiating between Zhunang and Keelung stations must wait for Express/Fast Local Train several times before it enters region between Zhunang and Badouzi stations. This number of waiting times is defined as "Number of waiting times before train enters region".
11	Waiting time before train initiates	Similar to situation explained in factor 8. A train not initiating between Zhunang and Keelung stations must wait for Express/Fast Local Train for a duration before it enters region between Zhunang and Badouzi stations. This waiting duration is defined as "Waiting time before train initiates".
12	Train-kilometers after train leaves region	Similar to situation explained in factor 8. A train not terminating between Zhunang and Keelung stations must run for a distance after leaving region between Zhunang and Badouzi stations. This distance is defined as "Train-kilometers after train leaves region".
13	Traveling time after train leaves region	Similar to situation explained in factor 8. A train not terminating between Zhunang and Keelung stations must run for a duration after it leaves region between Zhunang and Badouzi stations. This duration is defined as "Traveling time after train leaves region".
14	Number of waiting times after train leaves region	Similar to situation explained in factor 8. A train that does not terminate between Zhunang and Keelung stations must wait for Express/Fast Local Train several times after it leaves region between Zhunang and Badouzi stations. This number of waiting times is defined as "Number of waiting times after train leaves region".
15	Waiting time after train leaves region	Similar to situation explained in factor 8. A train that does not initiate between Zhunang and Keelung stations must wait for Express/Fast Local Train for some duration after it leaves region between Zhunang and Badouzi stations. This waiting duration is defined as "Waiting time after train leaves region".
16	Number of waiting times during operating (Tzu-Chiang Limited Express)	The total number of delays due to waiting for Tzu-Chiang Limited Express to pass while running between Zhunang and Keelung stations.
17	Number of passing trains during operating (Tzu-Chiang Limited Express)	Total number of passing Tzu-Chiang Limited Express trains while a train runs between Zhunang and Keelung stations.
18	Operating distance of awaited Express (Tzu-Chiang Limited Express)	Total distance which passing Tzu-Chiang Limited Express trains have run before Tzu-Chiang Limited Express passes a train that runs between Zhunang and Keelung stations.
19	Waiting time of a train (Tzu-Chiang Limited Express)	The total time a train must wait for Tzu-Chiang Limited Express to pass while running between Zhunang and Keelung stations.
20	Number of waiting times during operating (Puyuma/Taroko Express)	The total number of delays due to waiting for Puyuma/Taroko Express to pass while running between Zhunang and Keelung stations.
21	Number of passing trains during operating (Puyuma/Taroko Express)	Total number of passing Puyuma/Taroko Express trains while a train runs between Zhunang and Keelung stations.
22	Operating distance of awaited Express (Puyuma/Taroko Express)	Total distance passing Puyuma/Taroko Express trains have run before Puyuma/Taroko Express passes a train that runs between Zhunang and Keelung stations.
23	Waiting time of a train (Puyuma/Taroko Express)	Total time a train must wait for Puyuma/Taroko Express to pass while running between Zhunang and Keelung stations.
24	Number of waiting times during operating (Chu-Kuang Express)	Total number of delays while a train must wait for Chu-Kuang Express to pass while running between Zhunang and Keelung stations.
25	Number of passing trains during operating (Chu-Kuang Express)	Total number of passing Chu-Kuang Express trains while a train runs between Zhunang and Keelung stations.
26	Operating distance of awaited Express (Chu-Kuang Express)	Total distance the passing Chu-Kuang Express trains have run before Chu-Kuang Express pass a train that runs between Zhunang and Keelung stations.
27	Waiting time of a train (Chu-Kuang Express)	Total waiting time for a train while waiting for Chu-Kuang Express to pass while running between Zhunang and Keelung stations.
28	Number of waiting times during operating (Fast Local Train)	Total number of delays while a train must wait for Fast Local Train to pass while running between

		Zhunang and Keelung stations.
29	Number of passing trains during operating (Fast Local Train)	Total number of passing Fast Local Trains while a train runs between Zhunang and Keelung stations.
30	Operating distance of awaited Express (Fast Local Train)	Total distance which passing Fast Local Trains have run before Fast Local Trains pass a train that runs between Zhunang and Keelung stations.
31	Waiting time of a train (Fast Local Train)	Total waiting time for a train waiting for Fast Local Train to pass while running between Zhunang and Keelung stations.
32	Number of waiting times for awaiting one train	Total number of delays while a train must wait for any type of train while running between Zhunang and Keelung stations.
33	Number of waiting times for successively awaiting two trains	The total number of delays when a train must successively wait for two of any type of train while running between Zhunang and Keelung stations.

C. Build Preliminary GP Prediction Models

The GP technique is then applied to the prepared training and test data sets to build preliminary GP prediction models for the on-time percentages of trains. Here, the Discipulus GP software is used with its default parameter settings. The important parameters, including the population size, crossover rate, and mutation rate, are set as 500, 0.5, and 0.95, respectively. The GP software is repeatedly run ten times to select the optimal prediction model, and Table 3 summarizes the execution results. The fifth execution, denoted by a star (*), is selected as the best preliminary GP model since it can yield relatively low training RMSE and MAPE, as well as high training R² among ten runs. In addition, the coefficient of variance (CV) for the RMSE, R², and MAPE are 0.119279, 0.03201, and 0.14087, respectively, for the training data set. These values can be considered low enough. Therefore, the preliminary prediction models built through GP can be thought of as adequately robust.

Table 3. Execution results for building preliminary GP models

Execution No.	Training			Test		
	RMSE	R ²	MAPE	RMSE	R ²	MAPE
1	0.049533	0.71775	0.14033	0.047474	0.71523	0.11198
2	0.040292	0.70159	0.11509	0.056866	0.59651	0.12092
3	0.042408	0.69721	0.12253	0.057540	0.61642	0.13027
4	0.045794	0.69648	0.13509	0.055894	0.62280	0.11493
5*	0.033239	0.76070	0.10278	0.026404	0.81803	0.08709
6	0.044826	0.71107	0.13180	0.059220	0.62012	0.13106
7	0.036971	0.74628	0.14069	0.034618	0.75413	0.09325
8	0.037352	0.73539	0.13647	0.031590	0.79798	0.09385
9	0.046159	0.69462	0.13734	0.059839	0.60852	0.12654
10	0.041119	0.72748	0.12116	0.059599	0.58558	0.13238
Mean	0.041769	0.71885	0.12833	0.048904	0.67353	0.11423
Standard Deviation	0.004982	0.02301	0.01254	0.013078	0.08893	0.01720
CV	0.119279	0.03201	0.14087	0.267425	0.13204	0.13143

D. Feature Selection Based on Preliminary GP Models

The frequency of appearance for each independent variable while evolving in the building processes of the best GP model, i.e. the fifth model shown in Table 3, is summarized in Table 4. Based on Table 5, the independent variables #2 and #5 always appear, i.e. a 100% frequency, in

the evolving process of the best preliminary GP model. Therefore, these two independent variables are considered to be the most critical factors that can influence the dependent variable, i.e. on-time percentage of a train, and are thus selected as the featured variables in the TwoStep cluster analysis shown in the following Section.

Table 4. Frequencies of appearance of each independent variable while evolving the best GP model

Independent variable no.	1	2	3	4	5	6	7	8	9	10	11
Appearance frequency	50%	100%	50%	67%	100%	33%	17%	0%	17%	50%	50%
Independent variable no.	12	13	14	15	16	17	18	19	20	21	22
Appearance frequency	40%	30%	17%	50%	23%	10%	50%	83%	33%	50%	67%
Independent variable no.	23	24	25	26	27	28	29	30	31	32	33
Appearance frequency	23%	3%	60%	17%	37%	0%	3%	33%	33%	20%	13%

E. TwoStep Cluster Analysis for Dividing Data

The two independent variables identified in Section 4.D are fed into the TwoStep cluster analysis for further clustering the original data prepared in Section 4.B to yield data groups with enough diversity. SPSS software is used to implement the TwoStep cluster analysis where the distance between two items are evaluated by the likelihood measure, the BIC (Schwarz's Bayesian Criterion) clustering criterion is applied, and the optimal number of clusters is determined automatically by the SPSS process. Hence, the original data prepared in Step 1 with 278 rows are divided into four sub-groups. In addition, the data in each sub-group cluster are

further partitioned into training and test data groups based on a pre-determined 3:1 ratio as shown in Table 5. Table 5 also depicts the clustering centers of independent variables #2 and #5 for each group's data. From Table 5, the clustering centers of independent variables #2 and #5 in the first sub-group are 1.39 and 760.38, respectively. The second and fifth independent variables (factors) are "Operation days" and "Initial time" as shown in Table 2. In addition, the "Operation days" are coded as 1, 2, 3 and 4 for the trains running daily, on Sundays, daily except Sundays, and daily except Saturdays. Meanwhile, the values of "Initial time" are coded according to the time when the train sets out, e.g. coding 8:30

A Mixture of Expert-Based Prediction Approach by Using Genetic Programming and Clustering: A Case Study on Predicting the On-Time Percentages of Local TRA Trains in Taiwan

and 15:20 as 510 (minutes) and 920 (minutes), respectively. Therefore, the first sub-group's data mainly represent the Local Trains that run daily or on Sundays, and depart around 12:30. Similarly, the second sub-group's data primarily represent the Local Trains that run on Sundays and initiate

about 15:30. The major Local Trains that run on Sundays and start from their initiation station at about noon are partitioned into the third sub-group. Finally, the last sub-group's data mainly stand for the Local Trains with running days of Sundays, as well as initiating time of about 13:30.

Table 5. Four sub-groups' data obtained by the TwoStep cluster analysis

Sub groups	Quantity of data	Quantity of training data	Quantity of test data	Clustering center	
				Independent variable # 2	Independent variable # 5
1	76	57	19	1.39	760.38
2	88	66	22	1.70	925.12
3	52	39	13	1.83	732.77
4	62	47	15	1.68	806.85

F. Build GP models for sub-problems

The GP method implemented by the Discipulus software is utilized again for the training and test data of each sub-group shown in Table 5. For each sub-group's data, the GP software is run ten times, and Table 6 summaries the implementation results. Based on the RMSE, R^2 , and MAPE, the sixth, ninth, seventh, and second GP models in Table 6 are selected as the best prediction models for the four sub-groups' data and denoted by a star (*). From Table 6, it can be seen that the GP tool can build a superior model for predicting the on-time percentages of trains by the training data in the fourth sub-group. In other words, the on-time percentages of the Local Trains that mainly run on Sundays and set out at about 13:30 are relatively easy to predict. The first major reason we consider is that the time needed for getting on and off the trains can be significantly reduced since there are fewer working commuters at noon each day, especially on Sundays. The other reason is that most passengers on Sundays are travelers who start their travelling early in the morning and return home in the evenings or even very late. Hence, this situation further reduces the time for getting on and off the trains, especially around 13:30. Due to the above reasons, the on-time percentages of trains are higher, and thus are relatively easily predicted. Notably, the prediction performance via the tests MSE, R^2 , and MAPE are not always the best among all of the best GP models selected for each sub-group's data. However, the test R^2 still can attain a value of 0.89, which is considered high enough in the field of social science. On the other hand, the training MSE, R^2 , and MAPE in the second sub-group are worse among the best GP models selected for four sub-groups' data. That is to say that the on-time percentages of Local Trains that run on Sundays and mainly depart about 15:30 are relatively difficult to predict. The main possible cause is that the time 15:30 may be considered a rush time on Sundays since many travelers start their journey to the big cities for fun in the evening. In addition, there are more accidents at the level rail crossings since the road traffic is comparatively busy, thus delaying the running of trains and decreasing the on-time percentages of trains. Therefore, the on-time percentages of trains can vary greatly and are hard to predict. Besides, the MSE, R^2 , and MAPE for the test data in the second sub-group are worse among all of the best GP models chosen for each sub-group's

data. However, the R^2 for the training and test data can also reach 0.91 and 0.89, respectively, which can be regarded as high. The training and test MAPEs are all less than 15%, considered a difficult level that can be attained in dealing with the difficult prediction problems for on-time percentages of trains. In general, the GP technique can yield prediction models with sufficient quality since the lowest training and test R^2 s for all of the best GP models determined in each sub-group can be up to 0.91 and 0.89, respectively, and the largest training and test MAPEs are less than 0.051 and 0.15, individually. Furthermore, the superior training and test R^2 s can even achieve levels of 0.98 and 0.97, respectively. Based on the above information, the GP tools can produce a model with sufficient prediction performance, since GP can concentrate on constructing a prediction model for each sub-group's data where the data represent a particular, i.e. similar, situation of running and initiation for Local Trains.

G. Mixing Expert Models and Evaluating Prediction Performance

Each of the selected GP models can be thought of as an expert for predicting on-time percentages of trains for its corresponding sub-group's data. Hence, these selected models are merged to build an integrated prediction model to predict on-time percentages of trains, where which GP model should be applied for prediction is in accordance with which group a train's information is clustered in. Then, the criterion for evaluating prediction performance including MASE, R^2 , and MAPE is compared for the selected best preliminary GP model built in Section 4.C, denoted by GP_{All} , and the integrated one combined in Section 4.G, denoted by GP_{Int} . The comparison is provided in Table 7, in which we can see that the GP_{All} prediction model by mixing an expert in its professional field can provide less MSEs and MAPEs, and can yield higher R^2 s for both the training and test data compared to the GP_{All} model. In other words, the process for dividing the data by the critical factors, i.e. important feature variables for the dependent variable, into appropriate groups such that each group's data can be similar as much as possible is an effective method for improving the prediction performance of built GP models.

Table 6. GP Implementation Results for Each Sub-group in Table 5

Sub	Run No.	Training	Test
-----	---------	----------	------

group		MSE	R ²	MAPE	MSE	R ²	MAPE
1	1	0.003878	0.96666	0.03296	0.005185	0.90286	0.03582
	2	0.003746	0.96099	0.03253	0.005211	0.90142	0.03450
	3	0.004360	0.96633	0.03477	0.004385	0.91813	0.02577
	4	0.003718	0.96580	0.03199	0.004761	0.89982	0.03009
	5	0.003832	0.96298	0.02790	0.006025	0.88531	0.03625
	6*	0.002469	0.97867	0.02422	0.002547	0.96548	0.02284
	7	0.002990	0.97303	0.02711	0.007858	0.84500	0.04110
	8	0.002827	0.97185	0.02744	0.004092	0.93712	0.03122
	9	0.003543	0.96940	0.03173	0.007134	0.87358	0.03508
	10	0.004142	0.96866	0.03512	0.007286	0.87669	0.04222
	Mean	0.003551	0.96844	0.03058	0.005448	0.90054	0.03349
	Standard deviation	0.000602	0.00514	0.00364	0.001644	0.03413	0.00615
CV	0.169647	0.00531	0.11966	0.301858	0.03790	0.18366	
2	1	0.012843	0.90383	0.06209	0.058458	0.69453	0.15926
	2	0.014286	0.89381	0.06557	0.038088	0.82979	0.12412
	3	0.013526	0.88571	0.06027	0.043950	0.81992	0.13628
	4	0.013570	0.91142	0.06260	0.048038	0.80901	0.14837
	5	0.011713	0.89068	0.05353	0.045687	0.84975	0.14093
	6	0.013576	0.87850	0.06038	0.043168	0.86403	0.14026
	7	0.013256	0.90287	0.06194	0.026906	0.86186	0.11020
	8	0.011845	0.91103	0.05677	0.040638	0.74573	0.12779
	9*	0.009452	0.91965	0.05076	0.046447	0.89437	0.14449
	10	0.011334	0.89636	0.05632	0.046236	0.77562	0.14287
	Mean	0.012540	0.89939	0.05902	0.043762	0.81446	0.13746
	Standard deviation	0.001450	0.01275	0.00457	0.008012	0.06078	0.01378
CV	0.115611	0.01417	0.07739	0.183088	0.07462	0.10024	
3	1	0.005389	0.97013	0.03990	0.016691	0.96184	0.07177
	2	0.006853	0.96211	0.05026	0.019167	0.95092	0.07839
	3	0.004169	0.97596	0.03726	0.019048	0.96048	0.07374
	4	0.005335	0.97112	0.04741	0.020214	0.95013	0.06846
	5	0.007032	0.95792	0.05405	0.017480	0.96623	0.07762
	6	0.007537	0.95537	0.04837	0.026827	0.93644	0.08446
	7*	0.003667	0.97795	0.03509	0.013700	0.97373	0.04322
	8	0.004750	0.97135	0.03592	0.019509	0.94273	0.09212
	9	0.006794	0.96258	0.04597	0.028211	0.91973	0.09926
	10	0.007558	0.95729	0.04541	0.023546	0.93591	0.08203
	Mean	0.005908	0.96618	0.04396	0.020439	0.94981	0.07711
	Standard deviation	0.001426	0.00813	0.00653	0.004516	0.01643	0.01513
CV	0.241359	0.00841	0.14851	0.220948	0.01729	0.19617	
4	1	0.010539	0.94065	0.05874	0.024060	0.84735	0.08848
	2*	0.003375	0.98444	0.03584	0.011314	0.89106	0.05107
	3	0.006297	0.96767	0.04703	0.016395	0.84859	0.06743
	4	0.010592	0.94088	0.06043	0.023018	0.79539	0.08682
	5	0.009547	0.94642	0.05068	0.017296	0.82439	0.06940
	6	0.009285	0.95911	0.06679	0.014762	0.87253	0.06748
	7	0.006989	0.96676	0.05673	0.014304	0.85150	0.05809
	8	0.005367	0.96852	0.04227	0.017012	0.82408	0.07160
	9	0.005757	0.97611	0.05328	0.019887	0.83396	0.07857
	10	0.009723	0.96411	0.06983	0.011358	0.91817	0.05975
	Mean	0.007747	0.96147	0.05416	0.016940	0.85070	0.06987
	Standard deviation	0.002513	0.01474	0.01061	0.004360	0.03561	0.01211
CV	0.324319	0.01533	0.19585	0.25746	0.04186	0.17326	

Table 7. Comparison of Predicting On-time Percentages of Trains

GP model	Training			Test		
	MSE	R ²	MAPE	MSE	R ²	MAPE
GP _{All}	0.033239	0.76070	0.10278	0.026404	0.81803	0.08709

A Mixture of Expert-Based Prediction Approach by Using Genetic Programming and Clustering: A Case Study on Predicting the On-Time Percentages of Local TRA Trains in Taiwan

GP _{int}	0.005102	0.96273	0.03724	0.020551	0.89427	0.07267
-------------------	----------	---------	---------	----------	---------	---------

V. CONCLUSIONS

The on-time percentages of trains could significantly affect the operation of trains, as they are valuable information for arranging appropriate timetables, determining waiting of trains, deciding the tracks for running trains, and setting up adequate manpower etc. The prediction of the on-time percentages of trains, however, is a very complicated and difficult problem due to various factors that can influence the on-time percentages of trains and that are not easy to completely identify and consider. In addition, segmenting an original problem into several sub-problems, and creating an expert that has expertise in its specialized area by solving each sub-problem, called the mixture of experts, has been popular. Therefore, this study proposes a prediction procedure employing a mixture of experts by using genetic programming (GP) and clustering analysis. The usefulness and effectiveness of the proposed approach are verified through a case study predicting the on-time percentages of Local Trains operated by the Taiwan Railway Administration (TRA) in Taiwan. The implementation results show that GP can adequately construct an original prediction model for the whole dataset. Appropriate groups of data are then found by clustering with characteristics, i.e. features, variables that are identified through exploring the well-constructed GP model. An integrated prediction model can be obtained by mixing experts dedicated to predicting data in each cluster. A comparison shows that the integrated GP prediction model is much better than the original one based on the performance evaluation through MSE, R^2 , and MAPE. In other words, the process for building a distinct GP prediction model for each cluster's data, having similar characteristics, can indeed improve the prediction accuracy process. Therefore, our proposed approach can be considered a useful and effective tool for resolving a prediction problem in the real world.

ACKNOWLEDGMENT

The author thanks Jackson Liu in the Secretariat of Taiwan Railway Administration (TRA) in Taiwan, R.O.C. for fully supporting this study, as well as Minghsin University of Science and Technology, Taiwan, R.O.C., for partially supporting this study under Contract No. MUST-108BA-1.

REFERENCES

[1] N. Glisovic. (2016). A hybrid model for forecasting the volume of passenger flows on Serbian railways. *Oper. Res.* 16(2), pp. 271–285.

[2] Q. Lai, J. Liu, Y. Luo, M. Minshu. (2017). A hybrid short-term forecasting model of passenger flow on high-speed rail considering the impact of train service frequency. *Math. Probl. Eng.* Article ID 1828102.

[3] Z. Xie, L. Jia, Y. Qin. (2013). A hybrid temporal-spatio forecasting approach for passenger flow status in Chinese high-speed railway transport hub. *Discret. Dynam. Nature. Soc.* Article ID 239039.

[4] C. Liang, M. Ghazel, O. Cazier, E. El-Koursi. (2018). Developing accident prediction model for railway level crossings. *Safety. Sci.* 101, pp. 48–59.

[5] C. Z. Jiang, P. Huang, J. Lessan, L. P. Fu, C. Wen. (2019). Forecasting primary delay recovery of high-speed railway using multiple linear regression, supporting vector machine, artificial neural network, and random forest regression. *Can. J. Civil. Eng.* 46(5), pp. 353–363.

[6] M. Q. Xie, X. M. Li, W. L. Zhou, Y. B. Fu. (2014). Forecasting the short-term passenger flow on high-speed railway with neural networks. *Comput. Intel. Neurosc.* Article ID 375487.

[7] F. Dou, L. M. Jia, L. Wang, J. Xu, Y. K. Huang. (2014). Fuzzy temporal logic based railway passenger flow forecast model. *Comput. Intel. Neurosc.* Article ID 950371.

[8] P. Kecman, R. M. P. Goverde. (2015). Online data-driven adaptive prediction of train event times. *IEEE T. Intell. Transp.* 16(1), pp. 465–474.

[9] Z. Ding, Y. Zhou, G. G. Pu, M. C. Zhou. (2018). Online failure prediction for railway transportation systems based on fuzzy rules and data analysis. *IEEE T. Reliab.* 67(3), pp. 1143–1158.

[10] M. Yaghini, M. M. Khosraftar, M. Seyedabadi. (2013). Railway passenger train delay prediction via neural network model. *J. Adv. Transport.* 47(3), pp. 355–368.

[11] L. Oneto, E. Fumeo, G. Clerico, R. Canepa, F. Papa, C. Dambra, N. Mazzino, D. Anguita. (2018). Train delay prediction systems: a big data analytics perspective. *Big. Data. Res.* 11, pp. 54–64.

[12] Z. Y. Xie, Y. Qin. (2016). A passenger flow risk forecasting algorithm for high-speed railway transport hub based on surveillance sensor networks. *J. Sensors.* Article ID 564790.

[13] T. T. Tsai. (2014). A self-learning advanced booking model for railway arrival forecasting. *Transport. Res. C.* 39, pp. 80–89.

[14] K. P. Burnham, D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* New York: Springer-Verlag, 2002.

[15] H. Akaike. (1969). Fitting autoregressive models for prediction. *Ann. I. Stat. Math.* 21, pp. 243–247.

[16] J. H. Holland. *Adaptation in Nature and Artificial Systems.* Ann Arbor, MI: The University of Michigan Press, 1975.

[17] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* Cambridge, Mass: MIT Press, 1992.

[18] J. R. Koza, M. A. Keane, M. J. Streeter, W. Mydlowec, J. Yu, G. Lanza. *Genetic Programming: Genetic Programming IV: Routine Human-Competitive Machine Intelligence.* New York: Springer-Verlag, 2005.

[19] I. Cigliarić, A. Kidrić. (2006). Computer-aided derivation of the optimal mathematical models to study gear-pair dynamic by using genetic programming. *Struct. Multidiscip. O.* 32(2), pp. 153–160.

[20] J. R. Koza, M. J. Streeter, M. A. Keane. (2008). Routine high-return human-competitive automated problem-solving by means of genetic programming. *Inform. Sciences.* 178(23), pp. 4434–4452.



Chih-Ming Hsu Chih-Ming Hsu is currently a Professor in the Department of Business Administration at Minghsin University of Science and Technology, Taiwan. He holds a PhD in Industrial Engineering and Management from National Chiao Tung University, Taiwan. His present research interests include quality engineering, optimization methods in industrial applications and data mining applications in CRM.

柔性演算法於選擇權波動度預測之研究

徐志明^{1*} 彭莉雅²

¹ 明新科技大學企管系 (新竹縣新豐鄉新興路 1 號)

² 明新科技大學會計室 (新竹縣新豐鄉新興路 1 號)

*cmhsu@must.edu.tw

摘要

台灣經濟環境受國際情勢影響甚鉅，近年景氣擴張力偏弱，介於綠燈以及藍燈之間，顯示出景氣發展不如預期。政府提出經濟政策與各國間多重領域的合作，開啟多元化走向，尋找新的經濟發展，民眾則以投資理財賺得利益，市場上的投資商品多樣化，投資者可依自己的狀況作選擇，台灣年輕族群普遍低薪資存款少，因而選擇交易成本較低的衍生性金融商品進行投資，期望獲利高且損失輕微。本研究針對衍生性金融商品之選擇權為研究主軸，以 2011 年 1 月 3 日至 2016 年 10 月 28 日為研究區間，利用基因規劃法、倒傳遞類神經網路和基因演算法以建構波動度模型，以方法為基礎產生 20 個預測模式，從而精準預測隔日台灣加權股價指數波動度，進而可估計選擇權合理價格且得知風險大小。

實驗結果顯示，倒傳遞類神經網路模型預測能力最好，判定係數最好為 0.3，代表所建構的預測模型與選擇權市場的符合程度，比起過去研究來得高，並且瞭解到波動度的預測並非這麼簡單，提供投資者作為參考，提高報酬降低風險。

關鍵字(3~5 個字)：波動度預測、基因規劃法、倒傳遞類神經網路、基因演算法。

1. 緒論

台灣經濟景氣近年擴張力偏弱，可從景氣對策訊號解讀意義，目前介於綠燈及黃藍燈之間外溫內冷，政府若沒有解決薪資凍漲問題，難以促進消費意願，景氣回溫將成為困境。因此，台灣年輕人低薪資存款少，不進場買股票怕套牢，加上優質股門檻高買不起，所以選擇投資衍生性金融商品，以小

博大特色，具有避險、交易成本低、流動性高、放空較容易的優勢，已經在國際間存在許久，但台灣衍生性金融商品市場發展並沒有悠久的歷史，而是緊追國際步伐才有快速成長的跡象。金融市場變化多端且錯綜複雜，投資者藉由財報分析、歷年投資的經驗、電視媒體等獲得資訊及投資方向，但仍然有許多不確定的因素存在，影響著投資環境，有些投資者把畢生積蓄投入，大多投資者卻是不越雷池一步，但都期望獲利高且損失輕微。故本研究針對衍生性金融商品之選擇權為研究主軸，因虧損維持在權利金為門檻，風險相對於股票而言較低，對於資金不多的年輕族群可作為事前投資判斷之參考。

2. 文獻探討

依據研究之背景與動機瞭解台灣經濟環境與投資方向有連帶關係，資訊交易(Informed trading)會釋放出資訊提供投資者去推算未來股價的波動度，預測波動度的方法不只一種，學者努力建構出更準確之預測模型，使財務預測上有規模的成果。

2.1 波動度預測

翁煒翔(2014)提出 GARCH-M(GARCH in Mean)模型，以預測標準普爾 500 指數(Standard & Poor's 500 Index, S&P 500)，將選擇權市場因子隱含波動度價差(Implied Volatility Spread, VS)、波動度指數(Volatility Index, VIX)加入模型進行比較，模型 1 為 GARCH-M+VS；模型 2 為 GARCH-M+VIX；模型 3 為 GARCH-M+VS、VIX。實驗結果顯示，在概似比檢定(Likelihood Ratio, LR)，模型 1,2,3 皆優於 GARCH-M 模型；在均方根誤差(RMSE)，可知道隱含波動度價差(VS)因子

可降低實質波動度的誤差，使預測能力變更好。

鄭佩文(2014)利用澳幣、英鎊、加拿大幣、歐元、瑞士法郎和日圓六種外匯期貨選擇權隱含波動度偏斜，以多元迴歸式預測匯率走勢，然後，分為兩種狀況，一般情況下和美國五項重大總體經濟事件宣告下。實證結果顯示，一般情況下，隱含波動度偏斜主要捕捉負面資訊，無法準確預測匯率走勢，然而，在美國非農業就業人口、消費者物價指標及工產訂單的經濟事件宣告下，澳幣和加拿大幣期貨選擇權隱含波動度偏斜有顯著的預測能力，匯率變動呈現正向關係。

陳昌捷(2015)利用倒傳遞類神經網路模型，以預測台灣股市加權指數，首先以皮爾森相關係數篩選出輸入參數，為相關係數 r 值 0.7 以上的技術分析指標，然後，以 MATLAB 工具建構模型，將相關係數分為三組 r 值 0.7、0.8、0.9 進行預測。實驗結果顯示，三組中 r 值 0.8 的絕對誤差平均值 (MAPE) 最小，為 0.6315% 達到高度準確的預測能力，同時發現，預測漲跌方面，短期較具有高度準確性。

陳建名(2016)從台股現貨、期貨及選擇權市場中，挑選出三個變數，台股報酬率、台指期貨報酬率和波動率指數，建立向量自我迴歸模型 (vector autoregressive, VAR)，以預測股價波動率。利用 ADF 單根檢定確認資料為定態，將變數以赤池信息準則 (Akaike Information Criterion, AIC) 選定最適落後期，再透過 Granger 因果關係檢定，瞭解變數之間有緊密的互動關係，最後進行分析。實驗結果，任何波動率指數加入負向及端值效果，明顯地增加模型解釋力及預測準確性。

簡壬申(2017)利用倒傳遞類神經網路模型，以預測台灣 50 成分股報酬率，先運用移動視窗法，資料會隨時間變動，可避免造成預測上有偏誤，然後，輸入變數為 6 個基本指標和 6 個總經指標，產生輸出變數隔月報酬率，再加以建構出投資組合及策略，透過夏普指標、資訊比率指標、詹森指標、M2 指標之分析，結果顯示，投資組合優於台灣 50 指標及台灣加權股價指數。

Maciel *et al.* (2015) 利用線性迴歸 (Linear

Regression)、遞迴可能性模糊模型 (recursive Possibilistic Fuzzy Modeling, rPFM)、多層前饋神經網路 (multi-layer feedforward neural network, MLP) 和演化/神經模糊模型，以預測股票市場的波動度。實驗結果顯示，rPFM 是一個預測波動度的潛在工具，具有處理噪聲數據和異常值的能力，同時，結合學習模型結構和遞迴最小平方方法來估計參數，使模型獲得較佳的預測值，已經超越一些傳統的遞迴模糊和神經模糊模型。

Yaziz *et al.* (2015) 提出一個 Box-Jenkins-GARCH 混合模型，以預測黃金價格。使用 Box-Cox 將數據轉換，解決穩定性問題。第一階段 Box-Jenkins 模型處理線性資料，第二階段處理非線性資料，實驗結果顯示，Box-Jenkins-GARCH 混合模型內 ARIMA-GARCH 模型，預測黃金價格的波動能力較好，同時證明，GARCH 模型克服了 Box-Jenkins 模型中的非線性限制。

Atsalakis (2016) 認為歐盟排放交易體系 (European Union Emission Trading Scheme, EU ETS) 是一個以成本效益原則減少溫室氣體排放的工具。作者利用日常碳價格及時間序列資料，提出新的混合神經模糊控制器 (PATSSOS)、人工類神經網路 (ANN) 和調適性神經模糊推理系統 (Adaptive Neuro-Fuzzy Inference System, ANFIS) 三種智能技術，預測碳價格。結果顯示，PATSSOS 模型相對其他模型執行簡易，準確性較佳且及時預測，雖然有效地管理風險，但計算耗時是 PATSSOS 模型的缺點。

Barunik and Krehlik (2016) 提出三個不同的期間 (危機前、危機時危機後)，利用 GARCH 模型、部份整合型自我迴歸移動平均模型 (Autoregressive fractionally integrated moving average, IFMA)、異質自我迴歸模型 (heterogeneous autoregressive, HAR)、人工類神經網路 (ANN) 與 HAR-ANN 模型，以預測能源 (原油、加熱用燃油和天然氣) 價格。實驗結果顯示，所提出的方法，在金融危機時，對實質性結構斷裂是有效的，同時，使用高效率資料預測基本上優於 GARCH 模型。

2.2 Black-Scholes 評價模型

選擇權為權利契約的一種，可分為買權(Call Option)及賣權(Put Option)，買方支付權利金(Premium)給賣方，擁有未來某一段時間或特定日期的權利，可依契約上的履約價格向賣方買入或賣出標的資產(Underlying Asset)。

標的資產價格與履約價格之間的大小，可將選擇權分為三種型態，價內(In-the-Money)為標的資產價格大於履約價格之買權或標的資產價格小於履約價格之賣權，價外(Out-of-the-Money)為標的資產價格小於履約價格之買權或標的資產價格大於履約價格之賣權，價平(At-the-Money)為標的資產價格等於履約價格之買權和賣權，提供投資者遞延決策的功能。

1973年Fischer Black與Myron Scholes提出歐式選擇權Black-Scholes模型，廣泛運用於學界及業界，選擇權價值影響因素為履約價值和時間價值，連帶影響選擇權的市場價值，則影響市場價值(P)的五項變數：標的資產價格(S)、履約價格(K)、履約期間(T)、無風險利率(r)及標的資產價格波動度(σ)，方程式(1)。

$$c(P) = f(S, K, T, r, \sigma) \quad (1)$$

在Black-Scholes模型中，估計波動度是件困難的事，因此估計方法大致分為兩種：其一為歷史波動度(Historical Volatility)使用歷史資料估計波動度；另一種隱含波動度(Implied Volatility)將選擇權市場價值(P)直接代入Black-Scholes模型逆推估計波動度，本研究採用隱含波動度的估計方法，以標的資產價格(S)、履約價格(K)、履約期間(T)、無風險利率(r)及市場價值(P)五個已知變數，使用試誤法(Try and Error Approach)反覆求取，便可以精確估計標的資產價格波動度(σ)，方程式(2)，藉以提高投資者獲利機會。

$$\sigma = f(c(P), S, K, T, r) \quad (2)$$

2.3 基因演算法與基因規劃法

1975年Holland提出基因演算法(Genetic Algorithms, GA)，為一種最佳化隨機搜尋的方法，基於仿效生物界演化過程，透過自然法則中的物競

天擇，以參數編碼進行運算，可跳脫傳統搜尋單一節點的限制，選擇基因較好的母代，經過複製(Reproduction)、交配(Crossover)、突變(Mutation)三個主要運算子，來產生優於母代的下一代，重覆此方式直到目標值不再改變或達到設定的最大值為止，基因演算法演化過程中保留基因多樣性，廣泛應用於各種問題的解決上，包括經濟學、工業工程、機械自動化等多元領域。

1992年John Koza提出基因規劃法(Genetic Programming, GP)，與基因演算法(Genetic Algorithms, GA)的概念相同，具有染色體、適應函數及運算機制，但演化單位不同，基因演算法為0-1的二元字串及實數字串，基因規劃法為可變動大小、形狀及結構的分析樹(Parse Tree)表示染色體，每條染色體為一組程式碼，經過演化後，產出最佳的程式碼，以程式方式呈現，不需解碼並節省撰寫程式的時間。

分析樹中包含節點(Node)、節線(Edge)、層(Level)、終端集合(Terminal Set)和函數集合(Function Set)。上下兩層因有節線連結的關係，上層節點為父節點(Parent Node)，下層節點為子節點(Children Node)，如果不單單只有一層的關係，則稱祖父節點或子孫節點，同一層的節點且父節點相同稱為兄弟節點(Sibling Node)。終端集合以變數(Variables)和常數(Constants)所組成；函數集合以算數(+、-、×、÷)、邏輯(if、then)、比較(>、<、=)、布林(AND、OR)、程式，以及根據問題需要而定義出可運作之函數所組成。

2.4 類神經網路

1943年McCulloch與Pitts提出神經元數學模型，接著1949年Hebbian提出神經細胞的學習規則，建立神經科學及電腦科學有溝通的橋樑，發展類神經網路(Artificial Neural Network, ANN)。

人腦約有數百億個神經細胞(Nerve Cells)，神經細胞約有數千根突觸(Synapses)，與其他神經細胞連結，形成一個非線性且複雜無比的神經網路。類神經網路主要概念模仿人的神經系統，由大量互相連結的神經元(Neuron)或稱處理單元(Processing

Elements)所組成，以平行的方式進行運算，可同時分析大量資料。

類神經網路除了外形相似外，可利用樣本或資料來訓練，與人腦相同的三項重要能力，學習(Learn)、回想(Recall)及歸納推演(Generalize)，解決最佳化問題速度快速，且具有容錯性及平行處理能力，應用層面極為廣泛，有理工、商業管理及醫學領域等。

3. 研究架構

本研究針對臺灣證券交易所中之股價指數作為投資標的，以衍生性金融商品之選擇權 Black-Scholes 評價理論，進行資料蒐集及建構預測模型，使用基因規劃法、倒傳遞類神經網路及基因演算法預測多種模式，再對這些模式進行誤差分析，可得知波動度預測效果最佳之模式，研究架構圖如圖 1 所示。

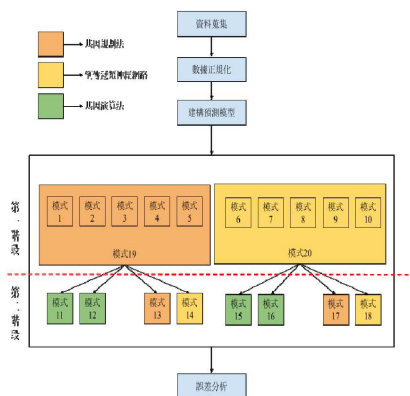


圖 1 研究架構

4. 實證結果與分析

本研究資料範圍自 2011 年 1 月 3 日至 2016 年 10 月 28 日之每日交易訊息共 1434 筆，交易訊息包含包含台灣加權股價指數收盤價、無風險利率、台指選擇權履約價格、履約期間以及收盤價 5 個輸入變數及 21 日波動度 1 個輸出變數，將資料正態化後介於零與一之間再分為兩部分，第一部份為 2011 年 1 月 3 日至 2014 年 12 月 30 日，先隨機排

序再以 3:1 比例分成訓練資料 742 筆及測試資料 248 筆，第二部分為 2014 年 12 月 31 日至 2016 年 10 月 28 日為預測資料 444 筆，以基因規劃法、倒傳遞類神經網路及基因演算法建構預測模型。

本研究採用 Discipulus 基因規劃軟體預測，參數設定為軟體本身內定值，其中重要參數有母體大小 500，表示每一族群有 500 個個體，交配率 0.5 為文獻中常見的值，突變率 0.95。

將訓練及測試資料放入類神經網路中，設定網路參數有些限制，隱藏層神經元的個數最多設定為變數個數兩倍，學習速率(η)過大或過小都會影響收斂速度，訓練次數可調整權重，但改善效果有限。

在基因演算法方面，本研究採用 Evolver 軟體，克服長時間計算，並解決類似自然界中進化的問題，操作簡單僅需要 Excel 即可。

將所蒐集的資料整理後，以基因規劃法建構模式 1 至 5，各模式重複預測 10 次，每次預測會得到訓練與測試均方誤差(MSE)，按照 3:1 計算加權平均，從 10 個加權平均數中挑選值最小的，模式 1 至 5 之 MSE 加權平均數如表 1 所示。

表 1 模式 1 至 5 之 MSE 加權平均數

MSE 次數	模式 1 價內一檔	模式 2 價內二檔	模式 3 價平	模式 4 價外一檔	模式 5 價外二檔
1	0.008716	0.008913	0.008959	0.008915	0.010904
2	0.009048	0.008209	0.009118	0.009803	0.010596
3	0.009618	0.011744	0.011118	0.011474	0.009773
4	0.009862	0.010848	0.010455	0.010888	0.010308
5	0.010485	0.011517	0.010624	0.011419	0.012248
6	0.011422	0.012366	0.011058	0.010672	0.010772
7	0.011754	0.011372	0.010650	0.009958	0.010769
8	0.011262	0.011080	0.012159	0.010773	0.010971
9	0.011348	0.010636	0.011692	0.011176	0.010463
10	0.009905	0.010016	0.011454	0.010782	0.011497

註：灰底代表最小值

模式 11 至 14 輸入變數為模式 1 至 5 的最佳輸出解，模式 11、12 以基因演算法重複預測 20 次，模式 11 為一階運算，權重設定 DV1 至 DV5 各 0.2，加總起來為 1，模式 12 為二階交叉運算，權重設定各為 0.2，兩模式經過 Evolver 軟體運算後，可得到各 20 次權重變化及均方誤差(MSE)，此外，模式 13 使用基因規劃法，模式 14 使用倒傳遞類神經網路，兩模式皆重複預測 10 次，每次預測會得到訓練與測試均方誤差(MSE)，按照 3:1 計算加權平均，模式 11 至 14 之均方誤差如表 2 所示。

表 2 模式 11 至 14 之 MSE

MSE 次數	模式 11 GA 一階	模式 12 GA 二階	模式 13 GP	模式 14 BPNN
1	0.007060649	0.006560862	0.006447	0.069875
2	0.007060586	0.006535856	0.006642	0.054350
3	0.007060663	0.006543208	0.006311	0.052325
4	0.007060608	0.006544964	0.006566	0.052150
5	0.007060572	0.006559647	0.007125	0.052025
6	0.007060599	0.006595211	0.006193	0.051825
7	0.007060593	0.006552124	0.006708	0.051050
8	0.007060609	0.006528559	0.006868	0.051575
9	0.007060585	0.006523067	0.007155	0.051300
10	0.007060588	0.006567225	0.006739	0.051200
11	0.007060573	0.006536567		
12	0.007060718	0.006613964		
13	0.007060572	0.006570082		
14	0.007060600	0.006554108		
15	0.007060815	0.006568434		
16	0.007060730	0.006535830		
17	0.007060581	0.006551528		
18	0.007060576	0.006567197		
19	0.007060573	0.006527135		
20	0.007060577	0.006533568		

註：灰底代表最小值

將所蒐集的資料整理後，以倒傳遞類神經網路建構模式 6 至 10，按照 3：1 計算加權平均，從 10 個加權平均數當中挑選值最小的，如表 3 所示。

表 3 模式 6 至 10 之 MSE 加權平均數

MSE 次數	模式 6 價內一檔	模式 7 價內二檔	模式 8 價平	模式 9 價外一檔	模式 10 價外二檔
1	0.081100	0.084500	0.080800	0.081600	0.083575
2	0.072450	0.073550	0.071550	0.071050	0.070925
3	0.060700	0.074375	0.072675	0.072275	0.072150
4	0.075725	0.076650	0.074875	0.074575	0.074725
5	0.075050	0.076150	0.074075	0.073575	0.073525
6	0.077475	0.078050	0.077000	0.077000	0.077250
7	0.074850	0.075700	0.074275	0.074175	0.074550
8	0.072000	0.073100	0.071025	0.070525	0.070475
9	0.074250	0.075275	0.062775	0.072800	0.072850
10	0.073675	0.074800	0.072600	0.072000	0.071925

註：灰底代表最小值

模式 15 至 18 輸入變數為模式 6 至 10 的最佳輸出解，模式 15、16 以基因演算法重複預測 20 次，模式 15 為一階運算，權重設定 DV1 至 DV5 各 0.2，加總起來為 1，模式 16 為二階交叉運算，權重設定各為 0.2，兩模式經過 Evolver 軟體運算後，可得到各 20 次權重變化及均方誤差(MSE)，此外，按照 3：1 計算加權平均，最後，各模式從中挑選均方誤差(MSE)值最小的，如表 4 所示。

表 4 模式 15 至 18 之 MSE

MSE 次數	模式 15 GA 一階	模式 16 GA 二階	模式 17 GP	模式 18 BPNN
1	0.012969321	0.012576306	0.011123	0.073125
2	0.012969315	0.012565471	0.010919	0.069725
3	0.012969251	0.012627444	0.011126	0.069575
4	0.012969308	0.012484024	0.011253	0.068950
5	0.012969259	0.012615942	0.010853	0.069250

MSE 次數	模式 15 GA 一階	模式 16 GA 二階	模式 17 GP	模式 18 BPNN
6	0.012969341	0.012517524	0.010693	0.070800
7	0.012969270	0.012537542	0.011047	0.071125
8	0.012969413	0.012616703	0.009713	0.069875
9	0.012969263	0.012749631	0.011207	0.070200
10	0.012969343	0.012595310	0.010768	0.070750
11	0.012969271	0.012503452		
12	0.012969291	0.012648022		
13	0.012969262	0.012676103		
14	0.012969264	0.012463153		
15	0.012969255	0.012599310		
16	0.012969298	0.012602390		
17	0.012969284	0.012434780		
18	0.012969335	0.012622248		
19	0.012969274	0.012614787		
20	0.012969268	0.012428207		

註：灰底代表最小值

模式 19、20 將所蒐集的資料一起預測，輸入變數為 17 個，模式 19 以基因規劃法重複預測 10 次，模式 20 以倒傳遞類神經網路預測 34 次，因隱藏層神經元最大個數等於輸入層神經元個數的兩倍，按照 3：1 計算加權平均，從中挑選最小值，如表 5 所示。

表 5 模式 19 至 20 之 MSE

MSE 次數	模式 19 GP	MSE 次數	模式 20 BPNN	
1	0.011475	1 18	0.073750	0.070775
2	0.012053	2 19	0.080625	0.070450
3	0.009413	3 20	0.073300	0.071800
4	0.011050	4 21	0.072600	0.070400
5	0.011301	5 22	0.072975	0.072075
6	0.011673	6 23	0.070425	0.069750
7	0.011288	7 24	0.071750	0.072375
8	0.011055	8 25	0.070750	0.073875
9	0.012007	9 26	0.074275	0.072150
10	0.010964	10 27	0.075350	0.069875
		11 28	0.071300	0.073125
		12 29	0.070750	0.069675
		13 30	0.069950	0.069725
		14 31	0.069575	0.069800
		15 32	0.070350	0.072450
		16 33	0.071800	0.071700
		17 34	0.072025	0.071175

註：灰底代表最小值

以同樣輸入變數作比較，價內一檔模式 6 優於模式 1；價內二檔模式 7 優於模式 2；價平模式 8 優於模式 3；價外一檔模式 9 優於模式 4；價外二檔模式 10 優於模式 5；以基因規劃法輸出為輸入之模式 14 優於模式 13；綜合以上可得知，倒傳遞類神經網路預測效果優於基因規劃法，唯獨以倒傳遞類神經網路輸出為輸入這組，預測效果基因規劃法優於倒傳遞類神經網路，如表 6 所示。

表 6 模式 1 至模式 20 結果分析

	訓練			測試			預測		
	R ²	RMSE	MAPE	R ²	RMSE	MAPE	R ²	RMSE	MAPE
模式 1	0.809821	0.026579	9.70%	0.707651	0.030919	10.45%	0.201068	0.058771	18.89%
模式 2	0.814121	0.026196	9.47%	0.744628	0.028940	9.56%	0.200608	0.052190	16.72%
模式 3	0.785768	0.028227	10.19%	0.764133	0.027760	9.67%	0.087875	0.070613	21.36%
模式 4	0.785238	0.028150	10.75%	0.766481	0.027714	9.87%	0.038011	0.060782	20.47%
模式 5	0.765122	0.029426	10.97%	0.742464	0.029163	10.49%	0.075751	0.057625	17.79%
模式 6	0.621168	0.038550	15.63%	0.599113	0.037368	14.44%	0.215573	0.042189	14.63%
模式 7	0.640995	0.036363	14.41%	0.627055	0.034958	13.25%	0.199844	0.046876	16.02%
模式 8	0.636820	0.036737	14.68%	0.612745	0.035734	13.65%	0.227432	0.043379	15.01%
模式 9	0.666652	0.035039	13.70%	0.649858	0.033856	12.72%	0.229873	0.045753	15.65%
模式 10	0.666942	0.035024	13.65%	0.652297	0.033741	12.64%	0.229472	0.045703	15.64%
模式 11	0.838713	0.024678	9.15%	0.801817	0.025770	9.11%	0.177903	0.052612	17.30%
模式 12	0.843941	0.023974	8.54%	0.824717	0.024027	8.20%	0.181985	0.059293	18.46%
模式 13	0.855820	0.023060	8.23%	0.820722	0.024284	8.38%	0.142681	0.059729	19.41%
模式 14	0.826656	0.025278	9.15%	0.804273	0.025340	8.78%	0.145954	0.060283	18.86%
模式 15	0.686533	0.034030	12.87%	0.663405	0.033198	12.02%	0.264022	0.045939	15.61%
模式 16	0.697235	0.033395	12.79%	0.682005	0.032243	11.72%	0.246388	0.046245	15.72%
模式 17	0.764421	0.029493	11.27%	0.750795	0.028595	10.09%	0.269457	0.044414	14.54%
模式 18	0.678750	0.034406	13.25%	0.660774	0.033335	12.45%	0.233945	0.044133	15.09%
模式 19	0.776271	0.028715	10.51%	0.740883	0.029113	10.24%	0.138654	0.055985	18.44%
模式 20	0.681899	0.034252	13.23%	0.638716	0.034390	12.80%	0.307007	0.037870	13.56%

5. 結論

本研究主軸為預測，是估計未來可能會發生的結果，以 Black-Scholes 選擇權評價模型作為研究理論，得知市場價值(P)會受到標的資產價格(S)、履約價格(K)、履約期間(T)、無風險利率(r)及標的資產價格波動度(σ)五項變數而影響，本研究則是以標的資產價格(S)、履約價格(K)、履約期間(T)、無風險利率(r)及市場價值(P)五個已知變數，預測一個未知變數標的資產價格波動度(σ)。針對台灣加權股價指數及台指選擇權 2011 年 1 月 3 日至 2016 年 10 月 28 日每日交易資料，使用基因規劃法、倒傳遞類神經網路與基因演算法以建構波動度預測模型，產生 20 個預測模式，從而預測隔日台灣加權股價指數波動度。實驗結果顯示，以同樣輸入變數作比較，倒傳遞類神經網路的預測效果最佳，同時發現，學習速率設定為 0.25 和動量設定為 0.95 的組合計算出來的訓練和測試的 MSE 加權平均值最小，也就是預測值與實際值之間誤差小，誤差小代表預測能力好，提供投資者買賣決策作為參考。本研究判定係數(R²)最好為 0.037870，代表預測模型的解釋能力有限，加以證實波動度難以預測，顯現出本研究的結果仍具有良好的參考價值。

誌謝

感謝明新科技大學校內專題研究計畫 MUST-108 企管-1 給予部分經費補助。

參考文獻

- 1 翁煒翔，選擇權資訊內涵對股價報酬波動度之分析與預測，東吳大學財務工程與精算數學系，碩士論文，2014。
- 2 陳昌捷，以倒傳遞類神經網路預測股市指數，國立宜蘭大學多媒體網路通訊數位學習碩士在職專班，碩士論文，2015。
- 3 陳建名，臺灣股價指數波動率與報酬率之關係，國立臺北大學經濟學系，碩士論文，2016。
- 4 鄭佩文，外匯期貨選擇權隱含波動度偏斜之資訊內涵與預測能力，國立中央大學財務金融學系，碩士論文，2014。
- 5 簡壬申，類神經網路在股票預測之獲利可能性研究—以台灣 50 成分股為例，國立雲林科技大學財務金融系，碩士論文，2017。
- 6 Atsalakis, GS., "Using computational intelligence to forecast carbon prices", *Applied Soft Computing*, Vol. 43, 2016, pp. 107-116.
- 7 Barunik, J., Krehlik, T., "Combining high

- frequency data with non-linear models for forecasting energy market volatility”, *Expert Systems with Applications*, Vol. 55, 2016, pp. 222–242.
- 8 Maciel, L., Gomide, F., Ballini, R., “Stock Market Volatility Prediction Using Possibilistic Fuzzy Modeling”, *Latin America Congress on Computational Intelligence (LA-CCI)*, 2015.
- 9 Yaziz, SR., Azizan, NA., Ahmad, MH., Zakaria, R., Agrawal, M., Boland, J., “Preliminary Analysis on Hybrid Box-Jenkins - GARCH Modeling in Forecasting Gold Price”, *The 2nd ISM International Statistical Conference 2014 (ISM-II)*, Vol. 1643, 2015, pp. 289–297.

The Study on Applying the Soft Computing in Forecasting the Options’ Volatilities

Chih-Ming Hsu^{1*} Li-Ya Peng²

¹Department of Business Administration (No. 1, Hsin-Hsing Rd., Hsin-Fong, Hsinchu, Taiwan)

² Account Office (No. 1, Hsin-Hsing Rd., Hsin-Fong, Hsinchu, Taiwan)

*cmhsu@must.edu.tw

Abstract

Taiwan’s economic environment affected by the international situation is quite large. In recent years, the intensity of the expansion is weak between the green light and the blue light, which shows the boom development not as good as expectation. Taiwan’s youth generation has low salaries and little deposits; Therefore, the investors choose the lower transaction costs of derivative financial products to invest in order to gain higher expected profits and slighter losses. The study focuses on the Taiwan index options in financial

derivatives. The data were collected from January 3, 2011 to October 28, 2016. We adopted the genetic programming, back propagation neural networks, genetic algorithms to construct the volatility forecasting model. Based on the construction, we created 20 prediction models. Thus, we can accurately predict the volatility of the closing price of the Taiwan Stock Exchange Capitalization Weighted Stock Index on the next trading day. in weighted index on the next day in Taiwan. Hence, we are able to estimate the reasonable price of the options and learn about the risk regarding the options’ investments.

The experimental results showed that the backpropagation neural network models can yield the best performance with the coefficient of determination 0.3. This implies that the constructed prediction models perform better than the models constructed previously, and can conform to the options’ markets. In addition, it revealed that tasks of forecasting volatilities are not simple works. This study can provide useful information regarding the options’s investments and lowering the investment risks for investors.

Keywords (3~5words) : volatility prediction, genetic programming, back propagation neural networks, genetic algorithms

Constructing an Optimal Portfolio by Using the MCDM and Krill Herd Optimization

Chih-Ming Hsu^{1*}

¹Department of Business Administration (No. 1, Hsin-Hsing Rd, Hsin-Fong, Hsinchu, Taiwan)

*cmhsu@must.edu.tw

Abstract

Making a profit via buying a stock and selling it at a low price and higher price, respectively, is intuitive, thus making stock investment is relatively easy while being compared with investing other commodities in the financial field. For an investor, it is very difficult task, however, to construct an optimal portfolio among many stocks. Therefore, this study intends to form an optimal portfolio through using the multiple criteria decision making (MCDM) and krill herd optimization techniques. Stocks in the finance and insurance subsector in Taiwan are used to illustrate our proposed procedure. Based on the experimental results, the average expected weekly returns on investment can attain 13.68%, which is much superior to the one-year certificate of deposit of about 1% in Taiwan. Hence, we can conclude that our proposed approach can fit the real-world stock market and can be used as a feasible, functional, and valuable tool for an investor.

Keywords : *Portfolio, multiple criteria decision making (MCDM), krill herd optimization*

1. Introduction

Relatively, investing stocks is easy while being compared to make an investment to other commodities in the field of financial investment. For an investor, however, it is an important and difficult work for choosing appropriate stocks to form an optimal portfolio. First, an investor must select stocks which are thought as potentially profitable from the

many kinds of stocks in the financial market. In the past, various multiple criteria decision making (MCDM) methods have been widely applied to evaluate the operating performance of corporations which issue the candidate investment stocks. For example, Wang *et al.*¹ had proposed a systematic method to assist organizations to determine the best sustainable development goals partner. They used a super slacks-based model, called super SBM, to rank companies, as well as to measure their efficiency, technical, and productivity changes in the past by using the Malmquist productivity index (MPI), and to forecast their future performance by using the GM (1,1) model. Then, the optimal green logistics partnerships and their competitiveness levels were identified through combining the data envelope analysis (DEA) and grey methods. Ghadikolaei *et al.*² had proposed a hybrid approach for evaluating the financial performance of automotive companies in the Tehran stock exchange (TSE) market. At first, they utilize the Fuzzy Analytic Hierarchy Process (FAHP) to find the optimal criteria weights. Then, these companies were ranked by simultaneously using the Fuzzy VIKOR, Fuzzy Additive Ratio Assessment (ARAS-F) and Fuzzy Complex Proportional Assessment (Fuzzy COPRAS) techniques. The experimental results indicate that the importance of economic value measures was higher than accounting measures in financial performance while these companies are evaluated.

After picking the investment targets from the

stock sea, investors must further determine the optimal portfolio by simultaneously maximizing the expected profit and minimizing the investment risk. Notably, the portfolio optimization problem is NP-hard. Therefore, it is a laborious task to obtain a hard solution, that is, a global optimum. Therefore, investors usually apply the soft computing techniques to acquire a soft solution, i.e. a near-optimal solution. For example, Kumar and Mishra³ presented a approach by mixing the co-variance principles with an artificial bee colony (ABC) to resolve the portfolio optimization problems algorithm with multiple conflictive objectives. Their method can provide various optimal trade-off solutions through simultaneously handling realistic constraints. Ni *et al.*⁴ revised the original particle swarm optimization (PSO) algorithm with dynamic random population topology abstracted into an undirected connected graph, which can be randomly generated based on some predefined rules and degrees. The implementation results showed that their proposed dynamic random population topology really can meliorate the calculation performance of the traditional PSO method significantly. Metawa *et al.*⁵ to propose a self-organizing method by applying the genetic algorithm (GA) for dynamically organizing bank lending decisions. The proposed model can consider the maximization of bank profit, as well as minimization of the probability of bank default to construct an optimal loan portfolio. Seyedhosseini *et al.*⁶ combined the harmony search (HS) and artificial bee colony (ABC) techniques to resolve a portfolio optimization problem by formulating a Markowitz mean-semivariance model. The efficiency and accuracy were demonstrated by comparing the efficient frontiers obtained by their proposed method to those provide by the HS and GA. According to the computational results, the proposed approach is better than HS and GA, and can yield the optimal solution.

Hajinezhad *et al.*⁷ had developed an artificial neural network (NN), called mixed Tabu machine (MTM), in which, the state transition mechanism is regulated by a Tabu search in both discrete and continuous search spaces for assisting the search process, and escaping from local minimum states of the energy, thus finding the global optimum. The Hang Seng in Hong Kong, DAX 100 in Germany, FTSE 100 in the UK, S&P100 in the USA and Nikkei 225 in Japan, are used to verify the efficiency of the MTM, and The experimental results revealed that the proposed MTM can yield excellent results within a very small CPU time.

2. Methodologies

2.1 Multiple criteria decision making

The analytic hierarchy process (AHP), introduced by Satty¹³, is a famous multiple criteria decision making (MCDM) tool. It's an effective structural tool for organizing and analyzing complex decision-making problems. Suppose a decision-making problem has m decomposed criteria and a decision maker has n alternatives. There are three steps in the AHP as follows. follows.

Step 1: Calculating the criteria vector

The AHP creates a criterion comparison matrix \mathbf{A} ($m \times m$) by comparing the relative importance of the criteria. The element a_{ij} in \mathbf{A} represents the importance of the i th criterion relative to the j th criterion. The i th criterion is more important than the j th criterion if $a_{ij} > 0$, and $a_{ij} < 0$ represents the opposite situation. The normalized criterion comparison matrix \mathbf{A}_{norm} ($m \times m$) is then calculated based on the following formula:

$$\bar{a}_{ij} = \frac{a_{ij}}{\sum_{k=1}^m a_{kj}}, i=1, \dots, m, j=1, \dots, m. \quad (1)$$

Finally, the criterion vector \mathbf{w} (m -dimensional column vector) can be obtained based on the following equation:

$$w_i = \frac{\sum_{k=1}^m a_{ik}}{m}, i = 1, \dots, m. \quad (2)$$

Step 2: Computing the score matrix

First, each alternative k is compared to the alternative l with respect to the i th criterion to form the element b_{kl}^i of the evaluation matrix $\mathbf{B}^{(i)}$ ($i = 1, 2, \dots, m$), which is an $n \times n$ matrix. The $b_{kl}^i > 1$ represents the k th alternative being better than the l th alternative. Notably, b_{kl}^i and b_{lk}^i must satisfy the constraints:

$$b_{kl}^i \times b_{lk}^i = 1, k = 1, \dots, n, l = 1, \dots, n \quad (3)$$

and

$$b_{kk}^i = 1, k = 1, \dots, n. \quad (4)$$

Next, the score matrix \mathbf{S} ($n \times m$) can be acquired as follows:

$$\mathbf{S} = [\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \dots, \mathbf{s}^{(m)}]. \quad (5)$$

Step 3: Alternative ranking

The global score vector \mathbf{v} ($n \times 1$) is calculated by

$$\mathbf{v} = \mathbf{S} \cdot \mathbf{w}. \quad (6)$$

where the k th element in \mathbf{v} indicates the global score that the AHP assigns to the k th alternative.

2.2 Krill Herd Optimization

In recent years, various nature-inspired metaheuristic approaches had been proposed to solve complex optimization problems. Usually, these algorithms are more powerful than the conventional methods which rely on the formal logics or mathematical programming Yang⁸. Among these approaches, the krill herd (KH) optimization algorithm Gandomi and Alavi⁹ is a very new metaheuristic relative to others. The KH is a bio-inspired method based on the simulation of the

herding behavior of the krill swarm in the nature. The time-dependent position of a krill individual is governed by three main actions including (1) movement induced by other krill individuals, (2) foraging activity, and (3) random diffusion. Hence, Gandomi and Alavi⁹ used the Lagrangian model to represent the herding process for the i th krill individual in an n dimensional decision space as follows:

$$\frac{dX_i}{dt} = N_i + F_i + D_i \quad (7)$$

where N_i is the motion induced by others; F_i is the foraging motion; D_i is the physical diffusion, regarding the i th krill individual.

According to Hofman *et al.*¹⁰, the krill individual tries to maintain its high density and moves due to their mutual effects. Therefore, the motion of the i th krill individual induced by other krill individuals can be defined as:

$$N_i^{new} = N^{max} \alpha_i + \omega_n N_i^{old} \quad (8)$$

where

$$\alpha_i = \alpha_i^{local} + \alpha_i^{target} \quad (9)$$

where N^{max} is the maximum induced speed that is usually taken as 0.01 (ms⁻¹) (Hofmann *et al.*¹⁰), ω_n is the inertia weight of the motion induced between 0 and 1, N_i^{old} is the last motion induced, α_i^{local} is the local effect provided by other krill individuals and α_i^{target} is the target direction effect given by the best krill individual. Notably, the local effect can be determined by assuming as an attractive tendency between the krill individuals. Define the following equations:

$$\hat{X}_{i,j} = \frac{X_j - X_i}{\|X_j - X_i\| + \eta} \quad (10)$$

$$\hat{K}_{i,j} = \frac{K_i - K_j}{K^{w\alpha\beta} - K^{b\alpha\beta}} \quad (11)$$

where X_i and X_j represent the position regarding the i th and j th krill individual, respectively

($i, j=1, 2, \dots, NN$); K_i and K_j represent the fitness (objective function value) of the i th and j th krill individual, respectively ($i, j=1, 2, \dots, NN$); K^{best} and K^{worst} are the best and the worst fitness values of the krill swarm so far. In addition, a small positive number, η , is added to the denominator to avoid 0; NN is the total number of the neighbor krill individuals. Notably, a krill individual is thought a neighbor of the i th krill individual if it locates within the sensing distance of the i th krill individual $d_{s,i}$ determined as follows:

$$d_{s,i} = \frac{1}{5N} \sum_{j=1}^N \|X_i - X_j\| \quad (12)$$

where N is the total number of the krill individuals. Hence, the local effect for the i th krill individual is determined only by its neighbors, i.e. the krill within the sensing distance of the i th krill, as follows:

$$\alpha_i^{\text{local}} = \sum_{j=1}^{NN} \hat{K}_{i,j} \hat{X}_{i,j} \quad (13)$$

For the effect of the best krill, i.e. the krill with the best fitness, on the i th krill individual is formulated as:

$$\alpha_i^{\text{target}} = C^{\text{best}} \hat{K}_{i,\text{ibest}} \hat{X}_{i,\text{ibest}} \quad (14)$$

where C^{best} is the effective coefficient of the best krill to the i th krill and is defined as:

$$C^{\text{best}} = 2(\text{rand} + \frac{I}{I_{\text{max}}}) \quad (15)$$

where rand is random value ranges from 0 to 1; I is the implementation iteration number; I_{max} is the maximum allowable implementation iteration number.

Next, the foraging motion contains two parts including the food location effect and the previous experience effect. First, define the center of food for each iteration as:

$$X^{\text{food}} = \frac{\sum_{i=1}^N \frac{1}{K_i} X_i}{\sum_{i=1}^N \frac{1}{K_i}} \quad (16)$$

Hence, the food attraction for the i th krill can be formulated as:

$$\beta_i^{\text{food}} = C^{\text{food}} \hat{K}_{i,\text{food}} \hat{K}_{i,\text{food}} \quad (17)$$

where C^{food} is the food coefficient of the i th krill and is calculated as:

$$C^{\text{food}} = 2(1 - \frac{I}{I_{\text{max}}}) \quad (18)$$

The effect of the best krill to the i th krill is then handled by the following equation:

$$\beta_i^{\text{best}} = \hat{K}_{i,\text{ibest}} \hat{K}_{i,\text{ibest}} \quad (19)$$

where the ibest is the best previously visited position of the i th krill. Therefore, the foraging motion can be formulated as follows:

$$F_i^{\text{new}} = V_f \beta_i + \omega_j F_i^{\text{old}} \quad (20)$$

where

$$\beta_i = \beta_i^{\text{food}} + \beta_i^{\text{best}} \quad (21)$$

and V_f is the foraging speed that is usually taken 0.02 (ms^{-1}) (Price¹¹), ω_j is the inertia weight of the foraging motion in the range [0, 1], F_i^{old} is the last foraging motion.

Finally, the physical diffusion of a krill can be considered to be a random process and be expressed by the following equation:

$$D_i = D^{\text{max}} \delta \quad (22)$$

where D^{max} is the maximum diffusion speed taken between 0.002 and 0.01 (ms^{-1}) (Morin *et al.*¹²); the δ is the random directional vector whose component's value lies between -1 and 1.

According to the above information, time, the position vector of a krill individual from the time t to the time $t + \Delta t$ can be formulated as follows:

$$X_i(t + \Delta t) = X_i(t) + \Delta t \frac{dX_i}{dt} \quad (23)$$

Notably, the Δt is a parameter that works as a scale factor of the speed vector and influences the searching space. The formula for determining the parameter Δt is as follows:

$$\Delta t = C_i \sum_{j=1}^{NN} (UB_j - LB_j) \quad (24)$$

where the NV is the total number of the decision variables, and LB_j and UB_j are lower and upper bounds of the j th variables ($j = 1, 2, \dots, NV$), respectively. The C_i is a constant ranging $[0, 2]$, and the lower the C_i , the more precisely the krill individual searches the solution space.

In addition, the genetic reproduction mechanisms, including the crossover and mutation, are also incorporated into the KH optimization algorithm to improve the searching performance. The crossover used in the KH is designed as follows:

$$x_{i,m} = \begin{cases} x_{r,m} & rand_{i,m} < Cr \\ x_{i,m} & else \end{cases} \quad (25)$$

and

$$Cr = 0.2\hat{K}_{i,best} \quad (26)$$

where $r \in \{1, 2, \dots, i-1, i+1, \dots, N\}$. The mutation method is defined by the following equations:

$$x_{i,m} = \begin{cases} x_{i,m} + \mu(x_{p,m} - x_{q,m}) & rand_{i,m} < Mu \\ x_{i,m} & else \end{cases} \quad (27)$$

and

$$Mu = 0.05\hat{K}_{i,best} \quad (28)$$

where $p, q \in \{1, 2, \dots, i-1, i+1, \dots, N\}$ and μ is a number between 0 and 1. Notably, both the crossover and mutation mechanisms converge to zero when the best fitness increases.

In conclusion, the general steps in the KH optimization algorithm can be depicted in Figure 1. (Gandomi and Alavi⁹):

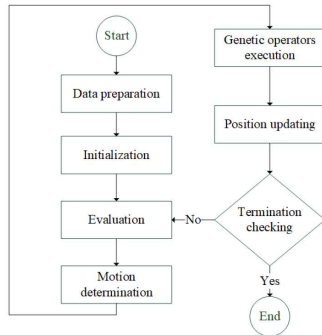


Figure 1. The general steps in the KH optimization.

3. Proposed Approach

This study develops a portfolio optimization procedure whose concept is briefly depicted in Figure 2.

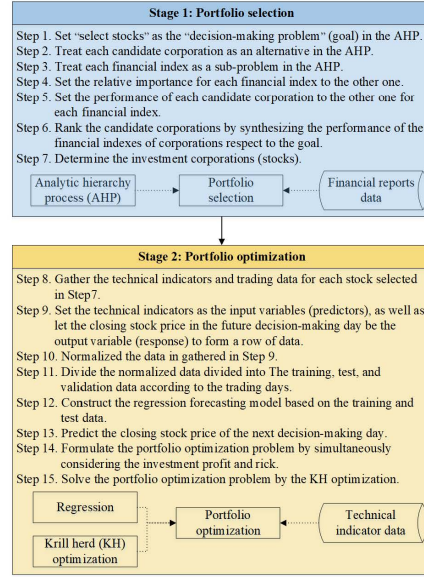


Figure 2. The proposed approach.

4. Case Study

In this study, the stocks in finance and insurance subsector of the Taiwan stock market are considered. The period of study is from 1 Jan. 2013 to 31 Dec. 2018. First, the financial reports announced at the end of the first quarter in 2018 were gathered from the TEJ (Taiwan Economic Journal) database for all the corporations which issue stocks in the finance and insurance subsector in Taiwan. Among the financial indexes, seven important indicators, including the earnings per share (EPS), return on assets (ROA), return on equity (ROE), gross profit margin (GPM), operating profit margin (OPM), debt ratio (DR), and price-earnings ratio (P/E) are chosen as the criteria in AHP. The selection of profitable stocks in a portfolio through the AHP then can be described as a

decision-making problem consisting of seven sub-problems. Each criterion (financial index) is considered as important as the other criterion (financial index). The corporations which are thought to have comparatively high operating efficiency then can be selected according to the scores obtained by synthesizing to the global goal, i.e. selecting stocks. Ten corporations are considered in this study through using the Expert Choice 11. The ten stocks include the stocks of represented codes F001~F010 according to their priorities to the global goal, are selected.

The technical indicators and trading data of the selected ten stocks were first collected from the C-Money and TEJ databases. Notably, there are 16 technical indicators, (1) the 10-day moving average, (2) 20-day bias, (3) moving average convergence/divergence, (4) 9-day stochastic indicator K, (5) 9-day stochastic indicator D, (6) 9-day Williams overbought/oversold index, (7) 10-day rate of change, (8) 5-day relative strength index, (9) 24-day commodity channel index, (10) 26-day volume ratio, (11) 13-day psychological line, (12) 14-day plus directional indicator, (13) 14-day minus directional indicator, (14) 26-day buying/selling momentum indicator, (15) 26-day buying/selling willingness indicator, and (16) 10-day momentum, considered in this study. For each trading day, the sixteen technical indicators along with the closing stock price of the trading day that is apart from the current trading day with a decision-making period are arranged into a row. These technical indicators and closing stock price serve as the input variables and output variable, respectively. Next, the data in each column are normalized into a range between 0 and 1 according to the maximum and minimum values of the corresponding column. Next, the arranged normalized data during 1 Jan. 2013 to 31 Dec. 2017, i.e. the model building period, to form the part (I) dataset. In addition, the part (II) dataset,

consisting of the arranged normalized data during 1 Jan 2018 to 1 Dec. 2018, i.e. the investment period, forms the validation data. The training and test data are generated by randomly dividing the part (II) data based on a proportion of 3:1. The SPSS regression tool is then applied to the training and test data to construct the forecasting model where the forecasting performance is evaluated through the cross-validation technique, and the parameters in the SVR are optimized by the grid-search approach.

Forecasting the closing stock price of the next decision-making (trading) day with a decision-making period, based on the technical indicators of the current decision-making (trading) day during the investment period, 1 Jan 2017 to 31 Dec. 2017, via the regression model, the KH, that is coded by C++, is implemented to resolve the portfolio optimization problem to determine the optimal capital allocation for each stock in the portfolio on the next decision-making day during the investment period. The KH procedure is implemented 10 times for each decision-making day.

According to the experimental results, the average expected weekly returns on investment is 13.68% for finance and insurance subsector. The interest rate for the one-year certificate of deposit in Taiwan is about 1%. Hence, the implementation results confirm that the proposed approach is a practical formatting a portfolio tool in the stock markets of the real world.

Acknowledgements

The author thanks the Minghsin University of Science and Technology, Taiwan, R.O.C., for supporting this study under Contract No. MUST-108BA-1.

References

1. C.-N. Wang, H.-X. T. Ho, S.-H. Luo and T.-F. Lin, An integrated approach to evaluating and selecting green logistics providers for sustainable development, *Sustainability-Basel*. **9**(2) (2017) Article ID 218.
2. A. S. Ghadikolaei, S. K. Esbouei and J. Antucheviciene, Applying fuzzy MCDM for financial performance evaluation of Iranian companies, *Technol. Econ. Dev. Eco.* **20**(2) (2014) 274–291.
3. D. Kumar and K. K. Mishra, Portfolio optimization using novel co-variance guided Artificial Bee Colony algorithm, *Swarm Evol. Comput.* **33** (2017) 119–130.
4. Q. Ni, X. Yin, K. Tian and Y. Zhai, Particle swarm optimization with dynamic random population topology strategies for a generalized portfolio selection problem, *Neural Comput.* **16**(1) (2017) 31–44.
5. N. Metawal, M. Elhoseny, M. K. Hassan and A. E. Hassanien, Loan portfolio optimization using genetic algorithm: a case of credit constraints, in Proc. 12th Int. Comput. Eng. Conf. (ICENCO), eds. IEEE (Cairo, Egypt, 2016), pp. 59–64.
6. S. M. Seyedhosseini, M. J. Esfahani and M. Ghaffari, A novel hybrid algorithm based on a harmony search and artificial bee colony for solving a portfolio optimization problem using a mean-semi variance approach, *J. Cent. South Univ.* **23**(1) (2016), 181–188.
7. E. Hajinezhad, S. Effati and R. Ghanbari, Mixed Tabu machine for portfolio optimization problem, *Int. J. Comput. Math.* **94**(6) (2017), 1089–1107.
8. X. S. Yang, Nature-inspired metaheuristic algorithms. Luniver Press, Frome, U.K. (2010).
9. A. H. Gandomi and A. H. Alavi, Krill herd: A new bio-inspired optimization algorithm. Communications, in *Nonlinear Sci. N. Simulation* **17**(12) (2012) 4831–4845.
10. E. E. Hofman, A. G. K. Haskell, J. M. Klinck and C. M. Lascara, Lagrangian modelling studies of Antarctic krill (*Euphasia superba*) swarm formation, *ICES J. Marine Sci.* **61**(4) (2004) 617–631.
11. H. J. Price, Swimming behavior of krill in response to algal patches: a mesocosm study, *Limnol. Oceanogr.* **34**(4) (1989) 649–659.
12. A. Morin, A. Okubo and K. Kawasaki, Acoustic data analysis and models of krill spatial distribution, Scientific Committee for the Conservation of Antarctic Marine Living Resources, Selected Scientific Papers Part I (1988) 311–329.

以多目標決策與磷蝦最佳化建構最

佳投資組合

徐志明^{1*}

¹ 明新科技大學企管系(新竹縣新豐鄉新興路1號)

*cmhsu@must.edu.tw

摘要

透過購買股票，並以低價買入和高價格出售以獲取利潤是一種直覺。因此，比起其他金融領域的投資商品，股票投資相對是較為容易。然而，對於投資者而言，在許多股票中藥構建一個最佳投資組合是非常困難的任務。因此，本研究欲透過以多標準決策和磷蝦最佳化技術以形成最佳投資組合。同時，並以台灣股票市場的金融保險類股票以說明我們提出的解題程序。依據實驗結果，平均預期每週投資回報率可以達到 13.68%，遠高於台灣一年期定期存款利率(約 1%)。因此，我們可以得出結論，我們提出的方法可以適應現實世界的股票市場，並可以作為投資者可行、實用和有價值的工具。

關鍵字：投資組合、多標準決策、磷蝦最佳化。