國立政治大學金融學系

碩士學位論文

機器學習結合波動聚集分析之投資策略實證研究

An Empirical Study of Investment Strategies Based on Machine Learning and Volatility Cluster Analysis

國立政治大學
National Chengchi University

指導教授：廖四郎 博士

研究生：李　強 撰

中 華 民 國 111 年 01 月

DOI:10.6814/NCCU202200358

摘　要

量化投資和機器學習在大數據時代充分展現了其獨特的優勢和魅力，兩者結合更是如虎添翼。機器學習不僅可以彌補量化投資的短板，還可以為量化投資的發展提供新的思路和方向。本文主要研究是否可以通過機器學習算法進行因子合成，以及該方法是否比傳統方法更有效。選取波動率、年化收益率、最大回撤、信息比率、夏普比率等評價指標進行因子分層回測結果的對比分析。

本文認為機器學習算法對量化投資和股票預測具有一定的重要性影響，為投資者的決策提供可行的解決方案。另外，本文詳細介紹了波動聚集現象和金融時間序列模型。然後本文將波動聚集性作為因子加入到策略的機器學習部分進行應用，以原策略為對照，並對回測結果進行詳細對比分析。

綜上可見，波動聚集現象的實際應用是值得研究的，這樣可以提高量化策略的性能和穩健性，同時對波動聚集現象的應用和發展提供了新想法。本文對 XGBoost 算法和量化投資中的波動聚集現象進行了研究和改進，為滬深 300 股票市場的量化投資者提供了新觀點。

關鍵字：機器學習，因子合成，波動聚集，量化投資

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202200358

Abstract

Quantitative investment and machine learning have adequately demonstrated their unique advantages and charms in this big data era and they become even more powerful by combination. Machine learning not only help to overcome the disadvantages of quantitative investment, but also provides new ideas and directions for its development. This paper mainly studies whether machine learning algorithm can be applied in factor composition, and whether its effect is better than the traditional method. Evaluation indexes such as volatility, annualized rate of return, maximum drawdown, information ratio and Sharpe ratio are selected to compare and analyze the results of factors' layered back test.

This paper points out that machine learning algorithm has guiding significance for quantitative investment and stock prediction, providing investors a feasible plan for decision-making. In addition, this paper introduces the phenomenon of volatility clustering and the financial time series model in detail. Then, this paper adds the volatility clustering as a factor to the machine learning part of the strategy, and compares and analyzes the back test results in detail by taking the initial strategy as a comparison.

We concluded that the practical application of volatility clustering is worth studying, which can improve the performance and robustness of quantitative strategy, and also provide a new idea for the application and development of volatility clustering. Our research and improvement on XGBoost algorithm and volatility clustering provide a new perspectives for quantitative investors in CSI 300.

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202200358

Keywords: Machine Learning， Factor Composition，Volatility Clustering，Quantitative Investment

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202200358

目　次

第一章　緒論.....7

第一節　研究背景.....7
第二節　研究動機.....7
第三節　研究目的.....8
第四節　研究架構.....9

第二章　文獻回顧.....12

第一節 國際相關文獻回顧.....12
第二節 中國相關文獻回顧.....14

第三章　研究方法.....17

第一節　集成學習算法.....17
第二節　XGBOOST 算法.....18
第三節　波動率模型.....19
第四節　本章小結.....23

第四章　因子合成與機器學習.....24

第一節　實驗架構設計.....24
第二節　資料來源與資料預處理.....25
第三節　不同算法的分析與結果.....27
第四節　本章小結.....40

第五章　機器學習與波動聚集.....41

第一節　研究設計.....41
第二節　機器學習與多因子策略的結合.....42
第三節　初始策略與波動聚集的結合.....46
第四節　本章小結.....58

第六章　結論與建議.....60

第一節　研究結論.....60
第二節　改進與展望.....61

參考文獻.....63

附錄.....66

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202200358

表　次

表 4-1　四種模型評價指標.....33

表 4-2　特徵排序及五分位組.....38

表 4-3　五組因子回測評價.....39

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202200358

圖　次

圖 3-1 集成學習演算法兩大分支.....17

圖 3-2 XGBoost 演算法流程示意圖.....18

圖 4-1 因子合成方案設計框架.....24

圖 4-2 分階段回測模型示意圖.....27

圖 4-3 左、右圖分別為隨機森林和 XGBoost 算法的執行結果.....28

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202200358

第一章　緒論

第一節　研究背景

隨著全球經濟的不斷發展，各國金融市場開放融入，各國的資本貿易日益頻繁便捷，金融全球化雖然帶來了經濟的高速發展，但不可避免地潛伏著巨大的風險。各個金融市場之間依賴性不斷地增強，金融體系穩定性就愈發衰弱，局部的經濟波動可能就會演變為席捲全球的金融危機。隨著世界經濟一體化進程的不斷推進，全球性金融危機便是接踵而來——80 年代的拉丁美洲債務危機、90 年代日本泡沫經濟崩潰、1997 年的東亞金融危機、2008 年美國次貸危機引爆的環球金融危機等等，無不伴隨著匯率動盪、股市暴跌、貨幣貶值等衰退現象。牽一髮而動全身的世界經濟，使得原本就多變的股票市場更加的變幻莫測。因此，科學地預測股市波動特徵，掌握系統性的投資方法對風險的降低和規避有著重要意義。

隨著中國經濟的快速發展，經濟全球化趨勢的加強，股票市場也是日益繁榮，股票投資已成為最具前景的理財方式的主要方式之一。然而福兮禍所倚，禍兮福所伏，巨大的收益總是伴隨著巨大的風險，而我們希望做到的，就是在獲得高收益的前提下，能最大程度地將風險降低到最小，或是在相同程度的風險中，取得最大的收益，最終實現投資效益的最大化。

第二節　研究動機

股市就好比是賭場，沒有人能百戰不殆，高風險高回報一直是股市的鮮明特點，不斷地吸引著投機者進入。變化莫測的股市能讓人一夜暴富，也能讓人一無所有，只有做一個理智的投資者，才能在股市中生存的更久。人的僥倖心理往往會阻礙投資者及時止損，從而造成了大量的損失，人類的情感往往會讓投資者失去理智。因此，我們需要冰冷的機器代替我們做決策，幫助我們成為一個“理智”的投資者。

伴隨著越來越多的上市公司湧入市場，如何在眾多股票中選擇優質股票是投

DOI:10.6814/NCCU202200358

資者面臨的又一個難題。為了降低投資風險以及投資成本的同時，保證更大的投資收益，選擇一個表現優秀的投資組合對投資者尤為關鍵。人腦處理資訊的能力是有限的，如果資本市場只有 100 檔股票，這對傳統投資是有利的，因為人腦可以深刻分析這 100 檔股票。但是面對日益擴大的市場交易量和擁有成千上萬檔股票的股市規模，讓人們很難依靠人腦的判斷得到一個有效的投資組合，這是一個長期的、複雜的工作。如何從股票海洋中挖掘出有潛力的股票，如何科學系統的選股，如何將這項複雜的任務化繁為簡，一勞永逸，也是一項意義重大的事情。

第三節　研究目的

量化投資是利用電腦技術並且採用一定的數學模型來實現投資理念和投資策略的過程。與國外成熟的市場相比，中國的市場仍屬於非有效或是弱有效市場，錯誤定價、非理性投資行為仍然屢見不鮮。因此對於量化投資的理論研究和實踐操作，構建系統科學的投資策略模型，通過量化投資策略提升廣大投資者的紀律性、系統性、及時性和準確性還大有可為，有著深遠重要的理論意義和現實意義。

但是，金融市場中蘊含著複雜的非線性行為，這給投資決策帶來了很大的困難，此外金融資料的非穩態問題，更是為投資者分析資料提高了難度，通過數學模型進行資料採擷實在是一項很複雜的工作。與此同時，巨大的交易資料、受限的電腦運算能力，若不合理的利用相關演算法，要從海量資料中進行挖掘，往往需要耗費大量時間。

而隨著大資料時代的到來，人們看到了將機器學習應用於量化投資中的優勢，機器學習不僅能夠提供非線性關係的模糊處理，彌補了人腦思維模式，還能利用這些演算法提高運行速度，大幅度的提高機器對於資料的學習效率，從眾多的金融資料中提取到有效地資訊，協助量化投資策略得出更好的投資組合。借助人工智慧，量化投資策略會變得更加豐富。

DOI:10.6814/NCCU202200358

儘管我們希望能通過量化投資來提高投資的有效性，但是在量化研究的過程也還存在各類的難題，其中一個比較現實的問題就是交易資料的獲取。要麼自己編寫爬蟲程式抓取交易資料，這種方式雖然成本較低，但是抓取的資料品質過於依賴爬蟲程式的優良性，同時還需要做大量的資料清洗工作，浪費了大量的時間成本。另外就是直接購買資料，但是從交易所購買資料金錢成本過高，而從資料服務商購買的資料雖然花費較少，但是品質良莠不齊。本文基於 JoinQuant 聚寬量化交易平臺[3]進行量化投資策略的研究，JoinQuant 量化交易平臺提供了全面且免費的回測資料，解決了當前量化投資研究中找資料難、回測難的現實問題。給廣大研究者在資料使用、模擬交易和策略回測等方面都帶來了便利，具有極大的實用價值。

第四節　研究架構

一、研究內容及框架

目前，機器學習已經應用在量化投資的許多方面，例如利用機器學習對 ST 股票進行分類，利用隨機森林演算法進行智慧選股，以及利用機器學習來預測股票的漲跌等等。大量的金融歷史資料為機器學習在量化研究的發展進程中提供了堅實的基礎和保障。通過文獻研究，主要歸納了以下幾點作為本文的研究基礎：其一，聚寬量化平擁有獨立安全的個人研究環境、免費使用的海量資料以及詳細精確的回測結果分析等優勢，學者們可以參考使用這樣的平臺進行量化研究。其二，經研究表明，XGBoost 演算法其預測精度略高於支援向量機演算法並且計算速度也比隨機森林更快。其三，波動聚集現象對投資有著重要的影響，當前大部分的研究主要還是對此現象進行描述和數量驗證，少部分研究者對波動聚集現象的成因進行了解釋以及度量了波動聚集現象的聚集程度，而將波動聚集現象加以應用的卻少之又少。在此基礎上，本文選用聚寬量化平臺作為研究平臺進行量化投資的學習研究。首先在聚寬量化平臺的研究環境中選取因子庫中的 30 個因子

DOI:10.6814/NCCU202200358

進行因子合成。然後採用機器學習構建一個多因子選股策略，並將波動聚集現象作為一個因子放入其中。最後對策略進行回測，針對回測結果對其進行優化調試並對回測的結果進行分析。

本文的章節安排主要分為以下五章：

第一章為緒論。主要介紹了本文的研究背景及研究目的，並對相關概念做了簡單的辨析。然後是研究設計部分，簡單地對本文的研究內容和框架以及本文的創新點進行了概括說明。

第二章為文獻回顧。主要對機器學習在量化上的研究應用以及波動聚集現象這些方面的文獻進行了歸納總結，為本文打下了夯實的研究基礎。

第三章為研究方法。主要就本文所涉及到的相關模型的理論知識做了具體的詳細介紹，主要包括了 XGBoost 演算法的理論知識、波動率模型尤其是金融時間序列模型的相關理論知識等，為後續的實證研究部分打下理論基礎。

第四章為機器學習與因子合成的結合。主要介紹了利用 XGBoost 演算法進行因子合成的具體步驟和方法。包括了資料的獲取、股票池的建立、資料預處理、模型的訓練以及優化調參等等步驟。其中還會對三種機器學習演算法進行對比，以及與傳統的線性模型進行對比分析。最後將所有因子按得分進行分組回測，驗證分析因子有效性，並進行總結歸納。

第五章為波動聚集與機器學習的結合。主要介紹了對於波動聚集現象創新性應用的研究設想和基本思路，並使用 GARCH 模型對滬深 300 指數的市場價格波動進行了擬合和預測，驗證了滬深 300 指數存在顯著的波動聚集現象。然後詳細地論述了 XGBoost 演算法結合多因子選股策略的構建過程，包括全域參數的設置、開盤前的局部參數設置和機器學習的訓練、開盤時的信號設置和倉位元調整，以及收盤時的成交記錄輸出。接著，創新地將波動聚集性作為一個因子加入到機器學習部分進行優化，對優化後的策略進行回測，與初始策略進行回測結果的對比分析，最後對實證研究結果進行總結歸納。

DOI:10.6814/NCCU202200358

第六章為總結與展望。對全文進行概括總結，簡述了全文的研究內容和實證研究方法，並歸納總結了本文的研究成果和結論，同時對本文的不足之處進行了展望。

二、論文創新點

本文的創新點主要有以下兩點：

一是將機器學習 XGBoost 演算法應用到因子合成中。傳統的方法是利用線性回歸進行因子合成，而引入 XGBoost 演算法後，可以將其過程看成一個映射函數，極大地改進了有效因子的篩選效果和效率。並且為了讓模型更貼合實際，學習到市場特徵變化，採用六階段滾動回測方法進行模型的訓練。

二是創新性地嘗試將波動聚集現象作為一個因子放入到機器學習中進行多因子選股。波動聚集現象對金融投資決策有著重大影響，是投資者不可忽視的關鍵因素。本文將其作為因子應用到智慧選股策略中，豐富了波動聚集現象應用方面的研究，為其實際應用提供了新的思路。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202200358

第二章　文獻回顧

量化投資按照交易策略的類型可分為量化選股、量化擇時、無風險套利、統計套利、高頻交易五大類。每一類在國內外學術界都有著大量的研究和成果。量化投資理論的基礎離不開數理金融的發展，很多金融模型都是量化投資的重要思想與核心。通過查閱資料和歸納整理，本文對量化投資以及機器學習在量化投資中的應用研究進行總結。

第一節　國際相關文獻回顧

Markowitz 在 1952 年提出了“均值—方差模型（Mean—Variance Model）”理論，第一次闡述了風險資產的收益率和風險之間的關係，為證券投資組合理論和量化投資理論奠定了堅實的基礎，也預示著量化投資時代的到來[5]。

隨著金融市場的不斷發展以及金融工具日益增多，金融衍生品的定價問題也隨之出現在研究學者的眼前。尤其是期權作為功能最多的金融工具，對於期權的定價問題，一直是金融數學中最複雜的問題之一。Black 和 Scholes 提出了期權定價模型（Black-Scholes Option Pricing Model），他們認為影響期權定價的因素分別是期權合約規定的合約時長、期權對應的股票價格、無風險利率以及期權合約的交割價格。巧合的是，Merton 也在近乎同一時間內發現了同樣的公式，並且還發現了很多涉及到期權方面的重要結論，因此此模型也可以稱為 Black- Scholes-Merton 定價模型[6]。

1986 年，Irwin 等人為了能夠使用該程式能夠發出有效的買賣交易信號，建立的自動化交易系統，為期貨基金的短線操作提供了有關材料幫助研究和瞭解[7]。Vapnik 於 1995 年提出了 SVM 模型，彌補了傳統線性預測方法對於線性系統分析的局限性[8]。Breiman 在 2001 年發表了一篇關於隨機森林的論文，隨機森林是在決策樹的基礎上發展而來，結合了集成學習原理，將許多的弱分類器通過並行的集合訓練，最終得到一個強分類器，這種集成學習的思想對模型性能有了大幅度的增強[9]。

Min 和 Lee 根據混合的特徵提取演算法來篩選股票指數的特徵，將該方法結合SVM 演算法來預測股票指數的趨勢。主要步驟是通過該篩選方法剔除掉不重要的特徵，根據得到的特徵向量構建一個預測模型，通過訓練的到的模型來預測股票指數，實驗結果表明，通過這種方法能夠取得較好的表現[10]。Kim 等

DOI:10.6814/NCCU202200358

人利用基於遺傳演算法的人工神經網路，建立了股票價格指數的預測模型。研究結果表明，該方法不僅可以達到降維的效果，同時還能夠提高預測精度[11]。Chen.Y.S 等人通過結合遺傳演算法來對 SVM 模型進行優化，不僅降低 SVM 模型的複雜性，並且達到了提高建模效率的目的，最後使用訓練得到的模型對股票價格走勢進行了預測，研究結果顯示，得到了該方法能夠取得優異的預測效果[12]。

Ming-Chi Lee 通過 SVM 演算法對股票進行分類，並使用機器學習演算法訓練模型，最後通過構建的模型預測股票的價格變化趨勢。研究結果表明，在預測股票價格走勢方面，SVM 演算法具有較高的預測準確率[13]。Majhi 等人為了預測股票指數趨勢，使用機器學習演算法中的回歸模型進行了模型的訓練，經過大量的實證研究分析，該方法在指數波動的短期和中長期趨勢的預測上取得了顯著效果[14]。

Ticknor 等人創新性的利用萊斯優化神經網路對證券市場行為進行了預測。通過貝葉斯正則化神經網路，將股票市場的價格資料以及一些技術指標資料作為訓練集進行模型訓練，預測未來的股票收盤價。研究結果表明，該模型不需要對資料進行預處理以及週期分析等大量的前提工作，最終也能取得一個較好的表現結果[15]。Somani 等人為了比較隱瑪律科夫（HMM）模型與傳統方法對股票的預測效果，通過大量的實證研究分析，使用隱瑪律科夫模型能夠對預測效果起到顯著的提升[16]。Poel 等人使用多種分類器演算法作為隨機森林演算法的弱學習器，基於歐洲的金融交易資料使用隨機森林演算法進行模型訓練來預測漲跌，最終取得了較高的預測精度[17]。

Tianqi Chen 和 Carlos 於 2016 年提出了 XGBoost 演算法，詳細地講解了XGBoost 的原理和數學邏輯，由於其出眾的效率、高度的靈活性和較高的預測精准度而逐漸風靡，取得了廣泛的關注和實用[18]。

對於波動聚集現象，最早是由 Mandlebrot 在 1963 年提出的，他發現資產價格的高波動率總是傾向於連續聚集的出現[19]。這種現象的存在就意味著計量經濟學中的許多時間序列模型就無法用來描述金融時間序列，因為多元線性回歸、自回歸移動平均（ARMA）模型都是基於同方差假設下的模型。之後就陸續地有學者提出各種方法試圖描述金融時間序列的這種波動特性。

直到 1982 年美國的經濟學家 Engle 首次提出了自回歸條件異方差（ARCH）模型，其基本思想是認為某時刻的誤差項是服從正態分佈的，資產收

DOI:10.6814/NCCU202200358

益的隨機擾動項是序列不相關的,但不是相互獨立的，它們的絕對值或平方值顯示出一定的正相關性。因此該模型主要應用於分析條件方差會隨著時間變化的時間序列[20]。自從 ARCH 模型被提出以來，在幾十年裡得到了迅速地發展和廣泛的應用，對於 ARCH 模型的應用研究不曾間斷過，是解釋預測金融市場波動以及研究金融時間序列的主流模型之一。Engle 也因 ARCH 模型的提出獲得了 2003 年的諾貝爾經濟學獎。

使用 ARCH 模型對金融時間序列進行擬合時，為了充分地描述資產收益的波動情況，往往需要很多的參數，這會導致擬合的效果降低。為了解決這個問題， Bollerslev 在 1986 年提出了在 ARCH 模型的基礎上增加了 p 階自回歸，得到了廣義自回歸條件方差（GARCH）模型，其基本思想是對 ARCH 模型的擴展，引入了一種允許條件方差轉化為一個 ARMA 過程的方法，這使 GARCH 模型特別適用於波動性的分析和預測，更好地刻畫了金融時間序列的波動聚集現象，對廣大的研究學者以及投資者具有指導性意義，因此 GARCH 模型在短時間內得到了大量的擴展研究以及廣泛的應用[21]。

之後，更多的學者對 GARCH 模型進行了擴展研究，提出了一系列 GARCH族模型。Zakoian 提出的 TGARCH 模型[22]以及 Nelson 提出的 EGARCH 模型[23],彌補了 GARCH 模型無法刻畫正負衝擊對金融資產收益和波動之間不對稱性的短板。此外，Bollerslev 又在 1996 年提出了 FIGARCH 模型用來刻畫金融時間序列的長期記憶性[24]。

2001 年，Michael 等人分別使用 ARCH、GARCH、APGARCH 模型對倫敦期貨交易所所有的期貨合約價格波動進行了研究，研究發現對於倫敦期貨交易所的期貨價格而言，GARCH 模型的擬合效果要優於其它兩種模型[25]。2007 年，Chin Wen 使用 GARCH 模型研究馬來西亞股票市場的價格波動特徵，研究發現馬來西亞股票市場的價格波動存在顯著的波動聚集性[26]。

第二節　中國相關文獻回顧

吳微等人研究分析了 A 股市場的特徵，利用 BP 神經網路對其進行分類，通過人工神經網路對滬市綜合指數的漲跌進行了預測。研究結果表明，人工神經網路是可以應用於股票市場投資的，並且具有較高的使用價值[27]。

彭麗芳等人使用機器學習演算法中的 SVM 演算法對股票收盤價進行預測，

DOI:10.6814/NCCU202200358

結合時間序列構建了回歸模型。實驗結果表明，與傳統的線性時間序列相比，結合機器學習演算法的預測模型擁有更高的預測精度[28]。

錢穎能[29]等人以及左輝[30]等人，基於不同的研究資料，都採用基於貝葉斯分類方法構建了選股模型，並通過模型獲得了超越基準收益率的超額收益。他們的研究表明，機器學習演算法能夠對投資策略起到較好的幫助，能夠幫助投資者更好的進行投資決策。

余樂安和汪壽陽為了對股票類別進行分類，獲取了滬深股票市場中的 20 支上市公司的股票資料，採用 K 均值聚類演算法對 20 支股票資料進行聚類，最終分析得到了 20 支股票的分類類別[31]。吳曼琪運用 K 均值聚類法對 ST 股票的分類進行了研究，選取了 A 股市場中 43 只 ST 股票樣本進行聚類並檢驗分類準確率，最終將 ST 股票分為四類，並對四類 ST 股票分別給出投資建議[32]。

劉毅為了研究因子選股模型，獲取了 25 個可能影響股票收益率的因子資料，用 2000 年到 2012 年的中國 A 股資料進行了實證研究，用其中表現較好的 8 個因子構建成最優因子組合，與幾種典型的投資方法構建的價值組合、成長組合等進行比較。研究結果表明，通過表現較好的 8 個指標構建出的最優因子組合在策略投資中的表現最好，優於其他幾種傳統策略[33]。

蘇治等人基於滬深股票市場的交易資料，利用核主成分遺傳演算法對基於 SVM 演算法的人工智慧選股模型進行了改進，通過分析訓練構建了改進的選股模型，並對其選股性能和預測精度進行了實證研究。研究結果表明，改進後的選股模型不僅提高了模型的預測精度，還獲得了更好的模型穩健性[34]。

曹正鳳等人基於 2012 年到 2013 年間的 363 檔股票資料，提出了以價值成長投資策略理念為基礎的選股模型，通過等頻演算法對樣本資料進行離散化處理，使用機器學習演算法中的隨機森林演算法進行選股模型的訓練，研究結果顯示，該方法可以提高股票分類的準確率[35]。

王淑燕等人選用了 2013 年 3 月的 200 檔股票交易資料，基於六因子模型提出了八因子選股模型，通過機器學習演算法中的隨機森林對模型進行了訓練，並對下個月的股票價格漲跌進行了預測，研究結果顯示，該模型能夠取得較高的準確率，另外從平均收益等方面對該模型與六因子模型的進行比較分析，證明了該模型在我國股票市場上擁有不錯的表現[36]。

王莉通過支援向量機演算法、深度神經網路演算法等人工智慧領域的演算

DOI:10.6814/NCCU202200358

法將傳統的預測價格漲幅的回歸問題，轉化為一個二類分類學習問題，減少了股票價格波動中高噪音、冗餘、非線性等問題因素所帶來的預測過程的複雜度和困難度。研究結果表明，該方法取得了非常理想的預測效果[37]。

張昊等人通過挖掘電子商務網站的使用者資訊，利用機器學習演算法構建分類模型，實現對使用者購買行為的預測分類。實驗過程中，還對 XGBoost 演算法、決策樹和隨機森林演算法進行了模型的對比，研究結果顯示，三種分類演算法中 XGBoost演算法預測效果最好，演算法運行速度也最快[38]。

淩筱玥採用 XGBoost 演算法對上證綜指、上證 50 指數以及標普指數的漲跌進行預測，實驗過程中對 XGBoost 演算法、SVM 演算法和決策樹模型進行了對比分析，研究結果顯示，XGBoost 演算法的預測效果最好，要優於 SVM 演算法和決策樹，並且表明使用該模型能夠使投資者獲得較高的超額收益[39]。

田浩以滬深 300 指數作為股票池獲取所需的歷史財務資料，利用 XGBoost 演算法進行模型訓練構建選股模型，選出了其中預測結果較好的 30 檔股票進行風險評估，並確定股票在投資組合中的最佳權重，最終結果表現不錯。實驗過程中，還對 XGBoost 演算法、隨機森林演算法、SVM 演算法以及 LR 演算法進行了比較，實驗結果表明 XGBoost 演算法在模型預測準確度是最高的[40]。

唐齊鳴等人使用 t 分佈 GARCH 模型對我國的滬深股票市場的價格波動特徵進行了研究，研究表明我國滬深股票市場的價格波動存在不對稱性、長期記憶性以及波動聚集性[41]。惠曉峰等人運用 GARCH 模型對我國人民幣的美元匯率進行了研究和預測，研究表明 GARCH 模型對於我國人民幣匯率的預測效果較好，並且匯率存在顯著的自相關性和異方差性[42]。曹野利用 GARCH 模型對我國現貨黃金價格的波動特徵進行了研究分析，研究表明我國現貨黃金價格的收益率具有尖峰厚尾的特徵以及存在顯著的波動聚集現象[43]。

周好文等人對波動聚集現象的產生原因進行了研究解釋，他們認為由於市場投資者過於關注已經發生的股票價格的波動噪音，並將這些波動噪音作為股票投資策略的一部分資訊進行了使用，因此在股價中會反映出前期的股價資訊，產生了波動聚集現象[44]。

DOI:10.6814/NCCU202200358

第三章　研究方法

第一節　集成學習算法

機器學習發展至今，演算法已經發展的較為成熟。比較常見的演算法有邏輯回歸、K-近鄰、SVM、隨機森林等等，都是各有各的優勢。2016 年陳天奇博士提出了XGBoost 演算法，可謂是異軍突起，其出眾的效率、高度的靈活性和預測精度，短短三年時間已經快速的佔領了市場，受到了廣大研究者的喜愛。經過大量研究者的研究發現，目前分類演算法中表現最好的為 SVM、隨機森林和 XGBoost 演算法。並且已經有大量的研究表明，XGBoost 演算法其預測精度和速度都是優於 SVM 和隨機森林的。因此，本文不再過多的對三種演算法進行比較證明，而是基於前人的研究結果，簡單地對三種演算法的精確度和 AUC 進行比較，並計畫決定採用XGBoost 演算法作為選股模型進行因子合成。

人多力量大，對機器學習模型來說也是如此。單個弱學習器的預測能力有限，但是將許多個弱學習器集成一個強學習器，就能“頂個諸葛亮”了。集成學習演算法主要有兩大類，如圖 3 - 1 所示，一類為 Bagging 系列，屬於並行方法，另一類為 Boosting 系列，是串列方法。對於多顆決策樹的演算法，如果用 Bagging 的方式組合起來，可以得到隨機森林演算法。如果用 Boosting 的方式組合起來，可以得到 GBDT 等方法。

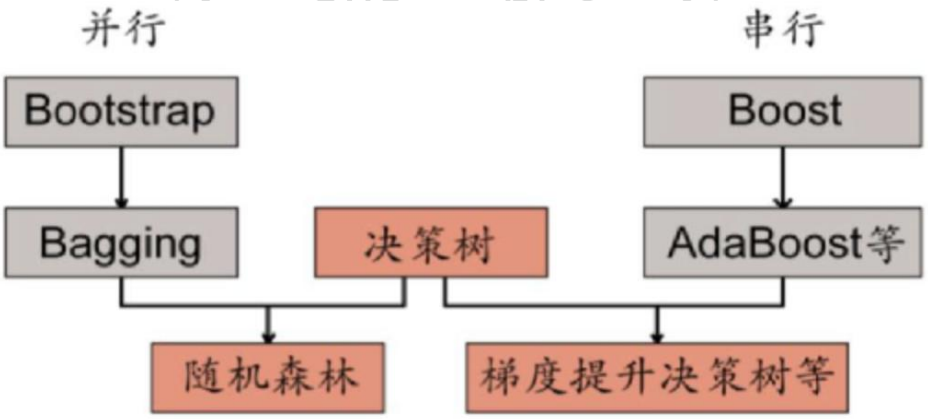


圖 3 - 1：集成學習演算法兩大分支

DOI:10.6814/NCCU202200358

第二節　XGBoost 算法

提升演算法 AdaBoost 演算法和梯度提升決策樹 GBDT 演算法在業界中已經得到了廣大的認可，具有良好的建模效果。而 XGBoost 演算法就是在這兩種演算法基礎上優化而來的，是對 GBDT 演算法的改進和提高。與傳統的 GBDT 演算法相比，XGBoost演算法在損失函數、切分點查找、正則化以及並行化設計等方面進行了提升和改進，使它在計算速度上更加快速。例如，對於 GBDT 演算法來說，在訓練第 n 棵樹時就必須用到第（n-1）棵樹的殘差，因此導致該演算法很難並行化。XGBoost 演算法對此進行了改進，通過對目標函數做二階泰勒展開，使得最終的目標函數只依賴每個數據點上損失函數的一階導和二階導，從而實現了並行，提升了演算法速度。如圖 3 - 2 所示，為 XGBoost 的演算法流程，它與 GBDT 在數學上的不同之處在於訓練每個弱學習器時的目標函數，其目標函數如下文所示，其中 $\Sigma_{i=1}^{n} L(y_i,\hat{y}_i)$ 為損失函數的二階泰勒展開，$\sum_{k=1}^{K} \Omega(f_k)$ 為正則項，控制模型複雜度。由於節點分裂，因此支援並行和特徵抽樣，每次分裂不需要窮舉掃描所有閾值，計算速度大幅度提升。

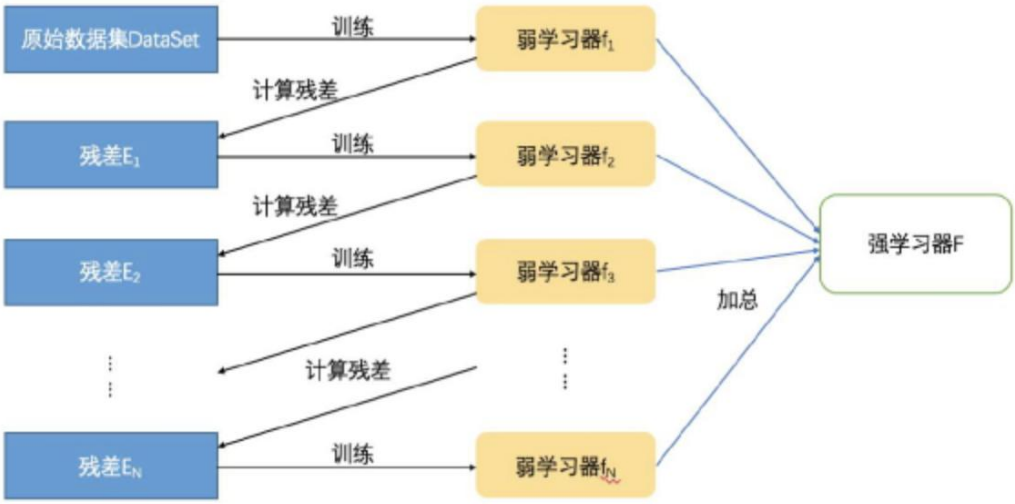


圖 3 - 2：XGBoost 演算法流程示意圖

目前主流的決策樹演算法包括 C4.5 和 CART 兩種，C4.5 每個節點可以分裂成多個子節點，但是不支援特徵的組合，只能用於分類問題。CART 樹每個節點只分裂成兩個子節點，但是能夠支援特徵的組合，可以用於分類和回歸問題。本文主要介紹以 CART 樹為基學習器的 XGBoost 演算法。其基本思想為，當任意給一個樣本點，可以計算得到該樣本在每棵 CART 樹上的得分，加

DOI:10.6814/NCCU202200358

總後就是該樣本的最終得分，由此為依據進行預測或是進行分類。XGBoost 模型的數學公式和目標函數為：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

$$Obj(\Theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

其中 n 為樣本個數，x_i為第 i 個樣本，y_i和$\hat{y}_i$分別為第 i 個樣本的真實值和預測值，K 為 CART 樹的個數，f_k 為第 k 棵 CART 樹，$L(y_i, \hat{y}_i)$為損失函數，$\Omega(f_k)$為正則項。在訓練第 t 棵 CART 樹時，相當於最小化第 t 棵樹的目標函數，而 XGBoost 算法的特別之處就是可以用損失函數的二階泰勒展開來近似原來的損失函數。

第三節 波動率模型

波動率無法直接從市場中觀測獲取，只能通過對已知的市場價格進行運算得出，可以分為歷史波動率和隱含波動率。大量的實證研究表明，波動率存在尖峰厚尾、波動聚集效應、波動率微笑等現象。通過 Black-Scholes 定價公式和市場上觀測到的期權價格，可以反解出參數 σ，這個結果就稱為隱含波動率，代表了在 B-S 模型的前提下，市場上蘊含的波動率資訊。隱含波動率不僅僅只是純粹的股價波動，而是包含了股價、利率、剩餘期限等等無法解釋的所有其它因素在內的資訊。

一、 BS 模型

1973 年芝加哥期權交易所成立後，對期權交易進行了統一的標準化，此後期權市場迅速發展。歐式期權形式簡單，又富有流動性，能夠為市場訊息校準。因此，期權定價研究首先是以歐式期權為目標。BS 期權定價模型一經提出就在理論研究和金融實踐中得到了巨大的關注和防範的應用。由於期權定價的前提假設是無摩擦市場，也稱完美資本市場。BS 模型有 7 個重要的假設條件：

（1）股票價格遵循幾何布朗運動；

（2）無風險利率是恒定的；

（3）不存在稅收和交易成本；

（4）在期權有效期內，金融資產無紅利；

（5）市場為風險中性；

DOI:10.6814/NCCU202200358

（6）不存在無風險套利機會；

（7）交易可以持續進行。

由假設中幾何布朗運動的隨機微分方程，結合風險中性原理，通過伊藤引理可以推導出歐式期權的定價公式如下：

$$C = S \cdot N(d_1) - e^{-rT} K \cdot N(d_2)$$

其中，

$$d_1 = \frac{\ln\left(\frac{S}{K}\right) + (r + \frac{\sigma^2}{2})(T)}{\sigma\sqrt{T}}$$

$$d_2 = \frac{\ln\left(\frac{S}{K}\right) + (r - \frac{\sigma^2}{2})(T)}{\sigma\sqrt{T}}$$

C 為期權的初始價格，S 為金融資產現價，K 為執行價格，T 為期權有效期，r 為連續複利形式無風險利率，N(·)為標準正太隨機變數的分佈函數，σ^2 為年度化方差。

BS 模型完美解決了歐式期權的定價問題，它同時取得了數學與實踐的巨大成功，榮獲諾貝爾經濟學獎。儘管它的嚴重缺陷已經眾所周知，但它依然在工業界衍生品市場上被廣泛的使用至今。

二、隨機波動模型

由於 BS 模型的假設過於理想，經歷了股災的研究學者為了掙脫 BS 模型的桎梏，對 BS 模型進行了修正。1993 年 Heston 根據 CIR 利率模型，採用了 Stein and Stein（1991）的建議，假定波動率服從 Ornstein-Uhlenbeck 過程，提出了隨機波動模型，即 Heston 模型。

假設股票價格過程S_t及其方差過程V_t服從以下擴散過程：

$$dS_t = \mu S_t dt + \sqrt{v_t} S_t dW_t^1$$

$$dV_t = \kappa(\theta - V_t)dt + \delta\sqrt{V_t} dW_t^2$$

$$dW_t^1 dW_t^2 = \rho_t dt$$

其中，μ是漂移項，θ為長期波動率期望，κ是對長期波動期望的調整係數，δ是方差變異係數，W_t^1和W_t^2是維納過程，ρ是維納過程的相關係數。

由以上三個公式和伊藤引理可得期權在 t 時刻的價值函數 $G(S,v,t)$ 滿足如下的偏微分方程：

DOI:10.6814/NCCU202200358

$$\frac{1}{2}\nu S^2\frac{\partial^2 G}{\partial^2 S}+\rho\delta\nu S\frac{\partial^2 G}{\partial S\partial\nu}+\frac{1}{2}\delta^2\nu\frac{\partial^2 G}{\partial^2\nu}+rS\frac{\partial G}{\partial S}-rG+[\kappa(\theta-\nu)-\Lambda]\frac{\partial G}{\partial\nu}+\frac{\partial G}{\partial t}=0$$

其中，$\Lambda=\Lambda(S,\nu,t)$是市場價格的波動率風險。且上式滿足以下邊界條件：

$$\begin{cases} G(S,\nu,t)=max(0,S-K) \\ G(0,\nu,t)=0 \\ \frac{\partial G}{\partial S}(\infty,\nu,t)=1 \\ rS\frac{\partial G}{\partial S}(S,0,t)+\kappa\theta\frac{\partial G}{\partial\nu}(S,0,t)-rG(S,0,t)+G_t(S,0,t)=0 \\ G(S,\infty,t)=S \end{cases}$$

記$C(S,\nu,t)$為歐式看漲期權的價格，由（2.15）和（2.16）式求解可得：

$$C(S,\nu,t)=SP_1-Ke^{-rt}P_2$$

三、時間序列模型

時間序列是金融研究中最為常見的資料類型之一，較為常見的研究時間序列的模型有 AR、MA、ARMA、ARIMA 模型等等。在大多數情況下，AR 模型和 MA 模型都很難單獨地描述以及預測一個時間序列，因此將 AR 模型和 MA模型相結合，改善了預測效果，是研究平穩時間序列的重要方法之一，推廣後的ARMA(p，q)模型的一般形式為：

$$r_t=\phi_0+\epsilon_t+\sum_{i=1}^{p}\phi_i r_{t-i}+\sum_{i=1}^{q}\theta_i\epsilon_{t-i}$$

其中，p 是 AR 面模型的階數，q 是 MA 模型的階數，誤差項$\{\epsilon_t\}$是均值為0，方差為 σ^2 的白色雜訊序列。

ARMA 模型主要是基於同方差的假設，但現實情況下，金融時間序列往往都存在著波動聚集現象，同方差假設一般是很難滿足的一種條件。因此考慮波動率的變化是必須面對的問題，而 ARMA 模型不足以擬合一般的金融時間序列。前人學者在此情況下引入了關於隨機擾動項方法的模型，即條件異方差模型（ARCH 模型）。

（一）ARCH 模型

1982年 Engle 提出了ARCH（Auto-Regressive Conditional Heteroskedasticity，ARCH）模型，他認為資產收益的隨機擾動項的條件方差存在序列相關性，該模型假設 t 時刻的隨機擾動項可以用其延遲值的簡單二次函數來描述，一個 q 階的 ARCH 模型形式如下：

DOI:10.6814/NCCU202200358

$$a_t = \sigma_t \epsilon_t$$
$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \alpha_2 a_{t-2}^2 + \cdots + \alpha_q a_{t-q}^2$$

其中，$\{\epsilon_t\}$是均值為0，方差為1的獨立同分佈隨機變量序列。σ_t^2為ϵ_t的條件方差，即波動率。$\alpha_0 > 0$，$\alpha_1,\ldots,\alpha_q \geq 0$，確保條件方差非負，$\sum_{i=1}^{q} \alpha_i < 1$確保該過程為平穩過程。從表達式中可以看出，過去的平方擾動$\{a_{t-i}^2\}$會影響方差項$\{\sigma_t^2\}$，即大的擾動項會傾向於影響下一個擾動項變大，也就是大的波動往往會緊跟著出現另一個大的波動。因此這個模型較好的刻畫了波動聚集現象。

（二）GARCH 模型

在許多學者的大量實證研究中可以發現，存在很多的金融時間序列，它們的隨機擾動項的條件方差往往是長期自相關的。在使用 ARCH 模型進行擬合時，為了充分地描述資產收益的波動情況，需要很多的參數，這會導致擬合的效果大大降低。為了解決這個問題，1986 年 Bollerslev 提出了在 ARCH（q）模型的基礎上增加了 p 階自回歸，得到了廣義的 ARCH（Generalized Auto-Regressive Conditional Heteroskedasticity，GARCH）模型，即 GARCH（p，q）模型來修正 ARCH 模型的不足，它可以用較少的參數來描述高階的 ARCH 模型。其具體形式如下：

$$a_t = \sigma_t \epsilon_t$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

其中，$\{\epsilon_t\}$ 是均值為 0，方差為 1 的獨立同分佈隨機變量序列。p 是 GARCH 模型的階數，q 是 ARCH 模型的階數。且$\forall_i > 0, \alpha_0 > 0, \alpha_i \geq 0, \beta_j \geq 0$，確保條件方差非負，$\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_j) < 1$ 確保該過程為平穩過程。若α_i和β_j都顯著大於零，則可以說明條件方差和前期確實存在正相關，即前期的波動會對未來的波動產生影響，可以驗證波動聚集現象。然而 GARCH 模型定階比較困難，在實際的應用中，一般只用到低階的 GARCH 模型，例如 GARCH(1,1)、GARCH(1,2) 以及 GARCH(2,1) 等。

DOI:10.6814/NCCU202200358

第四節　本章小結

本章主要對機器學習中的集成學習演算法做了簡單的介紹，重點介紹了集成學習演算法中的 XGBoost 演算法的相關理論知識。另外，詳細地介紹了幾個重要的波動率模型，包括 BS 模型、隨機波動模型以及幾個時間序列模型。主要目的是為下文的實證研究打下理論基礎。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202200358

第四章　因子合成與機器學習

基於學者們的文獻資料研究，傳統的因子合成是根據一元或多元線性回歸擬合得到，其主要思路是在橫截面上將股票收益對上期因子暴露進行回歸求解得到。目前最常用的因子合成方法有等權合成法、歷史收益率加權法、最大化 ICIR 加權法等。IC（Information Coefficient）為資訊係數，代表預測值與真實值之間的相關性，通常用於評價預測能力，IC 值越大，表示因子預測能力越好。IR（Information Ratio）為資訊比率，是超額收益的均值與標準差之比。可以根據 IC 近似計算，等於多個調倉週期的 IC 均值除以這些 IC 的標準差。

第一節　實驗架構設計

本文利用機器學習模型，將傳統的因子合成方法通過模型訓練來代替。主要方案設計框架如圖 4 - 1 所示，可以分為資料的獲取、資料的預處理、樣本內訓練、交叉驗證以及樣本外測試等。方案中採用隨機森林與 XGBoost 模型分別進行因子合成，通過準確率、AUC 等指標對模型進行分析比較，同時用傳統的線性模型做對比。並且，為了能讓模型可以學習到市場的特徵變化，本文在訓練模型時採用 6 階段滾動回測的方法。

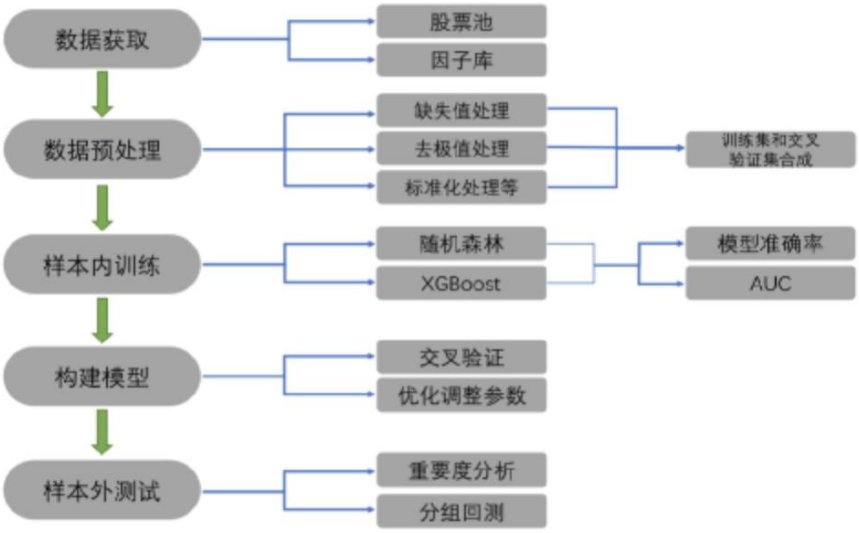


圖 4 - 1：因子合成方案設計框架

DOI:10.6814/NCCU202200358

第二節　資料來源與資料預處理

本文選取滬深 300 作為股票池。滬深 300 是上海證券交易所和深圳證券交易所在 2005 年 4 月 8 日聯合發佈的，由 300 檔股票構建而成，具有嚴格的選樣標準，每個成分股的權重是由分級靠檔法來確定的，成分股所處的行業分佈狀況與滬深兩市的行業分佈比例基本一致。滬深 300 指數每年對樣本股進行兩次調整，穩定性要求較高，並設置調整緩衝區，因此滬深 300 指數具有較高的代表性和穩定性，能夠反映市場主流的交易情況。滬深 300 指數盈利能力較強，成長性較高，聚集了大量的優質股票，是反映滬深兩市 A 股市場行情的綜合指數，也是反映我國股票市場整體形勢的“晴雨錶”。為了確保樣本穩定可靠，依然對股票池做一些過濾處理，例如剔除 ST 股票、每個截面期下一交易日停牌的股票以及上市 3 個月內的股票。

本文選取了聚寬因子庫中的一部分因子進行因子合成，其中包括了財務品質因子、成長因子、動量因子、風險因子、基礎因子、技術指標因子、情緒因子等。聚寬因子庫中共有 200 多種不同的因子，許多因子之間存在共線性，作用差異不大。同時，為了減小研究環境的運行負擔以及縮短執行時間，在本文中只選取其中 30 個在各類因子中具有代表性的因子作為研究對象，因子信息見附錄。其中 TTM 的全稱為 Trailing Twelve Months，是流量指標的一種滾動計算方法。

本文使用的資料都來源於聚寬量化平臺，因此免去了前期大量繁雜的資料清洗工作。據統計，資料處理工作費時費力，可能佔據了建模過程中 50%以上的時間，可以看出，專業可靠的資料能節省一個專案大量的寶貴時間。資料預處理主要包括異常值處理、缺失值處理、資料標準化等步驟。

異常值處理一般是資料處理的第一步，對後續的缺失值處理、資料標準化都會造成一定的影響。檢驗異常值的方法主要有 3σ原則和箱型圖法。3σ原則中認為，特徵的觀察值和平均值的偏差超過 3 倍標準差的，可以判定為異常值。箱型圖法認為，大於上邊緣或者小於下邊緣的值可以判斷為異常值。上邊緣等於上四分位數加上 1.5 倍的四分位間距，下邊緣等於下四分位數減去 1.5 倍的四分位間距，四分位間距等於上四分位數與下四分位數的差。異常值處理不是簡單的刪除異常值，因為刪除異常值不僅會減少樣本量，並且當資料量不多的時候，刪除異常值可能會影響資料的結構，對後續的建模造成影響。處理異常

DOI:10.6814/NCCU202200358

值需要考慮整體情況，結合實際考慮問題。本文使用聚寬量化平臺中的資料處理函數 winsorize 函數進行標準差倍數去極值的處理，最大限度的保留資料資訊。大於上界值的用上界值替代，小於下界值的用下界值替代。上界值等於中位數加上 5 倍的標準差，下界值等於中位數減去 5 倍的標準差。

資料缺失是樣本普遍存在的一種現象，任何資料都難免存在缺失值。當樣本缺失值過多時，可以考慮放棄該樣本，缺失值較少時，可以考慮對樣本進行補全。本文在對缺失值進行填補時，首先是使用行業平均值進行替代，但依然會存在一些缺失值無法補全，此時則使用所有行業的平均值替代。

資料標準化是機器學習建模前的基本工作之一，也是資料預處理較為重要的一步。不同特徵之間的量綱不同，單位不同，導致資料的數值差異巨大。例如成交額、成交量可能達到成百上千萬，但是漲跌幅只有百分之幾。如果不進行標準化處理直接建模的話，必然對結果產生巨大的錯誤影響。因此需要對資料進行標準化處理，將所有資料的數值按照一定比例進行縮放，常見的方法就是將每個值都映射到[-1,1]或者[0,1]內，這樣就達到了資料標準化的目的。常用的資料標準化方法有以下三種:

（1）最小-最大規範化

最小-最大規範化的原理是通過線性變化實現對資料的歸一化處理，將數值映射到[0，1]內。其轉換公式為:

$$x^* = \frac{x - min}{max - min}$$

其中 max 表示樣本中的最大值，min 表示樣本中的最小值，max-min 表示樣本極差。它的優點是在達到歸一化的前提下能夠保留資料之間的關係。而缺點就是如果資料較為集中且數值較大，則會導致標準化之後的值近乎於 0 且相差不大

（2）零-均值標準化

零-均值標準化也叫標準差標準化，其處理後的資料，均值為 0，標準差為 1，其轉換公式為:

$$x^* = \frac{x - \bar{x}}{\sigma}$$

其中 $\bar{x}$ 表示資料的均值，σ 表示樣本的標準差，是目前最為常用的標準化

DOI:10.6814/NCCU202200358

方法。當樣本數量較大時，標準化之後的資料會呈現為標準正態分佈。

（3）小數定標規範化

小數定標規範化的原理是通過移動樣本值的小數位數，將樣本屬性值映射到[-1，1]內，其轉換公式為：

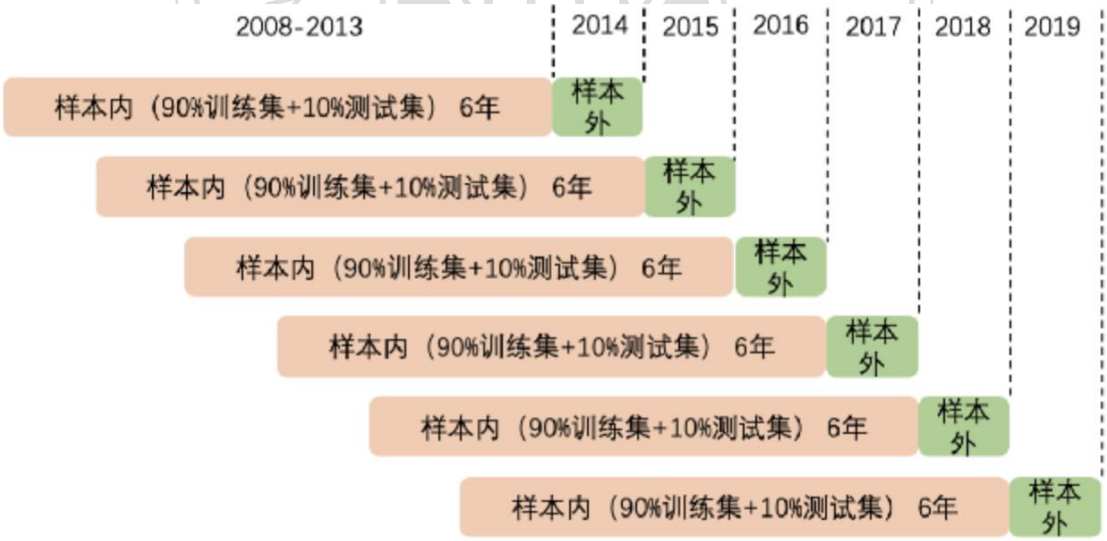
$$x^* = \frac{x}{10^k}$$

其中移動小數位數 k 由該屬性絕對值的最大值決定。

第三節　不同算法的分析與結果

一、 模型訓練與結果分析

根據上文的設計方案，本文將獲取 2008 年到 2020 年的資料進行訓練與回測。為了能讓模型更好地學習市場的特徵變化，本文在訓練模型時採用六階段滾動回測的方法。如圖 4-2 所示，2008 年 1 月到 2013 年 12 月共 6 年的資料作為第一個樣本區間，2009 年 1 月到 2014 年 12 月作為第二個樣本



區間，以此類推，2013 年 1 月到 2018 年 12 月作為第六個樣本區間。

圖 4-2：分階段回測模型示意圖

在獲取每個樣本區間的每檔股票（樣本）的因子資料的同時，在每個月末截面期獲取該檔股票的收益率，然後根據收益率對樣本進行標注。選取樣本收益率排名前 30%的股票作為正例，標注為“1”（label=1）；收益率倒數 30%的

DOI:10.6814/NCCU202200358

股票作為負例，標注為“-1”（label=-1），剔除收益率處於中間位置（未標記）的股票，降低雜訊影響。

將打上標籤的樣本進行合併，隨機選取 90%的樣本作為訓練集，其餘 10%的樣本作為交叉驗證集，進行樣本內訓練。由於計畫中將回測區間按年份劃分為 6個區間，因此需要對每個回測區間的不同訓練集進行重複訓練，確定模型的參數。模型訓練結束後，使用模型對交叉驗證集進行測試，選取其中 AUC 或平均 AUC最高的一組參數作為模型的最優參數。

（一）參數的優化調整

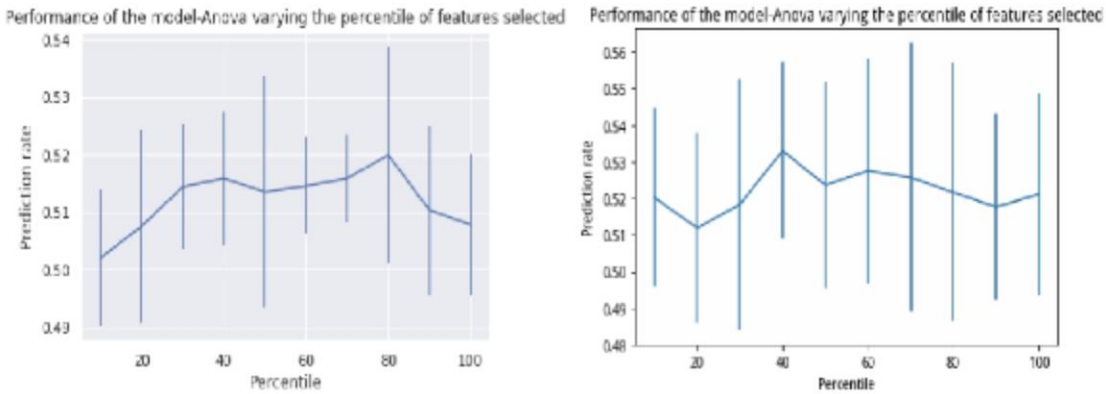


圖 4-3：左、右圖分別為隨機森林和 XGBoost 算法的執行結果

在建模過程中能夠發現，特徵數量對演算法是具有影響的，因此選擇合適的特徵數量能夠提升模型的預測能力。本文利用 pipeline（管道）將特徵的選擇與模型的構建相結合，形成一個新的物件，然後針對新物件選取不同的特徵選擇參數並對其做交叉驗證得到檢測結果，輔助最終模型得到最優的參數設置。本文使用 SelectPercentile（f_classif，percentile）函數來做特徵比例的選擇，其中 f_classif 項為特徵選擇的得分標準，percentile 項為特徵選擇比例。

由圖可知，隨機森林演算法隨著特徵比例的增加，預測效果越好；XGBoost 演算法則不太穩定，預測效果波動較大。而 XGBoost 演算法的特徵比例選擇 40%時最佳。儘管隨機森林演算法的預測效果會隨著特徵數量的增加而提升，但計算時間也會相應延長。因此結合效率綜合考慮，隨機森林與 XGBoost 演算法分別選擇 80%與40%作為最佳的特徵比例。

對於隨機森林而言，決策樹的個數直接影響到它的預測能力。決策樹過少，就無法發揮集成學習演算法的優勢，預測效果不佳；決策樹越大，預測效果越好，但是會大幅度增加計算量，隨著決策樹的增加，邊際效應會減小。本文通過多次交叉驗證，最終確定演算法的 n_estimators 參數（決策樹個數）為

DOI:10.6814/NCCU202200358

450-600 左右時，綜合效果較好，本文設定 n_estimators=500。另外，我們發現演算法的 max_depth（深度）參數，即每棵樹的複雜度，隨著深度的上升，F1-score 呈現上升趨勢，AUC 波動較大，模型不夠穩定。因此，結合訓練效果以及訓練時間進行綜合考慮，max_depth 選擇 3-5 較為合適，本文設定 max_depth=4。

對於 XGBoost 演算法，一般而言 max_depth 在 3-10 之間較為合適，但本文通過多次的交叉驗證，發現隨著演算法 max_depth 的提高，F1-score 和 AUC 都呈現下降趨勢，訓練時間也隨著增長。因此，針對本文這種情況，最終設定參數max_depth=1。subsample 參數為取樣速率，即選取訓練集參與模型訓練的比例，取值在 0-1 之間，可有效地防止過擬合問題。經過交叉驗證，隨著 subsample 取值的上升，F1-score 值呈現下降趨勢，模型的訓練時間有一定的縮短，綜合訓練效果和時間的考慮，本文設定 subsample=0.9，通常 subsample 值可在 0.5-0.9 之間。

由於本文主要目的是採用 XGBoost 演算法進行因子合成，因此這裡不對演算法的所有參數進行過多的優化調整，其餘參數都採用預設值進行訓練。

（二）樣本內訓練以及模型評價

在樣本內訓練中，確定了演算法參數之後，通過使用模型對交叉驗證集進行預測，選取交叉驗證集中 AUC 指標最高的一組參數作為模型的最優參數。確定最優參數之後，使用模型進行樣本外測試。如圖 4-3 分階段回測模型示意圖所示，取下一年所有樣本預處理後的因子資料作為模型的輸入，得到測試集的預測分數 ( ),將得到預測值 ( )視作合成後的因子值。

本文使用 SVM、隨機森林和 XGBoost 三種演算法模型對樣本進行訓練。由於已經有大量的前人學者對三種模型的優劣勢進行比較說明，本文不再進行重複說明，只是簡單地通過精確度和 AUC 指標對三種模型進行對比驗證，並與傳統的線性模型進行對比。如圖 4-4 至圖 4-6 所示，為隨機森林演算法在 6 個子區間內的模型預測效果；如圖 4-7 至圖 4-9 所示，為 SVM 演算法在 6 個子區間內的模型預測效果；如圖 4-10 至圖 4-12 所示，XGBoost 演算法在 6 個子區間上的模型預測效果。為了更加直觀地分析預測結果，將所有的預測精度和 AUC 指標繪製成折線圖，橫坐標為每個子區間的年月份，縱坐標為相應的預測指標數值，其中藍色實線為 AUC 指標，橙色虛線為模型精准度指標。

DOI:10.6814/NCCU202200358

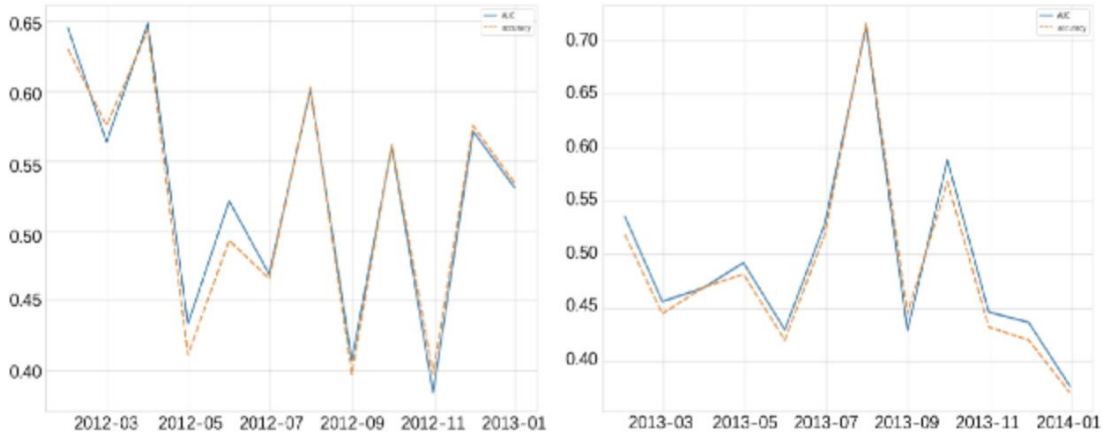


圖 4-4：第一、二子區間隨機森林算法預測精準度及 AUC 指標

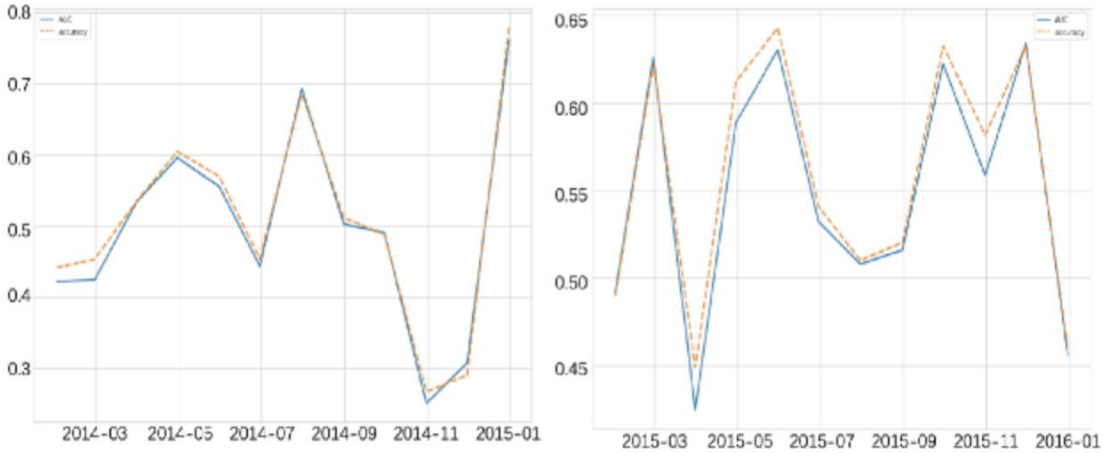


圖 4-5：第三、四子區間隨機森林算法預測精準度及 AUC 指標

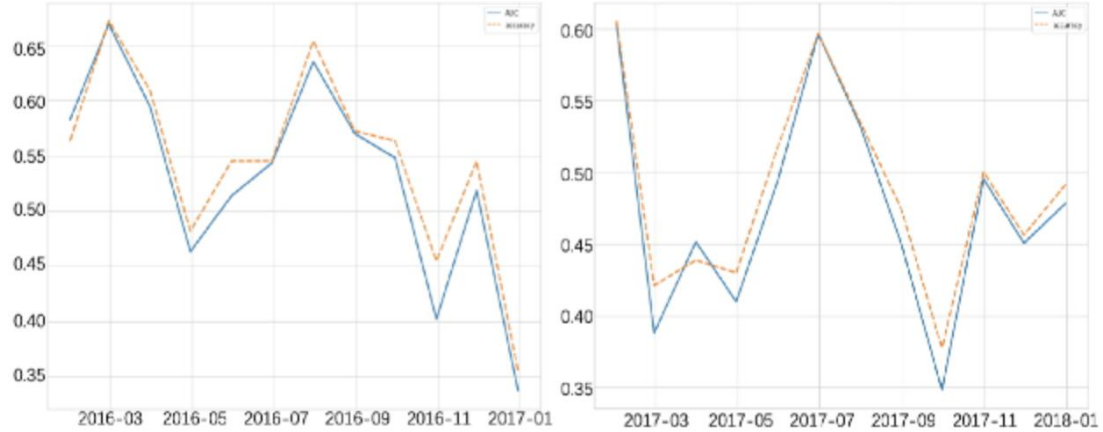


圖 4-6：第五、六子區間隨機森林算法預測精準度及 AUC 指標

DOI:10.6814/NCCU202200358

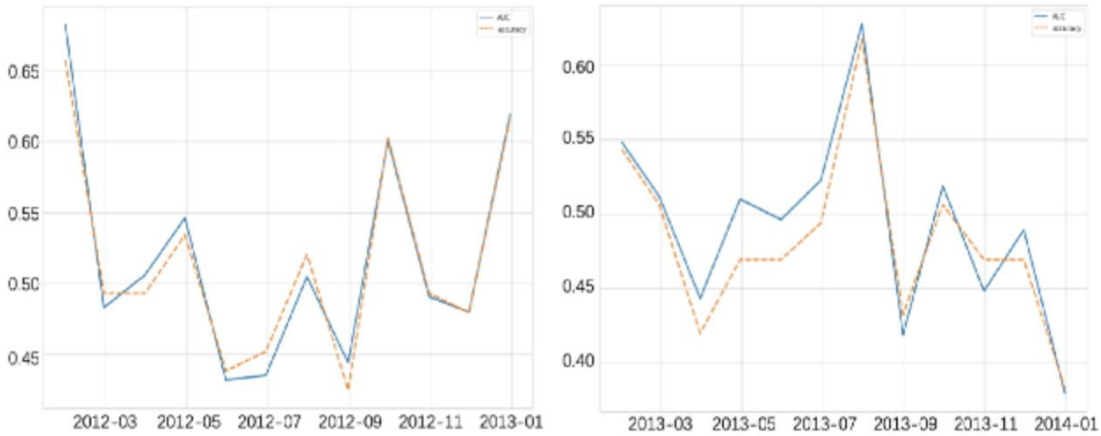


圖 4-7：第一、二子區間 SVM 算法預測精準度及 AUC 指標

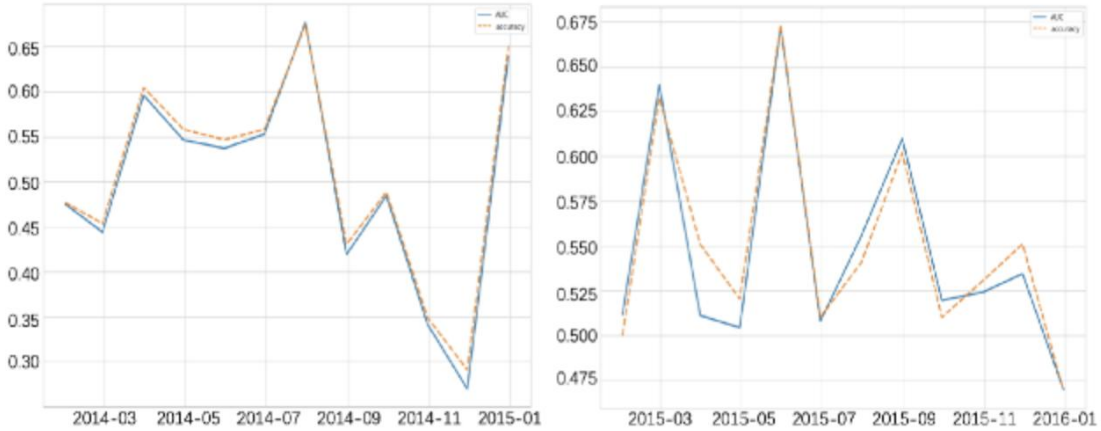


圖 4-8：第三、四子區間 SVM 算法預測精準度及 AUC 指標

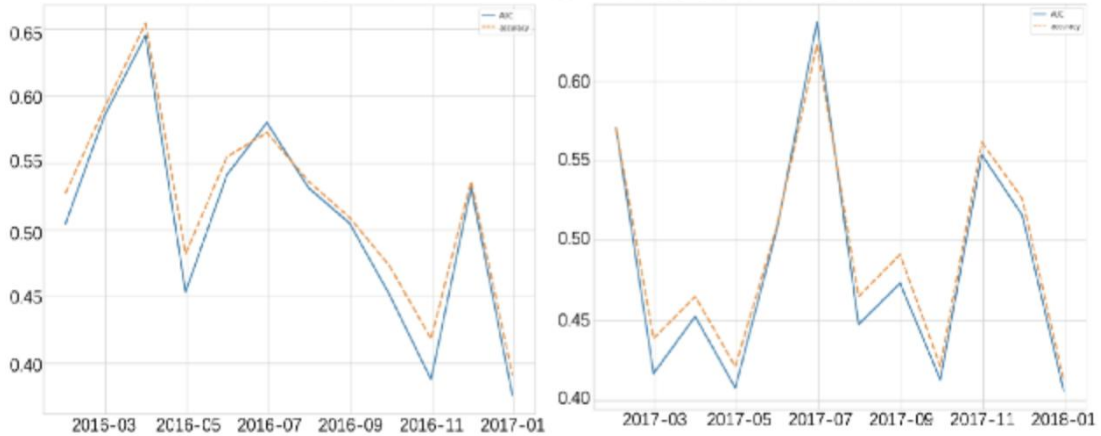


圖 4-9：第五、六子區間 SVM 算法預測精準度及 AUC 指標

DOI:10.6814/NCCU202200358

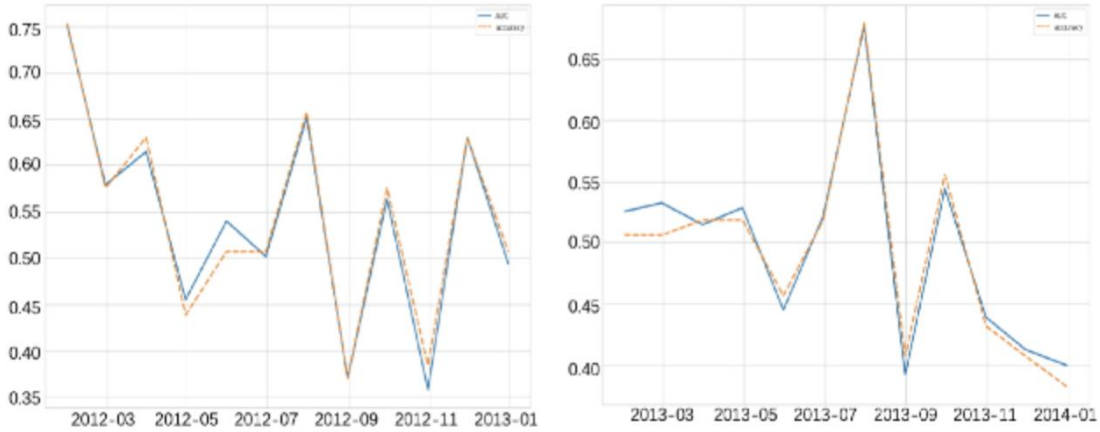


圖 4 - 10：第一、二子區間 XGBoost 算法預測精準度及 AUC 指標

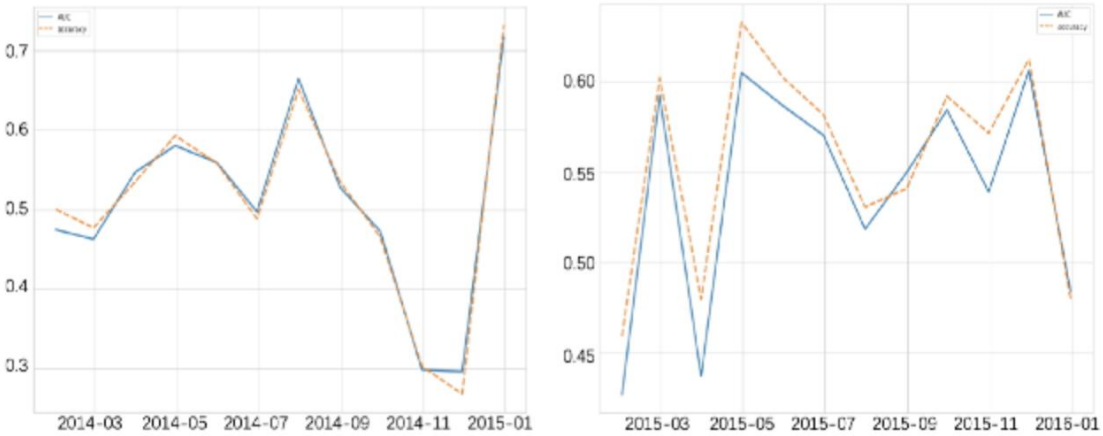


圖 4 - 11：第三、四子區間 XGBoost 算法預測精準度及 AUC 指標

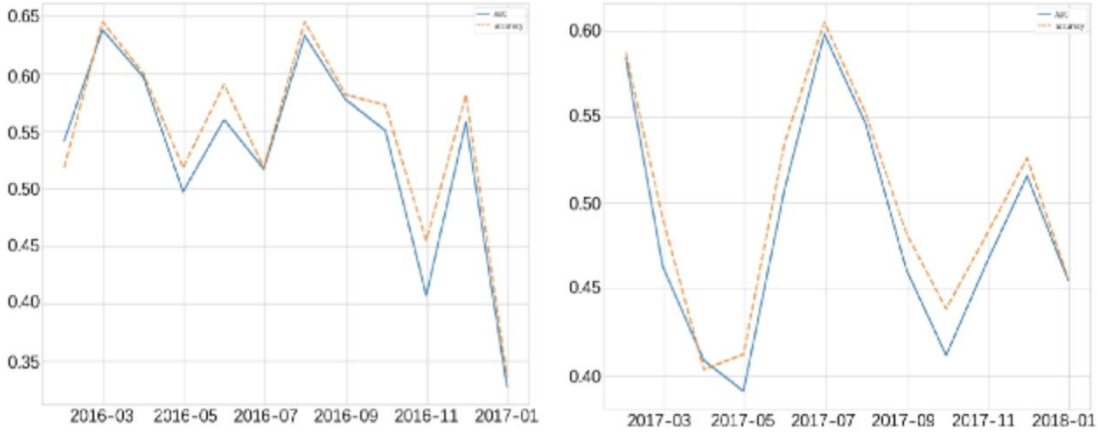


圖 4 - 12：第五、六子區間 XGBoost 算法預測精準度及 AUC 指標

從三種演算法的預測結果折線圖可以看出，三種演算法在每個子區間內的波動都比較相似，尤其是隨機森林演算法和 XGBoost 演算法的折線圖比較接近。但從數值上可以看出，XGBoost 演算法比隨機森林演算法略高。隨機森林演算法的預測準確度的最大值在第三個子區間中約為 0.75；SVM 演算法的預測準確度的最大值在第三、四個子區間內約為 0.675；XGBoost 演算法的預測準確度的最大值在第一個子區間內約為 0.75。從精准度和 AUC 指標的最大值看，

DOI:10.6814/NCCU202200358

隨機森林演算法和 XGBoost 演算法都比 SVM 演算法更高。另外，從三種演算法的總體預測趨勢來看，都是在第六個子區間的表現最差。此外，隨機森林演算法和 SVM 演算法的 AUC 指標和精准度平均值都是在第四個區間最高，XGBoost 演算法的 AUC 指標平均值在第一個子區間中最高，而精准度指標平均值在第四個區間內最高。

除了對三種機器學習演算法進行比較之外，實驗中還希望對非線性的因子合成方法與傳統的線性因子合成方法進行對比。而傳統的因子合成方法有等權合成法、歷史收益率加權法、最大化 ICIR 加權法等等，基本思想都是通過一定的方法賦予各個因子相應的權重。因為邏輯回歸（Logistic Regression，LR）演算法在本質上還是線性回歸，因此本文中使用邏輯回歸演算法訓練得到一個線性模型，近似地替代傳統的線性因子合成方法作為對照。

算法類型	年份 / 指標	08-13	09-14	10-15	11-16	12-17	13-18	max	avg	vol
隨機森林	AUC	0.527946	0.491503	0.498505	0.549096	0.531666	0.474702	0.549096	0.512236	2.83%
	Accuracy	0.523973	0.483539	0.506783	0.557823	0.54697	0.486842	0.557823	0.517655	3.08%
SVM	AUC	0.518686	0.492729	0.498505	0.546744	0.507644	0.4833	0.546744	0.507935	2.26%
	Accuracy	0.517123	0.481481	0.506783	0.54932	0.520455	0.491959	0.54932	0.511187	2.39%
XGBoost	AUC	0.542338	0.494363	0.508005	0.541629	0.533728	0.48421	0.542338	0.517379	2.53%
	Accuracy	0.544521	0.490741	0.508721	0.556973	0.54697	0.497807	0.556973	0.524289	2.85%
LR	AUC	0.539452	0.373203	0.758681	0.538978	0.383881	0.45205	0.758681	0.507708	14.25%
	Accuracy	0.547945	0.37037	0.744186	0.530612	0.409091	0.491228	0.744186	0.515572	13.16%

表 4-1：四種模型評價指標對比

獲取四種方法在每個子區間的預測值，分別計算其精准度與 AUC 指標的平均值，用平均值進行對比分析。將這些指標平均值繪製成表格以便對比分析，並從這四種方法 6 次訓練的平均最大值、平均值以及波動率進行分析，如表 4-1 所示。

表 4-1 為四種演算法在 6 個區間內的預測平均值資料，四種模型在本文資料的 6 個區間內的各進行了 6 次的訓練。由表中可知，其中線性模型在第三個子區間的精准度和 AUC 指標值是最大的，顯著地高於其它三種模型，而隨機森林、SVM、XGBoost 演算法 6 次訓練的最大值都相差不多。但線性模型的波動性也是顯著地高於其它三種模型，穩定性最差。最後從四種方法 6 次訓練的平均效果來看，四種模型也是有一定的差異，不管是精准度的平均值還是 AUC 指標的平均值，都是 XGBoost 演算法最優，隨機森林要優於傳統的線性模型，線性模型與 SVM 演算法相差不大。綜上所示，利用 XGBoost 演算法進行因子合成，不僅穩定性較高，而且綜合的預測效果最好。

二、重要度分析及因子五分位分組回測

DOI:10.6814/NCCU202200358

傳統的線性模型可以通過查看各項因子的權重來判斷線性模型中因子的重要度。Boosting 模型中也有類似的功能，它通過計算因子在決策樹中的表現，通過一定的運算可以計算出每個因子的最終得分，本文以此來作為分析因子重要度的關鍵指標。由於 SVM 演算法不屬於 Boosting 模型，因此沒有這個指標，之後不再提及。如圖 3.17、圖 3.18 所示，為隨機森林演算法以及 XGBoost 演算法在第 6 個區間中的特徵重要度指標。

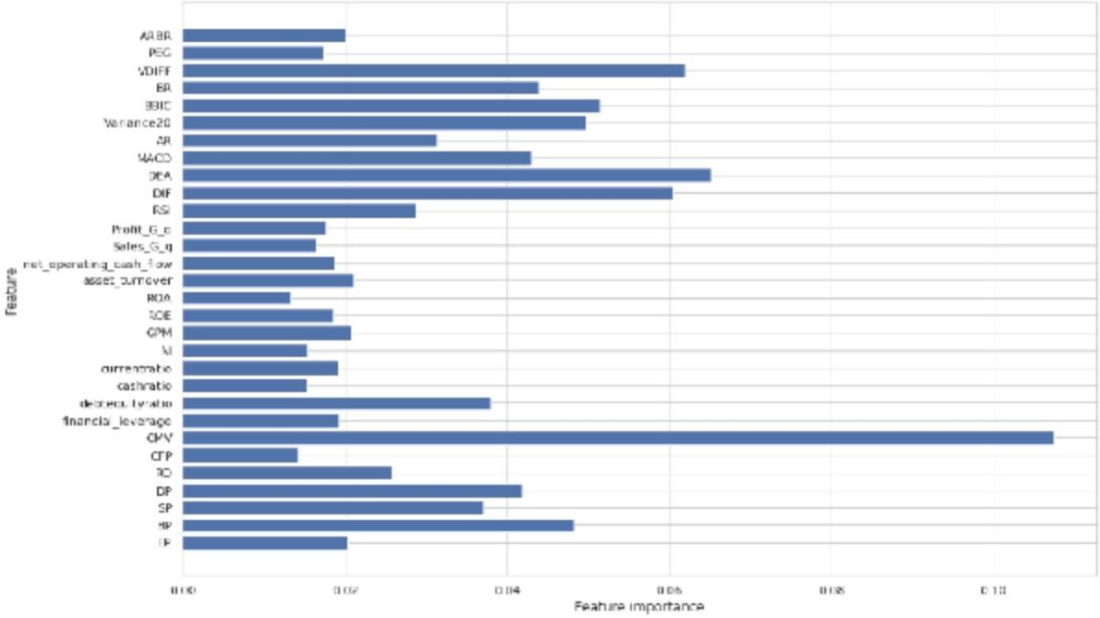


圖 3.17：隨機森林算法的特徵重要度統計圖

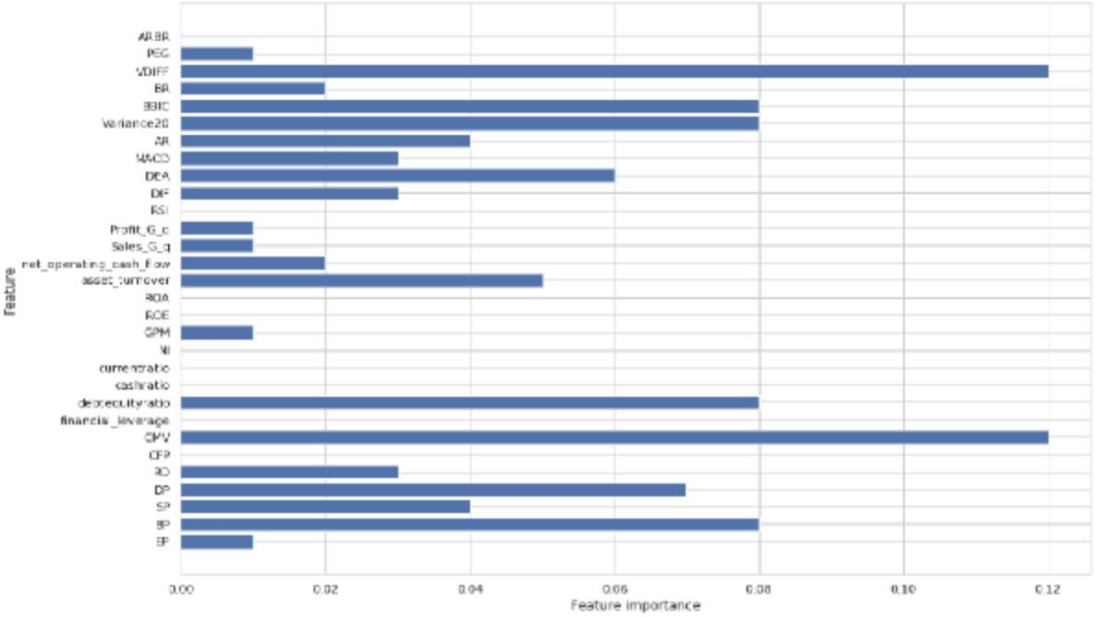


圖 3.18：XGBoost算法的特徵重要度統計圖

DOI:10.6814/NCCU202200358

從上圖可知，隨機森林演算法在訓練模型的過程中使用到了所有的特徵因子，即每個因子或多或少都佔據了一定的權重。而 XGboost 演算法會突出使用部分表現較好的特徵因子，並且會摒棄掉部分因子，因此 XGBoost 演算法的特徵重要度統計圖中，會有部分因子的特徵重要度指標為零，並且因子的重要度相差較大。如圖 3.19 至圖 3.24 所示，為 XGBoost 演算法在 6 個子區間中的特徵相關度熱力圖。

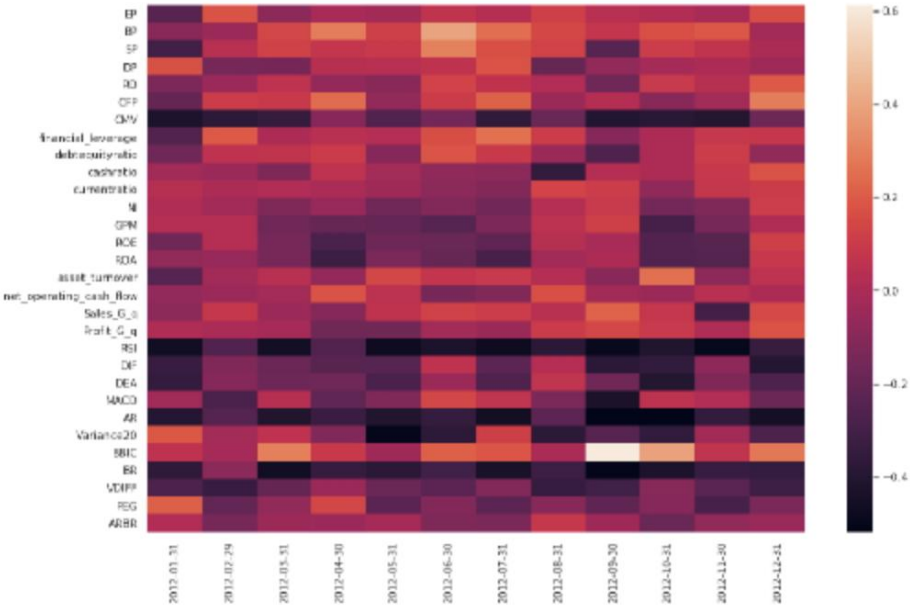


圖 3.19：第一個子區間XGBoost算法特徵相關度熱力圖

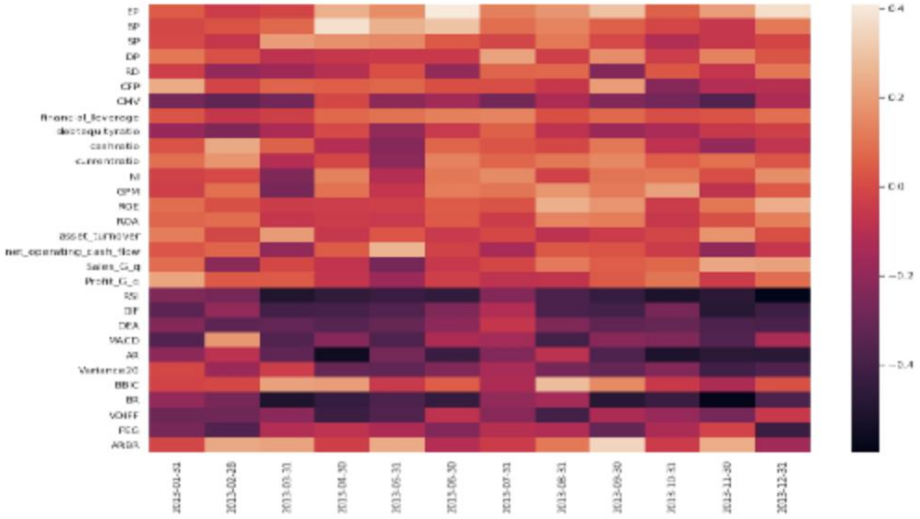


圖 3.20：第二個子區間XGBoost算法特徵相關度熱力圖

DOI:10.6814/NCCU202200358

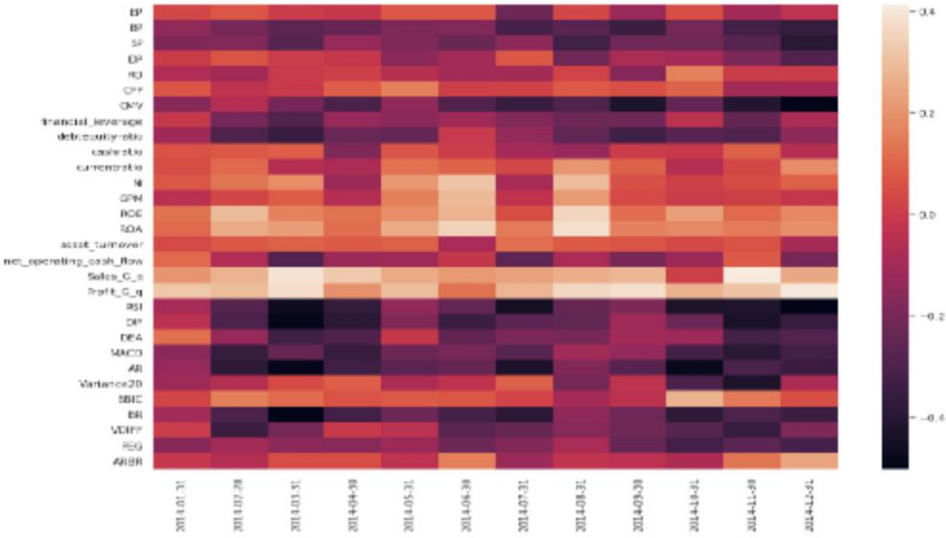


圖 3.21：第三個子區間XGBoost算法特徵相關度熱力圖

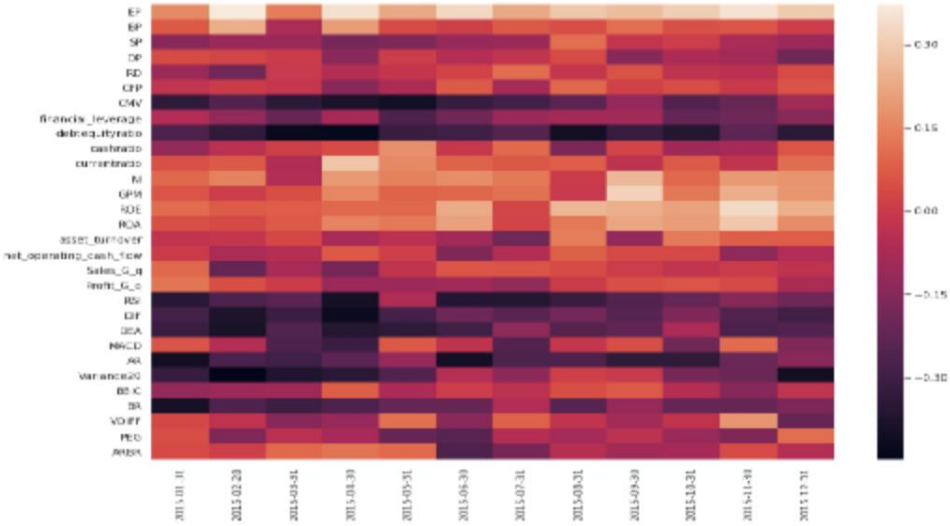
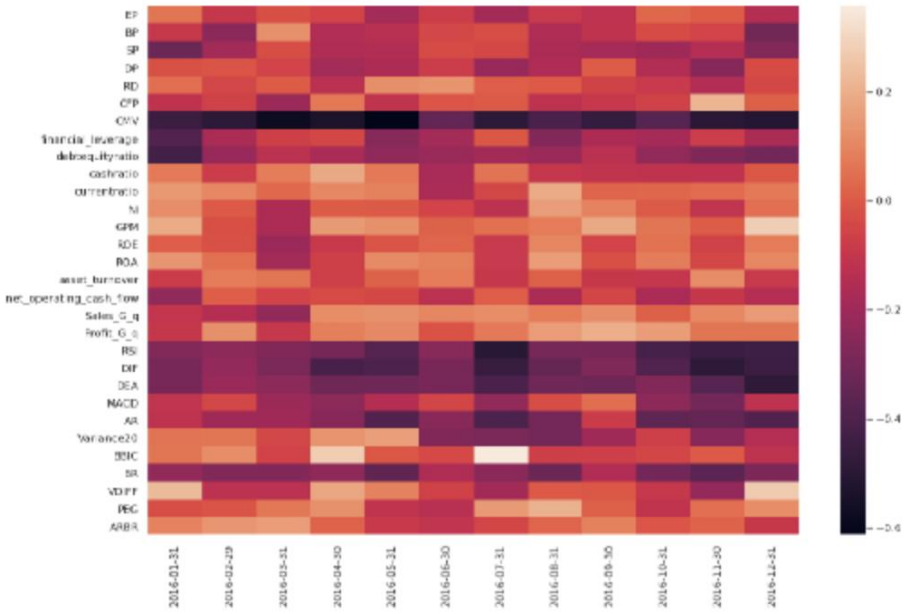


圖 3.22：第四個子區間XGBoost算法特徵相關度熱力圖



DOI:10.6814/NCCU202200358

圖 3.23：第五個子區間XGBoost算法特徵相關度熱力圖

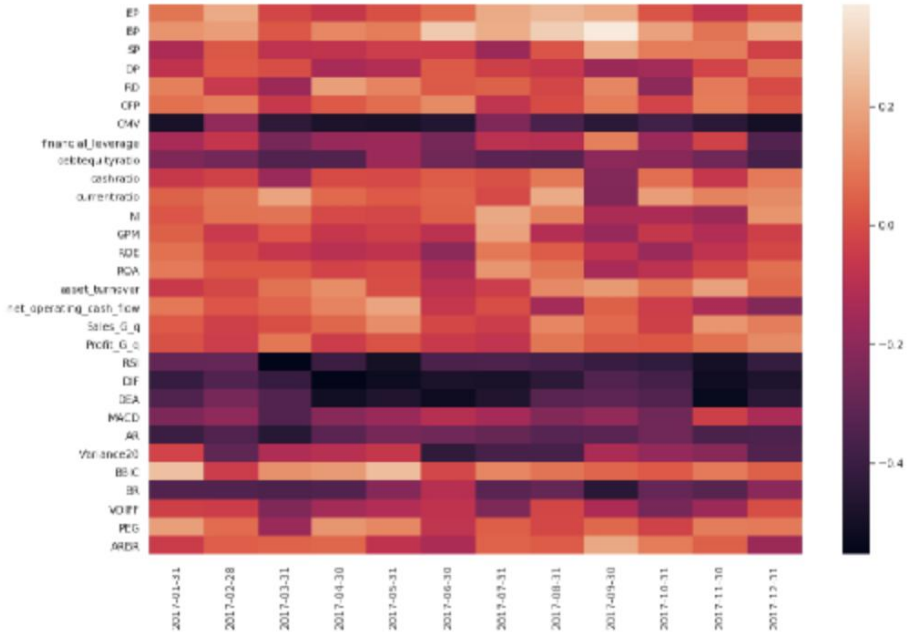


圖 3.24：第六個子區間XGBoost算法特徵相關度熱力圖

在圖 3.19 至圖 3.24 中，各因子的月度區塊顏色越趨近於深紅色，表明因子在該月份的預測中起到的作用越大；區塊顏色越趨近於白色，則說明該因子在當月預測模型中作用越小。特徵相關度熱力圖能夠非常直觀地展示了各個因子在預測月份的得分高低。從 6 次訓練的特徵相關度熱力圖可以看出，其特徵相關度分佈情況都大致相似，主要分佈在中下部以及中上部的幾個因子，中間的一批因子的相關度普遍都不太高。

由於單次訓練結果不具有代表性，為了平衡 6 次訓練預測結果，對 6 次特徵重要度指標取平均值，作為最終因子的重要度排序。同時，為了消除個別因子重要度過高以及過低的問題，使用排序法消除這種差異性。即首先對每個區間的重要度進行一次排序並重新賦值，30 為最大值，1 為最小值。然後對 6 次排序取平均值，將排序平均值作為最終指標進行重要度的排序。如表 4-2 所示。

DOI:10.6814/NCCU202200358

分组	feature	08-13年	09-14年	10-15年	11-16年	12-17年	13-18年	ave
1	VDFF	30	28	30	30	30	29	29.5
1	CMV	29	30	29	28	29	30	29.1666667
1	EP	24	27	28	29	22	14	24
1	debtequityratio	14	26	18	27	27	27	23.1666667
1	MACD	27	20	27	22	12	25	22.1666667
1	BP	9	8	23	21	21	28	18.3333333
2	DF	28	19	4	20	26	18	19.1666667
2	BBIC	20	18	8	23	24	17	18.3333333
2	Variance20	11	16	15	17	28	26	18.8333333
2	RD	18	5	22	19	25	19	18
2	CFP	23	29	21	13	9	9	17.3333333
2	financial_leverage	17	22	12	25	16	8	16.6666667
3	PEG	22	23	24	16	1	10	16
3	SP	5	7	26	15	17	21	15.1666667
3	DP	4	6	19	14	20	24	14.5
3	currentratio	8	17	25	7	23	6	14.3333333
3	DEA	15	10	3	18	14	23	13.8333333
3	NI	21	25	14	8	7	5	13.3333333
4	cashratio	13	11	17	24	8	7	13.3333333
4	RSI	25	24	20	2	2	2	12.5
4	AR	6	1	7	26	13	20	12.1666667
4	net_operating_cash_flow	16	14	13	4	4	16	11.1666667
4	BR	10	12	2	9	18	15	11
4	GPM	26	13	9	3	3	11	10.8333333
5	asset_turnover	1	3	11	5	19	22	10.1666667
5	Profit_G_q	3	15	6	12	11	13	10
5	Sales_G_q	7	2	10	10	15	12	9.33333333
5	ROE	2	21	5	11	6	4	8.16666667
5	ROA	12	4	16	6	5	3	7.66666667
5	ARBR	19	9	1	1	10	1	6.83333333

表 4 - 2：特徵排序及五分位分組

如上表 4-2 所示，將 30 個因子按排序平均值分為 5 組，每組 6 個因子，進行分組回測。回測區間為 2014 年 1 月 31 日至 2019 年 1 月 31 日，股票池為滬深 300 成分股，資料預處理方法與之前相同，剔除 ST 股票，剔除每個截面期下一交易日停牌的股票以及剔除上市 3 個月以內的股票。回測模型主要對五組因子組合年化收益率、夏普比率、資訊比率、最大回撤以及波動性等評價指標進行比較分析，如表 4-3 所示。五分位元數分組因子組合選股淨值走勢圖如圖 3.25 所示，橫坐標為回測年份，縱坐標為五組因子組合的回測收益率。

DOI:10.6814/NCCU202200358

	风险指标				纯股票多头时		对冲后		
分组	Alpha	Beta	夏普比率	信息比率	年化收益率	最大回撤	波动性	年化收益	最大回撤
1	0.175024	0.42178	0.675839	0.428083	0.23092088	0.446829	0.282495	0.18719	0.325916
2	0.120068	0.25097	0.401921	0.235232	0.16952726	0.494096	0.32227	0.115306	0.443201
3	0.07975	0.174116	0.3141	0.137126	0.12631223	0.51442	0.274792	0.062516	0.393674
4	-0.02339	0.096453	-0.06724	-0.15616	0.02024203	0.62801	0.293862	-0.08037	0.741357
5	0.030032	0.087463	0.118588	-0.01186	0.07332866	0.40379	0.281047	-0.00582	0.61333

表 4-3：五組因子回測評價指標

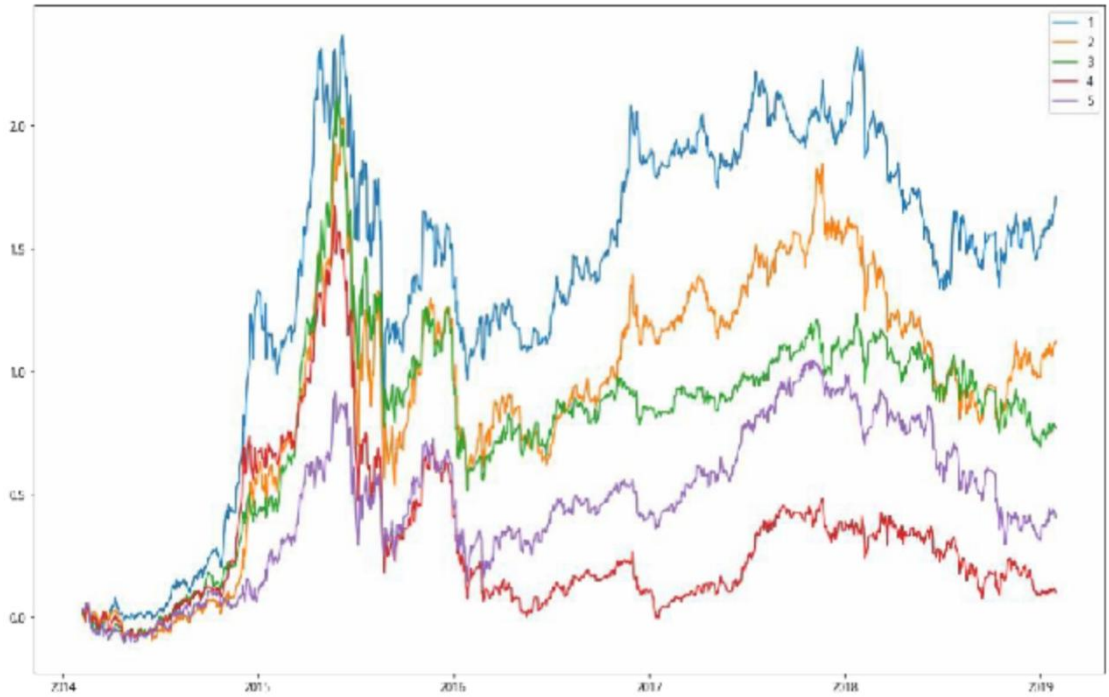


圖 3.25：五組因子組合選股淨值走勢圖

由上圖可以看出，第一組因子組合的收益率是最高的，始終保持著領先地位，顯著高於其它四組。第二組與第三組可以說是不相上下，在 2017 年到 2018 年之間以及 2019 年末這兩段時間內第二組比第三組的收益率略高一籌，其他時間基本上相差不大。第四組因子組合僅僅在 2016 年之前優於第五組，之後反而是第五組的收益率超過第四組。

對上述的現象進行分析總結，主要有以下幾種解釋：

其一，是因子的重要度並不能作為判斷因子效果好壞的唯一標準。部分因子之間存在相關性，可能會出現某兩個因子組合在一起之後造成效果的提升或者降低。例如，本文經過多次的回測，發現“CMV”和“DIF”這兩個因子組合在一起時，收益率會大大的降低。

其二，是排序方法造成的分組不明確。由於 XGBoost 模型在使用因子時，

DOI:10.6814/NCCU202200358

有部分的因子未被使用到，因此這些因子的重要度得分都為“0”，以及存在一些因子它們的重要度得分原本是相同的，但在經過排序之後就產生了差別。例如，若存在 6 個因子原本重要度相同，經過排序後，默認排序最高的就比默認最低的因子高出了 5 分。

其三，五分位數分界線導致得分相同的因子分為不同的兩組。經過評分排序再取均值之後，仍有部分的因子最終得分是相同的，如表 3.2 所示，一些因子的最終評分是相同的，但正好位於五分位數的分界線上，因此默認排序高的可能就分在上一組，而默認排序低的就分在下一組。

第四節　本章小結

本章主要介紹了利用 XGBoost 演算法進行因子合成的具體方法。包括了股票池的建立、資料的預處理、模型的訓練、演算法的優化調參以及分組回測等。簡單地將 XGBoost 演算法與 SVM 演算法和隨機森林演算法進行了對比，驗證了 XGBoost 演算法的預測精准度和 AUC 指標確實優於 SVM 和隨機森林。並且通過與線性模型的對比，可以看出使用 XGBoost 演算法進行因子合成是優於傳統的線性模型，提升了因子合成的效率。之後通過因子重要度分析對 30 個因子進行打分排序並將其分為 5 組進行分層回測。回測結果顯示，第四組與第五組的回測結果與預期結果有一定的出入，本文認為可能的原因是由於因子之間存在相關性，演算法提供的重要度指標並不能作為唯一的判斷依據，仍需要結合實際進行一定的優化調整。總體來看，將 XGBoost 演算法運用到因子合成當中，效果還是較為顯著的。可以看出，針對此資料集，經過本文的合成方法，無疑是可以得到一批表現較好的因子。

DOI:10.6814/NCCU202200358

第五章　機器學習與波動聚集

第一節　研究設計

在股票市場中，波動率一直是投資者和研究者關注的最重要的問題之一。對投資者而言，減少股票市場的不確定因素以及準確度量股票指數收益的波動率顯得尤為必要。筆者認為，若能對波動聚集效應加以應用，是能夠對投資起到指導性的意義。

假如投資者擁有預知未來的能力，那麼投資將變得無往不利，然而這是不可能的事情。但是，波動聚集效應的存在，為投資者帶來了一種新的投資思路。我們可能無法準確地預測一檔股票是漲還是跌，但是我們可以預測整個市場的波動率是小還是大。當投資者能夠充分地瞭解波動聚集現象時，風險厭惡型的投資者可以選擇退出股市，減少損失的風險性。而風險偏好型的投資者只要能冷靜充分地改進投資策略，也可以獲取到更高的收益。

然而金融資料太過龐大複雜，每日新增的金融資料資訊浩如星海，投資者想對所有資訊加以利用確實比較困難。而今，隨著量化投資和機器學習的興起，投資者對金融資料的處理研究愈加得心應手，越來越多的工作可以由機器幫助完成。

自然而然的，筆者想是否能夠將波動聚集效應加入到機器學習演算法中進行訓練？機器能否對波動聚集效應加以學習，並應用到投資策略中呢？通過大量的理論文獻研究，發現 GARCH 模型能夠較好地擬合市場中存在的波動聚集現象。因此，本文將 GARCH 模型的預測值加入到機器學習的訓練集中加以訓練，通過大量的訓練資料對模型進行訓練，期望機器能夠學習到波動聚集效應並對其加以應用。

為了驗證波動聚集現象的應用效果，本文通過兩群組原則的回測結果進行對比分析。同時為了控制無關變數，本文中對照群組原則與實驗群組原則的唯一差別，就是在實驗群組原則中的機器學習訓練集中加入了 GARCH 模型的預測值（後文將此值簡稱為波動聚集因子）作為研究變數。具體的實證研究步驟和結果會在後文中進行詳細的講述。

DOI:10.6814/NCCU202200358

第二節　機器學習與多因子策略的結合

量化策略是指利用電腦程式作為工具，通過一定的固有邏輯進行分析、判斷和決策來幫主投資者進行量化投資，量化策略既可以完全憑藉電腦自動運行，也可以人工作業執行。

一、策略執行思路及主要因子

本文策略的基礎股票為滬深 300 股票池。滬深 300 指數前文已經做過介紹，是滬深兩市中行業代表性好、股票流動性高、交易程度活躍的股票，能夠反映市場主流交易情況。因此在策略中將滬深 300 作為基準收益進行優劣對比。

本文策略的主要思路是基於滬深 300 股票池利用機器學習預測股票的漲跌從而進行選股換倉。基本方法是利用 XGBoost 演算法預測當日股票的漲跌，訓練集為前 200 天的因子數據。訓練集的長度過短，可能達不到預測效果；訓練集過長會增長策略的執行時間，時間對於量化投資而言是非常重要的，決定了能否及時買入股票，因此需要分秒必爭。一般訓練長度選擇 100 天左右較為合適，由於本文使用歷史資料進行回測，並不存在想買入的股票無法買入的情況，因此為了提高預測準確度，增長了策略的執行時間，將訓練集長度設置為 200 天。同時，策略會獲取當日的最新資料加入到訓練集中，每日都對訓練集進行更新，確保能夠學習到市場的最新變化，提高預測準確率。將預測為漲的股票按照持倉數目進行建倉，資金根據倉位元等額分配。當股票損失達到 8%時，進行止損，清倉賣出；當股票盈利達到 15%時，同樣進行賣出操作，進行止盈。當持倉數量小於最大持倉數時，進行買入操作，買入預測為漲的股票。倉數可以根據滬深 300 指數的價格進行減倉和加倉的調整。

本文策略中因子的選用是基於第二章節中結合 XGBoost 模型進行因子合成的方法，主要選取其中重要度指標較高的一批因子，再加入一些常用的技術指標因子。本文採用的因子分別是 VDIFF、CMV、EP、MACD、BP、KDJ、SMA、WMA、MOM 這 9 個因子，如表 4.1 所示。其中 VDIFF、CMV、EP、MACD、BP 為因子重要度五分位數分組中第一組中的 5 個因子，KDJ、SMA、WMA、MOM 為目前較為常用的技術指標。增加這 4 個因子的原因，是因為它們有一定的經濟學理論基礎，另外在聚寬量化平臺上能夠比較完整地獲取到這些因子的資料，有助於模型的訓練。因此，本文認為在原有因子的基礎上融

DOI:10.6814/NCCU202200358

入這 4 個因子所建立的策略，理論上是有助於緩解量化策略常常出現的過擬合問題。

因子类型	因子名称	因子描述	计算方法
情绪因子	VDIFF	计算 VMACD 因子的中间变量	EMA(VOLUME，SHORT)-EMA(VOLUME，LONG) short 设置为 12，long 设置为 26，M 设置为 9
杠杆因子	CMV	对数流通市值	流通市值的自然对数
估值因子	EP	净利润与总市值之比，市盈率（PE）的倒数	PE=1/（股权投资回报率-每股收益增长率）
估值因子	BP	净资产与总市值之比，市净率（PB）的倒数	PB=资产收益率/（股权投资回报率-每股收益增长率）
技术指标因子	MACD	平滑异移动平均线	MACD=(DIF-DEA)*2,COLORSTICK
技术指标因子	KDJ	随机指标（由 K 线、D 线、J 线组成）	K 线=SMA(RSV,M1,1)； D 线=SMA(K,M2,1)； J 线=3*K-2*D； RSV=(CLOSE-LLV(LOW,N))/(HHV(HIGH,N)-LLV(LOW,N))*100
技术指标因子	SMA	X 的 N 日移动平均,M 为权重	SMA(X,N,M)=(X*M+Y'*(N-M))/N
技术指标因子	WMA	X 的 N 日加权移动平均	WMA(X,N)=(1*X1+2*X2+...+n*Xn)/(1+2+...+n)
技术指标因子	MOM	动量线指标，当日收盘价与 N 日前收盘价之差	MOM=close - ref(close,n)

表 4.1：策略中的因子信息

ional Chengchi Univers

二、策略運行步驟

本文策略的運行是基於聚寬量化平臺的策略範本，總體上可以分為四個階段，分別是總體回測前階段、開盤前階段、開盤時階段以及收盤時階段。具體的階段步驟如圖 4.1 所示。

DOI:10.6814/NCCU202200358

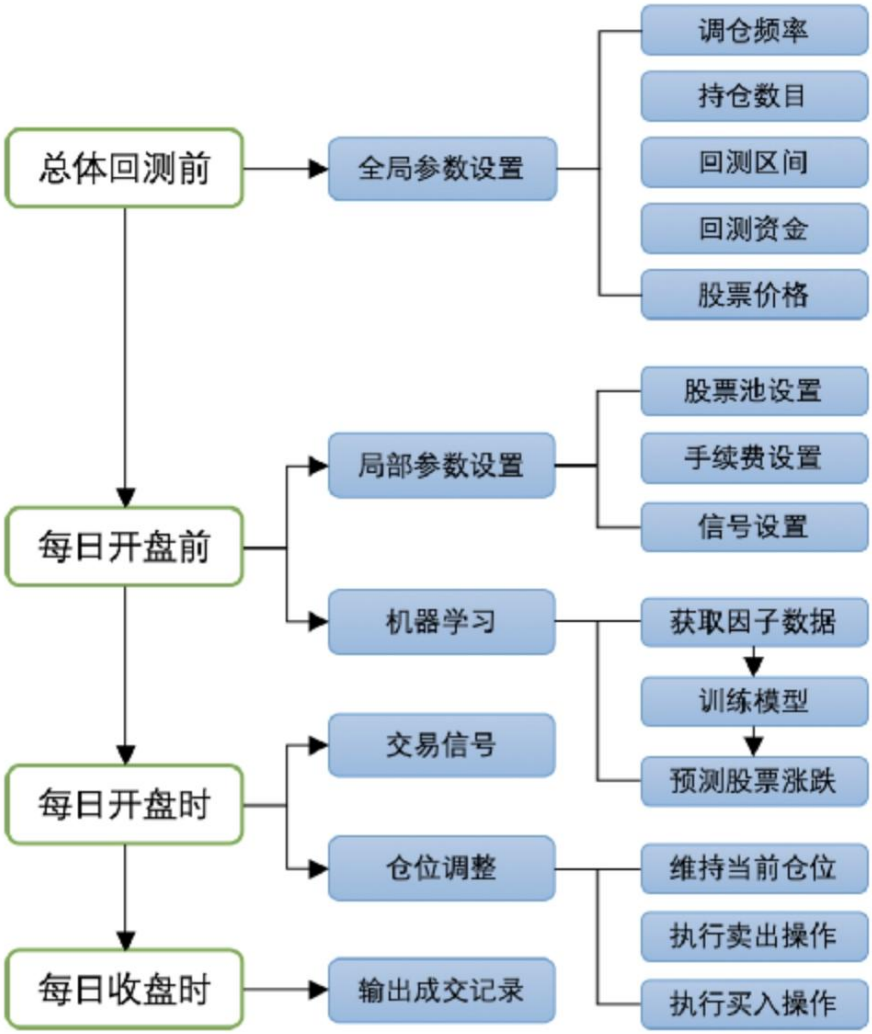


圖 4.1：策略運行步驟示意圖

第一階段，是在進行回測之前的一些全域參數的設置，其中包括了調倉頻率的設置、持倉數目的設定、回測區間的選取、初始回測資金的設定以及股票價格類型的選擇。調倉頻率可以設置為日級、分鐘級、tick 級，也可以設置成更長的時間段，例如 15 天或者 30 天。本文策略的調倉頻率為每天一次。持倉數目為可持有的股票數目，例如持倉數目為 20 時，同一時間內最多可持有 20 檔股票，本文策略的持倉數目設定為 10。本文選取的回測區間為 2012 年 1 月 1 日至 2019 年 12 月 31 日，初始回測資金設定為 10 萬。

本文的股票價格設置為動態複權模式的價格類型。因為公司除權除息會造成的股票價格以及股票指標的走勢發生變化躍遷，所謂的複權就是為了消除這種結果而採用的一種處理方式。通過複權的方法可以對股票價格進行修正，達到股票價格保持一種連續狀態的目的。目前主要的兩種複權模式為前複權和後複權。前複權是以最新價格為基準，將除權除息前的價格按照除權除息的係數

DOI:10.6814/NCCU202200358

進行縮減，因此是現有價格不變，以前的價格變小。優勢是現有單位的量與最新價格一致，有助於對接實盤交易，但是前複權回測使用到了未來資訊，對因子的回測有很大影響，可能造成欺騙性。後複權不僅在資料處理上更為簡單，並且肯定不會使用到未來資料，但是後複權後是以前的價格不變，現有價格增大，這會造成跟實盤的價格有較大差距，不利於實盤交易對接。而動態複權則是使用回測區間的截止時間所在時刻的複權係數進行複權，使用動態複權得到的該時刻的交易價格就是該時刻交易的真實價格，因此動態複權模型也叫真實價格模式。

第二階段，是在每日開盤前進行的局部參數的設置以及機器學習部分。局部參數設置首先是股票池的設置，本文依然是選取滬深 300 作為基準股票池。其次是手續費的設置，一般股票交易的手續費包括過戶費、交易所規費、印花稅、券商交易傭金這幾部分，一般過戶費的比例為本金的十萬分之六左右。有的證券公司會把交易所規費包含在傭金之中，統一上交到交易所或者證監會。印花稅對經濟活動和經濟交往中按交易金額大小計算的稅種，一般的收取比率為千分之一。券商交易傭金最高為成交金額的千分之三，最低為 5 元。2013 年之後，券商的資費大幅度降低，目前傭金基本為萬分之三，交易量較大時，傭金還能降低。因此，本文策略的手續費採用默認設置，即買入時收取傭金為目前券商收取的最大值萬分之三，賣出時傭金為券商收取的最大值萬分之三在加千分之一的印花稅，同時每筆交易傭金最低抵扣 5 塊錢。此參數可以更改，但目前不建議改動，因為這符合目前的現實情況，若後續有所變動，再進行調整。最後是本文的信號設置主要是對止盈參數以及止損參數的設置。當股票盈利或者損失達到參數值時，就會發出賣出信號進行賣出操作。本文策略的止盈比例為 15%，止損比例為 8%。

機器學習部分大體可以分為三步，主要是資料的獲取、模型的訓練以及股票的預測。資料獲取主要是獲取策略中所有股票樣本的因子資料和價格資料，以及幾個技術指標計算所涉及的所有相關資訊資料。訓練集每日進行滾動更新替換，確保訓練資料的可靠性，以達到讓模型能學習到市場最新變化的目的。模型演算法依然採用 XGBoost 演算法，參數設置與章節二中相同。最後使用模型對當日股票價格的漲跌進行預測，並進行標注，預測為漲的股票標籤為“True”，其餘為“False”。

DOI:10.6814/NCCU202200358

第三階段，是在每日開盤時進行的交易操作，主要包括了交易信號的獲取以及倉位元的調整。這一階段的主要任務首先是要判斷目前倉位元中的股票是否需要進行調倉操作。若目前所有股票未達到止盈止損的臨界值，則不需要進行賣出操作，同時若持倉數目已達到最大值，則也不需要進行買入操作，即維持當前倉位。當收到交易信號，存在股票達到止盈止損的臨界值，此時就進行倉位調整，執行賣出操作。賣出股票後，倉位元數目不飽和，再次收到交易信號進行倉位元調整，執行買入操作，買入標籤為“True”的股票，即開盤前預測為漲的股票。重複迴圈，直到收盤結束。

第四階段，即收盤階段，收盤後今日交易結束，輸出今日所有的交易記錄，以便後續的研究分析。

第三節　初始策略與波動聚集的結合

一、GARCH 模型預測波動率

本文構建 GARCH 模型過程中，均值方程採取 AR（4）模型。前文中已提及，與 ARCH 模型不同，GARCH 模型的階數很難確定，一般只用到低階的 GARCH 模型，例如 GARCH(1,1)、GARCH(1,2)、GARCH(2,1) 以及 GARCH(2,2)。本文波動率方程採用 GARCH(1,1)進行建模。以滬深 300 指數日漲跌幅資料為例，$p<0.05$，可以認為序列是平穩的，建立 AR(p)模型，判斷階數。經過試驗後得到最終模型 AR(4)-GARCH(1,1)。

通過多次調整優化，發現訓練的長度越長，模型擬合地越好，但由於聚寬量化平臺的資料最早只能到 2005 年，本文中的訓練資料的長度無法在增加了。圖4.2 是使用了 3500 天的訓練資料得到了訓練效果較好的模型,但從輸出結果我們可以看出，仍然有個別參數不是特別顯著，模型還有待優化提高。確定模型參數後，得到最終模型為：

$$r_t = 0.00055 + 0.02762a_t - 0.03162a_{t-1} + 0.02816a_{t-2} + 0.05639a_{t-3}$$

$$\sigma_t^2 = 0.000006 + 0.1a_{t-1}^2 + 0.88\sigma_{t-1}^2$$

分析 GARCH 模型的擬合結果，效果還是不錯的，其中 和 都顯著大於 0, 且 + 的值十分接近 1，表明了滬深 300 指數的波動是顯著存在波動聚集現象。輸出原始資料和擬合後的方差值以及預測值，如圖 4.3 所示，藍色為原始資料值，橙色為條件異方差擬合值，紅色為預測值。

DOI:10.6814/NCCU202200358

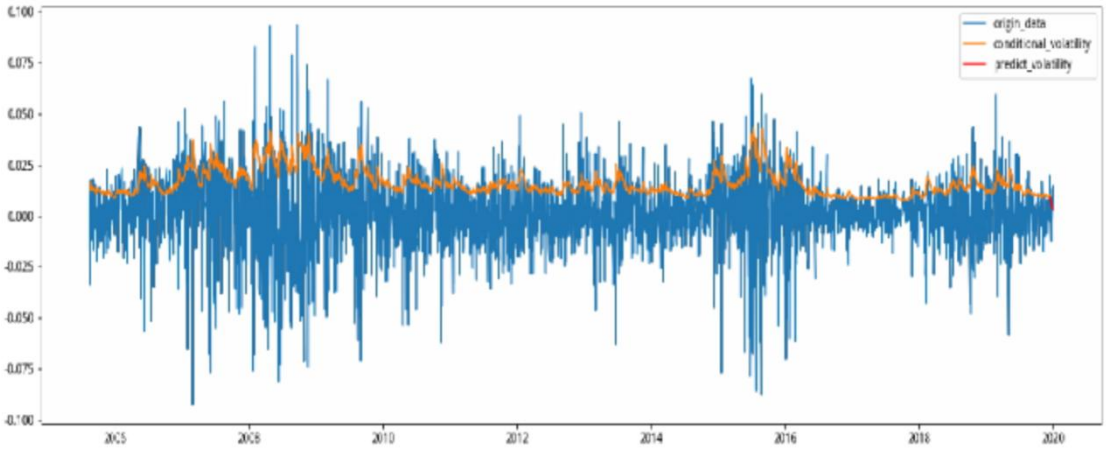


圖 4.3：模型效果輸出示意圖

可以看出，原始資料波動小的時候，方差線（橙色線條）較平穩；在 2007 年到 2010 年以及 2015 年到 2016 年後的這兩段時間，波動巨大且明顯有聚集現象，方差線同樣高居不下，方差擬合線基本符合真實資料，說明模型的擬合效果是不錯的。而預測線下降趨勢過大，預測效果有待加強。可能原因是預測時間過長，從模型的公式可以看出，預測值與前期資料高度相關，對於低階的 GARCH 模型，建議預測的時間長度越短越好，否則無法達到預測效果。由於從圖中預測時間過短無法看到預測效果，為了比較直觀地看到預測線，因此此處的預測時間長度為 10 天。

二、波動聚集的應用

在上文中已經具體講述了多因子策略與機器學習的結合應用。並且前文還介紹了使用 AR(4)-GARCH(1,1)模型對滬深 300 股指的市場波動進行了擬合以及對未來波動的預測，此模型可以很好描述刻畫波動聚集現象。

在波動聚集現象的研究中，大部分研究者都致力於對波動聚集現象的數量檢驗，少部分學者對波動聚集程度的度量方面以及波動聚集現象的產生原因進行研究，鮮少有學者將波動聚集現象加以實際應用。本文創新性地將波動聚集現象作為一個因子放入到機器學習中進行學習訓練，希望模型能夠學習到市場中的波動聚集性，以達到提高預測精度的目的和規避風險的目的。

經過驗證，已知滬深 300 指數的市場價格漲跌幅存在顯著的波動聚集效應，因此本文將 GARCH 模型的預測過程加入到多因子策略的開盤前階段，並將模型的預測值作為第 10 個因子加入到機器學習中進行訓練，預測股票的漲跌。最後，通過分析回測結果以及與原策略的回測結果進行對比分析，觀察波動聚

DOI:10.6814/NCCU202200358

集現象作為因子在其中起到的作用。為了控制實驗變數，實驗群組原則與對照群組原則的唯一區別就是實驗組的機器學習部分比對照組多了一個波動聚集因子，其餘代碼不變。

三、回測結果分析

（一）策略回測評估指標

策略回測的收益評估指標主要有策略收益、策略年化收益、基準收益、超額收益等。其中，策略收益的計算公式為：

$$Total\ Returns = \frac{P_{end} - P_{start}}{P_{start}} * 100\%$$

其中，P_{end}為策略最終股票和現金的總價值，P_{start}為策略開始股票和現金的總價值。策略收益是對一個策略回測的結果評估最直觀的指標之一。基準收益的計算方法與策略收益相同，主要作用是用來與策略收益進行對比。

策略年化收益的計算公式如下：

$$Total\ Annualized\ Returns = R_p = \left((1+P)^{\frac{250}{n}} - 1\right) * 100\%$$

其中，P 為策略收益，n 為策略執行的天數。策略年化收益是根據回測所得出的策略收益推算得到的年預期收益。一般而言，年化收益率高於 20%的交易策略較為優秀。

超額收益的計算公式如下：

$$超額收益 = \frac{策略收益 + 100\%}{基準收益 + 100\%} - 100\%$$

因為投資所產生的資產變動是具有複利效應的，所以淨值的變化過程是符合指數增長過程，因此在對於回測結果的超額收益計算過程中，使用減法代替除法顯得更加合理。並且通過觀察超額收益曲線，當超額收益曲線發生回撤時，就代表著在這段時間裡策略沒有跑贏基準，同理當超額收益曲線上漲了，就代表著這段時間裡策略優勝於基準。

而評估策略的風險指標較多，例如阿爾法指標、貝塔指標、夏普比率、索提諾比率、資訊比率等等。其中阿爾法代表了投資者面臨的非系統性風險，即投資者所獲得的與市場波動無關的回報，其計算公式如下：

$$Alpha = R_p - [R_f + \beta_P(R_m - R_f)]$$

其中，R_p為策略的年化收益率，R_m為基準年化收益率，R_f為無風險利率，

DOI:10.6814/NCCU202200358

通常默認為 0.04，β_P為策略的貝塔值。$Alpha$ 大於 0，代表策略相對於風險獲得了超額收益，$Alpha$ 小於 0 則說明獲得的收益要小於基準收益。

貝塔代表了投資的系統性風險，它能反映策略對於市場變化的敏感度。其計算公式為：

$$Beta = \frac{Cov(D_p, D_m)}{Var(D_m)}$$

其中，D_p為策略的日收益，D_m為基準的日收益。$Beta < 0$，通常代表策略與基準的走勢方向相反；$Beta = 0$ 代表兩者沒有相關性；$0 < Beta < 1$ 說明策略與基準的走勢方向相同但移動幅度小於基準；當 $Beta \geq 1$ 時代表策略與基準的走勢方向相同並且移動幅度大於等於基準。例如當一個策略的貝塔值為 1.6 時，每當大盤上漲 1%，策略預計能上漲 1.6%。

夏普比率的值代表了每承受一單位的風險，能夠產生的超額收益的大小。它是同時對收益和風險加以綜合考慮的三大經典指標之一，其計算公式為：

$$Sharpe\ Ratio = \frac{R_p + R_f}{\sigma_p}$$

其中，R_p為策略的年化收益率，R_f為無風險利率，通常默認為 0.04，σ_p為策略的收益波動率，在其他條件一致時，夏普比率的值越大代表了策略越優秀。

信息比率是以 Markwitz 的均異模型為基礎，代表了單位風險所帶來的超額收益。其計算公式如下：

$$Information\ Ratio = \frac{R_p + R_m}{\sigma_t}$$

其中，R_p為策略的年化收益率，R_m為基準年化收益率，σ_t為策略與基準每日收益差值的年化標準差。在相同條件下，信息比率越大，代表策略單位誤差所帶來的超額收益就越大。因此，在適當的風險承受範圍內，儘可能地提高信息比率，才是一個合格的投資者。

最大回撤是指投資過程中可能出現的最糟糕的情況，即最極端可能的虧損情況。其計算公式為：

$$Max\ Drawdown = \frac{Max(P_x + P_y)}{P_x}$$

其中，P_t為策略當日股票和現金的總價值，　>　。通常意義下，最大回撤代表著投資者對虧損的最大承受能力，理性的投資者應該盡可能地減小最大回

DOI:10.6814/NCCU202200358

撤值，儘量規避較長的回撤期以及較大的回撤率。

（二）策略回測結果對比

本文分別對未加入波動聚集因子的初始策略與加入波動聚集因子進行改良的策略進行了回測。策略內所有參數設置均保持一致，包括調倉頻率、最大持倉數、回測區間、初始資金等等。如圖 4.4 和圖 4.5，分別為未加入波動聚集因子的初始版策略以及加入了波動聚集因子的改良策略的回測結果。

圖 4.4：初始策略的回測結果示意圖

DOI:10.6814/NCCU202200358

圖 4.5：改良後悔策略的回測結果示意圖

從回測結果可以看出，加入了波動聚集因子的回測結果其策略收益有了一定的提高，從初始的 1564.09%增長到了 1737.01%，並且大部分的評價指標都有所提升。再對兩群組原則的收益曲線進行對比，結果如圖 4.6 所示。

圖 4.6：兩組策略收益曲線對比

通過對比，我們可以發現，在 2015 年之前，兩群組原則的收益率可以說是不相上下，但從 2015 年中旬開始，初始策略的收益開始高於改良後的策略。一直到 2019 年，改良策略實現了反超。從圖 4.4 以及圖 4.5 中的超額收益曲線也可以看出，從 2017 年初到 2018 年初，兩群組原則的超額收益都呈現下

DOI:10.6814/NCCU202200358

降趨勢，即策略收益都低於基準收益，沒有跑贏大盤。之後，初始收益的超額收益曲線基本為平穩狀態，也就是策略收益與基準收益相當。而改良後的策略，從 2019 年開始超額收益曲線有了上升趨勢，實現了對初始策略的反超。

如圖 4.7 以及圖 4.8，為兩群組原則的年度收益分佈圖，可以看出兩群組原則的最大收益都是在 2015 年，改進後的策略前期收益略低於初始策略，後期損失也更少。總體情況來看，改進後的策略收益波動率較低，有效地降低了風險，提高了總體收益。

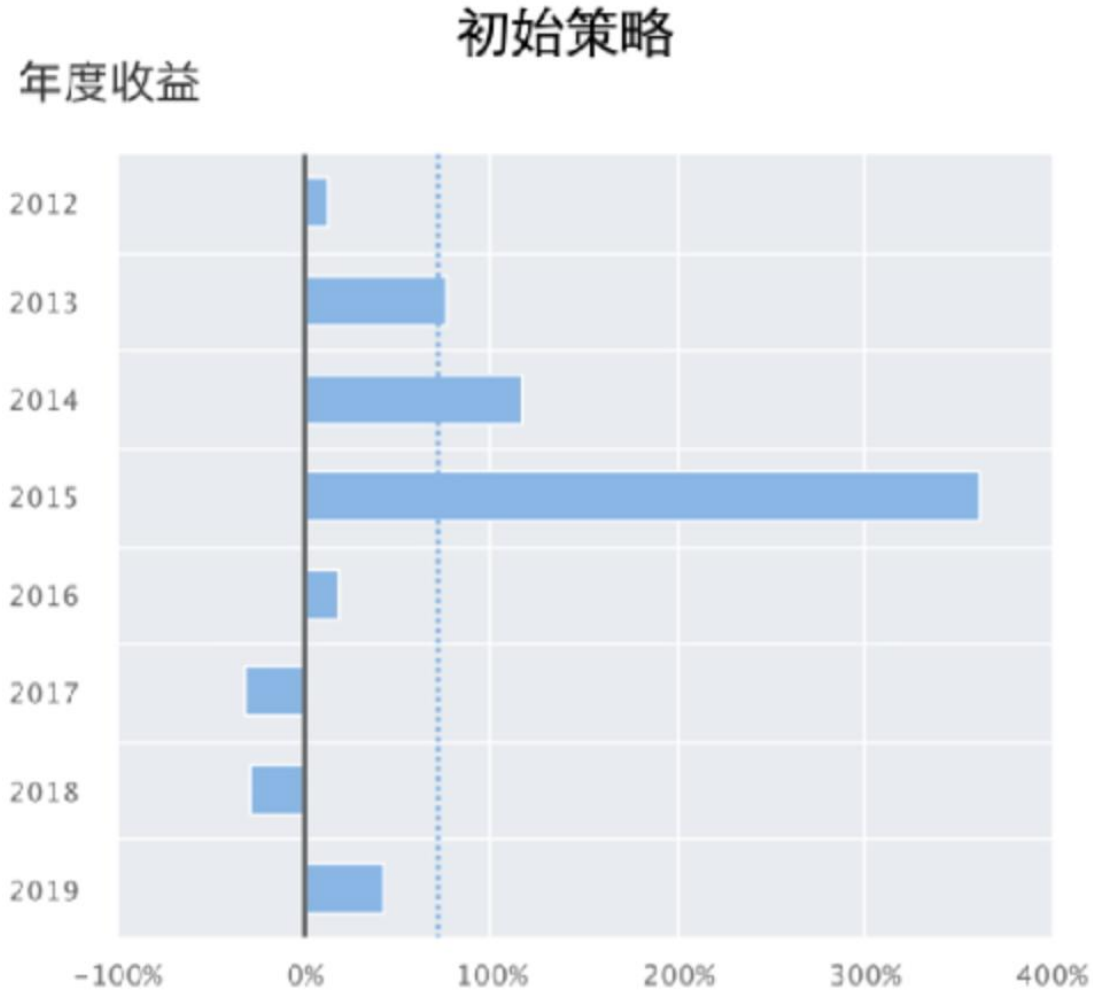


圖 4.7：初始策略的年度收益分佈圖

DOI:10.6814/NCCU202200358

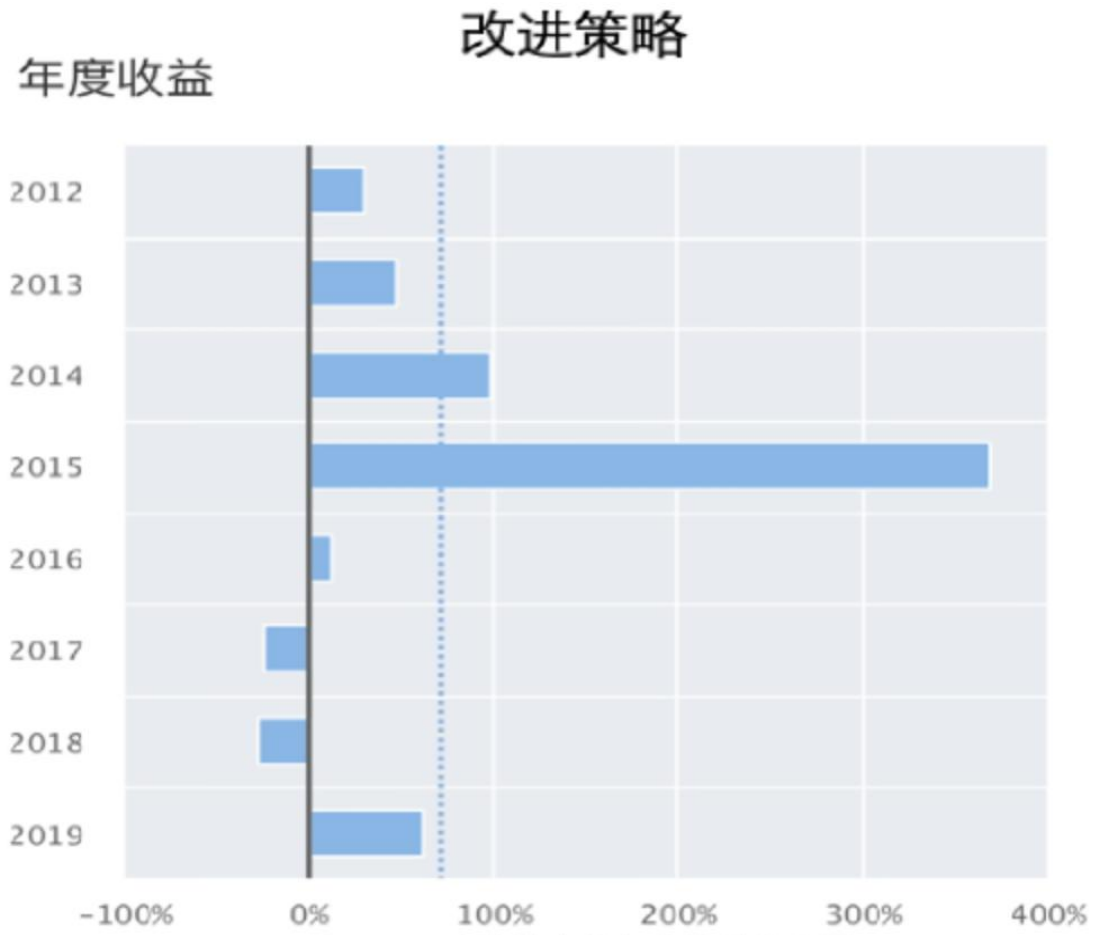


圖 4.8：改進策略的年度收益分佈圖

圖 4.9：初始策略的月度收益熱力圖

DOI:10.6814/NCCU202200358

圖 4.10：改進策略的月度收益熱力圖

上圖 4.9 以及圖 4.10 分別是初始策略以及改進策略的月度收益熱力圖，圖中區塊顏色越趨近于深紅，代表了當月月度收益越大；區塊的顏色越趨近于深綠，代表月度虧損越大；越趨近于純白，代表月度收益越接近於零。熱力圖較為直觀的展現了 2012 年到 2019 年期間每個月份的收益狀況。初始策略的最大收益月份為 2015 年的 5 月份，達到了 35.21%；改進後的策略最大收益月份為 2015 年的 3 月份，達到了 44.77%。虧損最大的月份都為 2016 年的 1 月份，均虧損了 32%左右。從兩群組原則的熱力圖對比中可以看出，兩群組原則的盈利月份大體上相同，改進後的策略盈利月份略多於初始策略，虧損月份略少於初始策略。而通過上文波動預測模型的資料輸出圖可知，2015 年正是市場波動較大的時期，說明策略主要在市場波動巨大時獲利。

如下圖 4.11 以及圖 4.12 所示，分別為初始策略以及改進策略的行業配置對比圖。經過策略篩選的股票主要分為十一類，分別為金融指數、工業指數、可選消費指數、材料指數、能源指數、醫療保健指數、資訊技術指數、日常消費指數、電信服務指數、公用事業指數以及房地產指數。基準配置即滬深 300 股指的行業配置主要以金融指數為主，占 37.76%，其次分別為工業指數、可選消費指數、材料指數、日常消費指數、資訊技術指數、醫療保健指數等。由圖中

DOI:10.6814/NCCU202200358

可知，兩群組原則的金融指數配置百分比都非常低，工業指數大幅增加，占比均超過 30%，其次是可選消費以及材料指數。

初始策略

行业名称	策略配置	基准配置	权重配置差
金融指数	0.96%	37.76%	-36.80%
工业指数	31.18%	14.30%	16.88%
可选消费指数	27.16%	10.35%	16.81%
材料指数	20.77%	7.84%	12.93%
日常消费指数	1.30%	7.62%	-6.32%
能源指数	0.97%	4.03%	-3.06%
医疗保健指数	7.92%	5.60%	2.32%
信息技术指数	5.27%	6.50%	-1.23%
电信服务指数	0.49%	1.12%	-0.63%
公用事业指数	2.42%	2.91%	-0.49%
房地产指数	1.42%	1.73%	-0.31%
现金	0.16%	0.00%	0.16%
其它	0.00%	0.05%	-0.05%

圖 4.11：初始策略的行業配置對比圖

DOI:10.6814/NCCU202200358

改进策略

行业名称	策略配置	基准配置	权重配置差
金融指数	2.66%	37.76%	-35.10%
工业指数	33.56%	14.30%	19.26%
可选消费指数	23.54%	10.35%	13.19%
材料指数	17.94%	7.84%	10.10%
日常消费指数	2.81%	7.62%	-4.81%
医疗保健指数	8.51%	5.60%	2.91%
能源指数	1.14%	4.03%	-2.89%
电信服务指数	0.00%	1.12%	-1.12%
房地产指数	0.65%	1.73%	-1.08%
公用事业指数	2.00%	2.91%	-0.91%
信息技术指数	7.08%	6.50%	0.58%
现金	0.11%	0.00%	0.11%
其它	0.00%	0.05%	-0.05%

圖 4.12：改進策略的行業配置對比圖

總結來說，本文兩群組原則的行業配置比例基本相同，是一個側重於工業、可選消費、材料、醫療保健以及信息技術產業為主的策略。

上文通過對比兩群組原則的收益曲線、年度收益分佈圖、月度收益熱力圖以及行業配置對比圖等等，基本瞭解了改進後的策略與初始策略在收益細節上的區別。為了更加全面細緻地對兩群組原則進行對比，本文將兩群組原則的回測評價指標放入表格中進行直觀的比較分析，具體的評估資訊對照表如表 4.2 所示。

DOI:10.6814/NCCU202200358

风险指标	XGBoost 算法 加波动聚集因子	XGBoost 算法 无波动聚集因子	差值 (改良-初始)
回測时间	2012-01-01 至 2019-12-31	2012-01-01 至 2019-12-31	
回測频率	每天	每天	
初始资金	100000	100000	0
策略收益	1737.01%	1564.09%	172.92%
策略年化收益	45.37%	43.54%	1.83%
基准收益	74.64%	74.64%	0.00%
Alpha	0.382	0.363	0.019
Beta	0.935	0.958	-0.023
Sharpe	1.258	1.189	0.069
胜率	0.606	0.572	0.034
盈亏比	1.819	1.574	0.245
最大回撤	52.90%	61.33%	-8.43%
日胜率	0.58	0.561	0.019
盈利次数	100	103	-3
亏损次数	65	77	-12
Sortino	1.504	1.433	0.071
信息比率	1.526	1.452	0.074
策略波动率	0.329	0.332	-0.003
基准波动率	0.231	0.231	0

表 4.2：兩組策略評估指標對比

從表 4.2 可以看出，改良後的策略除了貝塔係數略低於初始策略之外，其它的評估指標例如，阿爾法指標、夏普比率、索提諾比率、資訊比率、勝率以及盈虧比都較初始策略有所提升。另外，最大回撤、策略波動率都比初始策略要低，盈利次數和虧損次數也更少。綜上所述，加入波動聚集因子對策略的收益率以及魯棒性都有所提升。

為了排除實驗的偶然性，本文通過改變策略的演算法以及參數進行了多組對比實驗。其中一組對比實驗結果如表 4.3 所示，為隨機森林演算法的策略對比結果。

DOI:10.6814/NCCU202200358

风险指标	随机森林 有波动聚集因子	随机森林 无波动聚集因子	差值 (改良-初始)
回测时间	2012-01-01 - 2019-12-31	2012-01-01 - 2019-12-31	
回测频率	每天	每天	
初始资金	100000	100000	0
策略收益	1144.21%	1013.61%	130.60%
策略年化收益	38.27%	36.31%	1.96%
基准收益	74.64%	74.64%	0.00%
Alpha	0.311	0.291	0.02
Beta	0.921	0.941	-0.02
Sharpe	1.062	0.98	0.082
胜率	0.584	0.569	0.015
盈亏比	1.602	1.452	0.15
最大回撤	62.93%	62.38%	0.55%
日胜率	0.579	0.566	0.013
盈利次数	97	95	2
亏损次数	69	72	-3
Sortino	1.257	1.158	0.099
信息比率	1.266	1.161	0.105
策略波动率	0.323	0.33	-0.007
基准波动率	0.231	0.231	0

表 4.2：兩組策略（隨機森林算法）評估指標對比

可以看出,隨機森林演算法的策略收益與 XGBoost 演算法對比降低了許多,策略加入波動聚集作為因子之後，策略收益、年化收益、阿爾法指標、夏普比率、索提諾比率、資訊比率以及盈虧比同樣比初始策略有所提升。策略的波動率同樣減小，盈利次數增加，虧損次數減小，勝率有所提高。綜上所述，將波動聚集因子加入到隨機森林多因子策略中也有類似的提升效果。

第四節　本章小結

本章主要介紹了對於波動聚集現象創新性應用的研究設想和基本思路，並使用 GARCH 模型對波動率進行了預測，同時驗證了滬深 300 指數的市場價格具有顯著的波動聚集現象。之後詳細的介紹了本文基礎策略的幾個階段和具體步驟以及使用到的主要因子。因子主要包含了上一章節內通過 XGBoost 演算法進行因子合成中表現最優秀的 5 個因子以及較為常用的 4 個技術指標。接

DOI:10.6814/NCCU202200358

著將 GARCH 模型加入策略中進行改良優化，將其預測值作為波動聚集因子加入機器學習中進行學習訓練，對初始策略以及改進後的策略進行回測，並對兩群組原則的回測結果進行了分析對比。最後，為了排除對比實驗的偶然性，本文通過改變策略的機器學習演算法以及參數，進行了多組對比實驗，並對結果進行了分析比較。

總體來看，加入波動聚集性作為因子之後，其策略波動率都有所降低。其中個別實驗組改良策略的策略收益、策略年華收益率等指標要低於初始策略，但是改良後的策略勝率都有了提高，並且所有實驗組改良策略的策略波動率以及最大回撤都降低了。同時，改進後的策略，其虧損次數以及總交易次數都有小幅度的減少。另外，所有策略的最大收益月份基本都在 2015 年這段市場波動較大的時間段。並且，在使用 GARCH 模型對滬深 300 指數進行擬合時，可以看出市場收益率的波動聚集現象主要發生在 2015 年以及 2019 年這兩段時期。而加入了波動聚集因子之後，從月度收益熱力圖與年度收益分佈圖的對比中可以看到，模型在 2015 年以及 2019 年這兩段時期的收益都得到了提升，說明波動聚集因子在策略中起到了作用，策略對於波動聚集現象能夠加以學習以及應用，從中獲利。

綜上所述，本文的策略是一個收益比較穩定，並且收益趨勢與大盤較為一致，策略在股市震盪期間比較有優勢的策略。另外，通過對實驗組的歸納總結，可以得出，加入了波動聚集作為因子之後，儘管不能保證策略的收益百分之百會有所提高，但是策略的風險性會有所下降，改進後的策略會顯得更加穩健，魯棒性更強。

DOI:10.6814/NCCU202200358

第六章　結論與建議

第一節　研究結論

本文主要在聚寬量化平臺上研究了機器學習演算法和波動聚集對於量化投資的優化應用,將目前機器學習演算法中較為風靡的 XGBoost 演算法應用於量化投資的幾個方面。通過對前人學者的理論文獻研究，為本文的基礎理論知識奠定了基調，明確了研究方向。緒論部分主要介紹了文本的研究背景及研究意義、相關概念辨析、文獻綜述、研究設計框架以及本文的創新點。

本文首先介紹了集成學習演算法和 XGBoost 演算法的理論知識,對波動率模型做了相關的知識介紹，其中包括BS模型、隨機波動模型以及幾個時間序列模型。重點講述了目前較為風靡的 XGBoost 演算法以及能夠刻畫驗證波動聚集現象的 GARCH 模型。為下文的實證研究內容打下理論基礎。

然後本文詳細地論述了如何使用 XGBoost 演算法進行因子合成,包括資料的獲取、資料預處理、樣本內的訓練、模型的構建以及樣本外測試等步驟。模型構建的過程中，對 XGBoost 演算法、隨機森林演算法以及 SVM 演算法的精確度及 AUC 指標進行了對比分析，驗證了確實是 XGBoost 演算法要更優於其他兩種演算法。並與傳統的線性模型做了對比，證明了機器學習在因子合成方面的高效性及其優勢。為了提高模型的學習效果及綜合能力，將回測區間分為 6 個子區間,採用六階段滾動回測的方法訓練模型。接著通過 XGBoost 演算法中內置的因子打分機制獲得了所有因子 6 次訓練的重要度指標，對指標進行排序處理，利用五分位數分組法將因子分為 5 組進行分層回測。回測結果與預期結果基本一致，重要度指標越高的分組，回測收益率越高。基本驗證了通過本文的因子合成方法，可以獲得一些表現較好的因子。

接著本文主要圍繞著金融市場收益率的波動特性進行展開，講述了本文對於波動聚集現象創新性應用的研究設想和基本思路。接著使用第三章節實證結論中表現效果最優秀的一批因子進行多因子策略的構建,使用 XGBoost 演算法對股票價格進行預測，將此策略作為初始策略，即對照組，與改進後的策略進行對比研究。最後，本文通過 GARCH 模型對市場價格進行擬合，刻畫市場的波動聚集現象，創新地將模型預測值作為波動聚集因子加入到策略中進行策略

DOI:10.6814/NCCU202200358

的改進提升，與對照組進行回測結果的對比分析，觀察波動聚集現象的應用效果有哪些，是否對策略的提升有所幫助。通過多組實驗組的結果分析對比，我們發現加入波動聚集因子的策略，在穩定性上要強於初始策略，能夠提高穩定收益，達到了降低風險與策略波動率的目的。並且結合滬深 300 指數的市場收益率波動圖來看，在 2015 年和 2019 年這兩段有較為顯著的波動聚集現象的時間段上，改進後的策略比初始策略表現的更加優秀。實證研究結果顯示，通過機器學習對波動聚集現象加以學習應用，能夠使策略在股市震盪期更有優勢，並且能夠從總體上提高策略的穩健性，達到降低投資風險的目的。

第二節　改進與展望

關於本文的研究展望，可以拓展 VIX（Volatility of S&P 500 Index）相關方面的內容。VIX 是芝加哥期權交易所市場波動率指數，它誕生於 1993 年，現在用來衡量 S&P 500 指數期權所給出的隱含波動率，又稱“恐慌指數”。可以考慮將本文的研究方法應用於波動率指數 VIX 相關的衍生品投資。

此外，由於本文的演算法研究以及策略應用都是基於聚寬量化平臺進行的，並且作者本人是沒有經過系統的量化訓練的量化初學者。所以本文的重點在於利用機器學習演算法對量化投資的思路擴展和研究創新，本文使用的策略也是較為簡單和基礎的策略，還存在許多不足之處有待改進。

第一個有待改進的地方是因子合成部分的因子分組問題。由於一些因子之間會存在一些相關性，因此重要度指標並不能作為判斷因子好壞的唯一標準。如何加入新的考慮因素對因子進行更細化地優劣排序，這是一個可以提高改進的地方。

第二點是本文策略還有待進一步的優化。由於作者本人的程式設計能力以及量化策略研究能力不足，本文的策略還有很大的提升空間。具體來講，由於策略中使用的機器學習演算法是分類演算法，因此對股票進行預測時，僅僅區分為漲與不漲兩類，如果可以對預測為漲股票按照漲幅大小進行排序，預測漲幅更大的股票優先順序更高，執行買入操作時優先買入，相信可以提高策略的收益效果，這可以作為下一步的研究方向。

第三個值得考慮研究的方向是波動聚集現象的應用方法。目前對於波動聚集現象的應用仍十分稀少，本文的應用方法也比較單一。筆者認為還可以在其

DOI:10.6814/NCCU202200358

他方向加以應用，例如有些策略在股市波動較大時容易損失，則可以考慮將波動聚集作為一個指標進行動態調倉，當預測波動率超過一定數值時，進行減倉操作甚至是清倉退出市場，降低風險以達到提高勝率的目的。再例如本文中的策略，在波動較平穩時收益不高，甚至會造成損失，則可以設置為當預測波動率低於一定界值時，執行減倉操作，減少損失。筆者認為，對於波動聚集現象的實際應用這方面的研究還有待進一步深化，是一個十分有意義的研究方向。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202200358

參考文獻

[1] 肖晞暉，（2018）。基於大資料和機器學習的量化選股模型研究。華中師範大學。

[2] 鄒玉江，（2018）。基於機器學習的滬深 300 指數走勢預測研究。山東大學.

[3] 楊世林，（2018）。基於聚寬量化投資平臺的股票多因子策略應用。浙江大學。

[4] Harris M, Raviv A, (1993). Differences of opinion make a horse race. The Review of Financial Studies, 6(3):473-506.

[5] Harry Markowitz, (1952). Portfolio selection Journal of Finance. March, 7(1):77-91.

[6] Black Fischer, Myron Scholes, (1973). The Pricing of Options and Corporate Liabilities. Journal of Political Economy, 81(3):637-654.

[7] Lukac, Brorsen, Irwin, (1986). A comparison of twelve technical trading systems with market efficiency implications. Station bulletin - Dept. of Agricultural Economics, Purdue University, Agricultural Experiment Station (USA).

[8] V.N Vapnik, (1995). The Nature of Statistical Learning Theory. New York: Springer-Verlag.

[9] Leo Breiman, (2001). Random Forests. Machine Learning, 45(1):5-32.

[10] Min, Lee, (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications, 28(4):603-614.

[11] Kim K.J, 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.

[12] Chen.Y.S, Cheng.C.H, (2009). Evaluating industry performance using extracted RGR rules based on feature selection and rough sets classifier. Expert Systems with Applications, 36(5):9448-9456.

[13] MC Lee, (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications, 36(8):10896-10904.

[14] Majhi, Panda, Sahoo, (2009). Development and performance evaluation of FLANN based model for forecasting of stock markets. Expert Systems with Applications, 36(3):6800-6808.

[15] Ticknor, Jonathan, (2013). A Bayesian regularized artificial neural network for stock market forecasting. Expert Systems with Applications, 40(14):5501-5506.

[16] Somani, Talele, Sawant, (2015). Stock market prediction using Hidden Markov Model. Information Technology & Artificial Intelligence Conference.

[17] Ballings, Poel, Hespeels, (2015). Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 42(20):7046-7056.

[18] Chen, Carlos, (2016). XGBoost: A Scalable Tree Boosting System.

[19] Mandelbrot B, (1963). New Methods in Statistical Economics. Journal of Political Economy, 71(5):421-440.

[20] Engle R F, (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica,

DOI:10.6814/NCCU202200358

50(4):987-1007.
[21] Tim Bollerslev, (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3):307-327.
[22] Zakoian J M, (1994). Threshold heteroskedastic models. Journal of Economic Dynamics and Control, 18(5):931-955.
[23] Nelson, Daniel B, (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. Modelling Stock Market Volatility, 59(2):347-0.
[24] Bollerslev T, (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 74.
[25] Michael D. McKenzie, Heather Mitchell, Robert D, Brooks & Robert W, (2001). Power ARCH modelling of commodity futures data on the London Metal Exchange. The European Journal of Finance, 7(1):22-38.
[26] Chin Wen Cheong, Nor A H S M, Isa Z, (2007). Asymmetry and long-memory volatility: Some empirical evidence using GARCH. Physica A, 373(none):651-664.
[27] 吳微，陳維強，劉波，（2001）。用 BP 神經網路預測股票市場漲跌。大連理工大學學報，2001(01):9-15。
[28] 彭麗芳，孟志青，姜華，田密，（2006）。基於時間序列的支援向量機在股票預測中的應用。計算技術與自動化，2006(03):88-91。
[29] 錢穎能，胡運發，（2007）。用樸素貝葉斯分類法選股。電腦應用與軟體，2007(06):90-92。
[30] 左輝，樓新遠，（2008）。基於貝葉斯分類的選股方法。電腦知識與技術，2008(10):173-176+185。
[31] 余樂安，汪壽陽，（2009）。基於核主元聚類的股票分類。系統工程理論與實踐，29(12):1-8。
[32] 吳曼琪，（2010）。基於 K 均值聚類的 ST 股票分類研究及投資策略。中國城市經濟，2010(08):26+29。
[33] 劉毅（2012）。因子選股模型在中國市場的實證研究。復旦大學。
[34] 蘇治，傅曉媛，（2013）。核主成分遺傳演算法與 SVR 選股模型改進。統計研究，30(05):54-62。
[35] 曹正鳳，紀宏，謝邦昌，（2014）。使用隨機森林演算法實現優質股票的選擇。首都經濟貿易大學學報，16(02):21-27。
[36] 王淑燕，曹正鳳，陳銘芷，（2016）。隨機森林在量化選股中的應用研究。運籌與管理，25(03):163-168+177。
[37] 王莉，（2016）。基於人工智慧演算法的股票價格波動規律預測方法研究。吉林大學。
[38] 張昊，紀宏超，張紅宇，（2017）。XGBoost 演算法在電子商務商品推薦中的應用。物聯網技術，7(02):102-104。
[39] 淩筱玥，（2017）。基於 XGBoost 演算法的上證指數預測方案設計研究。上海師範大學。
[40] 田浩，（2018）。基於 XGBoost 的滬深 300 量化投資策略研究。上海師範大學。
[41] 唐齊鳴，陳健，（2001）。中國股市的 ARCH 效應分析。世界經濟，2001(03):29-36。
[42] 惠曉峰，柳鴻生，胡偉，et al，（2003）。基於時間序列 GARCH 模型的人民幣匯率預測。金融研究，2003(5)。
[43] 曹野，（2012）。基於 GARCH 族模型的黃金價格收益率及波動性研究。

DOI:10.6814/NCCU202200358

價值工程，2012(02):163-165。
[44] 周好文，徐守喜，謝金靜，(2012)。過度關注與噪音交易。雲南財經大學學報，28(3)。

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202200358

附錄

因子合成部分全部因子資訊表

因子類型	因子名稱	因子描述	計算方法
財務品質因子	current_ratio	流動比率(單季度)	流動比率=流動資產合計/流動負債合計
財務品質因子	net_profit_to_total_operate_revenue_ttm	淨利潤與營業總收入之比（NI）	淨利潤與營業總收入之比=淨利潤（TTM）/營業總收入（TTM）
財務品質因子	gross_income_ratio	銷售毛利率（GPM）	銷售毛利率=(營業收入（TTM）-營業成本（TTM）)/營業收入（TTM）
財務品質因子	roe_ttm	權益回報率 TTM（ROE）	權益回報率=淨利潤（TTM)/期末股東權益
財務品質因子	roa_ttm	資產回報率 TTM（ROA）	資產回報率=淨利潤（TTM）/期末總資產
財務品質因子	total_asset_turnover_rate	總資產周轉率	總資產周轉率=營業收入(ttm)/總資產
財務品質因子	net_operating_cash_flow_coverage	淨利潤現金含量	經營活動產生的現金流量淨額/歸屬於母公司所有者的淨利潤
財務品質因子	cash_to_current_liability	現金比率	期末現金及現金等價物餘額/流動負債合計的 12 個月均值
成長因子	operating_revenue_growth_rate	營業收入增長率	營業收入增長率=（今年營業收入（TTM）/去年營業收入（TTM））-1
成長因子	net_profit_growth_rate	淨利潤增長率	淨利潤增長率=(今年淨利潤（TTM）/去年淨利潤（TTM）)-1
成長因子	PEG	市盈率相對盈利增長比率	PEG = PE / (歸母公司淨利潤(TTM)增長率 * 100) # 如果 PE 或增長率為負，則為 nan
動量因子	BBIC	BBI 動量	BBI(3, 6, 12, 24) / 收盤價（BBI 為常用技術指標類因子“多空均線”）

DOI:10.6814/NCCU202200358

風險因子	Variance20	20 日年化收益方差	20 日年化收益方差
杠杆因子	CMV	對數流通市值	流通市值的自然對數
杠杆因子	financial_leverage	財務杠杆	總資產/淨資產
杠杆因子	debtequityratio	債資比率	非流動負債/淨資產
杠杆因子	cashratio	現金比率	現金比率=(貨幣資金+有價證券)/流動負債
杠杆因子	currentratio	流動比率	流動比率=流動資產/流動負債*100%
基礎因子	net_operate_cash_flow_ttm	經營活動現金流量淨額 TTM	計算過去 12 個月經營活動產生的現金流量淨值之和
基礎因子	net_profit_ttm	淨利潤 TTM	計算過去 12 個月淨利潤之和
基礎因子	non_recurring_gain_loss	非經常性損益	歸屬於母公司股東的淨利潤-扣除非經常損益後的淨利潤(元)
基礎因子	operating_revenue_ttm	營業收入 TTM	計算過去 12 個月的營業收入之和
技術指標因子	DIF	差離值	EMA(CLOSE,SHORT)-EMA(CLOSE,LONG)
技術指標因子	DEA	指差離值平均數（異同平均數）	EMA(DIF,MID)
技術指標因子	MACD	指數平滑異同移動平均線	(DIF-DEA)*2, COLORSTICK
技術指標因子	RSI	相對強弱指數	RSI=100*RS/(1+RS)
情緒因子	AR	人氣指標	AR=N 日內（當日最高價-當日開市價）之和 / N 日內（當日開市價-當日最低價）之和 * 100，n 設定為 26
情緒因子	BR	意願指標	BR=N 日內（當日最高價-昨日收盤價）之和 / N 日內（昨日收盤價－當日最低價）之和 ×100，n 設定為 26

DOI:10.6814/NCCU202200358

情緒因子	VDIFF	計算 VMACD 因子的中間變數	EMA(VOLUME, SHORT)-EMA(VOLUME，LONG) short 設置為 12，long 設置為 26，M 設置為 9
情緒因子	ARBR	ARBR	因子 AR 與因子 BR 的差
估值因子	EP	淨利潤與總市值之比，市盈率（PE）的倒數	PE=1/（股權投資回報率-每股收益增長率）
估值因子	BP	淨資產與總市值之比，市淨率（PB）的倒數	PB=資產收益率/（股權投資回報率-每股收益增長率）
估值因子	SP	營業收入與總市值之比，市銷率（SP）的倒數	PS=總市值/銷售額
估值因子	DP	股息率	DP=最近 12 個月現金紅利/總市值
估值因子	RD	開發支出與總市值之比	RD=開發支出/總市值
估值因子	CFP	經營現金流與總市值之比，市現率（PCF）的倒數	PCF=總市值/經營現金流

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202200358