國立陽明交通大學

資訊管理與財務金融系

課堂論文

多目標遺傳演算法優化 S&P 500 股票預測模型的特徵選擇

Multi-objective genetic algorithm for optimizing feature selection in S&P 500 stock prediction model

研究生:郭東輝

中華民國一零二年五月

摘要

這篇研究主要基於多目標遺傳演算法，用於選擇最佳的股票特徵指標並應用於美國股市預測。它透過演化過程找到同時最佳化波動度和報酬率的股票特徵，並將這些特徵應用於預測股票價格的波動性和報酬率，主要通過多目標基因演算法和機器學習模型，並對資料進行了特徵選擇和預測未來個股波動率和報酬率。最終，並篩選同時排名波動度前 30%低和報酬率前 30%高股票作為投資組合依據，並每四個月重新預測一次波動率與報酬率並重新建立投資組合。並觀察總體績效和風險，我發現在 2021 年至 2023 年間，由基因演算法篩選的股票特徵能有效建構低波動率和高報酬率的投資組合，在所有投資組合中計算了投資組合報酬率、最大回落比率和 Calmar 比率，大部分報酬率與風險皆優於標準普爾 500 指數，且在調整市場報酬後投資組合報酬率在統計檢定上顯著大於 0，且仍然保持了與市場整體趨勢的一致性。這種相關性的存在使得本研究股票選擇方法能夠更好地應對市場波動風險，並在市場上展現出較穩定的投資回報，總體來說，這個研究提供了一個基於多目標遺傳演算法的股票特徵選擇和投資組合建構方法，這對於投資者在股市中做出更明智的決策和配置資源非常有價值。

目錄

表目錄

圖目錄

第一章 緒論

在過去量化金融的研究中，依照研究方向大略可以區分為三大領域，第一為資產選擇，主要透過篩選潛在有成長機會股票並進行投資，第二為投資組合理論，將資金有效分配至不同資產上，透過理論上最佳化投資組合風險收益配置，例如報酬率或風險，第三為擇時，透過研究市場趨勢、技術指標和基本面因素，以確定最佳的進出點。本研究專注於資產選擇，細分來說，主要關注股票選擇，透過規則標記股票建立好股票與壞股票的分類，本研究考慮了類似 Sharpe ratio 的基本概念，綜合納入了收益及風險作為考量，建立了一套股票篩選機制，除此之外，建立 174 個股票特徵，包含價量資料、週期指標、數值變換、動量指標、圖表形態識別、價格變換、統計函數、波動性指標、交易量指標和重疊研究指標，本研究採用多目標基因演算法在所有可能空間中進行全局最優子集搜索，並使用篩選出的最佳特徵子集建構投資組合，並在後續表現中優於標準普爾 500 指數，為了衡量特徵子集的好壞，本研究將多目標基因演算法適應度定義為機器學習模型在對應的特徵子集上針對股票波動度和報酬率訓練時的均方誤差（Mean Squared Error），並設定同時最小化兩種適應度值。建立全局最佳特徵子集後，並透過機器學習模型以佳特徵子集各別建立每支股票預期波動度與報酬率，並篩選同時排名波動度前 30%低和報酬率前 30%高股票作為投資組合依據，並每四個月重新預測一次波動率與報酬率並重新建立投資組合，並觀察投資組合績效。本研究模型建構的投資組合在回測中優於市場平均水平。

第二章 文獻探討

股票市場的預測一直是投資者和研究人員關注的焦點之一。近年來，隨著技術分析和機器學習技術的發展，學術界開始探索使用股票特徵選股的方法，以提高股票預測模型的準確性和效能。主要的研究方法包括使用不同的模型，例如流行的機器學習模型或深度學習方法，以正確預測資本市場中的資產價格。然而，在資本市場中，不同資產具有多種性質，因此如何選擇適當的特徵進行分析成為主要的課題。過去的研究文獻中，Tsantekidis 等人（2017）利用卷積神經網絡（CNN）從限價單簿（limit order book）預測股票價格。研究結果顯示，使用 CNN 模型能夠在股票價格預測方面取得良好的準確性。Guo 等人（2019）提出了一種基於多目標深度強化學習的股票選擇方法。該方法在強化學習框架下設定多個目標並建立投資組合，以達到多個目標的最佳化。此外，Zhang 等人（2019）探討了股票市場分析中的特徵選擇和分類方法。他們提出了一種基於綜合權重的特徵選擇方法，並結合支持向量機（SVM）進行分類分析。這種方法能夠選擇具有關鍵影響力的特徵。Han 等人（2019）認為特徵選擇對於提高股票市場預測準確性具有重要作用，並且在不同市場條件下的表現存在差異。本研究試圖以 Sharpe ratio 為基本概念，建立衡量股票價值的方法，並透過多目標最佳化找出優秀的股票特徵因子。研究目標為找到影響標準普爾 500 指數成分股的關鍵因子，並利用這些關鍵因子建立一個優秀的投資組合。研究旨在提供投資者更具價值的股票選擇策略，並進一步優化投資組合的風險收益特性。這些研究結果對於投資者在股票市場中做出明智的選擇和決策提供了重要的參考依據。

第三章 研究方法

3.1 資料

本研究使用的股價資料為 yahoo finance 股價資料庫，總共抓取了 250 支股票，從 2019/01/01 至 2023/04/30 資料，資料分兩部分，2019/01/01 至 2019/10/30 資料為訓練模型與多目標演算法部分，剩下的部分為測試投資組合績效期。每一支股票，總共 250 檔，每一支股票都會計算 180 種股票特徵，建構股票特徵資料，股票特徵主要分為 10 種類別，分別為價量資料、週期指標、數值變換、動量指標、圖表形態識別、價格變換、統計函數、波動性指標、交易量指標和重疊研究指標，其中有些指標因為數值過大，並增加計算負擔，例如指數函數，故刪除，最後共選出了 174 個股票特徵作為篩選，在多目標基因演算法中進行全局最優子集搜索。

3.2 基因演算法

本研究主要透過多目標基因演算法，用於解決特徵選擇的最佳化問題。該問題的目標為從一組股票特徵中選擇最佳的子集，以用於股票價格預測。

3.2.1 初始化

本研究假設族群大小為 20，代表隨機生成 20 個以二進制為基礎的向量，每個基因由二項分布獨立同分布生成，總共生成長度 174，由 0 和 1 組成的基因序列。

3.2.2 適應度函數

為了衡量個體是否為優良的個體，需要建立一個適應度函數作為衡量個體優劣的標準，本研究設定小的適應度為較優秀的個體，代表他相較於其他個體優秀。我將個體的適應度函數分為兩個部分，第一個部分為機器學習模型在所有股票，共 250 支股票，對波動度預測訓練集上的均方誤差（Mean Squared Error），第二部分為機器學習模

型在每一支股票對報酬率預測訓練集上的均方誤差（Mean Squared Error），樣本期間由圖 3.1 所示。

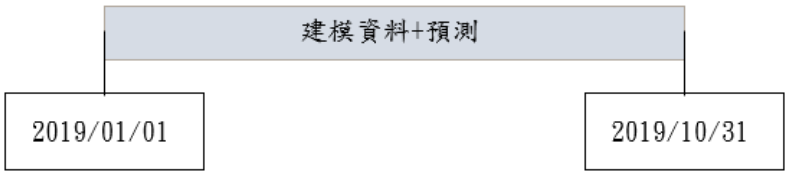

圖 3.1 適應度函數樣本期間

3.2.3 選擇與進化

在計算完兩部分的適應度函數後，就可以通過所有個體適應度的比較來進行適應度排序，依照由小到大，意味好到壞，並在每次演化中，先進行交配和突變操作生成子代(突變率為 0.5、交配率為 0.5)，然後計算子代的適應度值，並將適應度分配給子代個體。最後，根據其適應度大小選擇下一代的個體形成更新的族群，總共會經過 50 次的迭代，停止進化並選擇最佳個體作為最佳股票特徵子集。

3.3 機器學習

3.3.1 隨機森林

本研究選擇隨機森林作為適應度函數，透過訓練集建立多棵決策樹，對訓練集波動度和報酬率分別進行擬合，隨機森林為使用 bagging 加上隨機特徵產生出來的演算法，可以使結果不太容易過擬合，本研究總共有兩部分使用隨機森林模型進行擬合，第一部分為基因演算法中適應度函數的計算，第二部分為建立投資組合時預期波動度和預期報酬率的模型，第一部部分樣本期間由圖 3.1 所示，適應度函數為整個訓練期間的均方誤差（Mean Squared Error），第二部分則是將樣本 70%作為訓練集，剩下 30%作為測試集，本研究將每隻股票測試集平均預測波動度或報酬率作為建立投資組合之依據。樣本期間由圖 3.2 所示。

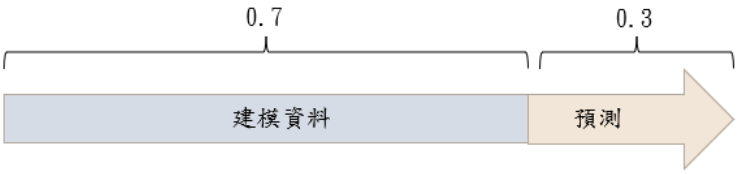

圖 3.2 樣本空間示意圖

3.4 投資組合

本研究試圖從大量的股票中，選擇低波動度的股票有助於降低整個投資組合的波動性和風險。且同時選擇高報酬潛力的股票有助於實現更好的投資回報。建構這樣的投資組合主要有一些優勢，透過選擇低波動度的資產，投資者能夠降低整體投資組合的風險水平。這是因為波動度較低的資產通常具有較穩定的價格波動，相對較少受到市場波動的影響，有助於保護投資組合免於劇烈波動和潛在損失，雖然低波動度資產的收益可能相對較低，但選擇具有高收益潛力的低波動度資產，投資者可以實現更好的收益-風險平衡。這些資產可能具有較穩定的長期增長潛力，並能夠提供相對穩定的投資回報。故使用多目標基因演算法對股票的特徵進行篩選，以找到最佳的股票特徵，並透過這些特徵對每支股票進行波動與報酬預測，並透過篩選的方式，建立低波動高收益投資組合。，具體而言，本研究只篩選同時排名波動度前 30%低和報酬率前 30%高股票作為投資組合依據，並每四個月重新預測一次波動率與報酬率並重新建立投資組合，最後比較投資組合績效。

第四章 研究結果

4.1 投資組合建構

投資組合建構期間為 2021 年 1 月 1 日到 2023 年 4 月 30 日，每四個月調整一次投資組合，總共有 7 個投資組合，每一個投資組合都會隨時間跟新預測波動度和報酬率樣本資料，確保能用到最新資料，關於衡量投資組合，本研究使用三個指標，為報酬率、最大回落程度(Max Drawdown)和 Calmar ratio，共同衡量投資組合報酬和風險程度，最後再與 sp500 指數比較績效，所有投資組合績效如表 4.1 所示，可發現在 7 個投資組合中，有六個投資組合能夠獲得正向的報酬，且有 5 個投資組合 Calmar ratio 皆大於一，代表投資組合在過去的表現中呈現較穩定的回報，且穩定且可靠的投資組合，因為它們具有較低的波動性和風險，但此之外還必須考慮相對於指數表現，故進一步考慮調整市場報酬後績效，由於投資組合是由 250 支 sp500 中股票，故選擇標準普爾 500 指數作為市場投資組合，調整後投資組合如表 4.2 所示，由表可發現，在經過調整後，7 個投資組合中仍有 6 個投資組合仍維持正報酬，代表表現超越了整個市場的平均表現，有可能這個投資組合在長期能夠獲得超額收益，且風險在大部份投資期間皆優於市場指數(Calmar ratio 在七個時間段有五段是優於市場指數)，且調整後報酬統計檢定 t 值為 4.929820066，在 1%信心水準下顯著大於零，代表整個投資組合是能獲得顯著超額收益，由觀察這 7 個投資組合與標準普爾 500 指數的走勢(圖 4.1 所示)，可以發現它們與標準普爾 500 指數之間存在高度相關性。這意味著即使在追求優異績效的同時，依然保持了與市場整體走勢的一定一致性。這種相關性的存在，使得本研究的股票選擇方法在市場波動期間能夠更好地應對風險，並且在市場上表現出較為穩定的投資回報。

表 4.1 投資組合績效

	報酬率	最大回落程度	Calmar ratio
投資組合	6.6%	3.71%	1.78
投資組合	10.54%	2.03%	5.2
投資組合	5.99%	5.37%	1.12
投資組合	5.12%	5.0%	1.02
投資組合	7.49%	11.91%	0.63
投資組合	-1.08%	18.35%	-0.06
投資組合	9.85%	8.66%	1.14

註:此表由上至下為 2021 年 1 月 1 日到 2023 年 4 月 30 日，每四個月調整一次投資組合，報酬率為未年化報酬率，以單利計算，最大回落程度為報酬率在當期回落最大程度，Calmar ratio 為報酬率除上最大回落程度

表 4.2 調整後投資組合績效

	報酬率	最大回落程度	Calmar ratio
投資組合	0.07%	3.71%	1.78
投資組合	4%	2.03%	5.2
投資組合	-3.63%	5.37%	1.12
投資組合	7.74%	5.0%	1.02
投資組合	3.66%	11.91%	0.63
投資組合	7.78%	18.35%	-0.06
投資組合	0.88%	8.66%	1.14

註:此表由上至下為 2021 年 1 月 1 日到 2023 年 4 月 30 日，每四個月調整一次投資組合，報酬率為經由市場報酬率調整後未年化報酬率，以單利計算，最大回落程度為報酬率在當期回落最大程度，Calmar ratio 為報酬率除上最大回落程度

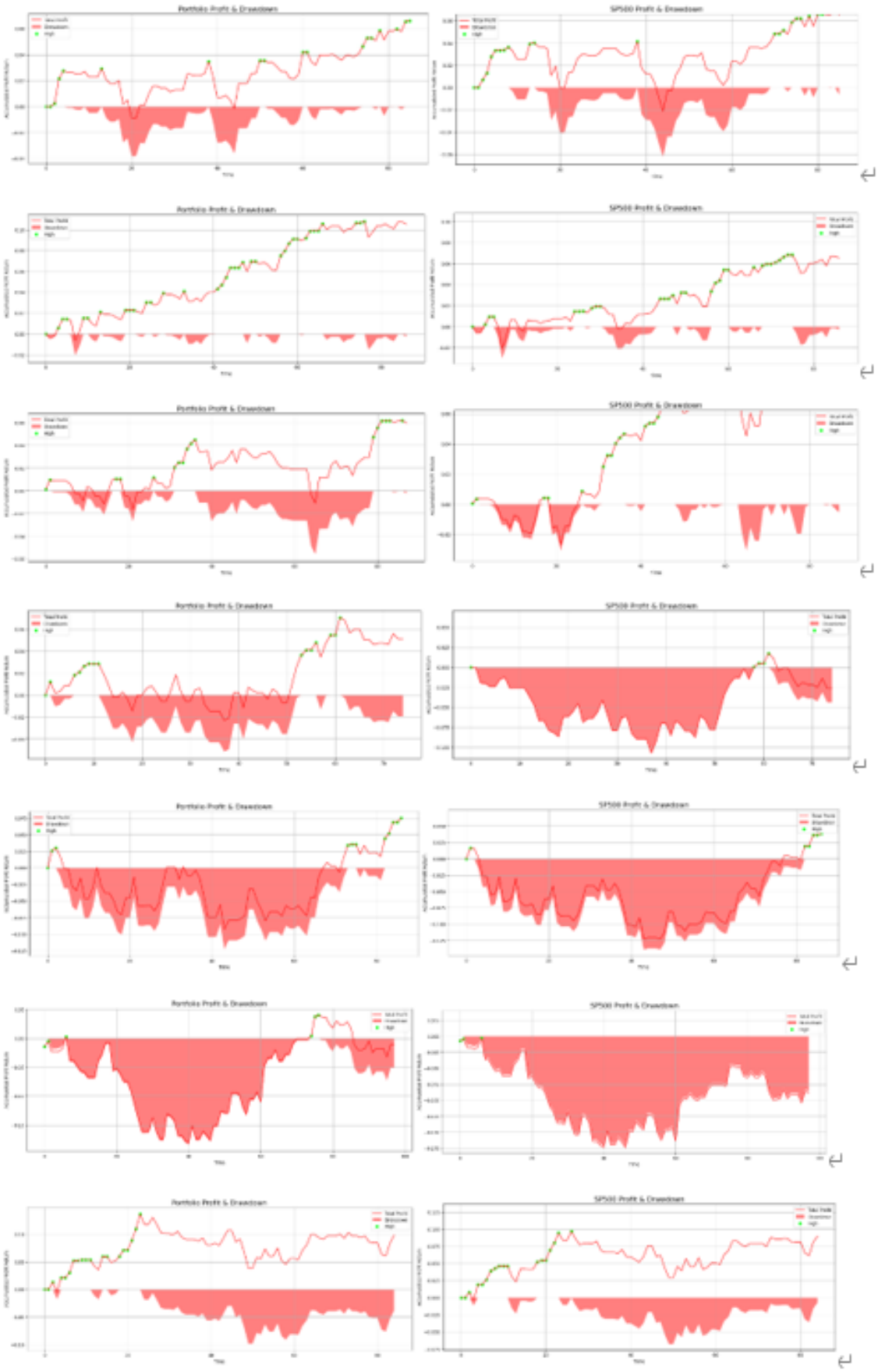

圖 4.1 投資組合與指數比較

第五章 結論

在大多數測試期間，投資組合在報酬率和風險方面表現優於標準普爾 500 指數。這顯示通過多目標最佳化策略，本研究成功地選擇了最佳的股票特徵並建立了一系列優秀的投資組合。這些投資組合能夠有效地過濾低報酬和高風險的股票，從而在大部分實證期間超越了指數。這一成就使得整體投資組合的報酬率更高，風險水平更低。除此之外，觀察這七個投資組合的走勢，可以發現它們與標準普爾 500 指數之間存在高度相關性。這意味著即使在追求卓越表現的同時，本研究仍然保持了與市場整體趨勢的一致性。這種相關性的存在使得本研究的股票選擇方法能夠更好地應對市場波動風險，並在市場上展現出較穩定的投資回報。

參考文獻

1. Tsantekidis, A., Passalis, N., Tefas, A., & Kanniainen, J. (2017). Forecasting stock prices from the limit order book using convolutional neural networks. IEEE Transactions on Neural Networks and Learning Systems, 29(11), 5447-5459.
2. Guo, Y., Feng, Y., & Yang, J. (2019). Stock selection based on multi-objective deep reinforcement learning. Neurocomputing, 335, 32-43.
3. Zhang, X., & Shen, Q. (2019). Feature selection and classification in stock market analysis. IEEE Access, 7, 135646-135658.
4. Han, Y., Wang, S., & Li, Z. (2019). Feature selection for stock market prediction: A comprehensive review. IEEE Access, 7, 77452-77467.