기술적 지표와 재무제표 기반의 시너지 알파 생성과 랜덤포레스트 앙상블을 사용한 중기 투자 기법

신홍기, 정석현, 윤민혁, 홍성호, 김의연, 조영진, 최용훈
광운대학교

{ghdrl95, jayze3736, gurals9368, gyogook608, dmldussla93}@gmail.com,
{yjaycho, yhchoi}@kw.ac.kr

Mid-Term Investment Technique Using Synergistic Formulaic Alpha Generation Based on Technical Indicators and Financial Statements with Random Forest Ensemble

Hong-Gi Shin, Sukhyun Jeong, Min-Hyeok Yun, Seongho Hong, Euiyeon Kim,
Youngjin Cho, Yong-Hoon Choi
Kwangwoon University

요 약

본 논문은 기술적 지표와 재무제표 기반의 synergistic formulaic alpha generation 과 랜덤포레스트 기반의 앙상블을 도입하여 alpha set 기반의 안정적인 주식 종목 선택 기법을 제안한다. 재무제표의 정보 업데이트 주기를 고려하여 formulaic alpha 가 목표한 주가 변동일을 90 일로 설정하고 이에 따라 학습과 검증 기간에 테스트 기간의 주가 변동을 관측하지 못하도록 여유 기간을 설정하였다. S&P500 데이터의 5 개의 서로 다른 기간으로 다양한 시장 상황에 안정적인 성능을 내는지 실험하였고, 기술적 지표 기반의 synergistic formulaic alpha 에 비해 평균적으로 더 높은 IC 와 Rank IC 를 달성하였다.

Ⅰ. 서론

퀀트 투자에서 alpha mining 은 금융 시장에서 초과 수익을 창출할 수 있는 요소인 alpha factor 를 발견하고 개발하는 과정이다. 이 과정에서 feature engineering 은 중요한 역할을 하며, raw feature 에서 유의미한 패턴을 추출하여 알파 팩터로 전환하는 작업을 포함한다.

formulaic alpha factor mining 은 인공 지능 연구의 한 분야로 주식 거래와 관련된 원시 특성에서 미래 수익과 높은 상관관계를 가진 공식을 생성하는 과정을 포함한다. 이는 다양한 연산자와 피연산자를 사용하여 생성된 수식적 팩터로 최근에는 기계 학습 모델을 사용하여 자동으로 이러한 표현식을 생성하는 방법이 연구된다 [1].

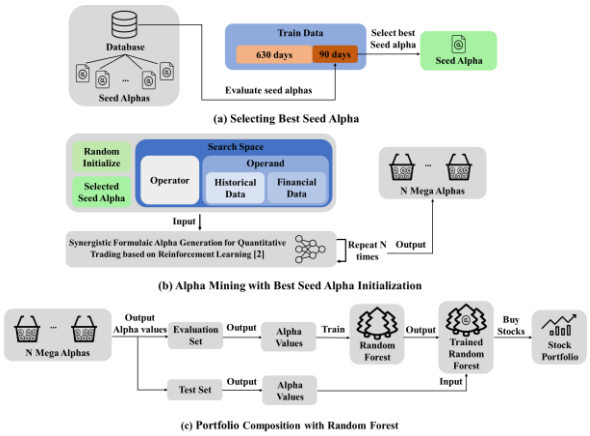

그림 1. 제안한 알파 마이닝 방법, (a)는 seed 알파 선택 방법, (b)는 seed 알파 초기화 및 재무제표를 이용한 알파 마이닝, (c)는 랜덤 포레스트 기반의 주식 종목 선택 방법

최근 formulaic alpha factor mining 동안 알파 팩터 조합의 성능 최적화에 대한 연구가 진행되고 있다 [2]. formulaic alpha factor mining 에 사용할 피연산자와 연산자를 확장한 검색 공간을 정의하고 사전 생성된 시드 수식 알파 세트로 초기화하는 향상된 초기화 방법을 제안함으로써, 강화 학습 기반 검색 알고리즘의 장점을 활용한다. 하지만 단일 데이터 소스만을 사용하는 접근법은 시장의 복잡성과 다양성을 완전히 포착하는 데 한계가 있다. 또한, 이러한 접근법은 높은 변동성과 안정성 부족의 문제를 내포하고 있으며, 이는 투자 결정의 효율성을 저하시킬 수 있다. 최근에는 단일 데이터 소스의 한계점을 극복하기 위해 기업의 재무 데이터를 함께 사용해 복잡성을 더 잘 포착할 수 있는 연구들이 진행되고 있다.

본 논문에서는 기술적 지표와 재무제표 데이터를 통합한 synergistic formulaic alpha generation 과 랜덤포레스트 기반의 앙상블 모델을 사용하여 안정적인 투자 신호의 생성 방법과 이를 활용한 투자 전략을 제안한다.

표 1. 제안한 방법에 대한 Test set 의 IC, Rank IC 평가

	Baseline [2]		Ours	
Phase	**IC**	**Rank IC**	**IC**	**Rank IC**
Phase 1	0.038 (±0.018)	0.034 (±0.024)	**0.121** (±0.001)	**0.103** (±0.006)
Phase 2	**0.157** (±0.080)	**0.134** (±0.079)	0.113 (±0.007)	0.128 (±0.010)
Phase 3	**0.063** (±0.037)	0.029 (±0.015)	0.015 (±0.002)	**0.043** (±0.001)
Phase 4	0.132 (±0.023)	0.234 (±0.048)	**0.155** (±0.045)	**0.245** (±0.051)
Phase 5	0.100 (±0.028)	0.017 (±0.028)	**0.167** (±0.027)	**0.240** (±0.009)
Phase 6	0.101 (±0.077)	**0.083** (±0.065)	**0.112** (±0.048)	0.079 (±0.036)
Average	0.098 (±0.039)	0.089 (±0.076)	**0.114** (±0.056)	**0.139** (±0.081)

Ⅱ. 본론

본 논문에서는 재무제표의 갱신 주기를 고려하여 90 일 뒤의 등락율과 정보 계수(Information Coefficient, IC)를 최대화할 수 있는 formulaic alpha 를 생성한다.

그림 1 은 제안한 알파 마이닝 방법과 주식 종목 선택 방법을 나타낸다. 시드 알파 초기화를 위해 90 일 간의 평가 세트로 알파들의 IC 를 계산하고, 가장 높은 IC 를 가진 알파를 선택한다. 선택된 알파를 시드 알파로 초기화하고 탐색 공간에 정의된 연산자와 피연산자를 사용해 모델을 훈련시킨다. 모델은 주어진 탐색 공간 내에서 뛰어난 성능을 가진 알파를 찾기 위해 알파 생성과 동시에 훈련을 진행한다.

평가 세트로 생성된 알파들의 IC 를 측정하며 가장 IC 가 높은 알파를 얻는다. 알파 생성 과정을 반복하여 N 개의 알파들을 생성하고 평가 세트를 사용하여 알파들에 대한 알파 값을 얻는다. 이 N 개의 알파 값을 피처로 사용하고 평가 세트 기간의 90 일 뒤 등락률 상승 여부를 라벨로 하여 랜덤 포레스트(Random Forest) 모델을 훈련시킨다. 알파 값을 훈련한 랜덤 포레스트 모델에 테스트 세트에 대한 알파들의 알파 값을 입력하여 주가 상승을 예측한 종목만을 매수한다.

본 연구에서는 S&P500 시장을 대상으로 생성한 시너지 알파의 성능을 평가한다. 2015 년 10 월 26 일부터 2021 년 3 월 1 일까지 총 1345 거래일로 구성된 주식 데이터를 사용한다. 실험에 사용하는 데이터셋은 다양한 시장 환경에서의 성능 검증을 위해 슬라이딩 윈도우 방법을 사용하여 구축한다 [3]. 슬라이딩 윈도우의 크기는 총 900 일로 설정했으며, 720 일의 훈련 기간, 90 일의 버퍼 기간, 90 일의 테스트 기간으로 구성된다. 버퍼 기간은 훈련 기간에서 등락률 계산 시 테스트 기간의 값을 관측하지 못하도록 방지한다. 슬라이딩 윈도우는 90 일 주기로 이동하여 총 6 개의 phase 로 데이터셋을 구성한다. 각 phase 기간 동안 평균적으로 534 개의 주식 종목이 존재하며, 각 종목의 시장 진입 시점부터 데이터를 참조할 수 있도록 설정하여 생존편향을 제거했다. 재무제표 데이터는 Financial Modeling Prep(FMP) [4]로부터 수집하였고 결측값이 있는 경우 과거 분기의 재무제표 데이터로 대체하여 전처리했다. 또한 신뢰성을 높이기 위해 총 5 개의 임의의 초깃값에서 시너지 알파의 pool size 를 10 으로 설정한 뒤 실험을 진행했다. 시드 알파 선택을 위해 훈련 기간의 마지막 90 일의 등락률과 IC 가 높은 알파를 선정하는 방식을 사용했다.

표 1 은 테스트 기간의 IC 와 Rank IC 비교 결과를 나타내며, 제안한 기법이 baseline [2]보다 평균적으로 높은 것을 알 수 있다. 표 2 는 baseline [2]과 제안한 기법의 투자 성과 지표인 Sharpe Ratio 를 나타내며 제안한 기법이 평균적으로 더 높은 것을 확인할 수 있다.

표 2. 제안한 방법에 대한 Test set 의 Sharpe Ratio 평가

Phase	**Baseline [2]**	**Ours**
Phase 1	**2.03**(±0.39)	1.18
Phase 2	**1.22**(±0.11)	0.51
Phase 3	-0.06(±0.08)	**1.37**
Phase 4	**1.80**(±0.30)	-0.26
Phase 5	0.29(±0.38)	**1.64**
Phase 6	0.42(±0.22)	**2.28**
Average	0.95(±0.82)	**1.12**(±0.81)

Ⅲ. 결론

본 논문은 기술적 지표와 재무제표를 혼합한 synergistic formulaic alpha generation 과 랜덤포레스트 기반의 앙상블로 주식 종목을 안정적으로 선택하는 방법을 제안하였다. 제안한 기법은 생존 편향이 제거된 기술적 지표 기반 synergistic formulaic alpha 에 비해 평균적으로 보다 높은 정보 계수(IC) 와 순위 정보 계수(Rank IC) 그리고 Sharpe Ratio 를 달성하였다.

향후 연구에서는 기술적 지표와 재무제표를 전처리한 operand 를 활용하여 수식의 복잡성을 줄인 formulaic alpha 생성과 연속된 alpha value 를 이용한 주식 종목 선택의 신뢰도를 높이는 방안을 탐구할 예정이다.

ACKNOWLEDGMENT

이 논문은 2024 년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(RS-2024-00406796, 2024 년 산업혁신인재성장지원사업)

참 고 문 헌

[1] Guo, J., Wang, S., Ni, L. M., and Shum, H. Y. “Quant 4.0: Engineering Quantitative Investment with Automated, Explainable and Knowledgedriven Artificial Intelligence,” arXiv preprint arXiv:2301.04020, 2022.

[2] Shin, H. G., Jeong, S., Kim, E. Y., Hong, S. H., Cho, Y. J., and Choi, Y, J. “Synergistic Formulaic Alpha Generation for Quantitative Trading based on Reinforcement Learning,” ICOIN 2024, The 38th International Conference on Information Networking. 2024.

[3] R. Sawhney, S. Agarwal, A. Wadhwa and R. R. Shah, "Spatiotemporal Hypergraph Convolution Network for Stock Movement Forecasting," 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, 2020, pp. 482-491, doi: 10.1109/ICDM50108.2020.00057.

[4] Financial Modeling Prep. (https://site.financialmodelingprep.com/)