PRESENTATION

Dynamic Reinforced Ensemble using Bayesian Optimization for Stock Trading

Paper Review

2025.06.03 | 박용성

Contents

USER EXPERIENCE

reo91004@gmail.com

01. Intro

Background

- 자동화된 주식 트레이딩 분야에서 심층 강화학습(DRL)은 시행착오를 통한 학습 능력 덕분에 매우 효과적인 기술로 부상했음.
- 하지만 단일 DRL 에이전트는 복잡하고 끊임없이 변화하는 시장의 동적 환경에 적응하는 유연성이 부족하여 최적의 전략을 내지 못하는 경우가 많음.
- 기존의 LSTM 같은 딥러닝(DL) 예측 모델들은 주식 시장 데이터의 노이즈가 심하고 비정상적인 특성 때문에 강건성이 부족하며, 복잡한 구조로 인해 과적합에 취약하다는 단점이 있음.

Purpose

변화무쌍한 시장 환경에 효과적으로 적응하고, 단일 강화학습 모델의 한계를 극복하여 안정적이면서도 높은 수익을 내는 자동화 주식 트레이딩 전략을 개발하는 것을 목표로 함.

reo91004@gmail.com

01. Intro

Contributions

DREB 제안

베이즈 최적화를 사용한 '동적 강화 앙상블(Dynamic Reinforced Ensemble, DREB)'이라는 새로운 자동 주식 트레이딩 전략을 제안함.

다양한 DRL 모델 활용

서로 다른 특성을 가진 5개의 모델-프리 DRL 기법(A2C, DDPG, TD3, SAC, PPO)를 기본 모델로 사용해 앙상블 구성함.

동적 가중치 및 최적화

최근 과거 데이터를 기반으로 각 모델에 대한 동적 가중치(time-varing weights)를 계산, 이 가중치 시스템의 하이퍼파라미터를 베이즈 최적화로 정밀 조정함.

reo91004@gmail.com

02. Related Work

기존 연구 및 관련 이론 정리

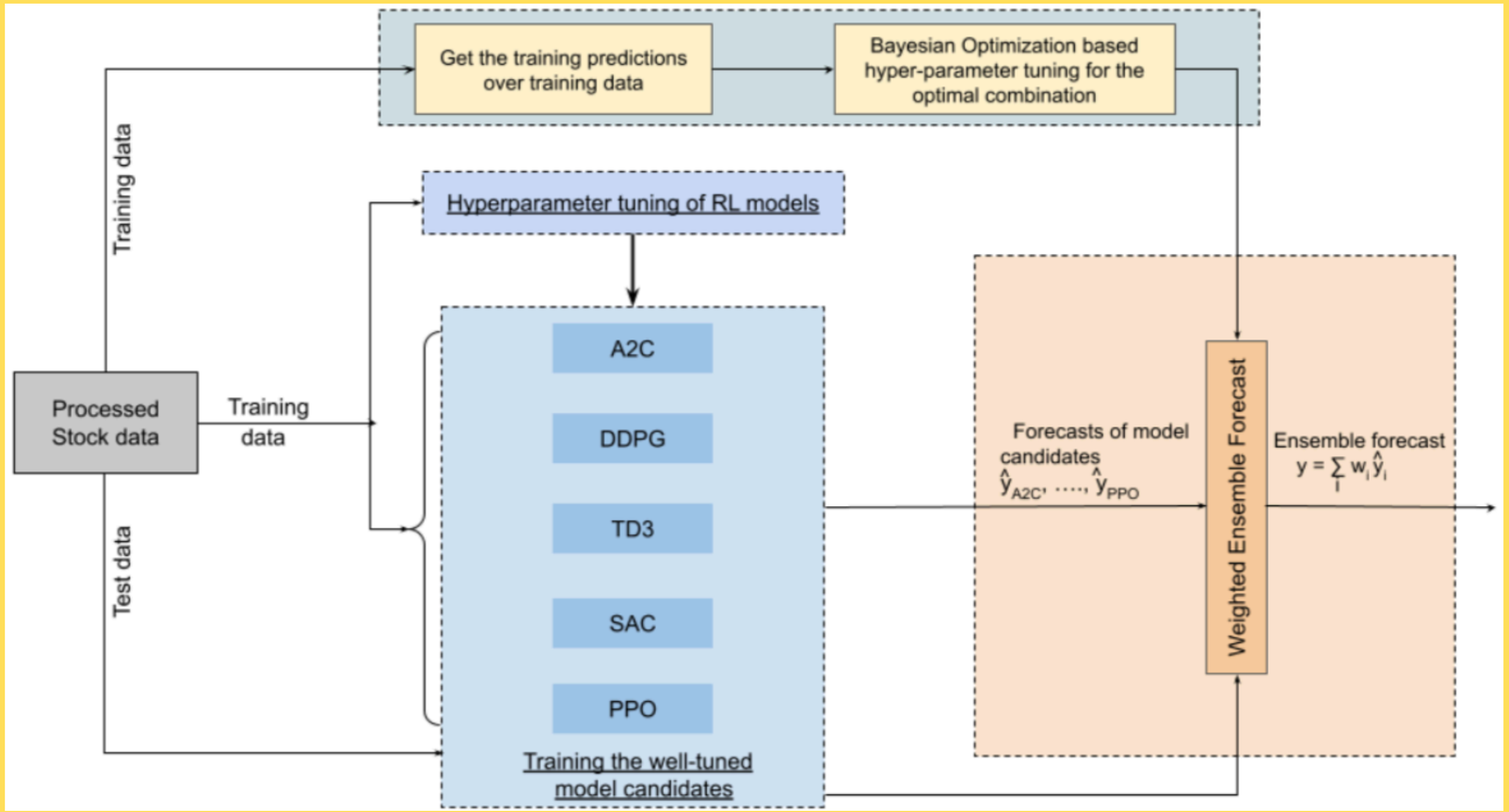
- 기존에 주식 트레이딩 문제를 마르코프 결정 과정(MDP)으로 모델링하고, 강화학습으로 해결하는 접근법이 활발히 연구되어 왔음.
- 단일 DRL 에이전트 연구로는,
  - LSTM과 Policy gradient를 결합한 외환 트레이딩 모델
  - GRU와 DRL(DQN, DDPG)를 결합한 적응형 트레이딩 시스템
  - 변동성 스케일링을 보상 함수에 통합한 트레이딩 전략
- 단일 DRL의 한계를 극복하기 위해 앙상블 기법이 제안됨.
  - 3개의 DRL 에이전트 주기적 재학습
  - 여러 DQN 인스턴스 학습 후 임계값 기반으로 최종 결정
  - 시장 상태에 맞춰 동적으로 에이전트 선택하는 중첩된 RL 방법론
- 시계열 예측을 위한 시변적응(time-varing adaptive weight) 방식

reo91004@gmail.com

03. Proposed Method

1. DREB

잘 튜닝된 여러 DRL 모델 후보들의 예측을 받아, 베이즈 최적화로 찾은 최적의 조합 방식을 통해 가중치를 매겨 최종 앙상블 예측을 수행함.



03. Proposed Method

2. MDP (마르코프 결정 과정)

다중 주식 트레이딩 문제를 마르코프 결정 과정으로 공식화함.

3. 트레이딩 환경 (FinRL 라이브러리 사용)

- **상태 공간**
  - n개의 주식을 거래할 때, 13n+1 차원의 벡터로 상태 표현. 여기에는 남은 현금 잔고, 각 주식 보유량, 각 주식 OHLC(시가, 고가, 저가, 종가) 가격, 그리고 8개의 기술적 지표가 포함됨.
- **행동 공간**
  - [-1, 1]사이의 연속적인 값으로 정규화된 공간 사용. 각 주식에 대해 -h(매도) 에서 +h(매수)까지의 주식 수 의미, 한 번에 거래 가능한 최대 주식 수 h는 100으로 제한
- **보상 함수**
  - 행동을 취한 후 포트폴리오 가치 변화량에서 거래 비용을 뺀 값

03. Proposed Method

4. 동적 가중치 할당 및 베이즈 최적화

- 특정 평가 기간동안의 모델 성능에 기반하여 각 모델의 가중치를 계산함. 이 가중치는 IMSE(역평균제곱오차)와 같은 방식을 통해 구해짐.

- 최종 가중치는 이전 시점의 가중치와 현재 시점의 가중치를 학습 파라미터 L을 이용해 부드럽개 갱신함.

$$w_{i_M}(t) = \begin{cases} 1 & \text{if } i_M = i_c \in N_m(t), i = 1 \\ l w_{i_M}(t-1) + (1-l)\hat{w}_{i_M}(t) & \text{if } i_M = 1, 2, \dots \end{cases}$$

- 이 동적 가중치 시스템의 하이퍼파라미터들(평가 기간, 가중치 업데이트 파라미터 L 등등)을 찾기 위해 베이즈 최적화, 그 중에서도 TPE(Tree-structured Parzen Estimator)기법을 활용함.

04. Exp. & Results

Table 1: Performance evaluation of the proposed approach against the base learners and the benchmarks for the DJI data.

Method/Benchmark	Cummulative Return	Annualized Return	Sharpe Ratio	Annualized Volatility	Maximum DrawDown	Average Profit per Trade
A2C [21]	10.05%	4.35%	0.34	1.56%	19.66%	501.99
DDPG [17]	12.95%	5.56%	0.40	1.67%	22.81%	564.88
TD3 [10]	9.29%	4.03%	0.31	1.72%	25.28%	553.97
SAC [11]	11.63%	5.01%	0.36	1.71%	25.29%	537.96
PPO [31]	13.46%	5.77%	0.43	1.52%	20.13%	519.72
DJI	8.81%	3.82%	0.32	1.47%	21.94%	-
Buy-hold	12.44%	5.34%	0.37	1.78%	26.79%	-
Mean Trading	11.98%	5.16%	0.37	1.73%	24.31%	435.08
Random Trading	5.06%	2.21%	0.85	**0.24%**	**2.76%**	127.60
Adaptive Ensemble [38]	10.79%	4.66%	0.35	1.69%	20.68%	520.83
MVO [20]	14.47%	6.19%	0.55	1.18%	14.67%	490.03
DREB (ours)	**15.24%**	**6.51%**	**1.05**	0.59%	6.56%	**569.49**

Table 2: Performance evaluation of the proposed approach against the base learners and the benchmarks for the Sensex data.

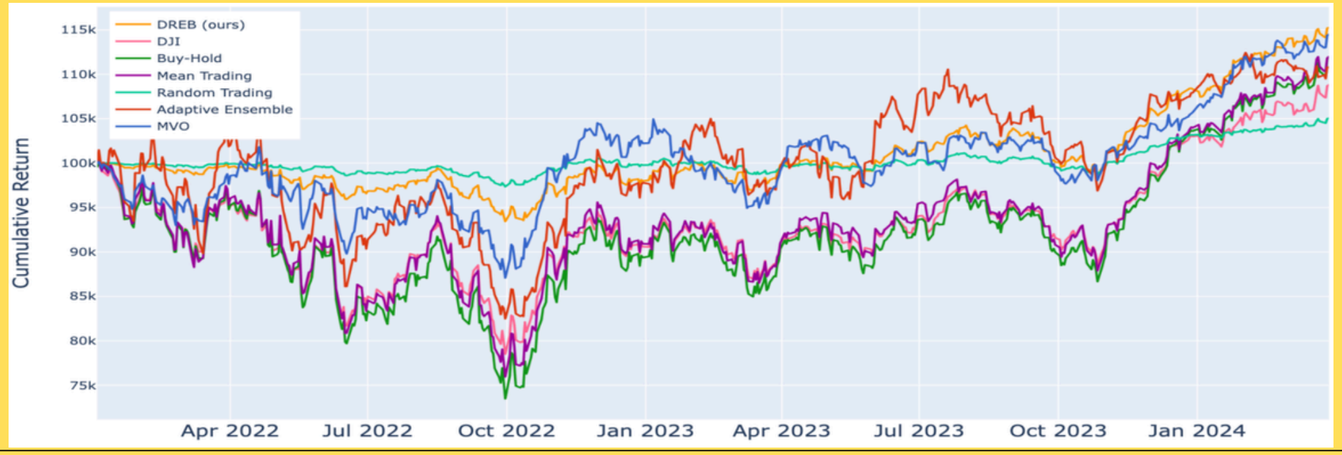
Method/Benchmark	Cummulative Return	Annualized Return	Sharpe Ratio	Annualized Volatility	Maximum DrawDown	Average Profit per Trade
A2C [21]	38.12%	15.43%	1.32	1.11%	14.53%	610.18
DDPG [17]	37.05%	15.04%	1.27	1.13%	15.36%	610.28
TD3 [10]	40.58%	16.35%	1.30	1.18%	13.96%	661.92
SAC [11]	35.23%	14.36%	1.19	1.15%	14.33%	650.98
PPO [31]	41.69%	16.75%	1.01	1.62%	22.18%	770.34
Sensex	26.67%	11.08%	0.85	1.31%	16.47%	-
Buy-hold	32.54%	13.34%	0.99	1.32%	18.76%	-
Mean Trading	28.83%	11.92%	0.89	1.35%	21.03%	648.86
Random Trading	6.27%	2.74%	1.50	**0.19%**	**1.52%**	91.89
Adaptive Ensemble [38]	38.91%	15.73%	0.91	1.73%	20.66%	680.19
MVO [20]	52.35%	20.57%	1.69	1.13%	11.05%	749.59
DREB (ours)	**56.05%**	**21.87%**	**1.82**	1.04%	9.54%	**791.84**

reo91004@gmail.com

04. Exp. & Results

Figure 2: Cumulative returns of the proposed DREB model and the base DRL agents for the DJI data over the entire trading period.

Figure 3: Cumulative returns of the proposed DREB model and the benchmark trading strategies for the DJI data over the entire trading period.



reo91004@gmail.com

05.
Conclusion

CONCLUSION

결론

베이즈 최적화를 활용한 동적 강화 앙상블 기법(DREB)를 성공적으로 제안함.

- 다양한 DRL 모델을 활용하고 동적 가중치를 할당함으로써 개별 모델의 오류에 강건하며, 시장 변화에 효과적으로 적응함을 실험을 통해 증명함.
- 개별 DRL 에이전트들은 특정 소수의 주식에만 편향되어 거래하는 경향을 보이는데, 이는 포트폴리오 다각화 측면에서 위험할 수 있음. 여기서 제안한 DREB 모델은 여러 에이전트의 행동을 종합하므로, 더 다양한 종목에 투자하게 되어 이러한 편향 문제를 완화하고 위험을 줄이는 효과가 있음.

향후 연구

- 앙상블의 포함될 기본 DRL 모델의 종류를 더 늘려 다양성을 확보하는 방안.
- 고도화된 데이터 전처리 기법이나 추가적인 기술 지표 통합.
- 샤프 지수나 CVaR같은 다른 형태의 보상 함수 탐색 → 시스템의 안정성 향상.

reo91004@gmail.com

06.
Limitations

장점

- **정교한 앙상블 구조**: 단순히 여러 모델을 섞는 것을 넘어, 각 모델의 시시각각 변하는 성능을 동적으로 추적하여 가중치를 부여하고, 이 가중치 시스템 자체의 파라미터까지 베이즈 최적화로 정밀하게 튜닝하는 접근 방식은 독창적인 것 같음.
- **뛰어난 안정성 및 수익성**: 실제 시장 데이터(DJI, Sensex)를 이용한 백테스트에서 하락장 방어 능력과 상승장 수익 창출 능력을 모두 보여주었음. 특히 샤프 지수가 다른 모든 모델 대비 월등히 높다는 점은 '위험 대비 수익률'이라는 실용적인 관점에서 큰 강점을 가짐.
- **편향성 해결**: 주식 선택 편향을 앙상블을 통해 자연스럽게 해결하고 포트폴리오를 다각화함.

reo91004@gmail.com

06.
Limitations

단점

- **'전문가 궤적'의 타당성 문제**: 동적 가중치를 계산하기 위해 '목표 수익률'이 필요한데, 이를 생성하기 위해 '전문가 궤적(expert trajectories)'이라는 것을 사용함. 그런데 이 전문가는 실제 사람이 아니라, 다음 시점(t+1)의 가격 변동률을 이용한 수식으로 정의됨. 이는 학습 데이터에 한해 일종의 '정답'을 미리 보고 학습하는 것과 같으므로 실제 미래를 알 수 없는 현실에서도 이 성능이 그대로 재현될지에 대한 의문이 남을 수 있음.
- **계산 복잡성**: 5개의 복잡한 DRL 모델을 각각 학습시키고, 그 위에서 또 베이즈 최적화를 수행하는 과정은 엄청난 계산 비용을 요구할 것으로 보임. 논문에서는 이 계산 복잡성에 대한 언급이 없어, 실제 트레이딩 시스템에 적용할 때의 현실적인 제약이 될 수 있음.
- **하이퍼파라미터 민감성**: 앙상블의 하이퍼파라미터는 베이즈 최적화로 찾았지만 , 그 기반이 되는 5개 DRL 모델 각각의 하이퍼파라미터 튜닝은 어떻게 이루어졌는지 상세한 설명이 부족하므로, 직접 구현 시 최종 성능이 이 기본 모델들의 튜닝 상태에 크게 좌우될 가능성이 높음.

reo91004@gmail.com

PRESENTATION

Q&A

Paper Review

2025.06.03