하이퍼 파라미터 최적화를 통한 머신러닝 기반 주가 예측 및 수익률 극대화 방법 연구

이지훈, 김재운*
순천향대학교

20201468@sch.ac.kr, *kimym38@sch.ac.kr

A Study on Stock Price Prediction and Maximizing Returns Using Machine Learning with Hyperparameter Optimization

Lee Jihun, Kim Jaeyun*
Soonchunhyang Univ.

요 약

본 논문은 머신러닝 기법을 활용하여 주가 예측을 통해 수익률을 극대화할 수 있는 방법을 제시하는 것을 목표로 한다. 주가 데이터는 Yahoo Finance 패키지를 이용하여 주식 시세를 수집하였으며, 예측 모델의 구축 과정에서 하이퍼 파라미터의 목적 함수를 이익 계수(Profit Factor)로 설정하여, 그 값이 증가할수록 더 높은 점수를 부여하는 방법으로 모델의 성능을 향상시켰다. 또한, 슬라이딩 윈도우 기법을 활용하여 최적의 과거 N 일의 입력변수를 선정함으로써 백테스팅을 실시한 결과, 제안된 모델이 기존 모델에 비해 매매 수익률을 더욱 극대화할 수 있음을 확인하였다.

Ⅰ. 서 론

정보통신 기술의 발전으로 빅데이터(Big Data)의 활용이 증가하면서 사회나 경제 및 금융 산업에서도 빅데이터 기술을 다각도로 적용하려는 시도가 활발히 이루어지고 있다. 특히, 머신러닝(Machine Learning)을 활용한 주가 예측 연구는 그 중요성이 점차 강조되고 있다.[1]

머신러닝을 통해 미래의 주가를 예측하기 위해서는 과거 데이터를 기반으로 한 학습 과정이 필수적이지만, 금융시장의 특성상 활용 가능한 데이터의 수가 제한적이기 때문에 일반적인 경우에는 주가의 비선형적 패턴을 효과적으로 학습하기에 어려울 수 있다.[2] 이에 따라, 기존 연구들은 ARIMA, RNN, CNN 등의 모델을 사용해 주가 예측의 정확도를 높이는데 초점을 맞추었다.[3] 그러나 이러한 연구들은 실제 거래에서 수익을 극대화하기에는 한계가 있었다. 주식 거래에서 효과적인 수익을 실현하기 위해서는 예측의 정확도를 높이는 것뿐만 아니라, 어느 시점에 효율적으로 매매를 하는 지가 중요하게 작용될 수 있다. 이에 따라, 본 연구는 모델의 하이퍼 파라미터 값을 전통적인 기준인 평균절대오차 (MAPE)나 제곱근 평균제곱오차(RMSE)가 아닌 이익 계수(Profit Factor)로 설정하여, 백테스트를 통한 수익률이 최대가 되는 설정 값을 도출하는 것을 목표로 하였다. 또한, 모델 내부의 주가 시계열의 연속성을 고려하기 위해 하나의 데이터 포인트에 과거 N 개의 시점이 입력되도록 설정하며 하이퍼 파라미터 최적화 과정에서 최적의 N 개를 탐색하도록 설계한다.

제안된 방법의 거래 성능 차이를 판단하기 위해 다양한 알고리즘을 트레이딩 시스템에 적용하여 비교 분석을 수행한다.

Ⅱ. 본론

본 연구에서는 하이퍼 파라미터 최적화 과정에서 목적함수를 총 손실 대비 총 수익의 비율인 이익 계수(Profit Factor)로 설정하였다. 최적화된 모델을 사용해 이후 테스트 데이터에 해당하는 각 시점에서의 포지션을 결정하여 거래 전략을 구성하였으며, 다양한 트레이딩 평가 지표를 통해 모델의 성능을 파악하였다.

연구에 사용된 모델은 랜덤 포레스트(Random Forest), 의사결정나무(Decision Tree), XGBoost, LGBM, 선형 회귀(Linear Regression)로, 각 모델의 성능 최적화를 위해 슬라이딩 윈도우 기법을 통한 성능을 비교함으로써, 모델 별로 최적의 입력 변수 개수(N 값)를 선정해 모델의 성능을 더욱 높이고자 하였다.

Ⅱ.Ⅰ 데이터

본 논문에서는 Yahoo Finance 라이브러리를 사용하여 2013 년부터 2024 년 6 월까지 시가총액 상위 50 위에 해당하는 종목들에 대해서 총 12 년 6 개월간의 데이터를 수집하였다. 이때, 설정된 기간 이후에 상장된 기업의 경우는 데이터가 존재하지 않기 때문에 제외하였으며, 총 39 개의 종목을 대상으로 선정하였다. 2023 년을 기준으로 학습과 테스트 데이터를 분류해 1 년 6 개월의 기간을 걸쳐 거래 성능을 분석하였다.

Ⅱ.Ⅱ 슬라이딩 윈도우 기법

데이터 전처리 과정에서는 종속 변수와 입력 변수를 생성하였다. 종속 변수는 다음날 종가로 설정하였으며, 이를 위해 수정 종가(Adj Close) 열을 한 칸씩 위로 이동시켜 다음날의 종가 데이터를 새로운 종속 변수로 추가하였다. 입력 변수는 각 날짜를 기준으로 과거 N 일 전의 수정 종가 데이터를 생성하여, $Adj\ Cbs\ e_t - 1$, $Adj\ Cbs\ e_t - 2$, …, $Adj\ Cbse\ \_t - (n - 1)$ 형식으로 구성되었다. 이러한 입력 변수는 점진적 학습 기법과 같은 타 기법과 달리 메모리 효율성과 학습 시간의 최적화에 특화된 슬라이딩 윈도우 기법을

사용하여 추가되었다. 그림 1 은 해당 기법의 예시를 보여준다.

각 모델에 적합한 입력 변수의 최적 개수를 찾기 위해 초기 N 값을 100 으로 설정한 후, N 값을 1 씩 증가시키며 시뮬레이션을 반복하였다. 각 N 값에 대해 Profit Factor 를 계산하여, 가장 높은 값을 보이는 N 값을 최적의 입력 변수 개수로 선정하였다.

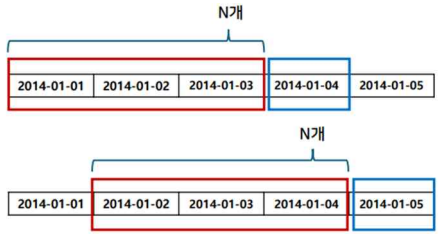

그림.1 슬라이딩 윈도우 예시

Ⅱ.Ⅲ 하이퍼 파라미터 최적화

선정된 최적의 N 값을 각 모델의 입력 변수로 사용한 후, 하이퍼 파라미터 최적화를 진행하였다. 이 과정에서는 그리드 서치(Grid Search)를 활용했으며, 목적 함수로 Profit Factor 를 설정하여 값이 높을수록 더 높은 점수를 부여하는 방식으로 최적화를 수행하였다.

XGBoost 모델의 경우, 트리 개수는 50, 100, 150 으로 설정하고, 최대 깊이는 3, 5, 7 로 탐색하였다. 이 외에도 학습률과 feature 샘플링 비율도 설정하여 탐색을 진행하였다. 최적의 하이퍼 파라미터는 교차 검증 후 반복 시뮬레이션을 통해 도출되었다.

Ⅱ.Ⅳ 트레이딩 시스템

모델의 예측 값을 바탕으로, T 시점에서의 예측 종가가 실제 종가보다 낮고, T+1 시점의 예측 종가가 실제 시가보다 높을 경우 매수(Buy) 포지션을, 반대로 예측 종가가 실제 값보다 높고, T+1 시점의 예측 종가가 낮을 것으로 예상될 경우 매도(Sell) 포지션을 취하였다.

학습된 모델에서 예상 값과 실제 값이 교차하는 시점이 잔차 산점도에서 이상치로 나타나며, 해당 전략은 이러한 현상을 활용할 수 있다는 데 이점이 있다. 이러한 이상치에 집중하여 매매를 진행함으로써 잠재적 수익을 최대한 활용하고자 하였다.

거래 전략의 성능을 평가하기 위해 평균 수익과 평균 손실의 비율(Payoff Ratio)과 총 수익과 총 손실의 비율(Profit Factor)를 트레이딩 평가지표로 사용하였다.

Ⅱ.Ⅴ 실험 결과

Table 1 은 하이퍼 파라미터의 목적 함수를 RMSE 로 설정했을 때와 Profit Factor 로 설정했을 때의 모델 별 평가지표 평균값을 제시한다. Profit Factor 를 목적 함수로 설정한 경우, 두 지표의 평균값이 기존 RMSE 를 기준으로 학습된 모델보다 우수한 성능을 보였으며, 각 모델 간 비교에서도 안정적으로 더 나은 성능을 확인할 수 있었다.

특히 RL 모델의 출력은 입력 변수들의 가중된 합으로 표현되며, 각 변수는 고정된 가중치 β에 따라 결과에 영향을 미친다. N 값이 낮을수록 해당 변수가 모델에서 더 높은 설명력을 가지므로, 적은 수의 입력 변수만으로도 더 우수한 성능을 발휘할 수 있다.

Table 1. 목적함수별 거래 성과 비교

		RMSE	Profit Factor

Model	N	Payoff ratio	Profit Factor	Payoff ratio	Profit Factor
LR	12	1.10 (0.61)	0.97 (0.53)	1.22 (0.43)	1.10 (0.51)
DT	19	0.86 (0.77)	1.02 (0.56)	0.83 (0.78)	0.91 (0.65)
RF	32	0.98 (0.68)	1.09 (0.72)	1.07 (0.66)	1.05 (0.52)
XGBoost	60	1.07 (0.64)	0.79 (0.74)	1.31 (0.54)	1.22 (0.62)
LGBM	58	1.06 (0.58)	0.83 (0.48)	1.26 (0.56)	1.05 (0.71)
평균		1.01 (0.66)	0.94 (0.61)	1.14 (0.59)	1.07 (0.60)

Ⅲ. 결론

본 연구에서는 수익률을 직접적으로 극대화하기 위해서 모델의 하이퍼 파라미터 최적화에 이익계수(Profit Factor)를 목적 함수로 설정하는 새로운 접근을 제안하였다.

실험 결과, Profit Factor 를 목적 함수로 사용한 모델이 전통적인 예측 정확도 기준(MAP, RMSE 등)을 사용한 모델보다 더 높은 수익률을 나타냈으며, 안정적인 성능을 보였다. 이는 실제 거래 환경에서 수익성을 높이기 위한 전략으로 매우 유용할 수 있음을 시사한다.

그러나 본 연구는 몇 가지 한계점을 가지고 있다. 첫째, 사용된 데이터는 특정 기간과 시장(코스피 지수 상위 50 개 종목)으로 한정되어 있어, 다른 시장 또는 다양한 경제 상황에서의 일반화 가능성에는 제한이 있을 수 있다. 둘째, 모델의 복잡성과 계산 비용이 증가할 수 있다는 점에서 실시간 트레이딩 환경에서의 적용 가능성에 대한 추가 연구가 필요하다.

ACKNOWLEDGMENT

본 연구는 2024 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구 결과로 수행되었음. (2021-0-01399) 또한, 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1A2C1092808).

참 고 문 헌

[1] Shin, Ha-Yan, and Sang-Hee Kweon. "An evaluation of determinants to viewer acceptance of artificial intelligence-based news anchor." *The Journal of the Korea Contents Association* 21.4 (2021): 205-219.

[2] Yuan, Yuyu, Wen Wen, and Jincui Yang. "Using data augmentation based reinforcement learning for daily stock trading." *Electronics* 9.9 (2020): 1384.

[3] Lee, Mo-Se, and Hyunchul Ahn. "A time series graph based convolutional neural network model for effective input variable pattern learning: Application to the prediction of stock market." *Journal of Intelligence and Information Systems* 24.1 (2018): 167-181