GAN 을 활용한 머신러닝 기반 한국 주가지수 트레이딩 시스템 개발

유성주, 박명석, 김재윤*
순천향대학교

sjyoo@sch.ac.kr, pmsk980122@sch.ac.kr, *kimym38@sch.ac.kr

Development of a Machine Learning-based Korea Stock Index Trading System

Yoo SungJu, Park Meyong-seok, Kim Jaeyun*
Soonchunhyang Univ.

요 약

최근 금융시장에서 머신러닝 기술을 접목하여 미래의 주가 방향을 예측하고자 하는 다양한 연구들이 이루어지고 있다. 그러나 금융시장의 특성상 대량의 데이터를 얻는데 시간과 비용이 많이 소요된다 이러한 문제점을 해결하기 위해 본 연구는 기하 브라운 운동 (Geometric Brownian Motion, GBM)을 이용하여 가상의 주가 데이터를 생성하며, GBM 모형을 활용 시 주가 데이터의 정규분포를 활용하는 대신 생성적 적대 신경망 (Generative Adversarial Network, GAN)을 사용하여 원본 주가 지수의 분포를 따라가는 난수를 만들어 사용하였다. 200 회의 시뮬레이션을 통해 생성된 가상의 데이터를 예측 모형의 학습데이터로 사용하여 트레이딩 시스템을 개발하였다. 트레이딩 성과평가 지표를 사용하여 비교한 결과, GAN 난수를 사용한 경과가 가장 우수한 성능을 보였다.

1. 서 론

최근 정보통신 기술의 비약적 발전으로 빅데이터(Big Data) 개념 또한 사회, 경제 분야에서 많은 관심을 받고 있는 추세에서 금융산업 또한 다각적 측면으로 머신러닝 (Machine Learning)을 적용시키고자 노력하고 있다.

머신러닝을 활용해 미래 주가의 방향성을 예측하려는 대표적인 연구로는 한국 주가지수 등락 예측을 위한 유전자 알고리즘 기반 인공지능 예측기법 결합모형 [1], 코스피 방향 예측을 위한 하이브리드 머신러닝 모델 [2], 호가창과 뉴스 헤드라인을 이용한 딥러닝 기반 주가 변동 예측 기법 [3] 등이 있다. 이처럼 머신러닝을 통해 미래주가의 방향성을 예측하고자 할 때 과거 데이터를 통해 학습하는 과정이 필수적으로 들어간다. 하지만 금융시장이라는 환경은 하루에 생산되는 데이터의 수가 한정되어 있기에 대량의 데이터를 수집하기 위해선 많은 시간과 비용이 소요된다는 한계점이 있다. 이러한 문제를 해결하고자 본 연구에선 기하 브라운 운동 (Geometric Brownian Motion, GBM)을 통해 가상의 주가 데이터를 생성하고자 한다. 이 때 GBM 모형 수식에 포함되는 정규분포 난수 대신 생성적 적대 신경망 (Generative Adversarial Network, GAN)을 사용해서 원 데이터의 분포를 따라가는 가상의 난수를 생성하여 대체한다. 이 과정을 200 회 시뮬레이션 하여 생성한 데이터를 토대로 기술적 지표 (technical indicators)들을 추출하고 이를 머신러닝의 학습 데이터셋으로 사용하여 미래 주가의 상승과 하락을 예측한다. 마지막으로 새롭게 생성한 데이터와 실제 데이터의 성능 차이를 판단하기 위해 트레이딩 시스템을 구현하여 비교한다.

2. 본론

본 연구에서는 6 년치 데이터를 GAN 난수를 사용한 GBM 모형에 적용한다. 그 후 GBM 모형의 결과로 얻은 주가 모형에 몬테카를로 시뮬레이션 기법을 적용하여 가상의 데이터를 생성한 뒤 이 데이터를 사용하여 기술적 지표들을 계산하여 학습 데이터 셋을 만든다. 이 데이터에 머신러닝을 사용하고 결과로 나온 예측 값을 바탕으로 각 시점의 포지션을 결정하여 거래 전략을 구성했고, 트레이딩 평가 지표를 사용해 모델의 성능을 파악하였다.

2.1 데이터

본 연구에서는 증권 데이터 수집 라이브러리인 FinanceDataReader 를 사용하여 2016 년부터 2022 년까지 7 년치의 코스피 지수 데이터를 가져왔다. 이중 2016 년에서 2022 년 데이터로부터 GBM 모형을 적용하기 위해 수익률의 평균과 표준편차를 추정하였고, 2022 년의 1 년치 데이터를 테스트 데이터로 사용하였다.

2.2 기하 브라운 운동 (Geometric Brownian Motion)

많은 금융공학 이론에서 주식 가격이 무작위적으로 상승하거나 하락하는 것이 아니라, 무작위적인 비율만큼 상승 또는 하락한다는 기법인 기하브라운 운동을 전제로 하고 있다. 이 식은 주가의 랜덤수익률의 평균과 표준편차를 포함하고 있기에 주가의 통계적 속성을 반영하는 대표적 모형으로 자리하고 있다. 본 연구에서는 코스피 지수의 6 년치 데이터를 사용해 주가의 기대수익률과 변동성을 추정하였고, 추가적으로 랜덤성을 나타내기 위해 사용된 정규분포 난수 대신 주가 수익률의 분포를 따라가는 GAN 난수로 대체하여 사용하였다.

2.3 생성적 적대 신경망 (Generative Adversarial Network)

GAN 은 생성자 (Generative) 와 식별자 (Adversarial)가 서로 경쟁하며 데이터를 생성하는 모델이다. 실제 데이터의 분포와 유사한 분포를 추정하기 위해 생성자와 식별자라는 두 모델을 적대적인 방식을 통해 학습하게 되는데 정규분포나 균등 분포로부터 난수를 가져와 처음 학습 데이터로 사용한다. 본 연구에서는 GBM 모형에 포함되는

정규분포 난수 대신 주가의 수익률 분포를 따라가는 GAN 난수를 생성하였다. Fig 2 는 Epoch 1500 회 일 때 원본 데이터 (주가 수익률) 와 생성 데이터 (GAN 을 통한 학습)의 분포를 비교한 것이다.

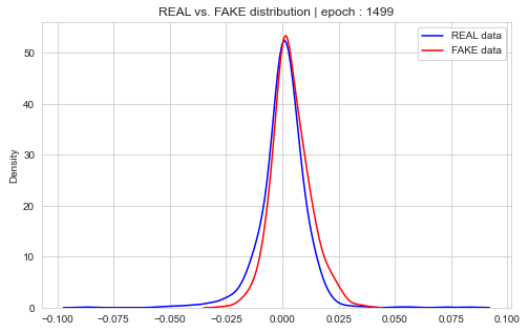

Fig 2. 원본 데이터와 생성 데이터 분포 비교

2.4 몬테카를로 시뮬레이션 (Monte Carlo Simulation)

몬테카를로 시뮬레이션이란 불확실한 사건의 가능한 결과를 예측하는 수학적 기법이다. 이를 통해 앞서 GBM 모형을 통해 파악한 주가의 진행 방향을 시뮬레이션 할 수 있다. 본 연구에서는 시뮬레이션 횟수를 200 회로 설정하여 가상의 주가 데이터를 생성하였다.

2.5 학습 데이터 생성 및 머신러닝 알고리즘

생성한 가상에 데이터를 학습 데이터로 사용하기 위해 기술적 지표들을 추출 하였다. Python 의 기술적 분석 라이브러리인 Ta-lib 패키지를 사용하였으며, 총 15 개의 기술적 지표들을 추출하였다. 라벨 데이터는 단순 가격 비교를 통해 내일의 종가가 오늘의 종가보다 크면 up, 작거나 같으면 down 으로 하여 생성하였다. 이를 바탕으로 머신러닝 학습을 하였고 사용한 모델은 Logistic regression (LR), Decision tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost)으로 총 4 가지 모델이며, 실제 데이터와의 비교를 위해 6 년치의 실제 코스피 지수에도 위 과정을 적용하여 비교했다.

2.6 트레이딩 시스템

모델이 예측한 결과값으로 T 와 T+1 시점에 따라 포지션을 결정하였다. T+1 시점의 예측 값이 T 시점에 비해 상승하거나 하락하였다면 Buy / Sell 포지션, T 시점이 Buy 나 Sell 포지션일 때 T+1 시점의 예측값이 상승이 지속되면 Holding, 하락이 지속되면 No action 으로 구분하였다. 이 거래 전략을 취했을 때 트레이딩 시스템의 성능을 파악하기 위해 트레이딩 평가지표를 사용하였다. 평균수익과 평균손실의 비율인 Payoff ratio 와 총 수익과 총 손실의 비율인 Profit Factor 2 가지의 지표를 사용하였다.

2.7 실험 결과

Table 1 은 코스피 지수로 트레이딩 전략을 취했을 때의 모델별 성과이다. Table 2 와 Table 3 은 각각 GAN 난수를 사용했을때와 정규분포 난수를 사용했을 때의 모델별 평가지표 평균값이다. GAN 난수를 사용했을 땐 두 지표의 평균이 실제 데이터보다 우수한 성능을 보였으며, 각 모델별로 비교해도 안정적으로 더 나은 성능을 보이는 것을 확인했다. 정규분포 난수를 사용한 데이터는 평균 Payoff ratio 지표가 높게 나타났지만 두 지표 모두 안정적으로 우수한 성능을 보이는 모델은 확인할 수 없었다.

Table 1. 실제 데이터를 사용한 트레이딩 성과

Model	Payoff ratio	Profit factor
LR	0.84	1.57
DT	1.06	1.02
RF	1.19	1.09
XGBoost	1.07	0.79
평균	1.04	1.12

Table 2. GAN 난수를 사용한 트레이딩 성과

Model	Payoff ratio	Profit factor
LR	1.25	2.92
DT	1.11	1.06
RF	1.26	1.08
XGBoost	0.85	1.52
평균	1.12	1.64

Table 3. 정규분포 난수를 사용한 트레이딩 성과

Model	Payoff ratio	Profit factor
LR	2.03	0.90
DT	0.82	1.13
RF	1.01	1.01
XGBoost	1.12	0.80
평균	1.24	0.96

3. 결론

본 연구는 금융 시장의 데이터 수집이 어렵다는 한계점을 보완하고자 GBM 모형을 통해 새로운 데이터 셋을 생성하였고 이를 통해 실제 데이터와의 성능을 비교해 보고자 하였다. 또 GBM 모형 수식에 들어가는 정규분포 난수 대신 GAN 을 사용해 주가의 수익률을 따라가는 난수로 대체하여 두 기법의 차이도 비교해 보았다. 실제 데이터와 생성 데이터의 평가 지표들을 비교해 본 결과 GAN 난수를 사용한 데이터에서 대부분의 모델에서 우수한 성능을 보였으며 특히 XGBoost, Logistic regression 에서 큰폭으로 성능이 향상 되었다. 이를 통해 생성한 데이터 셋이 학습에 사용할 수 있을 정도로 유의미한 데이터라는 결론을 얻었다. 본 연구의 한계점으로는 사용한 데이터가 코스피 지수만을 사용했다는 점과 트레이딩 시스템을 만들 때, 거래 수수료를 고려하지 않았다는 점이 있다. 보다 다양한 비교군이 필요하고 실제 거래에서는 거래 횟수에 따른 손실이 발생하기 때문에 다양한 종목들에 대해 거래 수수료까지 계산하여 평가하는 추가적인 연구가 필요하다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1A2C1092808).

참 고 문 헌

[1] 이형용. "한국 주가지수 등락 예측을 위한 유전자 알고리즘 기반 인공지능 예측기법 결합모형". Entrue Journal of Information Technology, vol. 7, pp.33-43, 2008.

[2] 황희수. (2021). 코스피 방향 예측을 위한 하이브리드 머신러닝 모델. 한국융합학회논문지, 12(6), 9-16.

[3] 류의림, 이기용, 정연돈. (2022). 호가창과 뉴스 헤드라인을 이용한 딥러닝 기반 주가 변동 예측 기법. 한국전자거래학회지, 27(1), 63-79.