기하 브라운 운동을 활용한 머신러닝 기반 주식 트레이딩 시스템 개발

유성주, 장주현, 김재윤*
순천향대학교

sjyoo@sch.ac.kr, kwack0202@sch.ac.kr, *kimym38@sch.ac.kr

Development of a stock trading system based on Machine Learning using Geometric Brownian Motion

Yoo SungJu, Jang Joo Hyun, Kim Jaeyun*
Soonchunhyang Univ.

요 약

머신러닝 기술의 발전으로 금융시장에서도 다각적 측면으로 머신러닝 기술을 접목시키려는 시도가 증가하고 있다. 하지만 금융시장의 특성상 대량의 데이터를 얻는데 많은 시간과 비용이 소요되는 문제가 발생한다. 이러한 배경속에서 본 연구는 코스피 상위 30 개의 종목에 대해 기하 브라운 운동 (Geometric Brownian Motion, GBM)을 적용해 주가의 통계적 특성을 따르는 데이터를 생성하고, 그 생성된 데이터를 머신러닝 알고리즘의 학습데이터로 사용한다. 학습된 예측 모형을 바탕으로 트레이딩 전략으로 구현하여 최종적으로 주가 트레이딩 시스템을 개발하였다. 실제 데이터와 GBM 을 이용하여 생성한 데이터를 통해 학습된 머신러닝의 트레이딩 성과를 비교 분석하였다. 트레이딩 성과를 비교한 결과, 통계적으로 유의미한 차이가 없었지만 평균적으로 실제 데이터를 학습한 경우보다 개선된 결과를 도출하였다.

Ⅰ. 서 론

최근 들어 빅데이터 (Big Data) 개념이 사회, 경제 분야에서 다양하게 활용되면서 세계적인 주목을 받고 있다. 이런 전세계적인 관심 속에서 금융산업 또한 다각적 측면으로 머신러닝을 적용하여 미래 주가의 움직임 방향을 예측하려는 연구들이 활발하게 이루어 지고 있다.

대표적으로 주가의 등락을 예측하는 연구로는 한국 주가지수 등락 예측을 위한 유전자 알고리즘 기반 인공지능 예측기법 결합모형 [1], 해외지수와 투자자별 매매 동향에 따른 딥러닝 기반 주가 등락 예측 [2] 텐서 회귀모형을 이용한 단기 평균 한국종합주가지수 (KOSPI) 예측[3] 등이 있다. 이와 같이 머신러닝을 사용해 미래주가의 방향성을 예측하는 대다수의 연구들은 과거 주가 데이터를 기반으로 한다는 공통점이 있다.

하지만 금융시장이라는 환경의 특성상 대량의 데이터를 수집하기 위해선 많은 시간과 비용이 필요하다는 한계점이 있다. 이러한 문제를 해결하고자 본 연구에서는 기하 브라운 운동 (Geometric Brownian Motion, GBM)을 통한 가상의 주가 데이터를 생성한다. 생성한 데이터를 토대로 기술적 지표 (technical indicators)들을 추출한 뒤, 이를 바탕으로 머신러닝 기반 주가 트레이딩 시스템을 개발하고자 한다. 특히 본 연구에서는 GBM 을 이용한 가상의 데이터가 실제 데이터를 대체하여 머신러닝 학습 데이터로 가능한지를 살펴보고자 한다.

Ⅱ. 본론

본 연구의 프레임워크는 Fig.1 과 같으며 연구 절차는 다음과 같다. 과거 5 년간의 데이터 (2016 ~ 2020 년, 일별)를 GBM 모형을 이용하기 위한 μ (the expected return) 와 σ (the standard deviation of returns)를 추정한다. 그 후 GBM 모형의 결과로 얻은 주가 모형에 몬테카를로 시뮬레이션 (Monte Carlo Simulation) 기법을 적용하여 가상의 데이터를 생성한다. 생성된 주가 데이터를 사용하여 기술적 지표들을 생성하고 이를 머신러닝 알고리즘의 입력변수로 사용한다. 학습된 머신러닝을 이용하여 트레이딩 전략을 구현하며, 트레이딩 결과를 측정하였다.

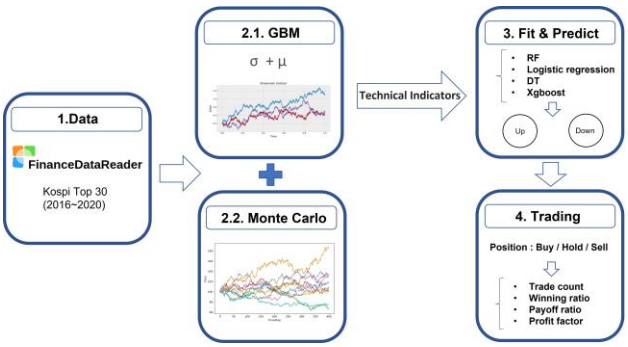

Fig.1 GBM 을 활용한 트레이딩 시스템 구현 프레임워크

2.1 데이터

본 연구에서는 국내 주식 및 해외 주식 데이터를 제공하는 파이썬 모듈인 FinanceDataReader 를 사용하여 일별 종가 데이터를 얻었다. 2022 년 8 월 기준 코스피 상위 30 개의 종목을 가져와 각 종목별로 2016 년부터 2020 년 까지 5 년치 데이터를 가져왔다. 이중 2016 년부터 2020 년의 5 년치 데이터에서 GBM 모형을 적용하기 위해 일별 수익률의 평균과 표준편차를 추정하였고, 2021 년의 1 년치 데이터를 테스트 데이터로 사용하였다.

2.2 기하 브라운 운동 (Geometric Brownian motion)

많은 금융 이론들에서 주가의 움직임은 앞으로의 방향을 예측할 수 없는 랜덤워크(Random Walk)와 같다고 전제하고 있다. 하지만 주가에는 변동성뿐만 아니라 추세도 포함되어 있기 때문에 주가의 움직임을 단순히 랜덤하게 움직인다고 설명하기엔 부족하다. 이를 보완하고자 주가의 기대수익률(추세) 와 변동성을 사용해 얻은 모형이 GBM 이다. GBM 모형을 통해 미래 주가의 방향성을 추정할 수 있기에 이 모형은 금융시장에서 주가의 통계적 속성을 반영하는 대표적 모형으로 자리하고 있다. 본 연구에서는 2022 년 8 월 기준 코스피 상위 30 개 종목 각각마다 5 년치 종가 데이터를 사용해 주가의 기대수익률과 변동성을 추정하였고, 여기에 GBM 모형을 적용하여 미래 주가의 등락을 예측하는데 사용하였다.

2.3 몬테카를로 시뮬레이션 (Monte Carlo Simulation)

GBM 모형을 통해 파악한 주가의 방향성을 토대로 미래주가의 전개과정을 시뮬레이션 해볼 수 있다. 본 연구에서 사용한 시뮬레이션은 몬테카를로 시뮬레이션으로 반복된 무작위 추출을 이용하여 함수의 값을 수리적으로 근사하는 알고리즘이다.

2.4 학습 데이터 생성 및 머신러닝 알고리즘

각 종목별로 생성된 데이터를 바탕으로 기술적 지표들을 추출하였다. Python 의 ta-lib 패키지를 사용하였으며, 종가데이터로 생성할 수 있는 15 개의 기술적 지표들을 사용하였다. 종속변수는 내일의 종가가 오늘의 종가보다 크면 up, 작거나 같으면 down 으로 하여 라벨을 만든 뒤 머신러닝을 적용하였다. 본 논문에서 사용한 모델은 Logistic regression (LR), Decision tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost)으로 총 4 가지 모델이며, 실제 데이터와의 비교를 위해 5 년치의 실제 종가 데이터에도 위 과정을 적용하였다.

2.5 트레이딩 시스템

모델이 예측한 결과값을 바탕으로 각 시점의 포지션을 결정하였다. T 시점의 예측 값이 down 이고 T+1 시점의 예측 값이 up 이면 Buy 포지션, T 시점의 예측 값이 up 이고 T+1 시점의 예측 값이 down 이면 Sell 포지션을 가졌고, Buy 포지션 상태에서 T+1 시점의 예측 값이 up 이 지속되면 Holding, Sell 포지션 상태에서 T+1 시점의 예측 값이 down 이 지속되면 No action 으로 거래 전략을 구성하였다. 이 거래 전략을 취했을 때 성능을 파악하기 위해 트레이딩 평가 지표를 사용하였다. 본 연구에서 사용한 트레이딩 평가 지표는 거래 횟수 (trade count), 승률 (winning ratio), payoff ratio, profit factor 로 총 4 개를 사용하였다.

2.5 실험 결과

Table 1 은 실제 데이터로 트레이딩 전략을 취했을 때의 모델별 평가지표 평균값이다. Table 2 는 GBM 을 통해 생성된 데이터의 모델별 평가지표 평균값이다. 생성된 데이터의 6 개의 그룹중 성능이 가장좋은 20 회일 때 평균값을 가져왔다. LR 모델 같은 경우는 모든 부분에서 실제 데이터보다 우수한 성능을 보여줬으며, 다른 모델의 경우도 대부분의 경우에서 비슷한성능을 보여 주었다. 전체 결과값의 평균을 비교해 보아도 실제 데이터와 생성 데이터의 결과가 유사한 것을 확인했다.

Table 1. 실제 데이터를 이용한 트레이딩 결과

Model	Trade count	Winning ratio	Payoff ratio	Profit factor
LR	44.06	0.45	1.01	0.91
DT	20.9	0.52	1.25	1.61
RF	39.43	0.46	1.01	0.93
XGBoost	36.17	0.47	1.11	1.07
Average	35.14	0.48	1.09	1.13

Table 2. GBM 을 이용한 트레이딩 결과

Model	Trade count	Winning ratio	Payoff ratio	Profit factor
LR	18.23	0.52	1.16	1.79
DT	30.47	0.6	1.09	1.71
RF	36.7	0.51	1.02	1.17
XGBoost	36.26	0.46	1.07	0.95
Average	30.42	0.52	1.09	1.41

Ⅲ. 결론

본 연구는 대량의 데이터 수집 시 드는 시간과 비용의 문제점을 보완하고자 가상의 데이터를 생성하고자 하였다. 이를 위해 주가의 통계적 속성을 잘 반영하는 GBM 모형과 시뮬레이션 기법을 접목시켜 데이터를 생성시켰으며, 생성된 데이터를 바탕으로 기술적 지표를 만들어 머신러닝 기반 주가지수 등락 예측 모형 개발에 필요한 학습데이터로 활용하였다. 또한 예측 값들을 바탕으로 트레이딩 시스템을 만들어 평가 지표들을 통해 성능을 측정하였다. 실제 데이터와 생성 데이터의 평가 지표들을 비교해 본 결과 두 그룹의 성능이 거의 유사하거나, 생성 데이터를 활용한 트레이딩 성과가 더 우수한 결과를 얻었다. 이는 주가의 통계적 속성이 반영되어 생성된 데이터셋이 학습에 사용할 수 있을 정도로 원본과 흡사하다는 의미로 해석할 수 있다.

본 연구의 한계점으로는 트레이딩 시스템 시뮬레이션 구현시 거래 수수료를 고려하지 않았다는 점이다. 거래 횟수에 따른 손실이 추가적으로 발생하기 때문에 각 종목별로 거래 수수료를 포함한 결과를 살펴볼 필요가 있다.

ACKNOWLEDGMENT

본 연구는 2021 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음 (2021-0-01399).

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] 김태승, 이수원. "해외지수와 투자자별 매매 동향에 따른 딥러닝 기반 주가 등락 예측". 정보처리학회논문지. 소프트웨어 및 데이터 공학, vol. 10, pp.367-374, 2021.

[3] 허진원, 고광이, 백장선. “텐서 회귀모형을 이용한 단기 평균 한국종합주가지수(KOSPI)예측”. 한국데이터정보과학회지, 33(4), 601-614, 2022.