머신러닝 트레이딩 시스템 개발을 위한 변동성 라벨링 방법

송영헌, 김재윤*
국민대학교, *순천향대학교

yhij82@gmail.com, *kimym38@sch.ac.kr

Volatility labeling method for the development of a machine learning trading system

Song Young Hun, Kim Jaeyun*
Kookmin Univ., * Soonchunhyang Univ.

요 약

금융산업 내 알고리즘 트레이딩 분야에서 머신러닝과 딥러닝을 활용한 트레이딩 시스템을 개발하는 연구들이 늘어나고 있다. 하지만 주가 데이터의 특성 상 학습이 쉽지 않은 한계점이 존재한다. 본 연구에서는 주가 데이터의 노이즈를 줄이고 비선형성과 비정상성의 문제를 개선하기 위해 변동성 라벨링 방법을 이용한 트레이딩 시스템을 제안한다. 제안한 트레이딩 시스템을 검증하기 위해 전통적인 방법인 Up-Down 라벨링 방법과 비교 분석했다. 비교 분석 결과, 제안한 변동성 라벨링 방법이 Up-Down 라벨링 방법보다 성능이 개선됨을 확인했다.

Ⅰ. 서 론

금융 분야에서는 투자, 리스크 관리, 포트폴리오 관리, 사기 탐지 및 재무 자문 등 많은 의사결정 문제들이 존재한다. 이러한 의사결정 문제는 순차적인 특성을 가지며 환경을 특정할 수 없기 때문에 해결하기가 쉽지 않다. 핀테크의 주요 분야인 알고리즘 트레이딩 역시 이러한 동일한 문제를 가지고 있다.[1] 알고리즘 트레이딩이란 수학적 또는 거래 규칙에 기초하여 거래 결정을 자동으로 내리는 접근법이다. 전통적으로 사람에 의해 거래 규칙을 알고리즘 트레이딩 시스템에 적용했지만 수많은 패턴이 존재하는 금융 시장에서 한계점이 존재했다. 따라서 최근에는 머신러닝 또는 딥러닝을 이용하여 사람이 발견하지 못한 패턴들을 발견 및 학습시켜 알고리즘 트레이딩을 구축하는 연구들이 늘어나고 있다.

Lee[2]는 주식 시장의 추세를 예측하기 위해 하이브리드 피쳐 선택 방식을 사용한 SVM (Support Vector Machine) 기반 예측 모델을 제안했다. 제안한 하이브리드 피쳐 선택 방법과 Wrapper 방법의 장점을 결합하여 최적의 피쳐를 추출했다. 제안한 모델의 성능을 검증하기 위해 Back-Propagation Neural Network 와 성능을 비교했다. 그 결과, 제안한 하이브리드 피쳐 선택 방식을 사용한 SVM 기반 예측 모델이 높은 수준의 정확도와 일반화 성능을 가짐을 확인했다. Chen 와 Liu[3]은 LGBM(Light Gradient Boosting Machine) 기반의 주가를 예측하는 금융거래 시스템을 구축하였다. 시스템의 정확성을 GLM (Generalized Linear Model), DNN (Deep Neural Network), RandomForest, SVM 모델들과 비교 진행했으며, 그 중 LGBM 의 모델 예측 성과가 가장 우수함을 확인하였다. 이와 같이 머신러닝 또는 딥러닝을 활용하는 선행 연구들은 주가 데이터의 특징을 고려하지 않는다는 한계점이 존재한다. 주가 데이터는 잡음 (noise)이 많고, 비정상성 (non-stationarity)과 비선형성 (non-linearity)의 특징을 가지기 때문에 학습이 쉽지 않다.[4] 따라서 본 연구에서는 주가 데이터의 노이즈를 줄이고 비선형성과 비정상성의 문제를 완화하기 위해 변동성 라벨링 방법을 이용한 머신러닝 트레이딩 시스템을 제안한다. 제안한 트레이딩 시스템을 검증하기 위해 기존 연구에서 진행되던 Up-Down 라벨링 방법과 비교 분석한다.

Ⅱ. 본론

2.1 변동성 라벨링

본 연구에서 제안하고자 하는 변동성 라벨링 (labeling)은 다음과 같다. 먼저 일별 수익률과 수익률의 표준편차를 계산한다. 계산된 수익률과 표준편차를 이용하여 수익률의 위치를 바탕으로 학습데이터의 라벨을 정의한다. 변동성 라벨링의 공식은 식 (1)과 식 (2)와 같다. α는 수익률 표준편차의 계수 (0.25, 0.5, 0.75, 1)를 의미하며, σ는 수익률의 표준편차를 의미한다. Fig 1. 수익률의 히스토그램을 보여준다.

$$\text{Up} = +\alpha\sigma < \text{Daily return (\%)} \quad (1)$$

$$\text{Down} = -\alpha\sigma > \text{Daily return (\%)} \quad (2)$$

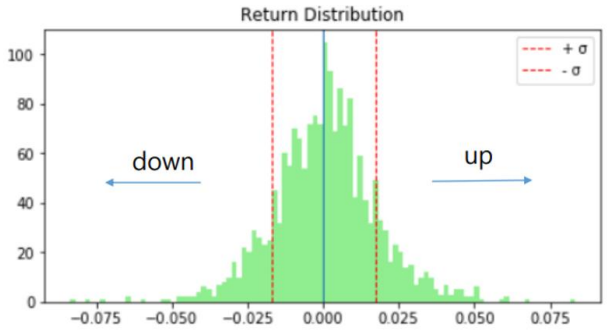

그림 1. Daily return histogram

2.2 Up-Down 라벨링

Up-Down 라벨링은 기존 연구에서 많이 사용되는 라벨링 방법이다. Up-Down 라벨링은 주식 가격 T 시점과 T+1 시점의 가격 차이를 바탕으로 라벨이 정의된다. 본 연구에서는 제안한 변동성 라벨링 방법과 비교를 위해 진행했으며 Up-Down 라벨링의 공식은 식(3)과 식(4)와 같다.

$$\text{Up} = \ Stock\ price_t < Stock\ price_{t+1} \quad (3)$$

$$\text{Down} = \ Stock\ price_t > Stock\ price_{t+1} \quad (4)$$

2.3 학습 및 테스트 데이터

데이터는 시가총액 100 위에 해당하는 종목들을 활용하였다. 학습 및 테스트 데이터를 구축하기 위해 필요한 입력 변수는 기술적 지표를 활용했다. 사용한 기술적 지표는 총 36 개이며 Python ta-lib 패키지를 활용해 추출했다. 학습 기간은 2009 년 ~ 2015 년 까지 총 7 년, 테스트 기간은 2016 년 ~ 2020 년까지 총 5 년이다.

2.4 머신러닝 알고리즘 및 트레이딩 시뮬레이션

구축된 학습 데이터를 바탕으로 머신러닝 알고리즘을 학습 시킨 후, 예측값을 바탕으로 트레이딩 시그널을 생성한다. 본 연구에서 트레이딩 시스템 구축을 위해 사용한 머신러닝 알고리즘은 총 6 개로 Logistic Regression, Decision tree, K-Nearest Neighbor (K-NN), Naïve bayes, Random forest, GBM (Gradient Boosting Machine)이다.

머신러닝의 예측값을 바탕으로 생성되는 트레이딩 시그널은 Table 1 과 같다. 또한 실제 투자와 유사한 결과를 확인하기 위해, 거래 수수료 0.015%를 부과하여 실험을 진행했다.

표 1. 트레이딩 시그널 예시

Date	Prediction	Trading signal
T	Up	Buy
T+1	Up	Hold
T+2	Down	Sell
T+3	Down	No action
T+4	Up	Buy

2.5 트레이딩 시뮬레이션 결과

트레이딩 시뮬레이션 결과를 평가하기 위해 트레이딩 평가 지표 사용한다. 사용한 평가 지표는 총 4 가지이며 거래 횟수, 승률 (=수익이 발생한 횟수 / 총 거래 횟수), Payoff ratio (=평균수익/평균손실), Profit factor (=총수익/총손실)이다. Table 2~3 은 종목별 트레이딩 시뮬레이션의 결과를 알고리즘별로 평균화하여 정리하였다. 본 연구에서 제안한 변동성 라벨링과 기존 연구에서 사용한 Up-Down 라벨링 방법을 비교한 결과, Payoff ratio 의 성능이 개선됨을 확인할 수 있다. 또한 제안한 변동성 라벨링 방법이 알고리즘별로 Profit factor 값이 큰 차이가 나지 않아 상대적으로 Up-Down 라벨링 방법보다 우수함을 알 수 있다.

이와 같은 결과는 변동성 라벨링 방법이 기존 Up-Down 라벨링 방법에 비해 트레이딩 타이밍을 잘 포착한다고 볼 수 있다. 더 나아가 이와 같은 결과는 제안한 변동성 라벨링 방법이 주가 데이터의 잡음을 제거하는 효과를 볼 수 있다고 판단된다.

표 2. 변동성 라벨링을 활용한 실험 결과

알고리즘	거래 횟수	승률	Payoff ratio	Profit factor
Logistic regression	188.47	0.45	1.28	1.11
Decision tree	281.71	0.46	1.18	1.04
K-NN	212.64	0.46	1.23	1.09
Random forest	206.57	0.45	1.25	1.08
Gradient boosting	268.01	0.46	1.19	1.03

표 3. Up-Down 라벨링을 활용한 실험 결과

알고리즘	거래 횟수	승률	Payoff ratio	Profit factor
Logistic regression	128.53	0.55	1.00	1.25
Decision tree	271.71	0.48	1.06	1.02
K-NN	277.75	0.47	1.08	1.00
Random forest	220.34	0.51	1.02	1.08
Gradient boosting	203.43	0.50	1.05	1.09

Ⅲ. 결론

본 연구에서는 대다수의 선행연구에서 사용되고 있는 Up-Down 라벨링 방법의 한계점을 개선하기 위해 변동성 라벨링 방법을 제안하였다. 실험결과 변동성 라벨링 방법이 주가 데이터의 노이즈를 줄이고 비정상성과 비선형성의 문제점을 완화시킴을 확인할 수 있었다. 트레이딩 시뮬레이션 결과, 제안한 변동성 라벨링 방법이 전통적인 라벨링 방법인 Up-Down 라벨링 방법보다 트레이딩 성과가 개선됨을 확인하였다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Théate, T., & Ernst, D. (2021). An application of deep reinforcement learning to algorithmic trading. Expert Systems with Applications, 173, 114632.

[2] Lee, M. C. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications, 36(8), 10896-10904.

[3] Chen, Y., Liu, K., Xie, Y., & Hu, M. (2020). Financial trading strategy system based on machine learning. Mathematical problems in engineering, 2020.

[4] Han, Y., Kim, J., & Enke, D. (2023). A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost. Expert Systems with Applications, 211, 118581.