앙상블 조합 방법에 따른 주가 예측 성능 비교

양현성[1], 박준[1], 소원호[2], 심춘보[1]
[1]순천대학교 IT-Bio융합시스템전공
[2]순천대학교 컴퓨터교육과
naiu8165@naver.com, cbsim@scnu.ac.kr

Comparison of Stock Price Forecasting Performance by Ensemble Combination Method

Huyn-Sung Yang*, Chun-Bo Sim*
*Interdisciplinary Program in IT-Bio Convergence System, Sunchon National University
**Department of Computer Education, Sunchon National University

요 약

본 연구에서는 머신러닝(Machine Learning, ML)과 딥러닝(Deep Learning, DL) 모델을 앙상블(Ensemble)하여 어떠한 주가 예측 방법이 우수한지에 대한 연구를 하고자 한다. 연구에 사용된 모델은 하이퍼파라미터(Hyperparameter) 조정을 통하여 최적의 결과를 출력한다. 앙상블 방법은 머신러닝과 딥러닝 모델의 앙상블, 머신러닝 모델의 앙상블, 딥러닝 모델의 앙상블이다. 세 가지 방법으로 얻은 결과를 평균 제곱근 오차(Root Mean Squared Error, RMSE)로 비교 분석하여 최적의 방법을 찾고자 한다. 제안한 방법은 주가 예측 연구의 시간과 비용을 절약하고, 최적 성능 모델 판별에 도움이 될 수 있다고 사료된다.

1. 서론

최근 퀀트 투자는 각종 커뮤니티와 매체, 포럼을 통해 대중들에게 친숙한 단어로 각인되고 있다. 퀀트 투자에서 퀀트(Quant)란 Quantitative와 Analyst의 합성어로써, 모형을 기준으로 금융상품의 가격을 산정하거나 투자를 하는 사람을 의미한다. 일반적으로 투자자들은 종목에 대한 기술적 분석을 통해 가치를 매기는 정성적인 투자법을 사용하는 데 반해 퀀트 투자는 수학과 통계를 기반으로 전략을 세우며 정량적인 투자법을 사용한다.

퀀트 투자자들이 정량적 투자를 위해 이용하는 데이터들은 본래 목적에 맞게 전처리가 필요하다. 적은 양의 데이터는 엑셀(Excel)을 이용해 간단한 백 테스트(Back-Test)가 가능하지만, 종목 수가 수천 종목을 넘고 특성 수가 방대해진다면 수작업은 사실상 불가능에 가깝다. 이러한 상황에서 머신러닝과 딥러닝을 이용하면 비용 효율적이다[1-2].

증권회사에서도 인공지능(Artificial Intelligence, AI) 기반의 주가 예측 프로그램을 개발하고 있다[3]. 주가 예측 프로그램 개발에 사용되는 머신러닝 알고리즘으로는 SVR(Support Vector Regression)[4]과 LightGBM(Light Gradient Boosting Machine)[5], 딥러닝은 알고리즘으로는 1DCNN(1D Convolutional Neural Network)[6]과 LSTM(Long Short-Term Memory)[7]이 있다.

머신러닝을 이용한 주가 예측 기존 연구들은 특정 모델을 사용하거나 여러 모델을 앙상블 했다. 앙상블이란 여러 개의 분류기를 생성하고, 그 예측을 결합함으로써 보다 정확한 예측을 도출하는 기법이다. 즉, 강력한 하나의 모델을 사용하는 대신 약한 모델 여러 개를 조합하여 더 정확한 예측에 도움을 주는 방식이다.

주가 예측 연구에서 어느 앙상블 구성의 방법이 더 나은지에 관한 연구는 아직 부족한 실정이다. 본 연구에서는 각 머신러닝, 딥러닝 모델들의 주요 하이퍼파라미터를 조정하여 훈련하고, 앙상블 하여 최적의 조합을 규명하는 연구를 하고자 한다.

2. 관련 연구

기존 머신러닝과 딥러닝 모델의 성능을 비교한 연구가 진행됐다. 실험결과 ARIMA(Auto-Regressive Integrated Moving Average) 모델보다 CNN의 예측 성능이 더 우수했다[8].

NARX(Nonlinear AutoRegressive with eXternal input) 모델을 활용한 연구에서는 코스닥(KOSDAQ)을 대상으로 단순 주식 관련 데이터뿐만 아니라 거

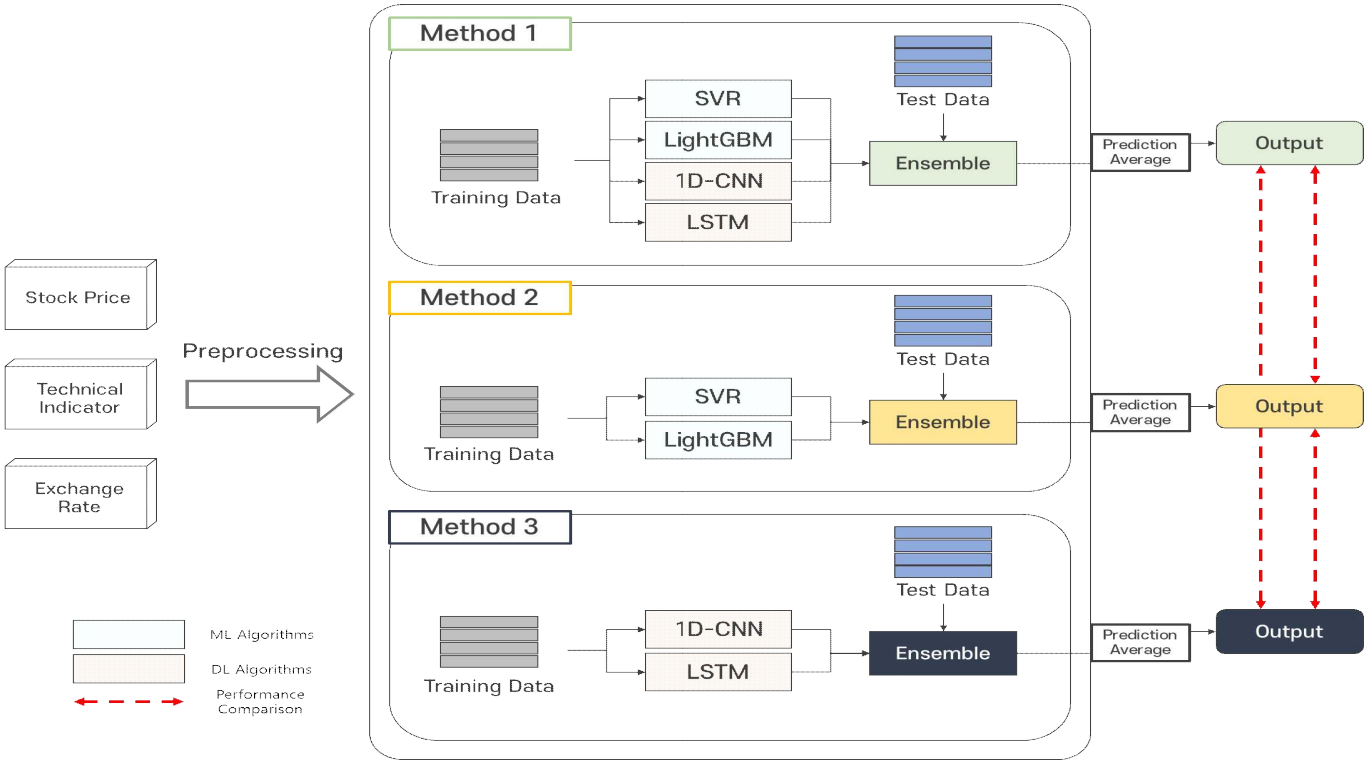

그림 1. 앙상블 방법에 따른 주가 예측 성능 비교 시스템 개요

시 경제적 지표 등을 활용하여 1년 치 종가, 외국인 비율, 금리, 환율 데이터를 다양하게 조합하였다. 연구 결과에 따르면 국내 통화 가치도 물론 중요하지만 주식 시장에서는 환율이 주식에 가장 큰 영향을 미친다. 실험결과 종가 데이터만을 사용했을 때의 오차가 가장 낮게 나왔다[9].

주가 예측 모델의 매개변수 조정 관련 연구에서는 LSTM 모델의 성능향상을 위해 크게 초기화(Initialization) 방법, 정규화(Regularization) 요소, 활성화 함수(Activation Function) 세 가지로 구분했다. 실험결과 예측 성능향상을

위해서 하이퍼파라미터 설정 방법에 따라 학습 성능이 달라질 수 있음을 확인했다[10].

주가 예측과 관련된 이전 연구들은 주가 데이터 및 보조 자료들을 이용해 진행됐다. 주가 예측에 있어서 국내시장의 흐름도 영향이 있지만, 국내 외국인의 비율이 상당한 만큼 환율에 대한 영향도 상당했으며, 다음 날의 주식 가격에 가장 큰 영향을 주는 종가로 예측했을 경우 예측 모델의 정확도가 높았음을 알 수 있다.

3. 제안하는 방법

방대한 양의 데이터를 처리할 수 있는 고사양 컴퓨터가 있더라도, 주가의 흐름을 예측하는 것은 힘들다. 단적인 예로 자연적 재해 같은 불가피한 상황들이 있기 때문에 이와 같은 외부 요인의 간섭을 최소화하고자 머신러닝, 딥러닝 모델에 다양한 특성을 추가하고, 다양한 방법으로 훈련을 진행한다.

본 실험에 사용할 주가 데이터는 코스피(KOrea Composite Stock Price Index, KOSPI) 상위 20개 종목을 대상으로 한 2011년부터 2021년 10년 치다. 주가 예측은 시가, 종가, 고가, 저가 종목의 4가지 특성과 외부 요인이 될 수 있는 KOSPI 지수, KOSDAQ(KOrea Securities Dealers Automated Quotation) 지수, S&P500(Standard and Poor's 500), 환율의 등폭을 추가로 사용한다. 10년간의 KOSPI, KOSDAQ, S&P500, 전처리 데이터의 80%를 훈련 데이터, 20%를 테스트 데이터로 나누고, 훈련 데이터의 20%는 검증 데이터로 사용한다. 머신러닝 모델은 SVR, LightGBM, 딥러닝 모델은 1D-CNN, LSTM을 사용한다. 주가 예측은 10일 치의 데이터를 이용해서 다음 날의 종가를 예측한다. 예측을 위해서 기존의 주가, 지수, 환율 데이터를 사용한다.

SVR은 SVM(Support Vector Machine)을 회귀문제에 이용한 모델로 종속변수가 연속형일 경우 사용할 수 있는 모델이다. SVM과 마찬가지로 마진(Margin) 밖에 있는 손실이 최소가 되도록 한다. 즉, 주가의 변동 폭 Margin에 최대한 많은 관측치가 포

함되도록 한다. SVR의 주요 하이퍼파라미터인 C를 1 ~ 10중 가장 좋은 값을 찾고, γ값을 0.001~1중 가장 좋은 값을 찾는다.

LightGBM은 기존의 Gradient Boosting 방식과 다르게 Leaf-Wise 방법을 사용한다. 주가 데이터가 Root-Node로 들어오면 다음 날의 값을 예측하기 위해 Child-Node가 생성되고, 더 정확히 예측한 Edge를 따라가는 방식으로 기존 Level-Wise 방식보다 손실을 줄일 수 있다. LightGBM의 하이퍼파라미터 중 Booster Method를 DART(Dropout Additive Regression Trees)로 설정한다. 주가 데이터가 10년치의 양으로 훈련 데이터에 과대 적합을 일으킬 수 있기 때문이다. 따라서 DART를 설정함으로써 신경망에 드롭아웃 레이어를 추가한다. 하이퍼파라미터 Maximum Depth는 -1로 설정하여 트리의 최하단까지 학습시킨다.

CNN은 주로 이미지를 분석하는 데 사용하지만, 1D-CNN은 주식 데이터와 같은 시계열 데이터 분석에도 사용한다. 합성곱 연산을 위한 커널과 대상 데이터의 모양이 1차원이면 1D-CNN을 사용하여 주가를 예측할 수 있다. 예측을 위해서는 이전 날의 데이터가 필요하다. 따라서, 이전 날을 의미하는 Kernel size를 10으로 설정한다.

LSTM도 시계열 데이터 분석에 적합한 모델이다. 시간의 경과에 따르는 주가 데이터를 입력으로 사용하고, 그 후의 값을 예측한다. DNN(Deep Neural Network)의 활성화 함수는 Tanh, Softsign, Softmax, Elu, Selu 등 다양하다. LSTM의 경우 활성화 함수로 Softsign을 사용했을 때 더 나은 성능을 보였기 때문에[11] 활성화 함수로 Softsign 함수를 사용한다.

그림 1은 앙상블 방법에 따른 주가 예측 성능 비교 시스템 개요다. 앙상블 기법으로는 랜덤 포레스트(Random Forest)를 사용한다[12]. 주가, 지수, 환율 데이터를 전처리 후 학습의 입력 데이터로 사용한다. 그림 1 상단의 Method 1은 학습 데이터로 머신러닝과 딥러닝 모델을 훈련하고 앙상블 했다. Method 2와 3도 같은 학습 데이터를 사용하여 머신러닝과 딥러닝 모델을 학습하고 앙상블을 했다. 앙상블 한 모델은 같은 테스트 데이터로 마지막 예측을 진행한다. 앙상블 결합에는 가장 우수한 성능을 보이는 곱 규칙(Product Rule)을 사용한다[13].

앙상블을 이용한 세 가지 방법의 성능을 평가하기 위해서 평균 제곱근 오차를 이용한다. RMSE는 모델의 예측값과 실제 주식 데이터의 가격의 차이를 다룰 때 자주 사용되는 평가지표로 식(1)과 같다. 연관성이 강한 특성의 개수가 다양할수록 오차율 또한 크다. 따라서 오차에 대해 큰 패널티를 주는 RMSE를 사용한다.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Predicted_i - Actual_i) LSUP2}{n}} \quad (1)$$

4. 결론

최근 퀀트 투자에 관한 연구는 꾸준히 발전하고 있다. 컴퓨팅 성능의 발전과 투자자들의 분석 고도화가 그 요인이다. 본 연구의 실험결과로 머신러닝과 딥러닝 중 어느 알고리즘이 월등하다고 단정 지을 수 없다. 실험에 이용한 모델과 데이터는 일부분이고 사람이 예측할 수 없는 변수들도 있기 때문이다.

본 연구에서는 이전 연구에서 발표되었던 다양한 모델들의 주가 예측 성능을 올리고자 특성의 개수를 추가하고, 특정 특성을 사용하여 전처리했다. 본 연구에서 사용한 모델들을 극히 일부분만을 예시로 언급했고, 추후에 더 나은 성능을 위해서 더 다양한 모델들의 활용과 하이퍼파라미터 조정 및 최적화 기법을 사용한다면 성능을 올릴 수 있을 것이라 기대한다.

참고문헌

[1] Sandeep Patalay and Madhusudhan Rao Bandlamudi, “Stock Price Prediction and Portfolio Selection Using Artificial Intelligence”, *Asia Pacific Journal of Information Systems*, Vol. 30, No. 1, pp. 31-52, 2020.

[2] Jingyi Shen and M. Omair Shafiq, “Short-term stock market price trend prediction using a comprehensive deep learning system”, *Journal of Big Data,* Vol. 7, No. 1, pp. 66, 2020.

[3] https://www.daishin.com/

[4] Harris Drucker, Chris J.C.Burges, Linda Kaufman, Alex Smola and Vladimir Vapnik, “Support Vector Regression Machines”, *Statistics and Computing*, Vol. 14, pp. 199-222, 2004.

[5] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, *Advances in neural information processing systems 30*, California, 2017, pp. 3146-3154.

[6] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince,Moncef Gabbouj and Daniel J. Inman, “1D Convolutional Neural Networks and Applications: A S

urvey“, *arXiv preprint, arXiv:*1905.03554, 2019.

[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory”, *Neural Computation*, vol. 9, no. 8, pp. 1735 - 1780, 1997.

[8] Hiransha M, Gopalakrishnan E.A, Vijay Krishna Men on and Soman K.P, “NSE Stock Market Prediction U sing Deep-Learning Models”, *Procedia Computer Sci ence*, Vol. 132, pp. 1351-1362, 2018.

[9] Min Jong Cheon, Ook Lee, “A Study on the stock pr ice prediction and influence factors throughNARX ne ural network optimization”, *Journel of the Korea Ac ademia-Industrial cooperation Society*, Vol 21, No 8, pp. 572-578, 2020.

[10] Jongjin Jung and Jiyeon Kim, “A Performance Anal ysis by Adjusting Learning Methods in StockPrice P rediction Model Using LSTM”, *Journal of Digital Co nvergence*, Vol. 18. No. 11, pp. 259-266, 2020.

[11] Xavier Glorot, Yoshua Bengio, “Understanding the d ifficulty of training deep feedforward neural network s”, *In Aistats*, Vol. 9, pp. 249-256, 2010.

[12] Leo Breiman, “Consistency For a Simple Model of Random Forests”, *Technical Report 670*, UC Berkele y, 2004.

[13] Sung-Wook Park, Jong-Chan Kim and Do-Yeon Ki m, “A Study on Classification Performance Analysis of Convolutional Neural Network using Ensemble Le arning Algorithm”, *Journal of Korea Multimedia Soc iety*, Vol. 22, No. 6, pp. 665-675, 2019.