https://doi.org/10.7840/kics.2023.48.6.751

Shallow CNN을 활용한 주가 예측 방법론

조 영 진•, 김 의 연*,
신 홍 기**, 최 용 훈°

Stock Price Prediction Methodology Using Shallow Convolutional Neural Network

Young-Jin Cho•, Eui-Yeon Kim*,
Hong-Gi Shin**, Yong-Hoon Choi°

요 약

딥러닝 기반으로 시계열 데이터 분석 및 예측 연구들은 최근에 활발하게 진행되고 있다. 이런 시계열 예측 딥러닝 모델들은 기본적으로 추세(trend)와 계절성(seasonality)과 같은 시계열의 특성을 학습하고 이를 기반으로 미래값을 예측하는 구조이다. 하지만 1~2일 정도의 단기 주가 시계열에서는 추세와 계절성 같은 특성을 발견하기 어렵다. 따라서 기존의 시계열 예측 딥러닝 모델을 통해서 주가를 예측하기는 어렵다. 본 논문에서 주가 시계열의 패턴을 얕은 CNN (Convolutional Neural Network)을 통해서 학습하여 1일 주식거래 분포를 예측하는 방법론을 제안한다. 1일 주식거래분포를 다양하게 표현할 수 있으나 본 논문에서는 박스플롯 예측 방법론을 제안한다.

Key Words : trend, seasonality, convolutional neural network (CNN), box plot, stock price prediction

ABSTRACT

Recently, research on analyzing and forecasting time series data using deep learning has been actively conducted. These time series forecasting using deep learning models learn the characteristics of time series such as trends and seasonality, and use them to predict future values. However, it is much more difficult to identify such characteristics in short-term stock price time series spanning 1 or 2 days. Therefore, it is difficult to predict stock prices using existing time series forecasting deep learning models. In this paper, we propose a methodology for forecasting daily stock trading distributions by training the pattern of stock price time series using shallow convolutional neural network (CNN). Although there are various ways to represent daily stock trading distributions, this paper proposes a methodology for forecasting box plots.

Ⅰ. 서 론

시계열 데이터 분석 및 예측 관련한 연구는 오랫동안 이루어져 왔고, 금융을 포함하여 여러 산업 분야에서 활용되어왔다. 전통적으로 AR, MA, ARMA 및 ARIMA와 같은 통계적인 시계열 예측 모형들이 있지만, 최근 들어 딥러닝 기술의 발전으로 딥러닝 기반 시계열 예측 연구들이 활발하게 이루어지고 있다. 그러나 최신 딥러닝 기반 시계열 예측 모델[1-5]을 이용하여 하루의 주가 변동을 예측하더라도 실제 거래에 활용할 수 있을 만큼의 유의미한 정확도를 얻지 못한다.

전통적인 통계적 시계열 예측 모델을 포함해서 딥러닝 기반 모델들은 추세 및 계절성 같은 시계열 데이터의 특성을 학습하고 이를 기반으로 예측을 진행한다. 하지만 다른 시계열 데이터와는 달리, 주가 시계열 데이터에서는 추세나 계절성 같은 특성을 발견하기 어렵다. 따라서 시계열 예측 분야에서 현재 가장 좋은 성능을 나타내는 모델들로 수행한 예측 실험에서도 유의미한 주가 예측 결과를 얻을 수 없었던 것으로 판단한다.

RNN (Recurrent Neural Network)은 딥러닝 모델 중에서 가장 기본적인 시퀀스 모델로 현재 시점의 데이터가 이전 시점의 데이터의 영향을 받는 시계열 데이터의 특성을 반영한 구조이다. RNN 모델은 깊은 신경망을 학습시키는 과정에서 기울기(gradient)가 소

※ 이 성과는 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1064080).
• First Author : (0009-0002-8810-6784) Division of Robotics, Kwangwoon University, yjaycho@gmail.com
° Corresponding Autho r (0000-0002-1460-0520) Division of Robotics, Kwangwoon University, yhchoi@kw.ac.kr, 교수, 종신회원
* (0009-0003-8853-3847) Division of Robotics, Kwangwoon University, dmldussla93@gmail.com
** (0000-0002-9349-1675) Division of Robotics, Kwangwoon University, ghdrl95@gmail.com, 학생회원
논문번호 : 202304-085-C-LU, Received April 21, 2023; Revised May 4, 2023; Accepted May 4, 2023

실되는 문제가 있어서, LSTM (Long Short Term Memory)과 GRU (Gated Recurrent Unit)와 같은 모델들로 발전되었다[1].

트랜스포머(Transformer) 모델은 자연어처리 분야에서 근간이 되는 시퀀스 모델이다. 서로 다른 시점의 정보들 간의 관계를 바탕으로 어텐션(attention)을 계산하여 학습하는 방식으로 동작한다. 초기에는 자연어처리 분야에서 응용되었지만, 시계열 데이터 예측, 컴퓨터 비젼 등 다양한 분야로 확장 연구되고 있다. 트랜스포머 기반의 시계열 데이터 예측 모델로 인포머(Informer)[2], 오토포머(Autoformer)[3] 등이 있다.

MLP (Multi-Layer Perceptron)만으로 구성된 비교적 단순한 구조를 가진 모델로서 대표적으로 N-BEATS[4]와 N-HiTS[5] 모델이 있다. N-HiTS 모델은 인포머, 오토포머와 같은 트랜스포머 기반 모델보다 약간 높은 예측 성능을 보여준다. N-HiTS 모델은 멀티 레이트 샘플링(multirate sampling) 기법을 통해 각 블록에서 시계열 데이터의 주요 주파수 특징을 학습하는 구조를 가지며, 단계적 보간법(hierarchical interpolation)을 통해 연산 효율을 높였다.

본 논문에서는 N-HiTS 모델을 이용하여 30분 주기의 1분 주가를 예측하는 실험을 수행하였다. 주가 시계열 특성상 날짜가 변경되는 시점에 데이터가 급격하게 변화하면서, 특정 시점의 정보는 이전 시점의 정보에 영향을 받게 되는 시계열 데이터의 특성인 연속성에 문제가 발생한다. 따라서 이런 문제점을 해결하기 위해서 입력구간과 예측구간에 날짜가 변경되는 데이터는 학습에서 제외시켰다. 이 실험에서 N-HiTS 모델은 주가 예측에 있어서 유의미한 정확도를 나타내지 못했다. 즉, N-HiTS 모델이 시계열 예측 모델에서 가장 뛰어난 성능을 보인 모델임에도 주가 예측에서는 뚜렷한 한계를 보였다. 이는 N-HiTS를 포함한 대다수의 딥러닝 기반의 시계열 예측 모델들이 추세와 계절성을 기반으로 예측하기 때문에, 단기 주가 시계열과 같이 추세와 계절성을 찾기 어려운 시계열에서는 유의미한 정확도의 예측을 기대하기 어렵다고 판단된다.

본 논문에서는 전일 장 후반 및 당일 장 초반 주식거래 시계열 데이터의 패턴을 얕은 합성곱 신경망(CNN: Convolutional Neural Network)을 통해 학습한 후, 최댓값, 최솟값, 3사분위값, 중위값, 1사분위값으로 구성된 1일 주식거래분포의 박스플롯을 예측하는 방법론을 제안한다.

Ⅱ. CNN 기반 주가 예측 방법론

박스플롯은 1일 주식거래분포(P_{max}, P_{Q3}, P_{Q2}, P_{Q1}, P_{min})를 예측하기 위해서 장 초반 1시간 주식거래분포를 기준으로 잡고, 이 기준값과 1일 주식거래분포의 차이인 변동값(ΔP)을 예측하여, 그림 1과 같이 최종적으로 1일 주식거래분포를 산출한다.

본 논문에서는 최근 2년간의 KOSPI 200 종목의 1분 거래 데이터를 학습데이터로 사용한다. 예측하고자 하는 당일 장 초반 1시간과 전일 장 후반 1시간의 1분 거래 데이타(시가, 종가, 저가, 고가, 거래량)를 표 1의 첫 번째 열에 나열된 총 6가지로 조합하여 입력 데이터로 구성한다.

표 1. 입력 데이터의 여러 조합에 따른 오차값
Table 1. Error values for different combinations of input

입력 데이터	사분위	MAE	RMSE
평균값 입력 (Case 1)	min	0.0576	0.0587
	Q1	0.0284	0.0329
	Q2	0.0261	0.0295
	Q3	0.0298	0.0338
	max	0.0082	0.0119
평균값, 거래량 입력 (Case 2)	min	0.0737	0.0747
	Q1	0.0714	0.0729
	Q2	0.0466	0.0489
	Q3	0.0314	0.0346
	max	0.0101	0.0122
종가, 시가, 고가, 저가, 거래량 입력 (Case 3)	min	0.0131	0.0182
	Q1	0.0208	0.0241
	Q2	0.1765	0.1864
	Q3	0.0902	0.0954
	max	0.0948	0.0967
전날 한 시간을 포함 평균값 입력 (Case 4)	min	0.0315	0.0367
	Q1	0.0620	0.0641
	Q2	0.0164	0.0208
	Q3	0.1146	0.1282
	max	0.0222	0.0260
전날 한 시간을 포함 평균값, 거래량 입력 (Case 5)	min	0.0246	0.0278
	Q1	0.1545	0.1551
	Q2	0.1147	0.1169
	Q3	0.0385	0.0415
	max	0.0613	0.0641
전날 한 시간을 포함 종가, 시가, 고가, 저가, 거래량 입력 (Case 6)	min	0.0513	0.0524
	Q1	0.0451	0.0470
	Q2	0.0835	0.0848
	Q3	0.0645	0.0675
	max	0.0160	0.0196

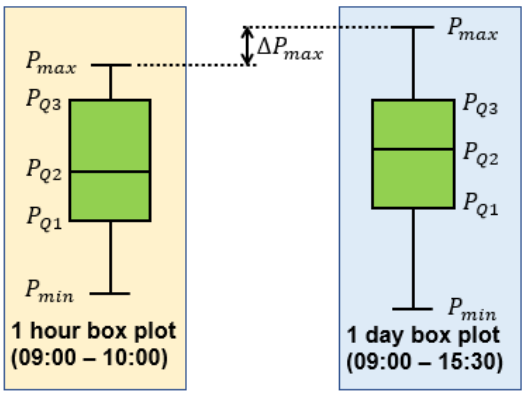

그림 1. 주식거래분포 박스플롯
Fig. 1. Box plot of daily stock trading distribution

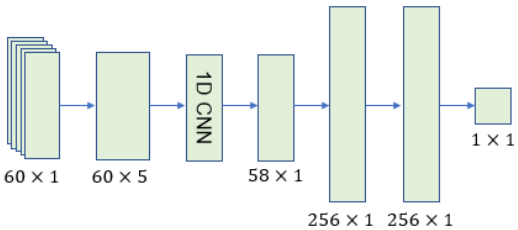

그림 2. 제안하는 Shallow CNN 모델
Fig. 2. Proposed shallow CNN architecture

5가지의 출력 ΔP_{max}, ΔP_{Q3}, ΔP_{Q2}, ΔP_{Q1}, ΔP_{min}을 예측하기 위해서 5가지 방법으로 얕은 CNN 모델을 학습시킨다. 이중 장 초반 1시간 데이터를 입력받아 당일의 박스플롯을 예측하는 모델의 구성은 그림 2와 같다. 모델 학습에 사용된 하이퍼파라미터들은 다음과 같다. 합성곱의 커널 크기는 3, 패딩은 0, 보폭은 1로 설정하였고, 모든 히든층의 활성화함수는 하이퍼볼릭 탄젠트를 사용하였다. 설정한 학습율은 0.001, 학습회수(epoch)는 30회, 배치크기는 1024이다.

Ⅲ. 실험 결과

본 논문에서는 실제값과 예측값의 차이를 관찰하기 위한 오차값으로 MAE (Mean Absolute Error)와 RMSE (Root Mean Square Error)를 관찰한다. 여러 조합의 입력 데이터로 학습된 모델들의 예측 오차값은 표 1과 같다. 장 초반 1시간의 1분 거래 시가와 종가의 평균값을 입력 데이터로 학습한 모델(Case 1)은 P_{max}, P_{Q3}를 가장 정확하게 예측하였다. 장 초반 1시간의 1분 거래 시가, 종가, 저가, 고가 및 거래량을 입력 데이터로 학습한 모델(Case 3)은 P_{Q1}, P_{min}을 가장 정확하게 예측하였다. 마지막으로 P_{Q2}를 예측하는 최적의 모델은 전일 장 후반 1시간과 당일 장 초반 1시간의 1분 거래 시가와 종가의 평균값을 입력 데이터로 학습한 모델(Case 4)이었다.

Ⅳ. 결 론

추세와 계절성과 같은 시계열 특성을 기반으로 학습되는 기존 딥러닝 예측 모델들은 이런 시계열 특성을 잘 띄지 않는 주가 데이터 예측에는 유의미한 예측 성능을 보이지 않는 것을 시험을 통해서 확인하였다. 따라서 본 논문에서는 금융 회사와 같은 주요 시장 참여자들이 활발히 거래하는 장 초반, 장 후반 주식 거래 데이터의 패턴을 학습하여 주식 거래 분포를 예측하는 방법론을 제안하였다. 1차원의 shallow CNN을 적용한 단순한 모델을 기반으로 다양한 입력 조건과 데이터를 구성하여 학습을 진행함으로써 최적의 모델을 찾을 수 있었다. 현재, 주가 변동의 일반적인 특성을 학습할 수 있는 모델 구조와 학습데이터 구조에 대한 연구를 수행중이다.

References

[1] D. Shin, K. Choi, and C. Kim, “Deep learning model for prediction rate improvement of stock price using RNN and LSTM,” *J. KIIT*, vol. 15, no. 10, pp. 9-16, Oct. 2017. (http://dx.doi.org/10.14801/jkiit.2017.15.10.9)

[2] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, “Informer: Beyond efficient transformer for long sequence time-series forecasting,” in *Proc. The Thirty-Fifth AAAI*, vol. 35, no. 12, pp. 11106-11115, 2021.

[3] H. Wu, J. Xu, J. Wang, and M. Long, “Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting,” in *Proc. NeurIPS*, vol. 34, pp. 22419-22430, 2021.

[4] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, “N-BEATS: Neural basis expansion analysis for interpretable time series forecasting,” in *Proc. Tenth ICLR*, 2020.

[5] C. Challu, K. Olivares, B. Oreshkin, F. Garza, M. Mergenthaler-Canseco, and A. Dubrawski,

"N-HiTS: Neural hierarchical interpolation for time series forecasting," *arXiv:2201.12886v5*, 2022.
(https://doi.org/10.48550/arXiv.2201.12886)