Paragraph Vector 와 GAN 을 활용한 KOSPI 일별 예측에 관한 연구

*박태준, *신은우, *정수환
서강대학교, 서강대학교, 서강대학교

taejoonparkwork@gmail.com, sew9869@gmail.com, soohwan34@naver.com

Daily KOSPI Stock Price Prediction Using Paragraph Vector and GAN

*Park Tae Joon, *Shin Eun Woo, *Jung Su Hwan
Sogang Univ., Sogang Univ., Sogang Univ.

요 약

본 논문은 일별 주식 데이터와 경제 뉴스를 입력으로 하고, LSTM 모델과 CNN 을 각각 Generator 와 Discriminator 로 하는 GAN 모델을 제시한다. 제시한 모델을 이용해 KOSPI 159 개 기업의 일별 주가의 시가, 종가, 고가, 저가를 예측하고자 하였다. 실험 결과, 경제 뉴스를 Paragraph Vector 로 변환해 사용할 경우 유의미한 차이가 있음을 확인하였다.

Ⅰ. 서 론

주식 가격에는 국가의 정책 방향, 세계 정세 등 사회의 모든 비가시적 정보 또한 주식 가격에 영향을 미친다. 일반적인 투자자들도 주식을 매매 시 주식의 가격만 보는 것이 아니라 기사, 세계 정세, 유가 등 다양한 정보를 얻고 가격을 예측하여 매매를 한다. 기존의 주식 가격을 예측하는 ML 모델의 경우에는 주식의 가격 데이터만을 입력으로 사용하여 예측하는 경우가 많았다.

주식의 가격 뿐만 아니라 비정형 데이터를 입력으로 사용한다면 더 높은 정확도를 기대할 수 있다. 이러한 방법을 이용하여 주식 가격을 예측하는 모델을 제안한 논문이 있다[1]. 해당 논문에서는 기사와 주식 가격을 모델의 입력으로 사용해 향상된 결과를 도출하였다.

많은 논문에서 시계열 데이터인 주식 가격을 LSTM 을 이용해 예측을 시도하였다. 앞서 설명한 논문 또한 기사와 주식 가격 데이터를 LSTM 모델을 사용하여 가격을 예측하였다. 그러나 최근 들어 그 한계가 명확해지고, GAN 을 통해 주식 가격을 예측하려는 시도가 있었다[2]. 본 논문은 기사와 주식 데이터를 사용하여 기존의 LSTM 모델을 Generator 로 사용하는 GAN 모델을 통해 정확도를 높인 모델을 제안한다.

Ⅱ. 본 론

본 논문은 일별 KOSPI 데이터를 예측하기 위해 기존의 주식 데이터와 비정형 데이터인 경제 기사를 사용하였다. 기사를 전처리하기 위해 [3]에서 제시하는 Paragraph vector 방식을 통해 기사를 벡터로 만들었다. 이를 주식 데이터와 연결하여 모델의 입력으로 하였다.

모델의 파이프 라인은 GAN 구조를 따른다. Generator 는 LSTM 을 사용하여 예측 주식 가격 $\hat{Y}$ 를 생성하고, Discriminator 는 CNN 과 FC layer 를 사용하여 $\hat{Y}$의 진위를 판별하였다. 위 과정을 반복하여 Generator 와 Discriminator 의 적대적 경쟁을 통해 Generator 의 예측 성능을 높였다.

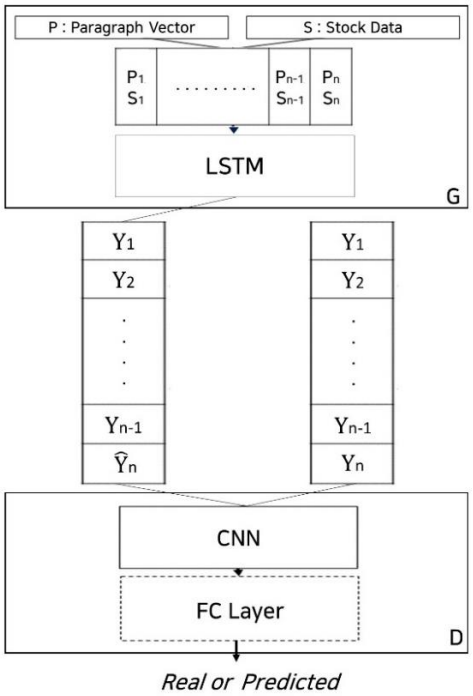

Figure 1. 제안한 KOSPI 가격 예측 모델

2.1 데이터셋

본 논문에서는 일별 주식 데이터와 뉴스 기사 데이터를 데이터셋으로 사용하였다. 주식 데이터로 KOSPI 200 의 기업 중 2010 년부터 거래 기록이 있는 159 개 기업의 일별 시가, 종가, 고가, 저가, 거래량을 사용하였다. 뉴스 기사 데이터는 네이버의 경제면 기사를 수집하여 paragraph vector 로 만들어 입력으로 사용하였다.

훈련 데이터는 2010 년부터 2018 년까지의 일별 주식과 기사 벡터 데이터를, 테스트 데이터는 2019 년의 일별 주식 기사 벡터 데이터를 사용하여 실험하였다. 따라서 총 2210 개의 훈련 데이터와 159 개의 기업의 주식 데이터를 사용하였다.

주식 가격은 기업마다 차이가 있어 각 주식 데이터는 전날 대비 등락률을 입력으로 사용하였다. KOSPI 시장은 최대

* These authors contributed equally.

±30%의 상,하한선이 있어 주가 등락률을 0.7 에서 1.3 사이로 정규화할 수 있다. 따라서 시가, 종가, 고가, 저가의 경우 아래와 같이 처리하였다.

$$Open_t = (Open_t)/(Close_{t-1})$$
$$Close_t = (Close_t)/(Close_{t-1})$$
$$Min_t = (Min_t)/(Close_{t-1})$$
$$Max_t = (Max_t)/(Close_{t-1})$$

거래량의 경우, 데이터의 실험적 결과에 따라 아래와 같이 변환하였다.

$$Volume_t = \begin{cases} 1.3 & (Volume_t)/(Volume_{t-1}) = 0 \\ 1.1 & (Volume_t)/(Volume_{t-1}) < 0.5 \\ 1.3 & (Volume_t)/(Volume_{t-1}) < 1 \\ 2 * \log(Volume_t) + 1.3 & (Volume_t)/(Volume_{t-1}) \geq 1 \\ 6.4 & (Volume_t)/(Close_{t-1}) \geq 6.4 \end{cases}$$

뉴스 기사는 약 1400 만개를 수집하였다. 기사를 [3]의 PV-DM 방식을 사용해 말뭉치를 학습한 후 paragraph vector 로 변화시켰다.

주가 데이터는 159 개 기업의 시가, 종가, 저가, 고가, 거래량의 5 개 정보를 포함하였다. 따라서 기사와 주식 정보를 동등하게 반영하기 위해 기사의 paragraph vector는 795 차원과 유사한 792 차원의 벡터로 변환되었다. 이후 생성된 두 벡터를 합쳐 1587 차원 데이터를 만들고, 이를 generator 의 입력으로 사용하였다.

2.2 GAN

Generator 는 time step 이 10 인 LSTM 으로, Many-to-One 방식을 사용해 11 번째 time step 의 기업 별 시가, 종가, 저가, 고가의 변화율을 예측하였다.

Discriminator 는 leaky ReLu 활성화 함수를 기반으로, 3 개의 Convolution Layer 와 2 개의 FC Layer 를 사용하였다. 상세 구조는 아래와 같다.

모델 레이어	변수
Convolution Layer	Filter 32, 4 * 2 * 1, leaky ReLU, Batch Normalization
Convolution Layer	Filter 64, 4 * 2 * 1, leaky ReLU, Batch Normalization
Convolution Layer	Filter 128, 4 * 2 * 1 , leaky ReLU, Batch Normalization
FC Layer	128, leaky ReLU
FC Layer	1, sigmoid
Optimizer: Adam, Batch Size: 40;	

Figure 2. Discriminator 모델 구성

2.3 모델 학습

GAN 모델은 Binary Cross-entropy 를 Loss 로 설정하여 학습하였고, learning rate 는 0.000001 로 설정하였다. batch size 는 40, epoch 는 10000 회로 실험하였다.

Ⅲ. 결 론

3.1 실험 결과

본 논문에서는 성능 측정을 위하여 MAPE (Mean Absolute Percentage Error) 를 사용하였다.

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

	ParagraphVector 사용	ParagraphVector 미사용
MAPE	1.1625 %	1.8 %

Figure 3. Paragraph Vector 사용에 따른 오차율

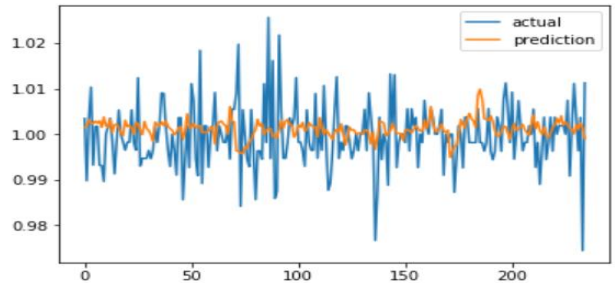

Figure 4-(a). KT 종가 변동률 예측 그래프

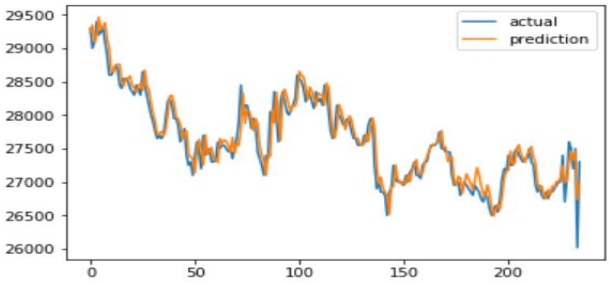

Figure 4-(b). 가격을 반영한 예측 그래프

종가 변동률 예측 그래프에서, 모델이 변동 방향을 예측하는 것을 확인할 수 있었다.

3.2 시사점 및 개선 방안

위 실험을 통해 Paragraph Vector 를 이용해 뉴스 기사를 벡터로 만들어 주가 예측에 사용할 경우 유의미한 차이를 확인할 수 있었다.

그러나 주가 변동 추세를 반영하지만 정확한 변동률을 예측하진 못하였으며, 이례적인 변동폭에 대해선 정확도가 떨어졌다. 이는 회귀 방식을 사용해 주가의 변동폭을 예측하려는 시도 대신 상승과 하락을 구분하는 분류 문제로 변환한다면 더 높은 성능을 낼 수 있을 것이라 기대된다.

Paragraph Vector 를 일별 예측에 적용할 때 약 2000 개의 일별 기사들을 단순 평균으로 하나의 벡터로 만드는 과정에서 정보 손실이 컸을 것이라고 예측된다. 약 2000 개의 기사를 평균으로 처리하기보다 다른 방식을 사용하면 더 유용할 것으로 판단된다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 연구결과로 수행되었음 (2015-0-00910)

참 고 문 헌

[1]R. Akita, A. Yoshihara, T. Matsubara and K. Uehara, "Deep learning for stock prediction using numerical and textual information," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, Okayama, 2016, pp. 1-6, doi: 10.1109/ICIS.2016.7550882.

[2] Zhou, Xingyu & Pan, Zhisong & Hu, Guyu & Tang, Siqi & Zhao, Cheng. (2018). Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets. Mathematical Problems in Engineering. 2018. 1-11. 10.1155/2018/4907423.

[3] Le, Q.V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *ArXiv, abs/1405.4053*.