딥러닝 기반의 주식 가격 예측을 위한 입력 특징의 비교 분석

김경중
충북대학교 컴퓨터공학과

Comparative Analysis on Input Features for Stock Price Prediction Using Deep Neural Network

Kim, Gyeong Joong
Chungbuk National University
E-mail : cjssla1@google.co.kr

요 약

이전부터 주식 시장을 예측하려고 하는 시도는 많이 있었으며 최근에는 딥러닝을 이용한 시도가 늘고 있는 추세이다. 딥러닝에서 중요한 것은 데이터라는 것에 누구가 공감할 것이며 어떤 데이터를 입력 특징으로 사용하는 가에 따라 모델의 성능은 천차만별이다. 본 논문에서는 기본 특징으로, 종가, 고가, 저가, 시가, 거래량을 이용하여 예측한 주식 가격과 변화율, 금값, 유가, 코스피, 원달러환율을 입력 특징에 추가하여 예측한 가격 및 실제 가격을 비교 분석하였다. 분석 결과 범용적인 주식 예측 모델 생성을 위해서는 추가 입력 특징없이 기본 특징만을 사용한 모델이 좋은 성능을 보였다.

1. 서론

이전부터 주식 시장을 예측하려는 시도는 많이 있었으며 최근에는 인공신경망 기반의 딥러닝을 이용한 시도가 늘고 있는 추세이다. 딥러닝을 이용하여 주식 시장을 예측하기 위해 사용한 특징들로는 시가, 저가, 종가, 고가, 거래량[1,2,3,4,5], 이동평균, CCI(Commodity channel index), RSI(Relative Strength Index), 원달러환율[3], 거래량 이동평균[4], KOSPI, 나스닥 지수[5]와 같이 여러 특징을 이용한 시도가 있었다.

이 논문에서는 참조한 논문들 중 공통적으로 나타난 종가, 고가, 저가, 시가, 거래량의 5개 특징을 기본 특징으로 선택하였고, 앞에 언급했던 논문들을 참조하여 KOSPI, 원달러 환율의 종가를 입력 특징으로 선택하였다. 종가를 선택한 이유는 다음날의 종가를 예측하는 모델을 설계하였기 때문이다. 또한 문헌[6]을 참조하여 국내에서 사용

하는 석유인 두바이 유가를 선택하였으며, 문헌[7]을 참조하여 런던 금시장의 금값을 선택하였다. 그 외에도 이번 논문에선 변화율이 입력 특징으로써 좋은 지 알기 위하여 변화율이란 특성을 선택하였는데, 변화율이란 전날 대비 종가의 변화를 퍼센트(%)로 나타낸 값이다. 예를 들어 전날 가격이 100이고 다음날이 101일때는 변화율은 +1%가 되며 전날 종가가 100일 때 다음 날의 종가가 99가 된다면 −1%로 표시되는 방식이다.

모델은 RNN(Recurrent Neural Network: 순환신경망)모델의 파생인 GRU(Gated Recurrent Unit)와 MLP(Multi−Layer Perceptron)을 쌓아서 구성하였으며 이전 7일의 데이터를 사용하여 다음날의 종가를 예측하도록 학습시켰다.

2. 본론

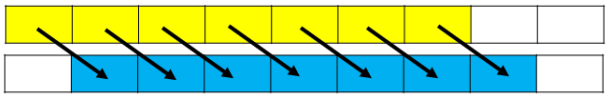
2−1. 데이터 구성

범용적인 종목들을 예측하기 위해 학습 데이터로는 삼성전자, SK하이닉스, 네이버, LG화학, 셀트리온, 현대차, 카카오, 삼성SDI, LG생활건강, 현대모비스, SK텔레콤, NC소프트, 기아차, 포스코, KB금융, SK, LG전자, LG, 한국전력의 2011.1.1 ~ 2020.8.1일의 데이터를 사용하였으며 검증 데이터로는 KT의 2001.1.1 ~ 2020.8.1일의 데이터를 사용하였다. 총 학습데이터는 44859개이며 검증데이터는 4841개이다.

데이터 수집에는 오픈 라이브러리인 Pandas Data Reader[9]와 Finance Data Reader[10]를 이용하였으며 원활한 학습을 위해 입력 특징들 별로 Min Max Scaling을 통하여 정규화 하였다. 지도학습을 하기 위하여, 학습 데이터 구성을 그림.1과 같이 설계하였다. 모델은 i−1번째 값으로 i번째 값을 예측하고 실제 i번째 값이 정답 레이블이 되어 지도학습의 방식으로 진행하였다.

$$x_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

(식 1. Min Max scaling)



(그림 1)

2−2. DNN(Deep Nueral Network) 모델

모델은 6층의 DNN이며 단방향 GRU(layer 4)와 배치정규화, 드랍아웃 이후 MLP(layer 2)로 구성되며, 그 외 주요 하이퍼 파라미터는 학습률은 0.001, 배치의 크기는 128, 히든 차원의 크기는 64, 학습 에포크는 30, 드랍아웃의 비율은 0.5, 옵티마이저는 Adam을 사용했다. 더 많은 하이퍼 파라미터나 각 레이어의 입력 차원과 출력 차원의 변화와 같이 더 자세한 구성을 확인하고 싶다면 문헌[11]를 참조하길 바란다.

2−3. 실험 과정

실험은 구글 colab 환경에서 이루어졌으며 프레임워크는 Pytorch를 사용하였고, 모델 생성시의 가중치가 항상 같도록 시드 값을 고정하였다.

학습은 한 종목의 데이터(2361개)를 30에포크 학습 후 다음 종목의 데이터(2361개)를 가지고 30에포크 학습하는 방식으로 진행하였으며 손실함수는 MSE(Mean Squared Error:평균제곱오차)를 사용하였다. 검증을 위해 MAE(Mean Absolute Error:평균절대오차) 값에 100을 곱한 값을 사용하였으며, 학습 도중 검증 데이터의 MAE값이 제일 작았을 때의 모델의 파라미터들을 가지고 비교한다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{true})^2$$

(식 2. MSE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_{true}|$$

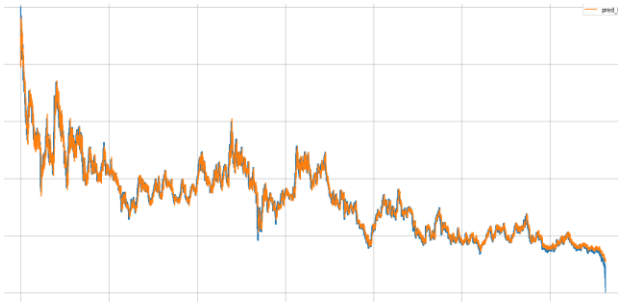
(식 3. MAE)

2-4. 실험 결과

(1) 시가, 고가, 저가, 종가, 거래량 5개의 특징
(2) (1)에 변화율 추가
(3.1) (2)에 KOSPI(종가) 추가
(3.2) (2)에 국제 금값(런던 금시장) 추가
(4) (3.2)에 국제 유가(두바이유가) 추가
(5) (4)에 KOSPI(종가) 추가
(6) (5)에 원/달러환율(종가) 추가

	Train	Validation
(1)	3.50	1.17
(2)	3.58	1.22
(3.1)	3.65	1.20
(3.2)	3.91	1.38
(4)	4.01	1.82
(5)	2.85	1.53
(6)	3.95	1.48

(표 1. 각 실험의 MAE)



(그림 2)

그림 2는 (1)의 모델을 가지고 검증 데이터(KT)의 예측한 값과 실제 값들의 분포를 볼 수 있다. 주황색 선이 예측 가격이며 파란색 선이 실제 가격으로 매우 유사한 결과를 얻을 수 있다.

그러나 검증데이터로 KT 한 종목만을 가지고 실험을 진행하였기 때문에 KT에 과적합되었을 뿐 실제 범용적인 예측능력이 떨어질 수 있다. 이를 확인하기 위해 CJ 대한 통운, CJ, LG이노텍, GS, 두산 중공업 5가지 종목의 2011.1.1 ~ 2020.8.1까지의 데이터를 가지고 실험을 진행하였으면 결과는 표 2와 같다. 모델은 표 1에서 사용한 모델들과 같다.

	대한통운	CJ	LG	GS	두산	평균
(1)	4.86	3.03	5.01	3.45	2.66	3.80
(2)	4.94	3.62	9.86	4.86	3.75	5.40
(3.1)	6.20	4.58	10.97	5.52	5.22	6.49
(3.2)	7.09	5.05	12.47	5.92	5.97	7.3
(4)	9.51	5.67	12.79	5.74	6.91	8.12
(5)	7.93	5.32	8.65	5.72	5.08	6.54
(6)	6.46	4.47	11.45	5.55	5.86	6.75

(표 2. 테스트 데이터의 MAE)

표 2의 실험결과를 참고하면 표 1의 결과와 같이 기본 5개 특징만을 이용한 모델이 제일 성능이 좋았음을 알 수 있다.

3. 결론

여러 종목을 예측하는 데 사용하기 위한 범용적인 모델을 생성하기 위해서는 기본적인 입력 특징으로 종가, 시가, 저가, 고가, 거래량만을 사용한 모델이 그 외 변화율, 유가, 금값, 환율, 코스피를 추가한 경우보다 성능이 좋다.

다만 주의할 점은 실제 사용에 앞서서, MAE의 값은 Min Max Scaling 된 데이터들의 결과물이기 때문에 MAE가 5~10라고 하더라도 실제 주식 가격과 예측의 차이가 5원에서 10원차이의 완벽에 가까운 모델이라고 할 수 없기에 다시 복원한 후 차이를 비교한 뒤 사용할 필요가 있다.

[참고문헌]

[1] 송유정, 이종우, “텐서플로우를 이용한 주가 변동 예측 딥러닝 모델 설계 및 개발”, 한국정보과학회 학술발표논문집 , 2017.06, 799-801 (3 pages)

[2] 주일택, 최승호, “양방향 LSTM 순환신경망 기반 주가예측모델”, 한국정보전자통신기술학회 논문지 11(2), 2018.4, 204-208(5 pages)

[3] 신동하, 최광호, 김창복, “RNN과 LSTM을 이용한 주가 예측율 향상을 위한 딥러닝 모델”, 한국정보기술학회논문지 15(10),

2017.10, 9-16(8 pages)
[4] 송유정, 이재원, 이종우, "텐서플로우를 이용한 주가 예측에서 가격-기반 입력 피쳐의 예측 성능 평가", KIISE Transactions on Computing Practices 23(11), 2017.11, 625-631(7 pages)
[5] 배지윤, 석준희, "입력 종목 수에 따른 LSTM 주식 가격 예측 모델 성능 비교", 한국통신학회 학술대회논문집 , 2018.1, 310-311(2 pages)
[6] 허은녕, 김지효, "국제 유가 변동이 국내 에너지 기업의 주가에 미치는 영향 연구", 한국신재생에너지학회 학술대회논문집 , 2008.05, 120-123(4 pages)
[7] 서지용, "국제 금 및 원유 선물시장의 거래정보는 글로벌 주식시장 수익률에 유의한 영향을 미치는가?", 대한경영학회지 24(1), 2011.2, 323-338(16 pages)
[8] idea Factory Kaist, 딥러닝 홀로서기, https://www.youtube.com/channel/UCTivi6Kji_93AjJu-7-osLQ
[9]FinanceData.KR , Finance Data Reader, https://github.com/FinanceData/FinanceDataReader
[10] Pandas, Pandas Data Reader, https://pandas-datareader.readthedocs.io/en/latest/
[11] 김경중, 주식 예측 모델, 깃허브 주소 https://github.com/cjssla1/projectJucker/blob/master/Code/%EA%B9%80%EA%B2%BD%EC%A4%91/GRU/GRU_Finance.ipynb
[12] 성노윤, 남기환, "온라인 뉴스 및 거시경제 변수를 활용한 주가예측", entrue journal of information technology December 2017 / Vol.16, No.2 (pp41-54)