대규모 외생 변수와 Deep Neural Network를 사용한 금융 시장 예측의 성능 향상에 관한 연구

천성길*, 이주홍*, 최범기*, 송재원**
*인하대학교 전기컴퓨터공학과
**(주)밸류파인더스
chkrdp@gmail.com, juhong@inha.ac.kr, bgchoi666@gmail.com,
jwsong@valuefinders.co.kr

A Study on Improving the Performance of Financial Market Forecasting Using Large Exogenous Variables and Deep Neural Network

Sung-gil Cheon*, Ju-Hong Lee*, Bumghi Choi*, Jae-Won Song*
*Dept. of Computer Engineering, Inha University
**ValueFinders Co., Ltd

요 약

시장예측 문제를 해결하기 위하여 과거부터 꾸준한 연구가 진행되어왔다. 하지만 금융 시계열 데이터에는 분산이 일정하지 않으며 Non-stationarity 등 예측을 하는 것에 있어서 여러 가지 방해 요인이 존재한다. 또한 광범위한 데이터 변수는 기존에 사람이 직접 경험적으로 선택하는 것에 한계가 있기 때문에, 모델이 변수를 자동으로 추출할 수 있어야 한다. 본 논문에서는 여러 가지 금융 시계열 데이터의 문제를 고려하여 타임 스텝 정규화를 제안하며 자동 변수 추출을 위해 LSTM 형태의 오토인코더 모델을 학습하였으며 LSTM 네트워크를 이용하여 시장 예측하는 모델을 제안한다. 해당 시스템은 실제 주식 거래나 시장 거래를 위하여 온라인 학습이 가능하며 긴 기간을 테스트 구간으로 실험한 결과 미래의 수익률을 예측하는 것에 있어서 우수한 성능을 보였다.

1. 서론

최근 딥러닝의 발달로 머신러닝 분야는 크게 주목받고 있으며 각종 의료, 헬스, 금융, 인터넷 분야에서도 좋은 성능을 보이며 많은 분야에서 딥러닝이 활용되고 있다[1].

금융 시장을 분석하기 위한 시도는 아주 오래전부터 진행되어왔었다. 시장예측을 한다는 것은 단순히 예측된 값을 보고 거래를 목적으로 하는 것이 아닌 투자자에게 투자의 방향성을 제시할 수 있는 중요한 자료가 되기 때문이다. 과거 전통적인 방법으로 통계 기반의 시계열 분석 방법이 많이 사용되었으며 최근 딥러닝의 발달로 ANN, CNN, LSTM등 다양한 방법으로도 시도되고 있다. 또한 기술 지표를 사용하는 방법뿐만 아니라 뉴스 크롤링과 시장 감정분석 등 여러 가지 방법을 이용하여 복합적으로 시도되고 있다[2].

하지만 많은 시도에도 불구하고 시장예측은 사람이 만족할 만큼의 결과를 주지 못하였다. 금융과 관련된 변수는 매우 다양하기 때문에 사람이 직접 수집하는 것 조차 쉬운 일이 아니며 전문적인 지식을 바탕으로 분석하는 것도 한계가 있다. 또한 금융 시계열 데이터에는 몇 가지 특징이 존재하는데, 분산이 일정하지 않은 Volatility, 기대치가 특정한 방향으로 증가 또는 감소하는 형태이거나 패턴이 명확하지 않은 Non-stationarity, 혹은 시차 간 상관관계가 존재하지 않은 Non-linearity, 과거의 정보가 미래의 정보에 영향을 미치는 장기 의존성 등의 특징들은 우리가 데이터의 trend와 noise를 다루는 것을 어렵게 만들며, 시장예측의 큰 방해 요인이 된다.

본 논문에서는 이러한 방해 요인들을 해결하기 위한 전처리 시스템, 모델이 자동으로 변수를 추출하는 시스템, 예측 모델 및 학습 시스템을 제안한다. Volatility 및 Non-stationarity를 해결하기 위해 타임 스텝 별로 정규화하고 target을 미래의 수익률로 정의한다. 정규화된 데이터는 오토인코더 모델을 학습시켜 자동 변수 추출 시스템을 구성하며 미래의 수익률을 예측하는 LSTM기반의 모델을 학습하고, 이 과정을 online 학습할 수 있는 시스템으로 구축한다.

2. 관련 연구

ARIMA모델은 시장예측 모델로서 많은 연구의 대상이 되었다[3]. 이 모델은 실제 다양한 응용에 대한 효과를 보였지만 비선형 관계를 제대로 모델링 할 수 없었으며 외생 변수를 입력으로 사용하기가 어려웠다. 이러한 문제점을 보완하기 위해 Nonlinear Autoregressive exogenous(NARX)와 관련된 모델들이 개발되었다[4]. 이 모델은 예측 시계열의 과거 데이터와 외생 변수 시계열 데이터를 사용하여 시계열을 예측할 수 있는 비선형 회귀 모델로 ARIMA모델의 단점을 어느 정도 보완을 해주었다. 하지만 장기 의존성 문제를 제대로 처리할 수 없었으며 외생 변수의 개수가 많아지게 되면 생기는 문제 또한 가지고 있었다.

한편 인공신경망을 사용하여 시장예측을 하는 연구로써, [5]는 MLP모델을 사용하여 도쿄 주식 시장의 지수를 예측하는 연구이고, [6]는 ANN과 SVM의 시장 예측 성능을 비교하여 ANN이 SVM보다 예측 능력이 더 뛰어남을 입증하였다. [7]는 DNN이 얕은 ANN보다 더 우수함을 보였다.

RNN과 LSTM은 과거의 특징을 추출하고 이를 기반으로 예측할 수 있는 모델이다. [8]은 LSTM을 사용하여 시장 수익률을 예측하였으며, [9]는 LSTM을 사용하여 시장 주식이 15분 뒤 상승할지 하락할지를 예측하여, LSTM이 MLP보다 성능이 더 좋음을 보였다.

3. LSTM 오토인코더 시장예측 시스템

3.1 전처리 시스템

예측 모델이나 분류 모델을 만들 때 학습의 효율을 높이기 위해 정규화된 데이터를 사용되며, z-변환이 자주 사용된다. z-변환은 다음 수식과 같다.

$$Z_t = \frac{X_t - \mu_X}{\sigma_X} \quad (1)$$

금융 시계열 데이터의 학습데이터와 테스트 데이터를 같이 정규화하면, 금융 데이터의 Non-stationarity 때문에 테스트 구간의 데이터 값이 제대로 정규화되어 있지 않은 문제가 발생한다. 즉 평균이 0이고 표준편차가 1이 되지 않으며 매우 큰 값이 존재하거나 매우 작은 값이 존재하므로, 테스트 데이터의 예측 실패율이 증가한다. 본 논문에서는 금융 시계열 데이터의 정규화 문제를 해결하고자 타임 스텝 정규화를 제안한다. 학습데이터 전체의 분포로 정규화를 하는 것이 아닌 LSTM 모델의 입력으로 들어가는 개별 과거 데이터의 분포로 해당 데이터를 정규화를 한다. 이렇게 하면, 모든 데이터에서 평균 0이고 표준편차가 1인 데이터가 생성된다.

3.2 오토인코더 차원축소 모델

본 논문에서는 가능한 많은 변수를 활용하여 차원 축소화 함께 피처를 잘 추출할 수 있는 오토인코더 모델을 제안한다. 이 오토인코더 모델은 many to many LSTM의 형태로 설계되었는데 이는 시장예측 모델이 LSTM기반이므로 입력 데이터의 형태를 고려하고 타임스텝간의 관계를 함께 모델링 할 수 있도록 하였다. 입력 데이터는 정규화된 과거 원가 데이터, 기술 지표, 경제 및 시장 등 다양한 변수를 갖는 데이터이며 일일 데이터를 사용한다. 이 시계열 데이터는 time×feature의 형태에서 batch×timestep ×feature 의 3D array로 변환한다. 여기서 timestep LSTM모델에 입력으로 들어갈 과거 history이다. 모델은 입력 데이터의 분포를 학습하여 생성하는 것이 목적으로 그 출력값 또한 입력 데이터가 되며 목적함수는 MSE(Mean squared error)를 사용한다.

3.3 시장예측 모델

본 논문에서 사용한 LSTM기반의 예측 모델은 Many to many LSTM의 형태로 구성하였으며 학습한 오토인코더의 인코더 출력을 입력 데이터로 batch×timestep×feature의 형태인 데이터를 사용하는데 feature의 수는 차원 축소된 데이터의 차원 수와 같다. 모델의 출력은 각 time별로 q일 후의 수익률로 수익률 변환 방법은 trend ratio를 사용하였다.

$$trendratio = \frac{price_{t+q} - price_t}{price_t} \times 100 \quad (2)$$

4. 실험 및 결과

4.1 실험 데이터

FnGuide에서 제공하는 각종 기술 지표와 경제, 시장, 환율 등 2331개의 변수를 갖는 데이터를 사용하였으며 예측하고자 하는 시장에 따라 해당하는 변수를 추가로 사용한다. 기업의 경우 주가 및 재무와 관련된 1659개의 변수가 추가되며 지수의 경우 323개의 주식지수 변수가 추가된다. 예측 시장은 삼성전자, POSCO, S-Oil, 나스닥 종합, 대한민국 KOSPI로 실험하였으며 데이터의 기간은 1997년부터 2019년까지의 데이터를 사용하였다.

4.2 축소 변수 크기별 실험

본 논문에서 실험은 추출된 변수의 개수에 따른 영향과 전반적인 시장예측 성능을 평가하게 된다. 평가 지표로는 MSE(Mean squared error), RMSE(Root Mean Square Error), MAPE(Mean absolute percentage error), 주가의 방향성에 대한 Accuracy를 기준으로 한다. MAPE는 예측 수익률에서 변환된 가격으로, 나머지는 예측된 수익률 값으로 계산하며 y는 실제 값(MAPE의 경우 실제 가격) $\hat{y}$은 예측 값(MAPE의 경우 예측 값으로 변환된 가격)이 된다.

$$MAE = \frac{1}{n}\sum|y-\hat{y}| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum(y-\hat{y})^2} \quad (4)$$

$$MAPE = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \times 100 \quad (5)$$

$$hit = \begin{cases} 1 & \text{if } y \cdot \hat{y} > 0 \\ 0 & otherwise \end{cases} \quad (6)$$

$$Accuracy = \frac{\sum hit}{n} \quad (7)$$

<표 1> 실험 파라미터.

파라미터	값
Lookback days	240
Time interval	2
Prediction day	60
Coding size	{128,256,512}

표<1>은 해당 실험에 필요한 파라미터이다. Lookback days는 t 시점으로부터 과거 t-240까지의 데이터를 사용하며 Time interval은 Lookback days의 시간 간격을 나타낸다. Prediction day는 t 시점으로부터 60일 뒤의 수익률 예측을 의미하고 Coding size는 비교실험을 위한 축소된 변수의 크기이다.

<표 2> 축소 변수 크기 별 실험 Accuracy.

	삼성전자	POSCO	S-Oil	나스닥 종합	대한민국 - KOSPI	평균
128	0.779	0.743	0.81	0.835	0.724	0.7782
256	0.817	0.761	0.821	0.856	0.722	0.7954
512*	0.809	0.782	0.818	0.842	0.746	0.7994
평균	0.8016	0.762	0.8163	0.8443	0.7306	0.791

<표 3> 축소 변수 크기 별 실험 RMSE.

	삼성전자	POSCO	S-Oil	나스닥 종합	대한민국 - KOSPI	평균
128	8.3956	8.7927	10.2672	5.1561	3.997	7.32172
256	7.6907	8.152	10.1084	4.9795	3.8547	6.95706
512*	7.6286	8.6154	10.1392	4.985	3.9067	7.05498
평균	7.904967	8.520033	10.1716	5.0402	3.919467	7.111253

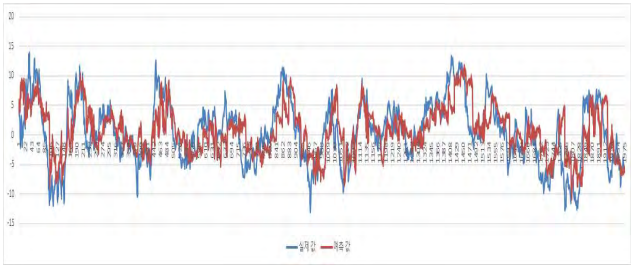
<표 4> 축소 변수 크기 별 실험 MAE.

	삼성전자	POSCO	S-Oil	나스닥 종합	대한민국 - KOSPI	평균
128	6.5677	6.8005	7.7102	3.8607	3.1942	5.6266
256	6.0151	6.3724	7.4256	3.681	3.0861	5.3160
512*	5.9852	5.7976	7.3871	3.6407	3.0641	5.1749
평균	6.1893	6.3235	7.5076	3.7274	3.1148	5.3725

<표 5> 축소 변수 크기 별 실험 MAPE.

	삼성전자	POSCO	S-Oil	나스닥 종합	대한민국 - KOSPI	평균
128	6.4208	6.8575	7.4835	3.7391	3.1699	5.5341
256	5.8455	6.4598	7.216	3.5808	3.0684	5.2341
512*	5.8149	5.8807	7.2436	3.5568	3.041	5.1074
평균	6.0270	6.3993	7.3143	3.6255	3.0931	5.2918

표(2)(3)(4)(5)에서 각 시장의 변수 크기별 결과가 조금씩 다르지만 대체로 512개의 변수로 줄였을 때가 가장 좋은 성능이 나왔다. 예측 테스트 기간 2012.01.02. ~ 2019.7.2.6에서의 Accuracy를 기준으로 평균 79%의 정확도를 보였다. 그림(1)(2)는 코스피의 예측 그래프를 보여준다. 파란 선은 실제 값을 나타내며 빨간 선은 예측 값을 나타낸다.



(그림 1) 코스피 예측 수익률 그래프.

(그림 2) 코스피 예측 가격 그래프.

5. 결론 및 향후 계획

타임 스텝 정규화를 하여 테스트 구간은 철저하게 가려진 채 모든 구간에서 정규화가 잘 이루어졌으며 또한 모델의 출력을 가격이 아닌 변환된 수익률로 함으로써 금융 시계열 데이터의 문제인 Non-stationarity를 전반적으로 잘 극복할 수 있었다.

본 논문에서는 금융과 관련된 많은 데이터를 수집하여 오토인코더 모델을 학습시켰고 이를 자동 변수 추출 모델로 데이터를 압축해서 표현하였다. LSTM 예측 모델은 긴 타임 스텝의 입력을 잘 처리하였으며. 장기 의존성 문제를 극복하고 추출된 변수의 패턴을 찾아 좋은 예측 결과를 보여주었다. 향후 자동 변수 추출 모델을 더욱더 고도화시키고 시장예측 모델에서 noise학습을 통한 trend 예측 연구를 진행할 예정이다.

6. Acknowledgement

본 연구는 2019년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임 [S2796242]

참고문헌

[1] Yann LeCun, Yoshua Bengio, “Deep learning”, Nature, 521, 436 - 444, 2015.

[2] Kartik Goyal1, “Stock Price Movement Prediction using Attention-Based Neural Network Framework”, ISSN, 2319-7064, 2017.

[3] Adebiyi A. Ariyo, “Stock Price Prediction Using the ARIMA Model”, 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge, UK, 2014.

[4] S. Chen, “Narx-based nonlinear system identification using orthogonal least squares basis hunting”, IEEE Transactions on Control Systems Technology, 16, 1, 78 - 84, 2008.

[5] T.Kimoto & K.Asakawa, “Stock Market Prediction System with Modular Neural Networks”, 1990 IJCNN International Joint Conference on Neural Networks, San Diego, CA, USA, USA, 1990.

[6] Yakup Kara, “Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange”, Expert Systems with Applications, 38, 5, 5311-5319, 2011.

[7] AH Moghaddam, “Stock market index prediction using artificial neural network”. Journal of Economics, Finance and Administrative Science 21, 41, 89-93, 2016.

[8] Kai Chen, “A LSTM-based method for stock returns prediction: A case study of China stock market”, 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 2015.

[9] David M. Q. Nelson, “Stock market's price movement prediction with LSTM neural networks”, 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017.