호가창(Limit Order Book)과 뉴스 헤드라인을 이용한 딥러닝 기반 주가 변동 예측

류의림*, 김채현**, 이기용**
숙명여자대학교 *빅데이터분석융합학과, **컴퓨터과학과
{euirimryoo, 7chaeny25, kiyonglee}@sookmyung.ac.kr

Deep Learning-based Stock Price Prediction Using Limit Order Books and News Headlines

Euirim Ryoo*, Chaehyeon Kim**, Ki Yong Lee**
*Dept. of Big Data Analysis Convergence, **Dept. of Computer Science
Sookmyung Women's University

요 약

본 논문은 어떤 기업의 주식 주문 정보를 담고 있는 호가창(limit order book)과 해당 기업과 관련된 뉴스 헤드라인을 사용하여 해당 기업의 주가 등락을 예측하는 딥러닝 기반 모델을 제안한다. 제안 모델은 호가창의 중기 변화와 단기 변화를 모두 고려하는 한편, 동기간 발생한 뉴스 헤드라인까지 예측에 고려함으로써 주가 등락 예측 정확도를 높인다. 제안 모델은 호가창의 변화의 특징을 CNN(convolutional neural network)으로 추출하고 뉴스 헤드라인을 Word2vec으로 생성된 단어 임베딩 벡터를 사용하여 나타낸 뒤, 이들 정보를 결합하여 특정 기업 주식의 다음 날 등락여부를 예측한다. NASDAQ 실데이터를 사용한 실험을 통해 제안 모델로 5개 종목(Amazon, Apple, Facebook, Google, Tesla)의 일일 주가 등락을 예측한 결과, 제안 모델은 기존 방법에 비해 정확도를 최대 17.14%, 평균 10.7% 향상시켰다.

1. 서론

주식 예측은 다양한 영역에서 계속해서 도전되어 온 문제이다. 하지만 기존의 전통적인 통계 기반 방법은 복잡하고 유동적으로 움직이는 주가의 패턴을 예측하기에 어려움이 있었다. 최근 인공지능을 활용한 연구가 활발히 수행되면서 머신러닝 및 딥러닝 기법으로 주식 가격을 예측하려는 시도가 다양하게 이루어져 왔다[1]. 지금까지의 딥러닝 기반 주가 예측 방법들은 주로 과거의 주식 가격 혹은 거래량을 사용하여 주가의 등락을 예측한다. 하지만 최근 들어 주식의 매매 주문 정보를 담고 있는 호가창(limit order book)을 이용하여 주가를 예측하려는 연구가 시도되고 있다[2][3]. 호가창은 주식 매수 및 매도 주문의 호가와 주문량 등 주문에 대한 상세한 정보를 담고 있기 때문에 보다 다차원적으로 주가를 예측할 수 있다는 장점이 있다.

본 논문에서는 어떤 기업의 호가창 정보뿐만 아니라 해당 기업에 대한 뉴스의 헤드라인까지 사용하여 해당 기업의 주가 등락을 예측하는 딥러닝 모델을 제안한다. 호가창의 최근 변화만을 보고 주가 등락을 예측하는 기존 연구와 달리 제안 모델은 호가창의 중기 변화와 단기 변화를 모두 고려하여 주가 등락을 예측한다. 호가창의 중기 변화는 주가의 전반적인 추세를 나타내고 호가창의 단기 변화는 주가의 바로 직전 추세를 나타내기 때문에, 이들을 모두 활용하면 예측 정확도를 더욱 높일 수 있다. 또한 주식 시장의 특성상 수치 데이터로만 주가를 예측하기 어렵다. 따라서 제안 모델은 같은 기간에 발생한 해당 기업에 대한 뉴스 헤드라인의 의미까지 예측에 고려함으로써 예측 정확도를 더욱 높인다. 본 논문의 제안 모델은 호가창의 변화에 대한 특징을 CNN(convolutional neural network)으로 추출하고 뉴스 헤드라인의 의미를 Word2vec으로 생성된 단어 임베딩 벡터를 사용하여 표현한 뒤, 이들 정보를 결합하여 특정 기업 주식의 다음 날 등락(상승, 하락, 유지)을 예측한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 간략히 정리하고, 3장에서는 주가 예측에 사용된 데이터를 설명한다. 4장에서는 제안하는 예측 모델을 상세히 설명하고, 5장에서 성능평가 결과를 제시한다. 마지막으로 6장에서는 결론을 맺는다.

2. 관련 연구

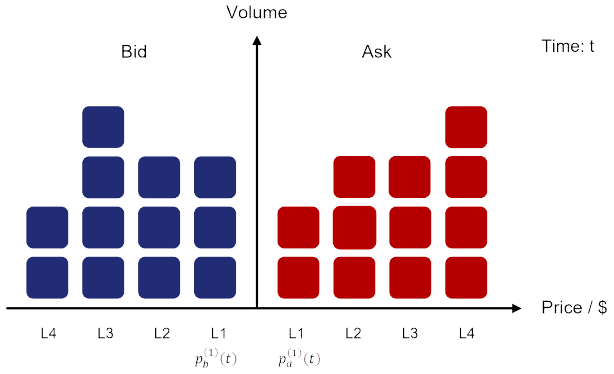
미래 주식 가격의 등락을 예측하는 것은 투자자들에게 매우 중요한 문제이기 때문에 주가 예측은 통계, 수학, 경제 등의 영역에서 활발히 연구되어왔다. 하지만 주식 가격은 매우 비선형적이고 유동적으로 움직이기 때문에 이러한 전통적인 방법들로 예측하기엔 한계가 있다. 따라서 최근 인공지능을 이용한 주가 예측 연구가 크게 인기를 끌고 있다. 최근에는 서론에서 언급한 바와 같이 호가창을 사용하여 주가를 예측하려는 연구가 진행되고 있다. [2]는 LSTM(long short-term memory)로 호가창 변화의 특징을 추출하여 주가를 예측하며, [3]은 CNN으로 개별 호가창 데이터의 특징을 추출한 뒤, 이들의 변화를 LSTM으로 학습하여 주가를 예측한다. 하지만 이들 모두 호가창의 최근 중기 변화만을 보고 주가 등락을 예측한다는 한계가 있다.

주식은 기술적인 지표와 정형적인 수치에 반응하기도 하지만 회사의 최근 정보를 포함하는 뉴스에도 영향을 받는다. 따라서 온라인 뉴스와 소셜 미디어의 텍스트를 활용한 주가 예측 연구들이 진행된 바가 있다. [4]는 과거의 주식 가격과 온라인 뉴스에 대한 워드 임베딩으로 주가의 움직임을 예측한 연구이다. [4]도 주가 관련 수치 데이터와 뉴스 텍스트 데이터를 모두 고려했다는 점에서 본 연구와 공통점이 있다. 하지만 본 논문은 호가창 데이터를 활용한다는 점에서 차별성이 있다.

3. 주가 예측에 사용된 데이터

3.1 호가창 데이터

호가창은 주식 거래를 위해서 제출한 매도(ask) 주문과 매수(bid) 주문의 수량을 호가별로 기록한 정보이다. 어떤 종목에 대한 호가창은 각 시간 t 시점에서의 매도 주문호가, 매도 주문량과 매수 주문호가, 매수 주문량을 포함하고 있다. (그림 1)은 어떤 시간 t에서의 호가창 데이터 일부를 나타낸다. 주문은 제출된 호가에 따라 여러 개의 레벨로 나뉘는데, (그림 1)에서 L1은 가장 낮은 가격의 레벨(레벨 1)을 뜻하며, L4는 가장 높은 가격의 레벨(레벨 4)을 뜻한다. 주어진 어떤 종목에 대해, 본 논문에서는 시간 t에서 레벨 l의 매도 주문호가와 매도 주문량을 각각 $p_a^{(l)}(t)$과 $v_a^{(l)}(t)$로 표시한다. 이와 유사하게 시간 t에서 레벨 l의 매수 주문호가와 매수 주문량을 각각 $p_b^{(l)}(t)$과 $v_b^{(l)}(t)$로 표시한다.



(그림 1) 호가창의 예

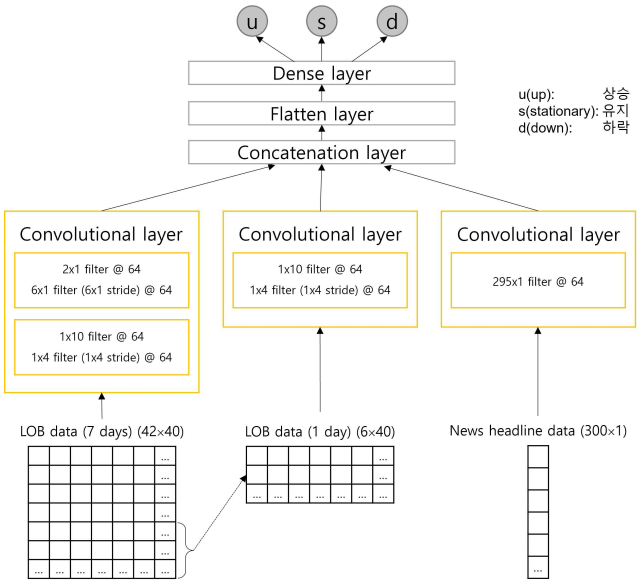
일반적으로 나스닥(NASDAQ)과 같은 주식 시장의 호가창 데이터는 나노초(nanosecond) 단위로 매우 빈번하게 변화가 기록된다. 하지만 본 논문에서는 그 정도까지의 상세한 정보는 필요하지 않다고 판단하여 1시간 단위로 호가창 데이터를 묶어 그들의 평균값을 주가 등락 예측에 사용하였다. 또한 레벨이 너무 높은 호가의 주문은 실제 거래로 연결되지 않기 때문에 레벨 1부터 10까지의 주문만 고려하였다.

3.2 뉴스 헤드라인

본 논문에서는 기업에 대한 뉴스를 검색하여 검색된 뉴스의 헤드라인을 모델 훈련에 사용한다. 이를 위해 본 논문에서는 뉴스 사이트에서 기업 뉴스를 검색하고, 검색된 뉴스의 헤드라인을 고정된 길이의 임베딩 벡터로 변환하였다. 검색된 뉴스의 헤드라인을 임베딩 벡터로 변환할 때는 뉴스 헤드라인을 단어들로 나누고 불용어를 제거한 뒤, 남은 단어들 각각을 Word2vec을 사용하여 임베딩 벡터로 변환하였다. 이후 이 임베딩 벡터들의 평균 벡터를 해당 뉴스 헤드라인을 나타내는 최종 임베딩 벡터로 사용하였다.

4. 제안 딥러닝 모델

본 논문에서 제안하는 딥러닝 기반 주가 등락 예측 모델의 구조는 (그림 2)와 같다. 제안 모델은 호가창의 중기 및 단기 추세를 모두 고려하기 위해 최근 7일 및 최근 1일의 호가창 데이터를 각각 입력으로 받는다. 또한 해당 종목에 대한 뉴스 정보를 반영하기 위해 바로 직전 1일의 뉴스 헤드라인 데이터를 입력으로 받는다. 제안 모델은 호가창의 중기, 단기 변화 및 뉴스의 특징을 각각 추출한 뒤, 이들을 결합하여 다음 날의 주가 변동(상승, 하락, 유지)을 예측한다.



(그림 2) 제안 주가 등락 예측 모델

4.1 호가창 변화 특징 추출을 위한 합성곱층

합성곱층(convolutional layer)은 입력 데이터를 스캔하면서 특징을 추출하는 층이다. 본 논문에서는 호가창의 변화의 특징을 합성곱층으로 추출하기 위해 호가창 데이터를 다음과 같이 표현하였다. 3.1절에서 설명한 바와 같이 본 논문에서는 각 날짜의 호가창 데이터를 1시간 단위로 묶어 그들의 평균값으로 매일 6개(6시간)의 호가창 데이터를 생성하였다. 각 시간 t의 호가창 데이터는 아래와 같이 레벨 1부터 10까지 각 레벨의 매도 주문호가($p_a^{(i)}(t)$), 매도 주문량($v_a^{(i)}(t)$), 매수 주문호가($p_b^{(i)}(t)$), 매수 주문량($v_b^{(i)}(t)$)을 나타내는 총 40개의 수치로 구성된다.

$$\left\{p_a^{(i)}(t), v_a^{(i)}(t), p_b^{(i)}(t), v_b^{(i)}(t)\right\}_{i=1}^{n=10}$$

따라서 1일치 호가창 데이터는 6(시간) × 40(개)의 크기를 가진다. 호가창의 매도 주문과 매수 주문 정보는 주식의 미래 가격을 결정하는데 매우 중요한 정보를 내포하고 있다. 본 논문에서는 6 × 40 크기를 가지는 1일치 호가창의 변화의 특징을 포착하기 위해 우선 첫 번째 합성곱층에서 크기가 1 × 4이고 스트라이드가 1 × 4인 필터를 64개 사용하여 시간별로 각 레벨의 특징을 추출하였다. 이후 이 결과에 다시 1 × 10 필터를 사용하여 시간별로 모든 레벨의 특징을 통합한 6 × 1 × 64 크기의 출력값을 얻는다.

반면 7일치 호가창의 변화의 특징을 포착하기 위해서는 먼저 7일치 호가창 데이터에 대해 1일치 호가창과 동일한 합성곱층을 적용하여 그 결과로 42 × 1 × 64 크기의 출력값을 얻는다. 이후 이 결과에 크기와 스트라이드가 6 × 1인 필터를 사용하여 날짜별로 통합된 특징을 추출하고, 마지막으로 크기가 2 × 1인 필터를 추가로 사용하여 최종적으로 1일치 호가창에 대한 합성곱층의 출력 결과와 크기가 동일한 6 × 1 × 64 크기의 출력값을 얻는다.

4.2 뉴스 헤드라인 특징 추출을 위한 합성곱층

3.2절에서 설명한 바와 같이 본 논문에서는 뉴스 헤드라인을 구성하는 각 단어를 Word2vec을 사용하여 임베딩 벡터로 변환한 뒤, 이들의 평균 벡터를 뉴스 헤드라인을 나타내는 임베딩 벡터로 사용한다. 본 논문에서는 300차원의 워드 임베딩 벡터를 사용하였으며, 따라서 뉴스 헤드라인에 대한 임베딩 벡터는 300 × 1 크기를 가진다.

본 논문에서는 뉴스 헤드라인에 대한 임베딩 벡터에 크기가 295 × 1인 필터를 64개 사용하는 합성곱층을 적용하여 특징을 추출하였다. 이를 통해 호가창에 대한 합성곱층의 출력 결과와 크기가 동일한 6 × 1 × 64 크기의 출력값을 얻는다.

4.3 최종 예측

앞서 설명한 합성곱층을 통해 7일치 호가창 데이터, 1일치 호가창 데이터, 뉴스 헤드라인 데이터의 특징이 각각 6 × 1 × 64 크기의 결과로 추출되면, 이들을 결합층(concatenation layer)을 통해 결합하여 18 × 1 × 64 크기의 데이터로 만든다. 이후 이 데이터를 평탄화층(flatten layer)을 통해 1152 × 1 크기의 벡터로 만든다. 마지막으로 밀집층(dense layer)은 이를 입력으로 받아 상승(u), 유지(s), 하락(d) 각각에 대한 확률값을 출력한다. 이를 위해 밀집층은 출력 함수로 소프트맥스(softmax)를 사용한다.

제안 모델의 학습을 위해 손실함수로는 범주형 크로스 엔트로피(categorical cross entropy)를 사용하였으며, 과적합(overfitting)을 피하기 위해 학습 과정에서 3개의 합성곱층 마지막 층 각각에 드롭아웃(dropout)을 적용하였다.

5. 성능 평가

5.1 실험 환경

실험에서는 제안 모델을 Python 3.8과 Tensorflow를 사용하여 구현하였으며, 실험은 Intel i7-6800K 3.40GHz CPU, 32GB 메모리가 장착된 Windows 10 환경의 PC에서 수행하였다. 모델의 성

비교 모델	Amazon	Apple	Facebook	Google	Tesla
제안 모델 (중기 LOB + 단기 LOB + 뉴스)	**70.22**	**64.90**	**64.43**	**75.17**	**64.40**
중기 LOB + 뉴스	64.92	64.46	59.98	69.75	60.80
중기 LOB + 단기 LOB	63.72	61.58	59.18	64.45	57.50
중기 LOB	59.82	58.35	53.90	58.03	55.08

<표 1> 제안 모델의 성능 (정확도) 평가 결과 (단위: %)

능은 10-겹 교차검증(10-fold cross validation)으로 측정하였으며, 성능 척도로는 상승, 유지, 하락 각각에 대한 예측 정확도를 구하여 이들의 평균 정확도를 사용하였다.

5.2 실험 데이터

실험에서는 나스닥(NASDAQ) 증권 거래소로부터 대표적 종목인 Amazon, Apple, Facebook, Google, Tesla 5개 종목의 2019년 7월 22일부터 2020년 7월 21일까지 1년 동안의 실제 호가창 데이터를 제공받아 활용하였다. 또한 나스닥 장 거래시간이 9:00부터 16:30인 점을 고려하여 매일 9:15부터 16:15까지 기록된 거래 데이터만 이용함으로써 장외 거래의 영향을 최소화했다.

뉴스 헤드라인 데이터로는 앞서 선정한 5개 종목(Amazon, Apple, Facebook, Google, Tesla)에 대한 2019년 7월 22일부터 2020년 7월 21일까지의 뉴스를 CNBC, Guardian, Reuters로부터 수집하여 이들의 뉴스 헤드라인을 훈련 데이터로 사용하였다.

5.3 실험 결과

<표 1>은 각 종목에 대한 제안 모델의 일일 주가 등락 예측 정확도를 측정한 결과이다. 제안 모델은 7일 간의 중기 호가창 정보, 1일 간의 단기 호가창 정보, 뉴스 헤드라인 정보를 모두 활용하여 주가 등락을 예측한다. 중기 호가창 정보만을 사용하는 기존 방법(중기 LOB)은 최소 53.90%, 최대 59.82%의 예측 정확도를 보였다. 하지만 제안 모델은 종목에 따라 최소 64.40%, 최대 75.17%의 예측 정확도를 보인다. 따라서 단기 호가창 정보와 뉴스 정보까지 사용하는 제안 방법이 보다 효과적임을 알 수 있다.

본 실험에서는 중기 호가창 정보에 단기 호가창 정보만을 추가로 사용하는 방법(중기 LOB + 단기 LOB)과 중기 호가창 정보에 뉴스 정보만을 추가로 사용하는 방법(중기 LOB + 뉴스)의 성능을 추가로 비교하였다. <표 1>에서 볼 수 있듯이 단기 호가창 정보를 추가로 사용하는 방법은 기존 방법(중기 LOB) 대비 정확도를 평균 약 4.25% 향상시키므로 유의미한 전략임을 알 수 있다. 또한 기존 연구에 뉴스 정보를 추가로 사용하는 방법 역시 기존 방법(중기 LOB) 대비 정확도를 평균 약 6.95% 향상시키므로 마찬가지로 효과적인 전략임을 확인할 수 있었다. 제안 방법은 단기 호가창 정보와 뉴스 정보 모두를 추가 활용함으로써 주가 예측 정확도를 최대 17.14%, 최소 6.55%, 평균 10.7% 향상시켰으며 비교 모델 중 가장 좋은 성능을 보였다.

6. 결론

본 논문에서는 중기 호가창 정보 외에도 단기 호가창 정보와 동기간 뉴스 헤드라인을 추가로 활용하여 정확도를 높이는 딥러닝 기반 주가 등락 예측 모델을 제안하였다. 제안 방법은 호가창의 중단기 정보와 뉴스 헤드라인의 특징을 각각 합성곱층으로 추출하고 이들의 결과를 종합하여 주가 등락을 예측한다. 실험 결과 제안 방법은 중기 호가창 정보만을 사용하는 기존 방법에 비해 예측 정확도를 최대 17.14%, 평균 10.7% 향상시킴을 확인하였다.

Acknowledgement

이 성과는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2021R1A2C1012543).

참고문헌

[1] W. Jiang, “Applications of deep learning in stock market prediction: Recent progress,” *Expert Systems with Applications*, Volume 184, 2021.

[2] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Using deep learning to detect price change indications in financial markets,” *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2511-2515, 2017.

[3] Z. Zhang, S. Zohren and S. Roberts, “DeepLOB: Deep Convolutional Neural Net- works for Limit Order Books,” *IEEE Transactions on Signal Processing*, vol. 67, no. 11, pp. 3001-3012, 2019.

[4] Y. Peng and H. Jiang, “Leverage financial news to predict stock price movements using word embeddings and deep neural networks,” *Proc. NAACL-HLT. San Diego, CA, USA: Association for Computational Linguistics*, pp. 374 - 379, 2016.