뉴스기사와 머신러닝을 활용한 암호화폐 가격 변화 예측에 관한 연구

최욱철[1], 구자환[1], 김응모[1]
[1]성균관대학교 소프트웨어융합대학
kikyo3536@g.skku.edu, jhkoo@skku.edu, ukim@skku.edu

A Study on Prediction of Cryptocurrency Price using News Articles and Machine learning

Uk-Cheol Choe[1], Jahwan Koo[1], Ungmo Kim[1]
[1]College of Software, Sungkyunkwan University

요 약

주식과 암호화폐 거래는 매매방식에 있어서 유사한 점이 있지만 기업의 사업분야, 자본금, 순이익 등의 경영현황과 미래가치에 영향을 많이 받는 주식과는 다르게 암호화폐는 실물 실체가 없으며 탈중앙화, 전산화된 데이터를 기반으로 하며 심리적인 요소가 크게 작용하여 단기적인 변동이 클 수 있다. 본 연구에서는 이러한 암호화폐 거래의 특성을 활용하여 특정 암호화폐에 관련된 뉴스기사들을 수집하고 그 암호화폐의 가격 변화 데이터와 연관되어 가격예측 딥러닝 모델을 생성하고 해당 암호화폐에 대한 신규 뉴스기사가 발생되었을 때 이를 이용하여 매수, 매도, 관망 등과 같은 매매 정보를 예측 적용할 수 있게 하였다. 첫째, 뉴스 기사에서 언급한 암호화폐를 매수, 매도, 관망 중 어느 편이 좋을 것인지 추천하는 알고리즘을 구현하였고, 둘째, 매수 이후 매매 차익을 위한 매도 시점이나 매도 이후 저가매수에 유리한 시점을 제안하는 알고리즘을 구현하였다. 또한, 실시간 뉴스기사 수집 및 예측한 매매 판단에 따라 매매 자동화 시스템을 구현하여 수익률을 직접 확인함으로써 그 유효성을 검증하였다.

1. 서론

국내외에서 주식에 대한 대중의 투자와 관심은 꾸준히 증가해왔으며, 이러한 양상은 주식과 유사한 성격을 상당히 가진 암호화폐로 이어졌다. 주식과 암호화폐는 매매방식에 있어 비슷한 점이 많지만, 큰 차이점들도 존재한다. 주식은 실제 존재하는 회사를 기반으로 하기에, 그 회사가 속한 사업 분야, 자본금, 순이익 등의 현황과 미래가치에 영향을 받는다. 그러나, 암호화폐는 실물 실체가 없으며 탈중앙화, 전산화된 데이터를 기반으로 한다. 또한 주식에 비해 현저하게 역사가 짧고, 아직 회의적인 의견과 논란이 많다. 제도와 규제도 명확하지 않은 상황이다. 즉, 그 회사의 실적과 가치 등이 반영되는 주식에 비해, 암호화폐는 심리적인 요소가 크게 작용할 수 있으며 단기적인 변동이 클 수 있다. 실제로, 소셜 미디어나 인터넷상의 유행에 따라 암호화폐가 급등 혹은 급락한 사례는 많다.

본 연구에서는 이러한 암호화폐의 특성을 잘 활용하여 특정 암호화폐에 관련된 지속적으로 생성되고 신뢰할 수 있는 정보를 포함하는 뉴스기사 데이터들을 수집하고 그 암호화폐의 가격 변화 데이터와 연관지어 딥러닝을 수행하는 가격예측 모델을 생성하고 분석하는 것을 목표로 하였다. 과거의 일정한 기간을 정해, 그 기간내 작성된 뉴스 기사들을 수집한다. 실제 암호화폐의 매매 가격 변동을 기록한 데이터는 그 기간을 전후로 하여 수집한다. 이 데이터들을 분석하여, 매매와 관련된 정보를 판단하여 이용자에게 매수/관망/매도 여부 및 그에 따른 미래의 매매에 대한 제안을 생성하는 것이다. 각 뉴스 기사들의 내용 텍스트들을 입력으로, 매매 제안 정보를 출력하는 딥러닝을 실행한다. 결과적으로, 현재 특정 암호화폐를 언급하는 뉴스 기사가 작성된 것을 실시간으로 감지하였다면, 그 기사의 내용을 앞서 생성한 딥러닝 모델에 주입하여 매매 정보를 예측, 이용자에게 제안할 수 있는 것이다. 여러 조건에서의 학습 정확도를 분석하며 적절히 학습할 수 있도록 하며, 실시간 뉴스 기사 수집과, 그것을 딥러닝 모델에

주입해 예측한 매매 판단에 따른 매매를 자동화하여 그 실적(수익률)을 확인함으로써 유효성을 검증하였다.

2. 선행연구

국내 주식 몇 개 종목을 정한 후, 각 종목들을 언급한 네이버 뉴스와 네이버 종목토론실의 글들을 수집하고, 그것들의 글 제목과 내용 텍스트, 찬성 반대 수 등을 데이터화한 후, 뉴스 기사들과 종목토론실 글들이 가지는 특성을 비교하고 딥러닝을 통해 주가 등락을 예측하는 연구[1]가 있었다. 또한, 빅데이터와 텍스트 마이닝의 중요성에 주목, 주식 커뮤니티에 작성된 게시글 데이터들을 수집하고 텍스트 마이닝 기법을 활용하여 7종의 감정지표를 산출 후 딥러닝을 실행해 KOSPI 200 선물 지수의 등락을 예측하는 연구[2]도 있었다. 암호화폐에 대해 기계학습을 수행하되 암호화폐 고유의 데이터인 온체인 데이터 기반 지표와 여러 경제 지표들의 활용 가능성을 실증, 딥러닝 기법 중 GRU 모형을 활용해 비트코인 가격 등락 예측 모델을 구축한 연구[3]도 있었다. 본 연구에서는 특정 암호화폐에 대해 여러 분야, 관점에서 작성한 대량의 텍스트 데이터을 쉽게 수집할 수 있고 그 정확성을 상당히 보장할 수 있는 뉴스 기사를 활용하였다.

3. 연구 설계

본 연구에서는, 일정 기간내 작성된 특정 암호화폐를 언급하는 뉴스 기사들과, 그 전후의 일정 기간 동안 그 암호화폐의 매매 가격 변동을 기록한 데이터를 수집한다. 그 데이터들을 이용하여 구현할 기능들은 다음과 같다. 첫째, 뉴스 기사에서 언급한 암호화폐를 매수, 매도, 관망 중 어느 편이 좋을 것인지 제안을 하는 것이다. 둘째, 현재 매수 혹은 매도하는 것이 좋을 것으로 판단했을 경우에 한해, 매수 이후 매매 차익을 위한 매도 시점이나 매도 이후 저가매수에 유리한 시점을 제안한다. 이를 위해 일정 기간동안 작성된, 특정 암호화폐와 관련된 뉴스들과, 그 기간 전후로 하여 그 암호화폐의 시세 변화의 기록을 수집하고 데이터화 할 필요가 있다. 대상 암호화폐는 시가총액이 크고 오랜 기간 관심을 받았으며 검색에 용이한 '비트코인', '이더리움', '에이다'로 선정하였다.

암호화폐의 과거 실제 시세 변화 데이터를 얻기 위해서, 국내에서 가장 거래량이 많은 암호화폐 거래소 중 하나인 '업비트(Upbit)'에서 제공하는 Open API를 사용하였다. 특정 시각 특정 시간 단위의 차트에서의 시가, 고가, 저가, 종가, 누적 거래 금액, 누적 거래량 등의 데이터를 요청할 수 있으므로, 필요한 과거 데이터 수집이 가능했다.

또한 각 암호화폐를 언급하는 뉴스를 수집하는 것은 국내 최대규모 포털 사이트인 '네이버(Naver)'의 뉴스 검색 기능을 활용하였다. 뉴스만을 검색되게 할 수 있으며, 매 검색마다 임의의 날짜 기간을 지정할 수 있기 때문이다. 파이썬의 BeautifulSoup, Selenium 등의 패키지들을 활용하여 URL 지정, 페이지 전환, 기사 링크 접속, 기사 작성 일자 및 제목, 내용 수집 등 순차적인 검색 과정을 자동화하였다.

암호화폐 시세 데이터와 뉴스 기사 데이터에 딥러닝을 적용해 매매와 관련된 판단을 만드는 데에 도입한 개념은 샘플링(sampling)이다. 시세 데이터의 분석에서 효율성을 높이기 위해 적용하였다. 매매 이후 가격 차트의 파동이 최종적으로 같은 등락률을 가지는 경우들을 비교하였을 때, 매매 이후 빠르게 등락하였거나 오른 가격이나 내린 가격 중 하나를 꾸준히 유지할수록 매수나 매도의 유불리 판단에 큰 의미를 갖게 된다. 이를 반영하기 위해 등비급수의 성질을 이용하였다. 뉴스 기사 작성 1분 후, 2분 후, 4분 후, 8분 후, 16분 후, ... 2^n분 후의 매매가를 조회해 등락률을 계산하는 것이다. 또한, 거래량이 클수록 더 유의미한 등락이므로 거래량도 고려해야 하는데, 암호화폐는 채굴로 인해 전체 발행량이 계속 늘어나 시가총액을 참조해야 하므로 "시가총액 회전율"을 계산한다. 따라서 n+1개의

$$\text{등락률}_k * \text{시가총액회전율}_k \quad (1)$$

을 구한 다음 평균을 구한다. 등락률은

$$2^{k-1}\text{분간 가격 증감} / \text{현재 가격} \quad (2)$$

으로, 시가총액 회전율은

$$2^{k-1}\text{분간 거래대금 합} / 2^{k-1}\text{분 후 시가총액} \quad (3)$$

으로 한다. 구한 평균값이 양수이며 클수록 매수가, 음수이며 작을수록 매도가 적절한 것으로 분석한다. 뉴스 기사들마다 이 값이 다를 것인데, 이 값이 양수인 것들과 음수인 것들로 분류한 후 각각 절대값이 가장 큰 값으로 모든 값들을 나누어, 결과값을 -1~+1사이로 정규화한다. +1에 가까울수록 강한 매수-매수-약한 매수, -1에 가까울수록 강한 매도-매도-약한 매도, 0 내외라면 관망으로 하도록 일정

한 범위들로 세분화한다. 그 후 뉴스 기사 내용을 입력값으로 하여 이 단계적 매매 제안을 예측하는 딥러닝을 실행한다. 뉴스 기사 내용 텍스트를 토큰화한 다음 임베딩과 LSTM(Long Short-Term Memory)기법을 적용하여, 각 기사 내용에 따라 매수, 관망, 매도 단계를 예측하는 딥러닝을 하도록 하였다. 학습셋과 테스트셋으로 나누어 정확도를 확인하였고 샘플의 개수를 11, 14, 16, 18로 바꿔가며 진행하였다.

우리가 제안한 방법은 원리가 간단하지만 여러 수치들을 대입하며 결과를 검토하기에 적절한 방법들도 추가 사용하였다. 뉴스 등록 시점을 기준으로 하여 일정 기간동안, 기준 시점과 비교한 등락률의 평균(차트의 이동평균선과 유사)을 구하는 방법과, 기준 시점보다 상승한 시간 대 하락한 시간의 비율을 구하는 방법이다. 일정 기간동안이라 함은 1분 후, 2분 후, 3분, ... , n분 후와 같이 n개의 시장가를 취한다는 뜻이다. 여기서 시간 단위를 1분 간격 뿐만 아니라 5분, 15분, 1시간, 6시간, 1일 간격으로 하고, 그때마다 n 값도 50, 100, 200으로 두고 각각 결과를 내어 다양한 결과를 비교할 수 있도록 한다. 두 방법 중 전자는 뉴스 등록 시점 기준 1분 후부터 n분 후까지 n개의

$$\frac{x\text{분 후의 매매가} - \text{뉴스등록시점 매매가}}{\text{뉴스등록시점 매매가}} \quad (4)$$

의 평균을 낸다. 그 다음 등비급수 방안에서와 같은 과정으로 뉴스 기사마다 다른 결과값들을 정규화한다. 후자는 뉴스 등록 시점 기준 1분 후부터 n분 후까지 n개의

$$x\text{분 후의 매매가} - \text{뉴스등록시점 매매가} \quad (5)$$

중 0 이상인 개수를 r, 0 미만인 개수를 f라 한다. 만약 r이 f보다 크거나 같다면 수식 (6)에, 그렇지 않다면 수식 (7)에 따라 계산한다.

$$2^{*}\frac{r}{r+f}-1 \quad (6) \qquad -2^{*}\frac{f}{r+f}+1 \quad (7)$$

이 값은 상승한 시간이 길수록 +1에, 하락한 시간이 길수록 -1에, 비슷하다면 0에 가깝게 된다. 등비급수 방법에서와 동일한 과정으로 매수, 관망, 매도 단계를 예측하는 딥러닝을 실행하였다. 최종적인 매매 제안 판단은, 등비급수 방법에서는 샘플의 개수 n=11일 때, 등락률 평균 방법과 상승 대 하락 시간 비율 방법에서는 양쪽 모두 n=50, 100, 200일 때 전부 6시간 단위로 하는 것이 대체로 정확도가 높다고 분석하였으므로, 그 조건들(총 7가지) 하에서의 매매 제안 결과의 평균으로 하였다.

다음으로, 매수나 매도를 제안하는 것으로 판단된 경우, 추후 매매차익 실현이나 저가매수에 적절한 시점을 제안하는 기능이다. 조사한 선행연구의 일부를 비롯하여 기존에 주식이나 암호화폐의 가격 차트를 분석하는 데에 널리 쓰이고 있는, '기술적 보조지표'들 중 스토캐스틱, MACD, RSI의 3종을 활용하였다. 파라미터는 (15, 5, 3), (12, 26, 9), 14로 하고, 1시간 단위로 계산하였다. 뉴스 기사마다, 만약 매수/매도하는 것이 좋겠다는 제안이 나왔다면, 뉴스가 등록된 시각 이후로 3종의 보조 지표가 모두 매도/매수 신호를 보내는 가장 가까운 미래 시점을 찾는다. 딥러닝에서 예측할 클래스이므로 적절히 카테고리화 하는 것이 좋다고 판단하여, 앞의 등비급수 방법과 같은 논리로 2^n분 단위로 시간대를 나누어 어느 시간대에 포함되는가로 하였다. 즉, 각 시간대 길이는 직전 시간대의 2배가 되며, 마지막 시간대는 '2^n분 초과'가 된다. 대부분의 데이터셋에서 2^{14}분을 초과하는 경우는 없는 것으로 밝혀져 시간대의 개수는 n=14로 하였다. 앞의 매수/관망/매도 제안 기능과 동일한 방법을 적용하여 딥러닝을 시행하였다.

양쪽 모두 epochs는 20으로, batch size는 100으로 임의로 정하였다. 또한 과적합을 막고 정확도 상승을 위해 여러 다양한 조건을 조성했다. 딥러닝 과정에 컨볼루션 신경망(CNN), 맥스 풀링, 드롭아웃 기법을 추가로 적용하였고, epoch의 테스트 결과에 따른 선택적 모델 갱신 및 조기 종료의 도입, 뉴스 기사를 수집할 기간을 1개월에서 3개월, 6개월로 확대, 비트코인 외 '이더리움', '에이다' 등을 시도하였다.

마지막으로, 비트코인과 관련된 뉴스를 실시간으로 수집해, 만들어둔 딥러닝 모델에 주입하고, 두 기능의 예측에 따라 비트코인을 자동으로 매매하는 프로그램을 구현해 수익률을 확인하였다.

4. 실행 결과 분석

'비트코인' 관련 기사 1개월, 3개월, 6개월 분량을 수집하였고 '이더리움', '에이다' 관련 기사 3개월 분량을 수집하였다. 2021년 1월 1일부터 1월, 3월, 6월 말일까지 작성된 기사를 대상으로 하였다. 비트코인 기사 1개월 분량은 175개, 3개월 분량은 543개, 6개월 분량은 1,198개가 수집되었으며, 이더리움과 에이다 기사 3개월 분량은 각각 207개와 64개가 수집되

었다. 과거 시세 및 시가총액 데이터는 2020년 12월 1일부터 2022년 1월 31일까지 수집한 것을 이용하였다. 즉, 612,166개의 1분봉 시세, 10,230개의 60분봉 시세, 426개의 일일 시가총액 데이터를 이용하였다.

비트코인 기사 1개월 분량에 임베딩과 LSTM만을 적용하였을 때에 비해, 앞서 언급하였듯 CNN, 조기 종료 등 다양한 단계 및 조건을 딥러닝 과정에 추가하거나 개량하였을 때, 전체적으로 정확도가 유의미하게 상승하였으며 학습 속도가 현저히 빨라졌다. 그 상태에서 비트코인 기사를 3개월 분량으로 확대하였을 때에도 학습 속도가 대체로 소폭 상승하였다. 그러나, 다시 6개월 분량으로 하였을 때에는 큰 차이가 없었다. 또한, 이더리움이나 에이다 관련 기사 3개월 분량으로 각각 학습한 결과도 비슷했다. 따라서 비트코인 관련 기사 3개월 분량으로 딥러닝을 실행하는 것이 가장 적절한 것으로 분석하였으며, 그 결과는 <표 1>에서부터 <표 4>까지와 같다. 정확도 값은 학습이 완료되었을 때 테스트셋의 예측 정확도이다. n은 시간 샘플의 개수, 시간 단위는 샘플 간의 시간 간격이다. 분석 기간이 길어져 과도한 자료 수집과 계산이 요구되는 조건은 제외하였다.

<표 1> 등비수열 방법을 적용했을 때의 정확도

	n=11	n=14	n=16	n=18
정확도	0.8834	0.6810	0.6319	0.4540

<표 2> 등락률 평균 방법을 적용했을 때의 정확도

	1분	5분	15분	1시간	6시간	1일
n=50	0.58	0.42	0.41	0.29	0.28	0.61
n=100	0.50	0.45	0.33	0.33	0.58	
n=200	0.42	0.34	0.29	0.44	0.63	

<표 3> 상승한 시각 대 하락한 시각 비율 방법을 적용했을 때의 정확도

	1분	5분	15분	1시간	6시간	1일
n=50	0.28	0.34	0.29	0.36	0.36	0.45
n=100	0.33	0.28	0.33	0.36	0.47	
n=200	0.29	0.30	0.36	0.29	0.45	

<표 4> 추후 매도/매수 시점 제안 기능의 정확도

	n=11	n=14	n=16	n=18
정확도	0.2454	0.5031	0.5031	0.5031

딥러닝이 예측하여야 하는, 데이터셋의 뉴스 기사 543개에 대한 실제 매수/관망/매도 분포는 [36, 216, 113, 125, 47, 4, 2] 와 같으며, 이는 [강한 매수, 매수, 약한 매수, 관망, 약한 매도, 매도, 강한 매도] 순으로 출력하였다. 관망을 예측한 기사 125개를 제외한, 나머지 418개 기사의 미래 매도/매수 시점 예측 분포는 [0, 0, 0, 0, 10, 14, 24, 45, 93, 89, 90, 49, 4, 0, 0]과 같으며, 이는 순서대로 [4분 미만, 4분 이상 8분 미만, 8분 이상 16분 미만, ... , 32,768분 이상]를 의미한다.

마지막으로, 비트코인 관련 기사를 실시간으로 수집해, 비트코인 기사 3개월 분량으로 학습한 딥러닝 모델에 주입하고, 그 예측(현 시점 매수/관망/매도, 그에 따른 미래의 매도/매수)에 따라 자동으로 매매하는 프로그램을 작성하였다. 예측한 매매 단계에 따라 한화와 비트코인 잔고에서 소정의 비율을 매매하도록 하여, 3월 25일부터 3월 31일까지 1주일간 실행하였다. 수익률은 +0.78%이며, 같은 기간 비트코인은 4.62% 상승하였다.

5. 결론

일정 기간내에 작성된 암호화폐와 관련된 뉴스 기사들과, 그 전후의 암호화폐 시세 변화 데이터를 수집하여, 토큰화하고 LSTM과 CNN, max pooling 등의 기법들을 통해 딥러닝을 실행하여 기사 내용에서 가격 변화를 예측하도록 하였다. 학습 정확도를 상당히 높이는 데에 성공하였으나, 한계가 있었다. 더 좋은 결과를 위해서는, 딥러닝 내적으로는 다른 기법도 도입하거나 레이어를 추가시키는 등의 방법이, 외적으로는 시세 데이터의 수집 및 분석에서 경제학적으로 더 유의미한 분석법이나 추가 데이터셋을 도입, 수집하는 등의 방법을 검토할 필요가 있을 것으로 보인다. 자동 매매에서는 잦은 매매와 시장가 매매로 인해 수수료와 호가 차이로 인한 손실이 큰 것으로 보여, 매도/매수 여부와 매매금액 규모의 판단을 보다 신중하게 할 필요가 있을 것으로 보인다.

참고문헌

[1] 이주화, "뉴스/소셜 미디어 텍스트 데이터를 활용한 주가 등락 예측 연구 : 딥 러닝 기법을 바탕으로", 성균관대학교 일반대학원 학위논문, 2021

[2] 전재헌, "딥러닝을 활용한 개인 투자자 감정지표 기반의 주가예측에 관한 연구", 광운대학교 일반대학원 경영학과 학위논문, 2019

[3] 방현기, "기계학습 모형을 활용한 암호화폐 가격 등락 예측에 관한 연구 - 비트코인과 GRU 모델을 중심으로", 동아대학교 일반대학원 디지털금융학과 학위논문, 2021