뉴스 데이터를 이용한 머신러닝 기반 비트코인 가격 등락 예측

강민규,김보선,신희종
고려대학교
{cxz3619,boseon12,tlswhd0215}@korea.ac.kr

Machine Learning based Prediction of Bitcoin Price Fluctuation Using News Data

MinGyu Gang, Boseon Kim, Hee-Jong Shin
Korea Univ

요 약

비트코인이 발표된 후, 비트코인은 현재 두 번째 열풍을 맞고 있다. 이러한 열풍과 관심에 비례하여 늘어나는 소셜 미디어, 뉴스 기사들은 많은 연구자들의 연구대상이 되었다. 기존의 연구들은 소셜 미디어와 뉴스 기사를 활용하여 비트코인의 가격 예측을 시도했으며, 많은 연구들이 감성분석을 이용했다. 감성분석을 이용한 방법은 감성분석 시 많은 데이터가 손실되고 감성 분석 사전이 비트코인 분야의 감성을 잘 읽지 못할 가능성이 있다. 본 논문은 감성분석 대신 뉴스 기사의 TF-IDF(Term Frequency-Inverse Document Frequency) 벡터 값과 뉴스 기사의 수를 이용해 다음날의 비트코인 종가를 예측한다. 예측 과정에서 8개 모델 간의 성능비교와 가격 데이터만 쓴 경우와 뉴스 데이터와 가격데이터를 함께 쓴 경우를 비교한다.

Ⅰ. 서 론

비트코인은 2008년 10월 31일 사토시 나카모토가 발표한 최초의 블록체인 기반의 탈중앙화 암호 화폐이다. 비트코인이 발표 된 후 비트코인의 거래량과 거래 가격은 점점 증가하여 2018년 1월 정점을 찍은 후, 2021년 3월에 다시 엄청난 열풍을 불러일으키고 있다. 이러한 열풍에 비례하는 관심만큼 소셜 미디어의 트윗과 비트코인 관련 기사 또한 증가하고 있다. 늘어가는 텍스트 데이터들을 활용하기 위해, 소셜 미디어나 비트코인 관련 기사들을 이용해 비트코인의 가격을 예측하는 많은 선행연구들이 수행되었다.

[1]은 뉴스 데이터를 Textblob 패키지를 이용해 감성분석 한 후, LSTM(Long Short-Term Memory)모델을 써서 비트코인의 가격등락을 예측했다. [2]는 Word2Vec을 이용해 주가 지수 사전을 구축한 후, 구축된 사전을 바탕으로 감성분석의 결과 값과 주가지수의 방향성을 비교했다. [3]은 뉴스 데이터를 CountVectorizer를 이용해 백터화 한 후,로지스틱 회귀,SVM(Support Vector Machine),나이브 베이즈 등의 분류기를 통해 비트코인, 이더리움, 라이트코인의 가격 변동을 예측했다.

뉴스 데이터를 이용해 주가를 예측하는 방법에는 주로 감성분석을 통해 뉴스 데이터를 수치화 하거나 분류한 후, 가격예측의 feature로 활용 한다. 하지만 뉴스 데이터를 감성분석 하는 경우, 많은 데이터의 손실이 있고 감성분석 사전이 비트코인 분야의 감성을 잘 분석하지 못할 가능성이 있다. 따라서 본 논문은 뉴스데이터를 감성분석하지 않고 TF-IDF 알고리즘을 이용해 나온 벡터와 뉴스 기사 수를 이용한 비트코인 가격 등락 예측 방법을 제안한다.

본 논문의 구성은 서론에 이어, 본론에서 데이터 수집, 데이터 전처리, 실험, 실험결과에 대해 설명하고, 마지막으로 결론 및 향후연구에 대해 언급한다.

Ⅱ. 본론

A.데이터 수집

[1]의 뉴스 순위에 따라 2013년에서 2020년 사이에 비트코인 관련 기사를 가장 많이 쓴 상위 4개의 언론사에서 기사를 수집했다. 수집된 기사들은 언론사, 날짜 ,제목 ,본문 4개의 열로 이루어져있으며 그림1과 같다.

	Press	Date	Title	Paragraph
0	BitcoinNews	2016-7-1	NewdatashowsChineseexchangeshaveaccountedfor42...	In the New York Times, writer Nathaniel Popper...
1	BitcoinNews	2016-7-1	OTCTradingForRegularPeopleSearchFortheUberofBi...	More and more people are looking to get starte...
2	BitcoinNews	2016-7-1	BitcoinContestTellIdeapodHowBlockchainCanChang...	The two influential thinkers, Don and Alex Tap...

그림 1.수집된 기사의 형태

B.데이터 전처리

기사들을 벡터화하기 앞서, 기사들은 불용어와 온점(“.”)등을 포함하고 있고 같은 형태의 복수형 또한 다른 단어로 인식하기에 벡터화 하는 것에 문제가 된다. 따라서 파이썬 *nltk* 패키지의 *word_tokenize*와 *stopword*를 활용해 토큰화 시킨 후, 불용어를 제거했다.

다음으로, 우리는 텍스트 데이터인 뉴스기사를 TF-IDF를 이용하여 뉴스 기사의 특징 벡터를 추출한다. TF-IDF는 정보 검색론 분야에서 사용하는 가중치를 구하는 알고리즘으로, 단어 빈도 수를 기반으로 가중치를 구한다.

$$tfidf(t,d,N) = tf(t,d)*idf(t,d) \quad (1)$$

$$idf(t,d) = \log\frac{N}{1+df(t)} \quad (2)$$

가중치인 $tfidf(t,d,N)$는 단어 빈도의 가중치인 $tf(t,d)$와 문서의 역 빈도 가중치인 $idf(t,d)$의 곱으로 이루어져 있다.$tf(t,d)$는 특정 문서 d에서 특정 단어 t의 등장 횟수를 의미하며, $idf(t,d)$는 특정 단어 t가 등장한 문서의 수의 역수 형태이다. $tfidf(t,d,N)$ 가중치는 단어의 빈도수가 높고, 전체 문장에서 적게 등장할수록 높아진다.

하루 단위로 합쳐진 기사의 제목과 본문에 각각 TF-IDF 알고리즘을 적용해 벡터화시켰으며 하루 단위의 기사 수 또한 추출했다. 추출된 데이터는 3개의 열[뉴스 기사 수, 제목의 TF-IDF값, 본문의 TF-iDF값]이며 행

은 950개의 일자이다. 데이터 예시는 그림2와 같다.

	date	num_news	paragraph_tfidf	title_tfidf
0	2018-08-01	15	[0, 0, 0, 0.287682, 0, 0.693147, 0.287682, 0, ...	[0, 0.182612, 0, 0.157864, 0, 0, 0, 0, 0.28768...
1	2018-08-02	32	[0, 0, 0.191147, 0, 0.575364, 0, 0, 0, 0.69314...	[0, 0.693147, 0.693147, 0, 0.575364, 0, 0, 0, ...
2	2018-08-03	43	[0, 0.693147, 0.693147, 0, 0.575364, 0, 0, 0, ...	[0, 0, 0, 0.287682, 0, 0.693147, 0.287682, 0, ...

그림 2. 데이터 전처리 결과

C.실험

본 절에서는 머신러닝 분류 분야에서 주로 쓰이는 8개의 분류모델을 이용했으며 *python*의 *scikit-learn* 라이브러리를 통해 제공되는 모델들과 파라미터들을 이용해 실험을 진행했다. 사용한 분류 모델들은 다음과 같다.

-로지스틱 회귀

로지스틱 회귀는 로지스틱 함수를 사용하여 데이터가 어느 1과0 사이의 값으로 예측하고 확률에 따라 가능성이 더 높은 범주로 분류해주는 모델이다.

-나이브 베이즈

나이브 베이즈는 조건부 확률을 이용한 나이브 베이즈 정리를 기반으로 하는 지도학습 알고리즘이며 분류하려는 대상의 긱 확률을 측정하여 분류하는 모델이다.

-KNN(K-Nearest Neighbors)

KNN모델은 새로운 데이터가 주어졌을 때, 기존 데이터 중 가장 가까운 k의 데이터를 바탕으로 새로운 데이터를 예측하는 모델이다.

-SVM(Suport Vector machine)

SVM 모델은 데이터 군들로부터 가장 먼 분류 경계면을 찾고 경계면을 기반으로 새로운 점이 어느 쪽에 확인하여 분류하는 모델이다.

-랜덤 포레스트

랜덤 포레스트는 여러 개의 피쳐 중 랜덤한 수의 피처를 선택한 후, 다수의 결정트리들을 학습하는 앙상블 기법을 이용한 모델이다.

-Extra Tree

Extra Tree는 랜덤 포레스트와 유사한 결정 트리이지만, 랜덤 포레스트와 다르게 피처 또한 무작위로 분할한 후 최상의 분할을 선택하는 앙상블 기법을 이용한 모델이다.

-AdaBoost

AdaBoost는 Boosting 알고리즘의 한 종류로서, 이전 약한 학습기 결과의 오차에 가중치를 두어 다음 약한 학습기의 입력으로 주어 학습하는 Boosting 모델이다.

-XGBoost

XGBoost는 Gradient Boosting을 개선시킨 알고리즘으로, 경사하강법과 병렬 처리 기법을 사용해 과적합 문제를 해결하고, 처리시간을 개선시킨 모델이다.

실험은 8개의 모델에 입력 값을 다르게 주는 방식으로 두 번에 걸쳐 비교했다. 첫 번째 입력 값은 ['전날의 종가']만을 넣은 방식이고, 두 번째 입력 값은 첫 번째 입력 값에 뉴스 데이터('기사 수','제목의 TF-IDF 값','본문의 TF-IDF 값')을 추가한 값이다. 이것은 뉴스 데이터를 이용한 예측이 가격 데이터만 이용했을 때와 비교해보기 위함이다.

D.실험 결과

모델 평가 결과, 입력 값으로 전날의 종가만을 준 모델들은 40% 후반 대에서 50% 후반까지 고르게 분포한 모습을 보여주었으며, XGBoost 모델이 가장 좋은 정확도인 60%를 보여주었다. 또한 입력 값으로 가격 데이터와 함께 뉴스 데이터(기사 수,제목의 TF-IDF값,본문의 TF-IDF값)를 넣어 주었을 때 나이브 베이즈를 제외한 모든 모델에서 precision, recall, f1-score, accuracy가 소폭 상승하거나 유지했다. 상승치가 없는 모델의 경우, 다른 라벨에 대해 precison과 recall이 0으로 잘 학습되지 못함 모델들이였고, 모든 수치가 크게 하락한 나이브 베이즈 모델 또한 잘 학습하지 못한 모델이었다. 등락을 잘 예측한 모델 중, accuracy를 기준으로 가장 큰 폭으로 상승한 모델은 랜덤 포레스트 모델이며, 가장 높은 accuracy를 기록한 모델은 62%를 기록한 adaBoost 모델이다.

자세한 수치는 표 3에 표시되며, 각각의 왼쪽 수치는 가격 정보만을 사용했을 때 수치이고, 오른쪽의 수치는 가격 정보와 뉴스 데이터를 함께 썼을 때 변화 폭이다.

	precision		recall		f1-score		accuracy	
로지스틱 회귀	58%	-	100%	-	73%	-	58%	-
나이브 베이즈	55%	-3%	100%	-46 %	73%	-	58%	-10 %
KNN	59%	**+5%**	56%	**+6%**	58%	**+5%**	52%	**+5%**
SVM	58%	-	100%	-	73%	-	58%	-
랜덤 포레스트	55%	**+5%**	51%	**+9%**	53%	**+8%**	48%	**+6%**
Extra Tree	55%	**+2%**	51%	**+3%**	53%	**+3%**	47%	**+3%**
AdaBoost	61%	**+2%**	85%	-	71%	**+1%**	59%	**+3%**
XGBoost	62%	-1%	79%	**+6%**	69%	**+1%**	60%	**+1%**

표 3.모델들의 성능 정보

Ⅲ. 결론

본 논문에서는 뉴스 데이터를 이용해 비트코인 가격의 등락을 예측하는 방법을 제시했으며, 8개 모델간의 성능을 비교하여 평가했다. 또한 정보만 쓰는 것 보다 뉴스 정보를 함께 사용하는 것이 비트코인의 등락을 예측하는데 도움이 된다는 것을 보여주었다. 그러나 모델들의 성능 향상에 있어 뉴스데이터의 존재는 그리 크지 않은 수준이다. 따라서 향후 연구로는 뉴스 데이터에서 조회 수, 감성분석 결과 등 더 많은 종류의 뉴스 데이터를 추가하고 모델의 최적화를 통해 성능을 고도화할 계획이다.

ACKNOWLEDGMENT

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2018R1D1A1B07045742)과 2020년도 산업통상자원부 및 한국산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임 (No. 20008902, IT비용 최소화를 위한 5채널 탐지기술 기반 SaaS SW Management Platform(SMP) 개발)

참 고 문 헌

[1]강민규, 김보선, 신무곤, 백의준, 김명섭, "LSTM 기반 감성분석을 이용한 비트코인 가격 등락 예측" 2020년도 한국통신학회 추계종합학술발표회, 온라인 개최, Nov. 13, 2020, pp. 1-2

[2] 김다예 · 이영인. "Word2Vec을 활용한 뉴스 기반 주가지수 방향성 예측용 감성 사전 구축" ,한국빅데이터학회지 제3권 제1호, 2018, pp. 13-20

[3] Connor Lamon, Eric Nielsen, Eric Redondo. "Cryptocurrency Price Prediction Using News and Social Media Sentiment" ,http://cs229.stanford.edu/proj2017/final-reports/5237280.pdf