J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

역순 워크 포워드 검증을 이용한 암호화폐 가격 예측

An Accurate Cryptocurrency Price Forecasting using Reverse Walk-Forward Validation

안 현[1] 장 백 철[1*]
Hyun Ahn Baekcheol Jang

요 약

암호화폐 시장의 규모는 날이 갈수록 커져가고 있으며, 대표적인 암호화폐인 비트코인의 경우 시가총액이 500조를 넘어섰다. 이에 따라 암호화폐의 가격을 예측하려는 연구도 많이 이루어졌으며, 이들은 대부분 주식가격을 예측하는 방법론과 유사성을 띄는 연구들이다. 하지만 선행연구를 비춰 봤을 때 주식가격예측과 달리 암호화폐 가격 예측은 머신러닝의 정확도가 우위에 있는 사례가 많다는 점, 개념적으로 주식과 달리 암호화폐는 소유로 인한 수동적 소득이 없다는 점, 통계적으로 시가총액 대비 하루 거래량의 비율을 살펴봤을 때 암호화폐가 주식 대비 최소 3배이상 높다는 점이 도출되었다. 이를 통해 암호화폐 가격 예측 연구에는 주식 가격 예측과 다른 방법론이 적용되어야 함을 본 논문에서 주장하였다. 우리는 기존에 주가 딥러닝 예측에 사용되던 워크 포워드 검증를 응용한 역순 워크 포워드 검증을 제안하였다. 역순 워크 포워드 검증은 워크 포워드 검증과 달리 검증 데이터셋을 테스트 데이터셋에 시계열상으로 바로 앞에 부분으로 고정시켜놓고, 훈련데이터를 훈련 데이터셋에 시계열상으로 바로 앞 부분부터 서서히 훈련 데이터셋의 크기를 늘려가면서 검증에 대한 정확도를 측정한다. 측정된 모든 검증 정확도 중 가장 높은 정확도를 보이는 훈련 데이터셋의 크기에 맞춰서 훈련 데이터를 절삭시킨 뒤 검증 데이터와 합쳐서 실험 데이터에 대한 정확도를 측정하였다. 분석모델로는 로지스틱 회귀분석과 SVM을 사용했으며, 우리가 제안한 역순 워크 포워드 검증의 신뢰성을 위해서 분석 모델 내부적으로도 L1, L2, rbf, poly등의 다양한 알고리즘과 정규화 파라미터를 적용하였다. 그 결과 모든 분석모델에서 기존 연구보다 향상된 정확도를 보임이 확인되었으며, 평균적으로도 1.23%p의 정확도 상승을 보였다. 선행연구를 통해 암호화폐 가격 예측의 정확도가 대부분 50%~60%사이에서 머무르는 걸 감안할 때 이는 상당한 정확도 개선이다.

☞ 주제어 : 암호화폐, 가격예측, 머신러닝

ABSTRACT

The size of the cryptocurrency market is growing. For example, market capitalization of bitcoin exceeded 500 trillion won. Accordingly, many studies have been conducted to predict the price of cryptocurrency, and most of them have similar methodology of predicting stock prices. However, unlike stock price predictions, machine learning become best model in cryptocurrency price predictions, conceptually cryptocurrency has no passive income from ownership, and statistically, cryptocurrency has at least three times higher liquidity than stocks. Thats why we argue that a methodology different from stock price prediction should be applied to cryptocurrency price prediction studies. We propose Reverse Walk-forward Validation (RWFV), which modifies Walk-forward Validation (WFV). Unlike WFV, RWFV measures accuracy for Validation by pinning the Validation dataset directly in front of the Test dataset in time series, and gradually increasing the size of the Training dataset in front of it in time series. Train data were cut according to the size of the Train dataset with the highest accuracy among all measured Validation accuracy, and then combined with Validation data to measure the accuracy of the Test data. Logistic regression analysis and Support Vector Machine (SVM) were used as the analysis model, and various algorithms and parameters such as L1, L2, rbf, and poly were applied for the reliability of our proposed RWFV. As a result, it was confirmed that all analysis models showed improved accuracy compared to existing studies, and on average, the accuracy increased by 1.23%p. This is a significant improvement in accuracy, given that most of the accuracy of cryptocurrency price prediction remains between 50% and 60% through previous studies.

☞ keyword : Cryptocurrency, Price prediction, Machine learning

1. 서 론

금융자산에 대한 가격예측에 관심이 많아지면서, 많은 사람들이 암호화폐 가격예측에도 관심을 가지기 시작했다. 회귀분석, SVM[1], Random forest 같은 전통적인 머신

1 Graduate School of Information, Yonsei University
* Corresponding author (bjang@yonsei.ac.kr)
[Received 11 July 2022, Reviewed 21 July 2022, Accepted 5 August 2022]

러닝 기법[2] 뿐만 아니라 ANN[3], RNN, LSTM[4]를 비롯한 딥러닝 기반 예측 기법도 활발하게 연구되고 있다. 기반이 되는 데이터 관련해서도 다양한 연구가 이루어지고 있는데, 단순 거래 데이터[5], SNS에서 사람들의 반응[6], 금 시세[7]를 독립변수로 넣어서 암호화폐 가격을 예측하는 연구가 이뤄졌었다. 그리고 이러한 연구 방향은 투자를 통한 이익 창출의 측면에서 주식과 유사한 전제를 가진 채 진행되어왔다. 문제는 이러한 암호화폐 가격예측 연구들의 전제가 암호화폐 가격 데이터의 변동성을 제대로 포함하지 못하고 있다는 것이다. 암호화폐 가격만 높고 봐도, 특정 암호화폐에 투자한 100만원이 약 14개월 뒤 5조 이상의 자산가치를 가진다는 건[1] 주식시장 역사상 최대 등락폭을 보였던 미시시피 회사와 남해해운을 기준으로 계산해도 불가능한 것이다. 현재까지 암호화폐 가격 예측에 대한 연구는 주식 가격 예측에 대한 연구와 그 궤를 같이 해왔지만, 현존하는 모든 연구성과를 뛰어넘기 위해서는 '주식 데이터와 구별되는 암호화폐 데이터만의 특징'을 잡아낼 필요성이 있다 판단되었다.

이에 본 논문에서는, 기존 주식시장의 데이터와 암호화폐 거래데이터의 통계 값 분석을 통해 암호화폐 데이터만의 특성으로 어떤 게 있는지를 가설로서 도출해낸 후, 해당 가설을 기존 방법론에 변용한 뒤 기존 방법론을 적용한 결과와 비교하였다. 통계적 기술과 개념적 분석을 통해 암호화폐의 가격변동의 휘발성이 주식보다 유의미하게 더 크다는 것을 밝혀냈다.

이러한 암호화폐 가격 데이터만의 특성을 반영하기 위해 우리는 새로운 검증 분할 방법론을 본 논문에서 제안하였다. 본 논문에서 우리는 주가 예측을 포함한 기존 시계열 데이터를 딥러닝을 통해 분석할 때 기용하던 워크 포워드 검증을 응용해서 역순 워크 포워드 검증을 새로 개발하였다. 역순 워크 포워드 검증은 주어진 시계열 데이터 중 학습데이터로 얼마나 오래된 데이터까지 쓸 지를 결정하게 해준다. 학습데이터의 크기를 다르게 해 가면서 학습데이터의 크기별 검증 정확도를 산출해 낸 뒤, 최고의 검증 정확도를 산출해낸 학습데이터의 크기로 실험데이터에 대한 모델을 학습시킨다.

역순 워크 포워드 검증이 암호화폐 가격예측에 있어서 정확도 향상을 가져오는지를 검정하기 위해 암호화폐의 일종인 이더리움의 1시간단위 가격변동 데이터를 사용하였다. 해당 시각의 가격 데이터부터 5시간 전의 가격데이터까지 독립변인으로 사용하였고, 분석 모델로는 로지스틱 회귀분석과 SVM을 기용하였다.

암호화폐 가격예측 선행연구는 딥러닝보다 머신러닝이 예측정확도의 측면에 있어서 다소 우세한 상태이고, 역순 워크 포워드 검증의 전신인 워크 포워드 검증이 딥러닝에서만 기용되는 방법론이므로 워크 포워드 검증을 적용한 예측정확도를 베이스라인(bassline)으로 적용하지 않고 훈련 데이터와 검증 데이터를 학습데이터로 사용했을때의 예측정확도를 베이스라인(bassline)으로 정했다. 그 결과, 모든 실험에서 베이스라인(bassline)과 비교했을 때 정확도가 상승하는걸 밝혀냈다.

위와 같은 결과를 통해 본 논문은 분석에 있어서 암호화폐 가격데이터와 주가 데이터의 차이를 밝혀냈으며, 역순 워크 포워드 검증의 원리적 범용성을 통해 이전 암호화폐 가격 예측 연구의 정확도를 끌어올리는 긍정적인 요소를 개발했음을 주장한다.

2. 관련연구 및 문제정의

2.1 암호화폐 가격예측

이하 선행연구들은 어떤 데이터를 사용했는지, 어떤 모델을 사용했는지를 기반으로 분류되었다. 분석 모델의 측면에서는 전통적인 머신 러닝을 이용한 것과 DNN/CNN/RNN으로 대표되는 딥러닝을 이용한 것들로 구분이 된다. 이를 표로 나타내면 표 1와 같다.

표 1의 전통적 머신 러닝 기법은 뉴런과 레이어로 이루어진 인공신경망을 이용한 기법을 제외한 모든 연구방법론을 의미한다. [2]의 경우 이더리움 암호화폐의 가격 데이터가 평균값 대비 상당히 큰 분산 값을 가지고 있다 판단하고, 특정 기간 동안의 가격 데이터만을 이용하였다. 이렇게 절삭 시킨 가격 데이터에서 1시간마다 1개씩, 하나의 데이터 안에 가격이라는 변수 하나가 존재하는 형태였는데, 여기에 해당시간에서 1~5시간 이전의 가격데이터를 넣어서 독립변수의 가격데이터를 6개로 증식시켰다. 종속변수는 해당 시간대의 암호화폐 가격 데이터와 1시간 이후의 가격데이터의 차이를 -1과 +1로 이진화시켜서 만들어냈다. 이렇게 만들어낸 데이터셋을 로지스틱 회귀분석, 나이브 베이즈 분류, SVM, Random forest, 자기회귀누적이동평균(Auto-regressive Integrated Moving Average;ARIMA), RNN을 통해 학습시켰다. 그 결과 자기회귀누적이동평균 모델이 61.17%의 정확도를 보

1 https://econoimst.co.kr/2021/10/31/stock/virtualCurrency/20211031184650977.html

(표 1) 기존 암호화폐 가격예측 연구의 분류
(Table 1) Related works of Crypto price prediction

	거래 데이터	거래데이터 및 외부 데이터 포함
전통적 머신러닝 기법	[1], [2]	[6], [8], [9], [10]
딥러닝 기법	[3], [4], [5]	[7]

여줘서 가장 성능이 좋음을 입증했고, 그 다음으로 로지스틱 이진 회귀 분석이 56.94%의 정확도를 보여주었다. 약 10개월에 달하는 데이터셋에, 다른 연구들이 대부분 하루단위로 거래데이터를 정리한 것과 달리 1시간 단위로 데이터가 기록이 된 만큼 데이터의 수가 딥러닝을 이용하기에 충분했을 것이다. 그럼에도 불구하고 인공 신경망을 이용하지 않은 전통적 머신 러닝 기법이 정확도가 더 높게 나왔고, 이는 기존 주식가격 데이터에 존재하지 않는 암호화폐 가격 데이터만의 고유한 특성이 있음을 방증한다. 이 외에도 [1]에서는 2012년부터 2018년까지 비트코인 가격 데이터를 일단위로 정리한 데이터를 SVM과 로지스틱 회귀분석을 통해 분석했으며, 분석과정에서 데이터의 특정 부분으로부터의 의존성을 낮추기 위해 10겹 교차 검증(10-fold cross-validation)을 분석과정에 적용하였다.

이와는 달리, 암호화폐의 거래기록 외의 데이터까지 이용해서 가격을 예측하려는 시도도 있었다. [6]의 경우 구글 트렌드, 트위터 게시글, 비트코인 가격을 분석에 사용하였다. 이들은 트위터 게시글을 전처리 후 감정분석을 진행하였으며, 일별 트위터 게시글 수와 구글 트렌드로 독립변수를 추출해냈다. 예측하는 날짜를 t라 가정하면, 전술한 과정을 통해 t-l(l은 임의의 숫자)일부터 t-1일까지 정리된 데이터를 독립변수로 최종 지정하였다. 이렇게 정리된 독립변수와 비트코인 가격이라는 종속변수를 기반으로 그레디언트 부스팅(Gradient boosting) 알고리즘을 학습시켰고, l=3일 때 57.84%의 정확도를 산출해냈다. [8]의 경우에도 이와 유사하게 트위터의 게시글 수 및 감정분석 결과와 비트코인 가격 데이터를 Autoregressive Intergrated Moving Average Exogenous Model(ARIMAX)와 LSTM 기반 인공신경망으로 학습시켰으나, ARIMAX가 LSTM 기반 인공신경망보다 더 정확한 예측을 하였다고 발표하였다. [9]의 경우 감정분석보다는 트위터/뉴스 헤드라인 속 특정 단어의 빈도에 따른 가격예측을 로지스틱 선형 회귀분석, 선형 SVM, 나이브 베이즈 모델을 통해 시도하였다. 총 세 종류의 암호화폐에 해당 예측연구를 실시한 결과 비트코인 및 라이트코인에서 로지스틱 회귀 분석이 가장 좋은 퍼포먼스를 보였다.

이처럼 트위터 게시글의 내용을 통해 암호화폐 가격을 예측하는 연구도 있었던 반면, 반대의 의견을 개진한 선행연구도 있었다. [10]에 따르면 감정분석에 기반한 연구들은 암호화폐 가격이 상승하던 시기와 맞물려 진행된 것들이 많으며, 암호화폐에 대한 트위터 게시글의 전체적인 경향은 언제나 긍정적이라는 점을 지적했다.

전통적 머신러닝을 이용한 암호화폐 가격 예측 연구의 상당수가 암호화폐 거래 데이터외의 외부데이터(주로 트위터 및 구글)을 기반 데이터셋으로 기용했다면, 딥러닝을 모델로 기용한 선행연구(또는 딥러닝이 가장 좋은 정확도를 보인 연구)의 경우 상당수가 암호화폐 거래 관련 데이터만을 이용하여 연구되었다. [5]의 경우 암호화폐 가격 뿐만 아니라 시간당 거래횟수, 해당 시점에 채굴된 비트코인의 개수, 시간당 신규 생성된 암호화폐 주소의 수 등을 독립변수로 이용했으며, 로지스틱 회귀분석, SVM, 인공신경망을 통해 한 시간 이후의 암호화폐 가격을 예측했다. 그 분류의 정확도에 있어서 2개의 은닉층을 이용한 인공신경망이 55.1%로 가장 높은 예측정확도를 보여주었다. [3]의 경우에도 2개의 은닉층을 이용해서 비트코인/비트코인캐시/대시코인의 가격을 일 단위와 1시간 단위로 예측하였다. [4]의 경우에는 위의 두 연구보다 훨씬 시계열 데이터에 적합한 것으로 알려진 LSTM을 RNN과 혼합해서 학습모델로 기용하였다.

외부데이터를 포함한 딥러닝 학습을 통해 암호화폐 가격예측을 하려는 시도도 있었다. [7]의 경우 비트코인 거래 데이터와 동일시기 금 거래 데이터, 트위터 게시글의 감정분석을 CNN/LSTM/RNN으로 학습시키는 연구를 진행했다. 회귀 예측 결과 금 거래 데이터를 포함한 분석은 암호화폐 거래 데이터만을 사용한 경우보다 회귀예측의 정확도가 떨어지는 걸로 밝혀졌고, 트위터 게시글에 대한 감정분석 결과를 포함할 경우 암호화폐 거래 데이터만으로 학습시켰을 때 보다는 회귀예측의 정확도가 높은 걸로 나타났다.

위의 내용과 더불어 가격변동의 방향 예측의 정확도가 주로 50%~60%사이의 값으로 산출된다는 점 또한 주목할 필요가 있다. [2], [11]의 연구결과를 살펴보면 여타 분야의 이진 분류문제의 정확도와 달리, 암호화폐 가격예측의 이진분류 정확도는 대부분의 모델에서 60%이하의 값을 가진다. 물론 [12]처럼 이진 분류 정확도가 60%이상을 상회하는 경우도 많지만, 이 경우 여러 분석모델을 혼합시킨 앙상블 모델을 기용했을 때 60%를 넘긴것이지, 각 모

델단위로 분석결과를 살펴보면 딥러닝 기준 58.84%, 전통적 머신러닝 기준 59.45%의 최대치를 가질 뿐이었다. 이처럼 암호화폐 가격 예측 선행연구들은 분석 대상이 사회경제적인 영역에 있음을 감안해도 주가 예측 연구보다 떨어지는 예측 정확도를 보인다. [3]에 따르면 암호화폐 가격에 영향을 주는 사회경제적 요인이 주가에 영향을 주는 요인보다 많기 때문으로 볼 수 있다. 주가 예측 연구 대비 더 많은 사회경제적 요인들이 암호화폐 가격에 영향을 주기 때문에 독립변인을 어떻게 설정해도 동일한 종류의 독립변인으로 분석된 주가 예측 연구보다도 정확도가 떨어질 수 밖에 없다.

2.2 주식 가격예측

주식 가격예측연구도 암호화폐 가격예측연구와 마찬가지로 어떤 데이터를 사용했는지, 어떤 모델을 사용했는지를 이용해서 분류가 가능하다. 이를 이용하여 선행연구들을 분류하면 표 2와 같다. 후술할 주식 가격예측 선행연구들을 보면, 암호화폐 가격예측 연구와 방향성/데이터 수집방식/학습모델 구축의 측면에서 선후관계 또는 유사성이 있음을 확인할 수 있다.

(표 2) 주식 가격예측 선행연구의 분류
(Table 2) Related works of stock price prediction

	거래 데이터	거래데이터 및 외부 데이터 포함
전통적 머신러닝 기법	[13], [14]	[15]
딥러닝 기법	[16]	[17]

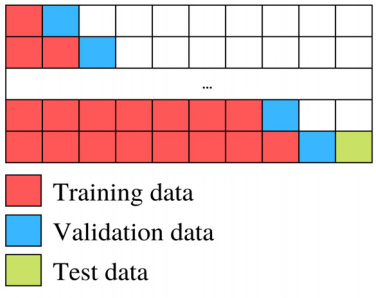
[13]의 경우, 로지스틱 회귀분석을 기반에 둔 모델을 이용해서 다음달 심천 개발 주식 A(Shenzhen Development stock A)의 가격을 예측하였다. [14]에서는 SVM을 다중변수를 적용할 수 있는 모델로 확장시킨 구조적 서포트 벡터 머신(Structural support vector machine)을 이용해서 매달 가격을 예측하여 투자 한 결과 10%에서 17%의 수익률을 산출해냈다. 단순히 지표를 따라 투자한 경우 -17%~9%의 수익률을 낸다는 점과 비교하면 충분히 의미있는 결과인 것이다. [15]에서는 이러한 데이터 분석에 트위터 게시글에 대한 감정 분석 데이터를 포함시켜서 연구했다. 나이브 베이즈 분류모형과 Random Forest로 트위터 게시글의 감정분석을 진행시킨 후, 분석된 데이터를 기반으로 인도네시아 주식 시장에 상장된 회사들의 주가를 일단위로 예측하였다.

딥러닝 학습모델을 기용한 선행연구들의 경우 대부분 베이스라인(Bassline)을 자기회귀누적이동평균 기법으로 정했다. [16]의 경우 인도니프티50에 포함된 세 회사(Infosys, TCS, Cipia)의 주가를 세 학습모델 모두 오류율의 측면에서 자기회귀누적이동평균 모델보다 뛰어난 정확도를 보여주었다. [17]의 경우 여기서 한 발 더 나아가서 265000개의 경제 신문기사를 포함한 가격예측을 시도하였다. RNN을 학습모델로 기용하여 분석한 결과 자기회귀누적이동평균 모델을 이용한 분석보다 평균절대비오차(Mean absolute percentage error; MAPE)가 상당히 감소함을 확인할 수 있었다.

상술한 학습모델, 방법론, 데이터 수집 방법등은 이전에 서술한 암호화폐에서도 유사하게 이용되었다.

2.3 Walk-forward Validation

주식 가격 예측을 포함한 시계열 데이터 분석에 있어서 워크 포워드 검증 방법론은 고려할 만한 선택지 중 하나로 인지되어왔다[18, 19, 20]. 그림 1에 나온 것처럼 워크 포워드 검증 방법론은 하나의 데이터에서 여러 개의 훈련-검증 데이터쌍을 만들어서 분석모델로 하여금 누적된 학습을 하도록 만들었다. 여러 개의 검증 데이터를 하나의 분석모델에 적용하기에 딥러닝 분석 모델에 쓰였다.



(그림 1) 워크 포워드 검증 방법론
(Figure 1) Walk-forward Validation

특히 [20]에 따르면, 데이터 분석에 있어서 일반적으로 사용되던 교차 검증 방법론은 시계열 분석에 있어서 데이터의 앞뒤순서를 반영하지 못한다는 한계에 부딪혔었다. 시간순서에 따른 데이터의 변화가 시계열 데이터에서는 중요시 되는데, 교차 검증 방법론을 쓰게 되면 미래의 데이터를 기반으로 과거의 데이터를 예측하게 되는 상황이 생기기 때문이다. 이 때문에 딥러닝을 이용한 시계열

(표 3) 암호화폐, 한국 주식시장(KOSPI), 미국 주식시장(NASDAQ)거래소의 시가총액 1/2위의 하루 거래량, 시가총액 그리고 총액 대비 하루 거래량

(Table 3) The Daily trading volume, market capitalization and ratio of daily trading volume per market cap on cryptocurrency, the Korean stock market(KOSPI), and the U.S. stock market(NASDAQ) exchanges.

대분류	암호화폐		KOSPI		NASDAQ	
소분류	비트코인	이더리움	삼성전자	SK 하이닉스	APPLE	Microsoft
하루 거래량 (백만 원/달러)	27138	12252	1012938	36299	13939	8655
시가 총액 (백만 원/달러)	723039	336857	454160000	79350000	2580000	2130000
비율(%)	3.75	3.64	0.22	0.46	0.54	0.41

데이터 분석에 있어서는 워크 포워드 검증 방법론이 기용되게 되었다.

2.4 문제 정의

전술한 주식/암호화폐 가격 예측 연구들의 공통된 방법론적 전제는 '과거의 데이터를 통해 미래를 분석한다'는 점이다. 그리고 더욱 정확한 분석을 위해 대부분의 암호화폐 선행연구들은 상당히 긴 시계열 데이터를 학습용 데이터로 기용하였다. 하지만 [17]에서 '주가가 저조하거나 변동성이 높은 시기에는 예측결과가 제대로 부합하지 않는 경우가 있다'라고 보고된 점이 전술된 거대한 학습 데이터 기용의 반론으로 작용할 수도 있다. 즉, 암호화폐 가격예측에 있어서 '가격이 저조하거나 변동성이 높은 시기'는 극단치(outlier)로 작용하는 것이 아닌 데이터 전체에 일반적으로 자리잡은 경향으로 봐야 할 수도 있다. 주식 가격 예측분야에서는 딥러닝 기반 학습모델의 성능의 비교대상으로 기재되면 자기회귀누적이동평균 모델이[16,17] 암호화폐 가격 예측연구에서는 오히려 딥러닝 기반 학습모델보다 뛰어난 정확도를 보이는 경우가 많았던 것이 이를 방증한다. 암호화폐는 주식과 달리 가격변동에 있어서 경향의 휘발성을 가지며, [2]에서 주장된 '암호화폐 가격데이터의 분산 값이 평균에 비해 지나치게 높은 것'도 이를 뒷받침하는 근거 중 하나이다.

암호화폐 가격변동 경향의 휘발성은 개념적인 부분에서도 근거를 가진다. 주가예측의 대상인 주식의 경우 상당수의 회사가 주기적인 배당을 시행하고 있다. [21]를 비롯한 수많은 선행연구에서 은퇴 후 적절한 배당수익을 얻기 위한 주식 포트폴리오의 구성을 설명하고 있다는 점에서 주식의 배당을 통한 수익이 시세차익을 통한 수익보다 사람들의 관심이 적지 않다는 점을 방증한다. 하지만 암호화폐의 경우에는 이러한 정기적인 배당수익이 존재하지 않는다. 그나마 배당과 가까운 수익이라고 한다면 포크, 스테이킹 정도가 있지만 포크는 정기적으로 이루어지는 것이 아니며, 포크를 통해 새로 생성된 암호화폐의 가치가 오른 뒤 매각해서 수익이 실현된다. 스테이킹 또한 비정기적으로 이루어지며, 배당과 달리 자산소유자가 직접 신청해야 한다는 차이를 가진다. 정리하면, 주식은 시세차익 외에도 배당을 통해 수익을 낼 수 있지만, 암호화폐를 통한 수익창출은 시세차익에 집중 되어있다.

이러한 가격변동의 휘발성은 통계 값 분석을 통해서도 입증할 수 있다. 표 3은 각 금융자산의 시가총액 대비 하루동안의 거래량을 달러로 표현한 것이다. 암호화폐와 비교할 주식으로는 한국 KOSPI 시가총액 1/2위인 삼성전자와 SK 하이닉스, 미국 NASDAQ 시가총액 1/2위인 APPLE과 Microsoft이다. 물론 위의 통계값에는 많은 외재변수가 있다. 우선, 주식시장과 달리 코인 거래소는 24시간 연중무휴로 거래가 이루어진다. 또한 이더리움의 경우 작년부터 NFT열풍으로 인해 그 이용도가 높아진 것 또한 사실이다. 하지만 이러한 외재변수를 전부 고려해도 암호화폐 자산의 총액 대비 거래량이 최소 2배이상 많다고 볼 수 있다, 이러한 많은 거래량은 [17]이 주식시장에서 가끔 발생하는 환경으로 지적한 '가격이 저조하거나 변동성이 높은 시기'중 '변동성이 높은 시기'가 암호화폐 거래 시기의 대다수를 차지하는 근거로 볼 수 있는 것이다.

3. Method

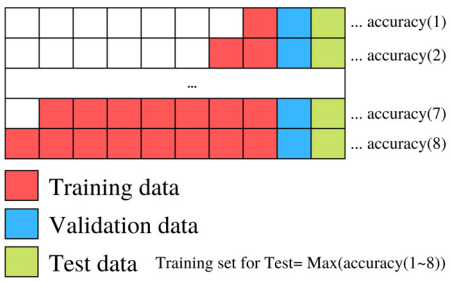
3.1 Reverse walk-forward validation

전술했듯이 이전부터 시계열 데이터 분석에 있어서 훈련과 검증을 점진적으로 진행하는 워크 포워드 검증기법이 사용되어왔다[18, 19, 20]. 하지만 이 기법은 그림 1에 있는 것처럼 검증 데이터가 일정 위치에 고정되는 것이 아니라 계속 바뀌는 것이기에 최종적으로 예측해야 하는 부분인 실험데이터의 가격변동 경향과 다를 확률이 너무 높다. 전술한 가격변동 경향의 휘발성은 독립변인으로 사용된 데이터의 시간대와 종속변인으로 사용된 데이터의 시간대가 차이가 날수록 설명력이 감소함을 의미한다. 이를 기반으로 그림 1의 워크 포워드 검증 방법론을 보면, 훈련 데이터의 시작점이 실험데이터와 시간상으로 너무 떨어져 있는 것도 문제가 되며 해당 훈련 데이터에 대해 검증의 역할을 해주는 검증 데이터가 시간상으로 떨어져 있다는 것도 문제가 된다. 검증 데이터를 통해 실험데이터에 적용했을 때의 예측력이 향상되었는지를 확인해야 하는데, 검증 데이터가 실험 데이터와 시간상으로 거리가 있게 되면 그만큼 실험 데이터의 가격변동 경향을 검증 데이터가 제대로 포함하지 못할 확률이 높다.

뿐만 아니라 워크 포워드 검증 방법론은 딥러닝에서만 쓸 수 있다는 점도 암호화폐 가격 예측 연구에 있어서는 문제가 된다. 선술했듯이 암호화폐 가격 예측 연구에서는 딥러닝을 이용한 연구보다 전통적 머신러닝을 이용한 연구가 더 높은 정확도를 가지는 경우가 많은데, 이 경우에 워크 포워드 검증 방법론을 이용하기 요원해지기 때문이다.

암호화폐 가격예측 연구에 있어서 워크 포워드 검증 방법론이 가진 위의 두 문제점을 개선하기 위해 우리는 워크 포워드 검증 방법론을 응용한 역순 워크 포워드 검증 방법론을 개발하였다. 그림 2에 있는 것처럼 검증 데이터를 실험 데이터의 바로 앞부분으로 고정시키고, 훈련 데이터를 검증 데이터 앞부분에서부터 서서히 늘리는 방식으로 분석을 진행하려 한다. 모든 검증 정확도를 확인 후 가장 검증 정확도가 높게 나오는 훈련-검증 데이터쌍을 훈련 데이터로 선정한 뒤 실험 데이터를 예측할 것이다. 이를 통해 어느 시간대까지 훈련 데이터로 써야 예측 정확도가 높게 나오는지를 알아낼 수 있다.

역순 워크 포워드 검증 방법론는 기존 역순 워크 포워드 검증 방법론의 암호화폐 연구로의 적용에 있어서 지적된 문제점을 보완할 수 있다. 워크 포워드 검증 방법론의 문제점 중 하나는 훈련 데이터와 검증 데이터가 실험



(그림 2) 역순 워크 포워드 검증 방법론
(Figure 2) Reverse Walk-forward Validation

데이터와 시간대상으로 떨어져 있게 되어서 실험 데이터의 가격변동 경향이 반영되어있지 않을 공산이 크다는 것이었다. 하지만 역순 워크 포워드 검증 방법론은 검증 데이터를 실험 데이터 바로 앞에 배치시킴으로 인해 검증 데이터와 실험 데이터의 가격변동 경향의 차이를 최소화시켰다. 뿐만 아니라 워크 포워드 검증 방법론에서는 시간상으로 가장 예전에 생성된 데이터부터 훈련 데이터로 추가시키면서 학습을 진행했지만, 역순 워크 포워드 검증 방법론의 훈련 데이터는 실험데이터와 시간상 인접한 순으로 데이터를 추가시키면서 학습을 하기에 어느 시점까지의 훈련 데이터가 예측에 도움이 되는지를 이전보다 정확히 알아낼 수 있다. 마지막으로 역순 워크 포워드 검증 방법론은 워크 포워드 검증 방법론과 다르게 하나의 훈련-검증 데이터 쌍을 누적해서 학습하는 것이 아니라, 각 훈련-검증 데이터 쌍의 검증 정확도를 비교하여 최고의 검증 정확도를 보인 훈련-검증 쌍 하나만을 최종 실험 데이터에 대한 정확도 도출에 사용한다. 이를 통해 기존 워크 포워드 검증 방법론과 달리 딥러닝을 이용한 분석뿐만 아니라 전통적 머신러닝을 이용한 분석에도 충분히 이용될 수 있게 되었다. 이는 암호화폐 가격 예측에서 전통적 머신러닝 기법이 여전히 의미있는 예측정확도를 보인다는 점에서 더더욱 의미있다.

4. 실 험

본 논문에서는 [2]와 유사한 실험설계를 한 뒤 정확도가 얼마나 증가했는지를 통해 암호화폐 가격변동 경향의 휘발성이 존재함을 증명하려 한다. 아래 서술된 실험은 Google Colaboratory를 통해 진행되었다.

4.1 데이터셋

캐글(Kaggle)에 있는 이더리움 가격 데이터셋[2]을 통해 데이터를 수집하였다. 해당 데이터셋은 1시간을 단위시간으로 가지며, 해당 시간 내에 시작가격(Open), 마감가격(Close), 최고가격(High), 최저가격(Low)이 기입되어있다. 이러한 데이터셋을 불러온 뒤 2018년 2월 27일 00시부터 2018년 12월 3일 00시까지의 데이터만 사용하였다. 시간 내에 가격데이터 중에서는 시작가격 데이터만 사용한다.

4.2 데이터 분할

총 4종류로 데이터를 분할하였다. 8:1:1의 비율로 훈련, 검증, 실험 데이터를 1차적으로 분할하였고, 최종적으로 실험데이터를 검증하기 위해 훈련 데이터와 검증 데이터를 합친 제 2의 훈련 데이터를 추가로 저장해놓았다. 제 2의 훈련데이터는 실험데이터의 정확도를 검증하기 위한 학습데이터로 사용된다.

4.3 특성 공학(Feature Engineering)

[2]에서 데이터를 절삭시킨 뒤 연구를 진행했던 이유는 평균 대비 분산값이 너무 높아서 이를 안정화시키기 위함이었다. 하지만 절삭시킨 뒤에도 평균 220, 표준편차 122로 그다지 안정적인 통계값을 보여주지 못하였다. 이에 우리는 모든 가격 데이터를 이전 단위시간 대비 변화율로 변환하였다.

이후에는 [2]과 동일하게 이전 단위시간의 가격데이터를 입력값에 추가시켰다. 기존에 특정 단위시간에 존재하는 가격 데이터를 P_t라 하고, 여기서 t가 해당 단위시간을 의미하는 것이라 가정하면, 해당 단위시간의 입력값으로 P_t뿐만 아니라 P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_{t-5}까지 들어간 것이다.

그리고 마지막으로 단위시간 직후의 가격변동률을 이용해서 라벨(Label) 데이터를 생성했다. 라벨(Label) 데이터가 후술할 분석에서 종속변인으로 작용할 것이다.

4.4 분석 모델

4.4.1 로지스틱 회귀분석

로지스틱 회귀분석은 전통적 머신러닝의 방법론 중 하나로, [2]에서도 딥러닝을 포함한 모든 분석모델 중 예측율 2위를 기록하였다. 로지스틱 회귀분석은 독립변수를 시그모이드 함수를 거치게 하여서 0과 1사이 값을 만들어내며, 해당 값에서 0.5이상일 경우 1, 아닐 경우 0의 라벨 값을 예측해내도록 설계되었다. 해당 로지스틱 회귀분석은 사이킷 런(Scikit-Learn) 라이브러리를 이용하였다.

4.4.2 Support Vector Machine(SVM)

로지스틱 회귀분석과 마찬가지로, SVM은 전통적 머신러닝의 방법론 중 하나이다. [2]에서는 딥러닝을 포함한 모든 분석모델 중 예측율 3위를 기록하였다. SVM은 학습데이터를 기반으로 각 분류값의 결정경계를 학습하여 추후 들어오는 데이터를 결정경계를 기준으로 구분해준다. 해당 SVM은 사이킷 런(Scikit-Learn) 라이브러리를 이용하였다.

4.5 결과 비교방법

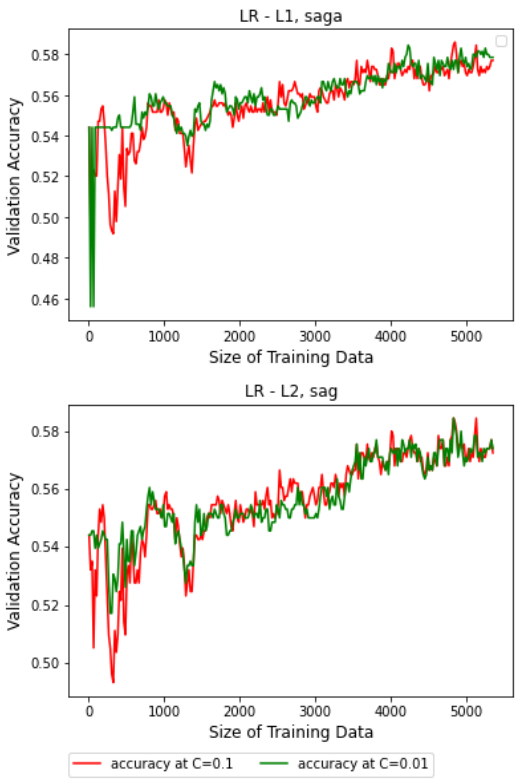
라벨을 이진변수로 지정했기에 예측정확도(Prediction accuracy)에 기반해서 결과를 비교할 것이다. 가격이 다음 단위시간에 오르는 비율과(라벨 기준 1)
내려가는 비율(라벨 기준 0)이 크게 비대칭을 이루고 있지 않기 때문이다. 비교대상이 되는 베이스라인(Bassline) 정확도는 각 모델의 제 2의 훈련 데이터를 절삭시키지 않고 온전히 학습에 이용했을 때의 정확도로 정하였다.

5. 실험결과

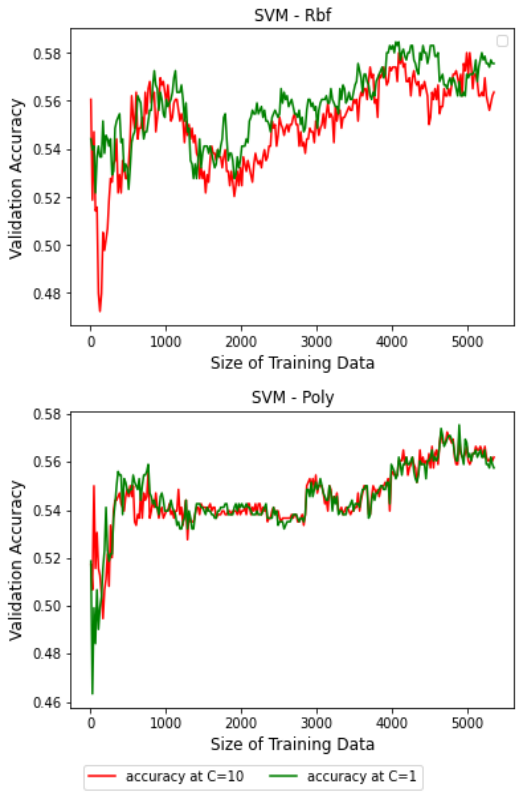
그림 3은 세부조건을 다양하게 분화시킨 상태에서 훈련 데이터의 크기에 따른 로지스틱 회귀분석 검증 정확도를 나타낸 것이다. 비교의 객관성을 위해 Penalty 파라미터로 L1, L2를 모두 사용했으며, L1 Penalty의 최적화 알고리즘으로는 saga, L2 Penalty의 최적화 알고리즘으로는 sag를 기용하였다. 또한 정규화 강도를 조절하는 C 파라미터로는 L1-saga와 L2-sag에 동일하게 0.1과 0.01을 모두 적용하였다. 그림 3의 위의 그래프는 L1 Penalty, saga 알고리즘을 이용했을 때의 그래프이고 밑에 그래프는 L2 Penalty, sag 알고리즘을 이용했을 때의 그래프이다.

그림 4는 세부조건을 다양하게 분화시킨 상태에서 훈련 데이터의 크기에 따른 SVM 검증 정확도를 나타낸 것이다. 비교의 객관성을 위해 Kernel 파라미터로 rbf, poly를 사용했다. 또한 정규화 강도를 조절하는 C 파라미터로는 rbf와 poly에 동일하게 10과 1을 모두 적용하였다.

2 https://www.kaggle.com/datasets/prasoonkottarathil/ethereum-historical-dataset?select=ETH_1H.csv



(그림 3) 훈련 데이터 크기에 따른 검증 정확도(로지스틱 회귀분석)
(Figure 3) Validation accuracy on the size of Train data(Logistic regression)



(그림 4) 훈련 데이터 크기에 따른 검증 정확도(Support Vector Machine)
(Figure 4) Validation accuracy on the size of Train data(Support Vector Machine)

(표 4) 예측 결과 정확도
(Table 4) Predicted Results Accuracy

모델	세부 파라미터1	세부 파라미터2(C)	베이스라인(Bassline)	Result of RWFV
로지스틱 회귀분석	L1 + saga	0.1	57.0%	58.1%
		0.01	55.9%	57.0%
	L2 + sag	0.1	57.4%	58.5%
		0.01	57.4%	58.8%
Support Vector Machine	rbf	10	56.5%	57.5%
		1	57.0%	58.9%
	poly	10	55.0%	56.0%
		1	55.0%	56.2%

그림 4의 위의 그래프는 rbf Kernel, 아래의 그래프는 poly Kernel일 경우의 그래프이다.

그림 3과 그림 4를 보면 모든 방법론에서 훈련 데이터의 크기가 커질수록 검증 정확도가 높아지는 것을 볼 수 있다. 하지만 훈련 데이터의 크기가 가장 큰 상태가 가장 높은 검증 정확도를 보이지는 않는다. 또한 대부분의 그래프에서 중간에 검증 정확도의 상승이 정체되거나, 아예 하락하기도 한다. 이는 문제정의에서 언급한 '가격변동 경향의 휘발성'으로 인해 바뀌는 경향이 아예 새로운 것이 아닌, 정해진 경향이 주기적으로 주 경향으로 상승했다가 다른 경향에게 자리를 양보하는 패턴의 반복으로 해석할 수 있다.

예를 들어, 가격변동 경향이 2개여서 훈련 데이터가 중간크기일 때에는 검증 데이터의 경향과 아예 다른 경향의 데이터가 학습데이터로 추가되어 검증 정확도가 정체 및 하락한 것일 수 있다. 또한 훈련데이터가 중간보다 더 큰 크기를 가질 경우 검증 데이터의 경향이나 훈련 데이터 초반부의 경향과 유사한 경향의 데이터가 학습데이터로 추가되어 검증 정확도가 상승한 것으로 해석할 수 있다. 물론 이러한 경향이 2개 이상일 수도 있고, 한 번에 여러 개의 경향이 동시에 적용되었을 확률도 있다. 하지만 그 어떤 상황에서도 경향은 주기성을 가지며, 특정 시기에 반영된 가격변동 경향은 언젠가는 다시 가격변동에 영향을 미치게 된다는 것을 알 수 있다.

그림 3과 그림 4를 통해 각 모델과 세부 파라미터 값에서 검증 정확도가 최고치인 훈련 데이터 크기를 도출해냈으며, 해당 크기만큼 절삭시킨 제 2의 훈련 데이터와 절삭시키지 않은 제 2의 훈련 데이터의 실험 데이터에 대한 정확도를 표 4에서 비교해보았다.

[2], [11], [12]에서 제시된 것처럼 암호화폐 가격예측 연구의 정확도는 대부분 60%를 넘기기 힘들다. 즉, 그 안에서 조금이라도 정확도의 상승이 있다면 작은 수치여도 충분히 의미가 있다고 판단하였다. 또한 이러한 정확도 상승이 우연이 아님을 증명하기 위해 하나의 모델 안에서 4개의 세부 파라미터 조건을 설정하였고, 총 8종류의 실험에서 정확도가 전부 상승했음을 알 수 있다. 이를 통해, 본 논문에서 제시되었던 역순 워크 포워드 검증 방법론이 유의미한 예측 정확도 상승을 가져온다고 볼 수 있다.

6. 결 론

본 논문에서는 시계열 데이터 분석으로서 주가 예측 연구와 암호화폐 가격 예측 연구들이 비슷한 방법론, 비슷한 틀 아래 연구되어왔다는 점을 지적하였다. 통계적, 개념적인 비교를 통해 주식과 암호화폐의 차이점을 찾아내었고, 이를 '암호화폐 가격변동 경향의 휘발성'이라 명명하였다. 그리고 이러한 휘발성을 분석 방법론에 반영하기 위해 기존 시계열 분석에서 쓰이던 워크 포워드 검증 방법론을 응용한 역순 워크 포워드 검증 방법론을 개발하여 제안하였다. 역순 워크 포워드 검증 방법론은 암호화폐의 가격변동 경향은 시간의 흐름에 따라 바뀐다는 것을 전제로 만들어졌다. 분석대상이 되는 실험 데이터로부터 얼마나 시계열상으로 떨어진 데이터까지를 훈련 데이터로 쓸 지를 실험 데이터 바로 앞 시간대에 배치된 검증 데이터로 검증해낸다. 이를 통해 어떤 시점까지 훈련 데이터로 기용하는 것이 더 좋은 예측 정확도를 보이는지를 알아낼 수 있게 되었다.

새로 개발한 역순 워크 포워드 검증 방법론을 검증하기 위해 로지스틱 회귀분석과 SVM을 분석 기법으로 기용하였으며, 해당 방법론의 정확도 향상이 일반적임을 증명하기 위해 각 분석기법마다 4개의 세부 파라미터 설정값을 실험모델로 채택하였다. 암호화폐 가격 예측 연구에서는 여전히 전통적 머신러닝이 예측 정확도 측면에 있어서 딥러닝보다 다소 우세하기에 전통적 머신러닝을 분석 기법으로 채용한 것이다. 이로인해 워크 포워드 검증 방법론을 이용한 예측정확도가 베이스라인(Bassline)이 아닌 시계열상으로 선행되는 90%의 데이터를 학습데이터로 한 예측정확도가 베이스라인(Bassline)이 되었다. 전술했듯이 워크 포워드 검증 방법론은 딥러닝에 적용할 것을 전제로 한 방법론이기 때문이다.

그 결과 총 8개의 실험모델에서 모두 베이스라인(Bassline) 대비 정확도가 상승한 것으로 나타났다. 특히 SVM에서는 최대 1.9%p의 예측 정확도 상승을 보여주었다. 암호화폐 선행연구들이 대부분 50%~60%사이의 예측 정확도를 보인다는 점을 감안했을 때, 이는 상당한 수준의 예측 정확도 개선이라 볼 수 있을 것이다.

Acknowledgement

This work was supported by the National Research Foundation of Korea Fund of NRF-2022R1F1A1063961.

참고문헌(Reference)

[1] Karasu, S., Altan, A., Saraç, Z., & Hacioğlu, R., "Prediction of Bitcoin prices with machine learning

methods using time series data.", 2018 26th signal processing and communications applications conference (SIU), pp. 1-4. 2018.
https://doi.org/10.1109/SIU.2018.8404760

[2] Chen, M., Narwal, N., & Schultz, M., "Predicting price changes in Ethereum", International Journal on Computer Science and Engineering (IJCSE) ISSN, 0975-3397, 2019.
http://cs229.stanford.edu/proj2017/final-reports/5244039.pdf

[3] Almasri, E., & Arslan, E., "Predicting cryptocurrencies prices with neural networks.", 2018 6th International Conference on Control Engineering & Information Technology (CEIT), pp. 1-5, 2018.
https://doi.org/10.1109/CEIT.2018.8751939

[4] Tandon, S., Tripathi, S., Saraswat, P., & Dabas, C., "Bitcoin price forecasting using lstm and 10-fold cross validation.", 2019 International Conference on Signal Processing and Communication (ICSC), pp. 323-328, 2019.
https://doi.org/10.1109/ICSC45622.2019.8938251

[5] Greaves, A., & Au, B., "Using the bitcoin transaction graph to predict the price of bitcoin", 2015.
http://snap.stanford.edu/class/cs224w-2015/projects_2015/Using_the_Bitcoin_Transaction_Graph_to_Predict_the_Price_of_Bitcoin.pdf

[6] Huy, N. H., Dao, B., Mai, T. T., & Nguyen-An, K., "Predicting cryptocurrency price movements based on Social Media.", 2019 International Conference on Advanced Computing and Applications (ACOMP), pp. 57-64, 2019.
https://doi.org/10.1109/ACOMP.2019.00016

[7] Aggarwal, A., Gupta, I., Garg, N., & Goel, A., "Deep learning approach to determine the impact of socio economic factors on bitcoin price prediction", 2019 Twelfth International Conference on Contemporary Computing (IC3), pp. 1-5, 2019.
https://doi.org/10.1109/IC3.2019.8844928

[8] Serafini, G., Yi, P., Zhang, Q., Brambilla, M., Wang, J., Hu, Y., & Li, B. "Sentiment-driven price prediction of the bitcoin based on statistical and deep learning approaches", 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2020.
https://doi.org/10.1109/IJCNN48605.2020.9206704

[9] Lamon, C., Nielsen, E., & Redondo, E., "Cryptocurrency price prediction using news and social media sentiment.", SMU Data Sci. Rev, 1(3), 1-22, 2017.
http://cs229.stanford.edu/proj2017/final-reports/5237280.pdf

[10] Abraham, J., Higdon, D., Nelson, J., & Ibarra, J., "Cryptocurrency price prediction using tweet volumes and sentiment analysis", SMU Data Science Review, 1(3), 1, 2018.
https://scholar.smu.edu/datasciencereview/vol1/iss3/1/

[11] McNally, S., "Predicting the price of Bitcoin using Machine Learning", Diss. Dublin, National College of Ireland, 2016.
http://norma.ncirl.ie/2496/

[12] Mallqui, Dennys CA, and Ricardo AS Fernandes. "Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques.", Applied Soft Computing Vol. 75, 596-606, 2019.
https://doi.org/10.1016/j.asoc.2018.11.038

[13] Gong, J., & Sun, S., "A new approach of stock price prediction based on logistic regression model.", 2009 International Conference on New Trends in Information and Service Science, pp. 1366-1371, 2009.
https://doi.org/10.1109/NISS.2009.267

[14] MacKinnon, R. K., & Leung, C. K., "Stock price prediction in undirected graphs using a structural support vector machine.", 2015 IEEE/WIC/ ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) ,Vol. 1, pp. 548-555, 2015.
https://doi.org/10.1109/WI-IAT.2015.189

[15] Cakra, Y. E., & Trisedya, B. D., "Stock price prediction using linear regression based on sentiment analysis.", 2015 international conference on advanced computer science and information systems (ICACSIS), pp. 147-154, 2015.
https://doi.org/10.1109/ICACSIS.2015.7415179

[16] Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P., "Stock price prediction using LSTM, RNN and CNN-sliding window model.", 2017 international conference on advances in computing, communications and informatics (icacci), pp. 1643-1647,

2017.
https://doi.org/10.1109/ICACCI.2017.8126078
[17] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C., “Stock price prediction using news sentiment analysis.”, 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), pp. 205-208, 2019.
https://doi.org/10.1109/BigDataService.2019.00035
[18] Yoon, H. S., “Time Series Data Analysis using WaveNet and Walk Forward Validation. Journal of the Korea Society for Simulation”, 30(4), 1-8, 2021.
https://doi.org/10.9709/JKSS.2021.30.4.001
[19] Tran, T. N., & Phuc, D. T., “Grid search of multilayer perceptron based on the walk-forward validation methodology.”, International Journal of Electrical and Computer Engineering, 11(2), 1742, 2021.
http://doi.org/10.11591/ijece.v11i2.pp1742-1751
[20] Börjesson, Lukas, and Martin Singull. ″Forecasting financial time series through causal and dilated convolutional neural networks.″, Entropy 2020, 22(10), 1094, 2020.
http://dx.doi.org/10.3390/e22101094
[21] Butt, A., Khemka, G., & Warren, G. J., “What dividend imputation means for retirement savers”, Economic Record, 95(309), 181-199, 2019.
https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-4932.12468

◐ 저 자 소 개 ◑

안 현(Hyun Ahn)
2022년～현재 연세대학교 비즈니스빅데이터분석학과(석사과정)
관심분야 : 빅데이터, 인공지능.
E-mail : haka4700@yonsei.ac.kr

장 백 철(Beakcheol Jang)
2009년 North Carolina State University 컴퓨터과학과(공학박사)
2021년～현재 연세대학교 정보대학원 교수
관심분야: 인공지능, 빅데이터분석, 자연어처리, 무선네트워크
E-mail : bjang@yonsei.ac.kr