JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 1, pp. 17-26, January 2023
https://doi.org/10.9708/jksci.2023.28.01.017

A Comparative study on smoothing techniques for performance improvement of LSTM learning model

Tae-Jin Park*, Gab-Sig Sim**

*Instructor, College of Liberal Arts, Pukyong National University, Pusan, Korea
**Professor, Dept. of Human Health Care, Gyeongsang National University, Jinju, Korea

[Abstract]

In this paper, we propose a several smoothing techniques are compared and applied to increase the application of the LSTM-based learning model and its effectiveness. The applied smoothing technique is Savitky-Golay, exponential smoothing, and weighted moving average. Through this study, the LSTM algorithm with the Savitky-Golay filter applied in the preprocessing process showed significant best results in prediction performance than the result value shown when applying the LSTM model to Bitcoin data. To confirm the predictive performance results, the learning loss rate and verification loss rate according to the Savitzky-Golay LSTM model were compared with the case of LSTM used to remove complex factors from Bitcoin price prediction, and experimented with an average value of 20 times to increase its reliability. As a result, values of (3.0556, 0.00005) and (1.4659, 0.00002) could be obtained. As a result, since crypto-currencies such as Bitcoin have more volatility than stocks, noise was removed by applying the Savitzky-Golay in the data preprocessing process, and the data after preprocessing were obtained the most-significant to increase the Bitcoin prediction rate through LSTM neural network learning.

▸ **Key words**: LSTM/GRU learning model, Filters:Savitky-Golay/single exponential smoothing/weighted moving average, time series data, pre-processing

[요 약]

본 연구논문에서는 LSTM 기반의 학습 모델 적용과 그 효용성을 높일 수 있도록 몇 가지 평활 기법을 비교, 적용하고자 한다. 적용된 평활 기법은 Savitky-Golay, 지수 평활법, 가중치 이동 평균 등이다. 본 연구를 통해 비트코인 데이터에 LSTM모델 적용 시 보여준 결과 값보다 전처리 과정에서 적용된 Savitky-Golay 필터가 적용된 LSTM 알고리즘이 예측 성능에 유의미한 좋은 결과를 보였다. 예측 성능 결과를 확인하기 위해 비트코인 가격 예측에 따른 복잡 요인을 제거하는데 사용된 LSTM의 경우와 Savitzky-Golay LSTM 모델에 따른 학습 손실율과 검증 손실율을 비교하고 그 신뢰성을 높일 수 있도록 20회 평균값으로 실험하였다. 그 결과 (3.0556, 0.00005), (1.4659, 0.00002)의 값을 얻을 수 있었다. 결과적으로는 비트코인과 같은 암호화폐가 주식보다 더한 변동성을 가지는 만큼 데이터 전처리 과정에서 평활 기법(Savitzky-Golay)을 적용하여 잡음(Noise)을 제거하였으며, 전처리 후의 데이터는 LSTM 신경망 학습을 통해서 비트코인 예측률을 높이는데 가장 유의미한 결과를 얻을 수 있었다.

▸ **주제어**: LSTM/GRU학습모델, 필터:Savitky-Golay/지수 평활법/가중치 이동 평균, 시계열 데이터, 전처리

• First Author: Tae-Jin Park, Corresponding Author: Gab-Sig Sim
*Tae-Jin Park (csptj2@naver.com), College of Liberal Arts, Pukyong National University
**Gab-Sig Sim (gssim@gnu.ac.kr), Dept. of Human Health Care, Gyeongsang National University
• Received: 2022. 11. 07, Revised: 2022. 12. 22, Accepted: 2022. 12. 28.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

경제 규모가 거대하고 시장 수요에 영향을 주는 암호화폐들은 블록체인으로 구성되어 있으며, 개인의 거래 내용이 담긴 소규모 데이터 블록(Block)을 사슬(Chain) 형태로 무수히 엮은 것을 블록체인이라 정의하고 있다[1]. 이와 같은 블록체인 기반의 온라인 암호화폐는 금융 회사의 중앙 집중형 방식이 아닌 P2P(Peer to Peer) 방식 거래를 지향하는 탈중앙화 거래 방식이다. 여러 암호화폐들은 비트코인을 기반으로 파생된 기술인만큼 비트코인의 가격에 영향을 받을 수밖에 없다. 또한 시장 수요가 큰 비트코인은 시간의 흐름이 고려된 시계열 특성을 가지므로 시장에서의 가격을 예측하는데 여러 모델을 적용하거나 개선하여 활용할 수 있다. 현재 머신러닝을 중심으로 하는 다양한 예측 모델을 연구, 적용하고 있으나 시장에서 예측할 수 없는 외부적 데이터 변수가 많으며, 변동성이 매우 큰 만큼 정확한 가격 예측이 어렵다. 또한 다양한 예측 모델을 적용하는 과정에서 기존 시계열 데이터가 가지는 특성을 분석 또는 새로운 변수들을 대입할 수 있으며, 관련된 기법을 글로벌 시장 이슈에 맞추어 해석할 수도 있다[2-3]. 또한 비트코인 가격 예측 시 사용된 시계열 데이터에서의 기울기 소실을 효율적으로 개선하거나 복잡한 변수를 신경망 모델 알고리즘에 대입, 학습시킴으로써 비트코인의 가격 변동에 따른 예측 성능을 높이는데 그 의미를 두고 있다[4, 8]. 뿐만 아니라 비트코인 투자자들이 딥러닝 모형을 적용함으로써 투자전략에 따른 수익률 개선 효과가 있음을 보여주고 있다. 즉 LSTM 예측 가격을 이용한 이동평균선 교차전략을 통해서 기존 투자 전략(Buy & Hold) 또는 전통적인 통계적 이동평균선 교차전략의 수익률보다 더 나은 개선 효과를 보인다는 것이다. 결과적으로 비트코인 시장의 평균 수익을 초과하는 투자 전략에 따른 존재 가능성을 제시하는데 의미가 있다[5]. 현재 머신러닝을 중심으로 하는 다양한 예측 모델을 연구, 적용하고 있으나 시장에서 예측할 수 없는 외부적 데이터 변수가 많으며, 변동성이 매우 큰 만큼 정확한 가격 예측이 어렵다. 본 연구 논문에서는 신경망 학습 모델의 결과를 반영하기 전 전처리 과정에서 수행된 필터 기법들의 적용이 미래 가격 예측에 얼마나 영향을 줄 수 있는지, 적용된 필터 중 가장 의미 있는 효과를 보인 평활 기법은 무엇인지 실험을 통해서 비교 분석하는데 중점을 두었다. 즉 비트코인과 같은 암호화폐가 주식보다 더한 변동성을 가지는 만큼 데이터 전처리 과정에서 평활 기법을 적용했을 때 시계열 데이터 분석에 미치는 정도를 평가하도록 한다. 본 연구 논문에서는 첫째, 서론에서 연구의 목적과 실험적 의미를 작성하였으며, 둘째, 관련성 연구와의 차이점을 살펴본 후 셋째, 시계열 데이터 분석을 위해 의미있는 칼럼의 확인과 평활 기법의 적용, LSTM 모델을 통한 학습 결과를 보이도록 한다. 끝으로 평활 기법이 적용된 데이터가 LSTM 알고리즘의 적용으로 비트코인의 미래 가격 예측 성능을 높이는데 얼마나 기여하는지 그 효과성을 확인하는데 있다. 더불어 LSTM 학습 알고리즘의 효용성을 더욱 높일 수 있도록 평활 기법을 적용했다. 대표적인 평활 기법 중 Savitzky-Golay 필터, 지수 평활법, 가중치 이동 방식을 적용했다. 즉 평활의 유무가 가격 예측에 어떤 영향을 미치는지와 주어진 평활 기법을 적용함으로써 실제값 및 예측값 비교를 통한 정확성과 신뢰성 여부, 그리고 성능 평가를 높이는데 유용한 각 알고리즘의 특징을 비교하였다. 적용된 비트코인 데이터는 Investing.com 사이트(플랫폼)를 활용했다[6].

II. Preliminaries : Related Theory and Background

1. LSTM(Long Short-Term Memory) Model

LSTM은 순환 신경망(Recurrent Neural Network; RNN)의 한 종류로써 과거의 상태가 현재 학습에 영향을 미치지 못하는 순환 신경망의 기울기 소실에 따른 약점과 학습 속도의 더딘 단점을 극복하고자 Hochreiter와 Schimidhuber에 의해 소개되었다[7]. 즉 이와 유사한 학습 모델 셀 중의 하나인 RNN 신경망 모듈을 반복 연결시킨 체인과 같은 형태를 하고 있다.

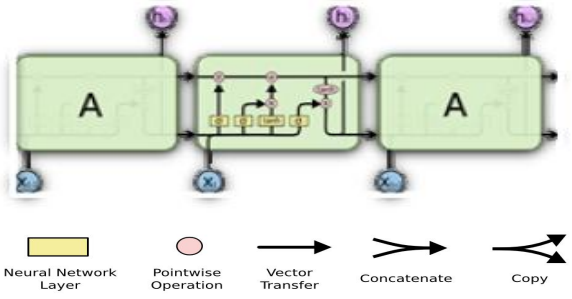


Fig. 1. Repeating Module of LSTM for Layer Interaction

LSTM 모델은 Fig. 1과 같이 입력에 대한 출력과 다음 단계의 입력을 계산하기 위해 사각형 모양의 셀 구조로 표시되며, 하나의 셀이 가지는 장기 기억 상태는 입력에 대한 출력을 계속해서 누적하여 값을 저장하고 단기 기억 상

태는 한 단계에 대한 출력 값을 저장하고 값을 반환한다. 또한 기울기 소실 문제를 해결하기 위해 Fig. 2와 같은 4개의 레이어 및 3개의 게이트를 만들고 각각의 게이트는 과거의 값을 얼마나 사용할지, 현재의 값을 얼마나 사용할지, 출력을 얼마나 내보낼지에 대한 연산으로 기울기 소실 문제를 해결한다[7-9].

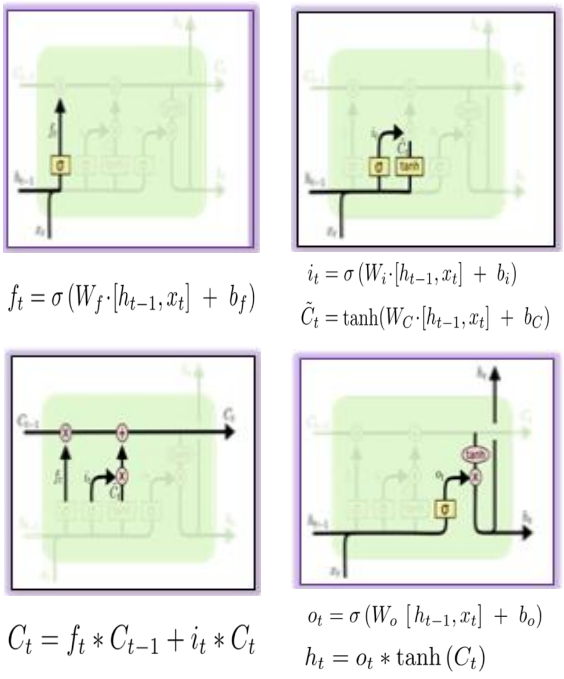


Fig. 2. Cell Status and Movement for Each Step

(a) Forget Gate Layer (b) Input Gate Layer
(c) Cell State Update (d) Cell State Output

시그모이드 계층(Sigmoid Layer)에 의해 결정되는 LSTM의 첫 단계는 셀 상태(Cell State)로부터 정보의 유용성을 확인하는 것으로 h_{t-1}과 X_t를 받은 후 0과 1 사이의 값을 C_{t-1}에 전달하게 되며, 그 값이 '1'이면 '정보 보존', 그렇지 않으면 '0'이 된다(Fig. 2(a)). 다음 단계는 입력되는 새로운 정보 중 갱신될 값을 시그모이드 계층에서 정하게 되며, 새로운 후보 값인 C_t벡터를 생성한다. 이와 같은 값들은 또 다른 갱신된 값을 만들게 된다(Fig. 2(b)). 이전 셀 상태인 C_{t-1}을 갱신 후 새로운 C_t를 생성, f_t와의 곱을 통해서 불필요한 값의 소멸과 새로이 갱신된 값만큼의 스케일을 정한다(Fig. 2(c)). 시그모이드 계층에 입력된 데이터와 필터처리 후의 출력 셀 상태를 정하고 계산된 결과 값(Sigmoid Gate)으로 유용한 값만을 취할 수 있다(Fig. 2(d)).

2. Smoothing Techniques

2.1 Savitzky-Golay Filter

Savitky-Golay 필터는 단위 시간당 진동 횟수로 표현되는 주파수에 포함된 많은 잡음을 제거함과 동시에 신호를 평활화하는데 효과적이며, 파형이 가지는 모양 및 높이 등 신호에 포함된 유용한 정보를 효율적으로 유지할 수 있다. 즉 Savitky-Golay 필터는 N-차수의 다항식 근사치에 기초하여 잡음 제거 효과를 높일 수 있도록 최소 자승법(LSM : Method of Least Squares)을 적용한 것이며, 차수를 높일수록 오류를 상당히 낮출 수 있는 만큼 효율적인 차수의 다항식을 찾는 것이 중요하다. 물론 차수를 높일수록 오류를 줄일 수 있으나 과적합의 문제가 발생할 수 있으며, 효율성을 고려해서 2차 다항식을 최소자승법에 적용하는 것이 바람직하다.

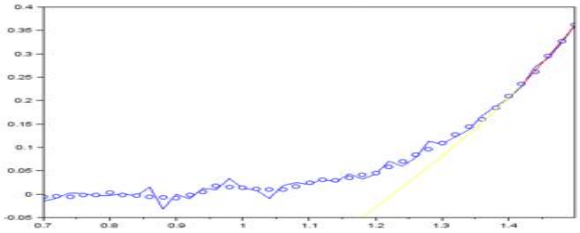


Fig. 3. The Smoothing Process of Signals Using a Multinomial Regression Model

Savitky-Golay는 어떤 데이터를 적용하더라도 계수가 동일하다는 점, 회귀선의 차수에 따른 계수를 정리함으로써 계산 시간을 단축시킬 수 있다는 점과 신호의 정보를 유지하면서 잡음 제거 효과에 탁월한 성능을 보인다는 것이다. 식 (1)에서와 같이 N차 회귀모델 $p(n)$은 평활화된 신호를 나타낼 수 있도록 식 (2)와 같이 표현된다. 즉 왼쪽 $-M$에서 오른쪽 $+M$개의 신호를 획득, $2M+1$길이의 신호에 대해서 N차 회귀모델 $p(n)$으로 대체하게 된다. 결과적으로 $p(n)$은 원 신호와의 오류를 줄일 수 있도록 계수 $a_k\,(k=0,1,2,...,N)$들로 구성되며(식 (3)), 식 (1 - 3)을 적용, 임의의 이산신호의 샘플 $x[n]$이 가지는 시간의 흐름에서 $n=0$ 중심의 입력 샘플 그룹에 대한 평균-제곱 근사 오차를 최소화한다[10-11].

$$p(n) = \sum_{k=0}^{N} a_k n^k \tag{1}$$

$$p(n) = a_0 + a_1 n + a_2 n + \ldots + a_N n \tag{2}$$

$$\epsilon_N = \sum_{n=-M}^{M} (p(n) - x[n])^2 \tag{3}$$

$$= \sum_{n=-M}^{M} (\sum_{k=0}^{N} a_k n^k - x[n])^2$$

식 (1) ~ 식 (3)에서의 수식 기호 $p(n)$← N차 회귀모델, n← 시간 샘플, $x[n]$← 임의의 이산신호: $-M \leq x \leq M$, a_k← 다항식 계수, M← 근사치 간격:half width로 나타낸다.

2.2 Simple Exponential Smoothing

시계열 분석을 위하여 불규칙적 변동 또는 추세 및 계절성을 포함하는 경우를 고려해볼 수 있다. 비트코인과 같은 시계열은 추세적인 요소 또는 시간 흐름에서의 평균값과는 무관하게 불규칙 변동에 따른 다양한 외적 변수가 큰 영향을 주므로 정해진 패턴을 탐색하는 것은 무의미하다. 더불어 급격한 변동성을 가지는 시계열 특성에서 여러 요인을 분석하고 비트코인의 미래 가격을 예측하는 것 또한 쉽지 않다. 따라서 급격한 변동성이 가지는 잡음을 제거할 수 있도록 평활을 진행하게 되며, 이 때 적용되는 기법이 단순 지수 평활법이다. 단순 지수 평활법은 최근의 데이터들에 대해 더 많은 가중치를 부여함으로써 새롭게 추가된 데이터가 계산에 반영될 수 있도록 한다. 이는 이동 평균과 같이 시계열 데이터 변화에 의한 평균값을 반영할 수 없다는 단점과 방대한 데이터가 계산에 포함되어야 한다는 단점을 보완한 것이다. 결과적으로 단순 지수 평활은 시계열 값들에 대한 가중 평균 평활값으로 미래 가격을 예측할 수 있다.

$$
\begin{aligned}
F_{n+1} &= \alpha Z_n + (1-\alpha)F_n \\
&= \alpha Z_n + (1-\alpha)[\alpha Z_{n-1} + (1-\alpha)F_{n-1}] \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2 F_{n-1} \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2[\alpha Z_{n-2} + (1-\alpha)F_{n-2}] \\
&\quad \cdots\cdots\cdots\cdots \\
&\quad \cdots\cdots\cdots\cdots \\
&\quad \cdots\cdots\cdots\cdots \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + \alpha(1-\alpha)^2 Z_{n-2} + (1-\alpha)^3 F_{n-3} + \cdots
\end{aligned}
\tag{4}
$$

시점 t에서 관측된 값은 단순 지수 평활 계수 α를 0~1 사이의 값으로 정한 뒤 예측값 F_1에서부터 F_{n+1}까지 계산을 수행한다. 식 (2)와 같이 계산된 결과는 시계열 관측값들에 대한 가중 평균값으로 미래 가격을 예측할 수 있으며, 지수 형태의 증가 또는 감소는 지수 평활 계수에 따라서 영향을 받는다. 여기서 F_{n+1}← t시점에서 추정한 시점인 $n+1$의 예측값, α← 지수 평활 계수, Z_n← n시점의 관측값을 나타낸다.

2.3 Weighted Moving Average

이동평균은 식 (5)에서와 같이 전체 데이터 집합의 여러 하위 집합에 따른 일련의 평균을 만들어 데이터 요소를 분석, 계산하는 방법이며, 산출 방식에 따라 단순이동평균, 지수이동평균, 가중이동평균 등으로 분류할 수 있다.

$$
\begin{aligned}
&(n\,day)\text{가중이동평균} = \\
&\frac{(n)\times P_0 + (n-1)\times P_{-1} + \ldots + (1)\times P_{-(n-1)}}{n + (n-1) + \ldots + 1}
\end{aligned}
\tag{5}
$$

P_0 : 당일의 시장 가격
P_{-n} : n일 전의 시장 가격

따라서 이동평균은 주식의 추세를 파악하는데 자주 사용되며, 주식 흐름의 일정 기간 중 주가에 따른 평균을 계산하기 때문에 무작위성의 제거 및 추세를 파악할 수 있는데 도움을 준다. 물론 주식뿐만 아니라 비트코인 데이터에도 적용을 시도할 수 있을 것이다. 가중이동평균이 가지는 장점은 최근 데이터에 더 높은 가중치를 둠으로써 단순 이동평균 기법들에 비해 최근의 시장 분위기를 잘 반영할 수 있다는 특징을 가진다.

III. Experimental Methods and Procedures

1. Selection and Method of Experimental Elements

본 연구 논문에서는 비트코인 가격 예측의 정확성을 높이기 위해 전처리 과정에서 3가지 평활 기법을 적용했으며, 데이터 학습에 사용된 모델은 RNN 기반의 GRU 및 LSTM 알고리즘이다. LSTM 및 GRU 모두 기울기 소실의 문제를 해결하기 위한 목적으로 개발되어 졌으나 데이터를 유지하는 방법에서는 차이점을 가지고 있다. LSTM에서는 '망각 게이트(Forget Gate)', '입력 게이트(Input Gate)', '출력 게이트(Output Gate)', 그리고 은닉층(Hidden Layer) 및 셀 상태(Cell State)라고 하는 계층을 통해서 오래된 정보를 효율적으로 보존한다. 즉 시그모이드 계층의 출력과 곱셈 연산을 통해 정보의 추가 또는 제거 기능을 수행한 후 업데이트될 값을 정하고 벡터를 생성, 기존 셀 상태에 추가함으로써 이전 정보와 신규 정보를 결합하고 새로운 셀 상태를 만든다. 이제 출력값은 새로운 변환값에 대한 곱셈 연산을 적용한 후 필터된 값으로 완성된다. GRU는 셀 상태와 출력 게이트를 포함하지 않기 때문에 적용되는 파라미터 또한 적으며, LSTM보다 학습 속도가 빠르다고 볼 수 있다. 또한 GRU 구조 측면에서 '리셋(Reset)'과 '업데이트(Update)'라고 하는 다른 게이트를 가지고 있다. 두 게이트는 이전 은닉층이 무시될지의 여부, 새로운 은닉 상태에 업데이트 될지 여부, 은닉 상태

가 무시될지의 여부, 업데이트에서 새로운 은닉 상태에 대한 반영 비율과 더불어 업데이트에 사용되기 위한 많은 정보량의 유지 또는 무시할지 여부를 결정하기 위해서 일시적 은닉층을 생성하게 된다. 또한 학습될 가중치 파라미터와 노이즈 벡터를 포함, 이 가중치에 따라서 은닉층 사이의 가중 평균을 계산한다. 이를 통해서 이전 정보의 반영 정도와 새로운 정보의 반영 정도를 결정하게 되는 것이다. 본 연구 실험에서는 비트코인이 가지는 시장 가격의 변동성과 미래 가격에 대한 예측 신뢰성을 높이는데 있으므로 학습 속도의 성능 향상보다 전처리 과정에서 평활 기법을 적용, 변동성을 낮추는데 있다. 본 연구 실험을 위해서 적용된 신경망 알고리즘은 학습 속도가 빠른 GRU보다 시계열 데이터가 가지는 특성과 비트코인이 가지는 정량적, 정성적 평가 변수 및 큰 폭의 거래량과 가격 변동에 따른 가중치를 고려해서 LSTM 모델을 선정했으며, 평활 기법이 적용된 시계열 데이터가 두 학습 모델에 적용되었을 때 미래 가격 예측에 따른 학습 훈련 효과를 비교해 본다. 본 실험에 사용된 데이터는 Investing.com에서 제공하는 비트코인의 값으로 일별 종가, 시가, 고가, 저가, 거래량, 변동 등 전체 7개의 칼럼으로 구성되어있다. 입력된 데이터는 전처리 수행 후 실험에 적용된 칼럼간의 상관관계를 통해서 주어진 데이터의 신뢰성을 제공하도록 했다. 다음은 전처리 후 적용된 데이터에서의 필터 적용 및 필터 성능을 검증할 수 있도록 평활 기법의 적용과 LSTM 모델 학습을 수행함으로써 학습의 신뢰성 및 예측값을 높이는데 중점을 두었다.

2. Data Processing and Correlation

실험 데이터의 기간은 비트코인의 첫 거래가 기록되어 있는 시점부터 최근 연구 시점까지의 데이터(4310×5)를 활용했다. 단계별 전처리 후의 실험 데이터(Fig. 4)는 아래와 같다.

【단계:1】 기존 csv파일에 대한 데이터프레임 생성 후 비트코인 특성상 '거래량' 칼럼을 제거한다.
【단계:2】 문자형으로 구성된 모든 칼럼 내 데이터 중에서 날짜 데이터가 가지는 문자를 제거한 후 시계열 데이터 타입으로 변경한다.
【단계:3】 숫자값으로 표시된 다른 칼럼 역시 데이터 내 문자를 제거한 뒤 실수 타입의 데이터로 변경한다.
【단계:4】 칼럼명은 데이터의 이름을 영문으로 변경한다.

(a) Data Frames before Preprocessing

	날짜	종가	오픈	고가	저가	거래량	변동 %
0	2022년 04월 12일	40,073.0	39,507.0	40,678.0	39,293.0	581.36M	1.46%
1	2022년 04월 11일	39,497.0	42,144.0	42,418.0	39,202.0	608.38M	-6.27%
2	2022년 04월 10일	42,138.0	42,760.0	43,421.0	41,884.0	255.83M	-1.47%
3	2022년 04월 09일	42,767.0	42,275.0	42,809.0	42,129.0	165.16M	1.16%
4	2022년 04월 08일	42,275.0	43,450.0	43,979.0	42,113.0	467.83M	-2.70%
...	...	...	...	...	...	...	...
4282	2010년 07월 22일	0.1	0.1	0.1	0.1	2.16K	0.00%
4283	2010년 07월 21일	0.1	0.1	0.1	0.1	0.58K	0.00%
4284	2010년 07월 20일	0.1	0.1	0.1	0.1	0.26K	0.00%
4285	2010년 07월 19일	0.1	0.1	0.1	0.1	0.57K	0.00%
4286	2010년 07월 18일	0.1	0.0	0.1	0.1	0.08K	0.00%

4287 rows × 7 columns

(b) Data Frame after Preprocessing

Date	Closings	Open	High price	Low price	Fluctuation
2010-07-18	0.1	0.0	0.1	0.1	0.00
2010-07-19	0.1	0.1	0.1	0.1	0.00
2010-07-20	0.1	0.1	0.1	0.1	0.00
2010-07-21	0.1	0.1	0.1	0.1	0.00
2010-07-22	0.1	0.1	0.1	0.1	0.00
...	...	...	...	...	...
2022-05-01	38461.0	37642.0	38676.0	37397.0	2.15
2022-05-02	38514.0	38472.0	39134.0	38061.0	0.14
2022-05-03	37718.0	38515.0	38647.0	37513.0	-2.07
2022-05-04	39688.0	37717.0	40021.0	37660.0	5.22
2022-05-05	36844.0	39686.0	39833.0	36560.0	-7.17

4310 rows × 5 columns

Fig. 4. Bitcoin Data

분석에 사용된 칼럼은 Closings(종가), Open(시가), Low price(저가), High price(고가)로 구분하였으며, 데이터 분석에 사용될 핵심 칼럼을 피어슨 상관 계수를 통해 확인하였다. 데이터 간의 상관관계를 통해 계산된 최솟값이 Fig. 5에서와 같이 적어도 0.99 이상인 만큼 적용 데이터에는 큰 편차가 없음을 알 수 있다. 이 상관관계를 토대로 본 실험에서는 Closings(종가)를 핵심 분석 데이터로 선정하였다.

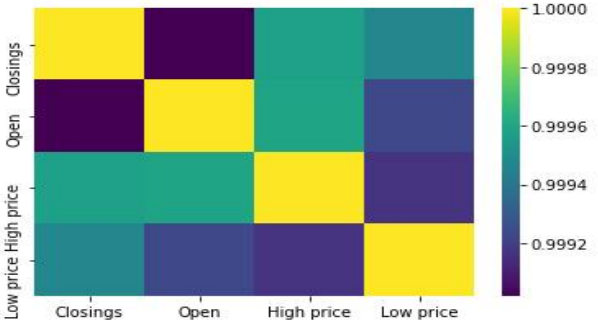


Fig. 5. Pearson Correlation for Selecting Analytical Data

3. Filter Application and Performance Validation Using LSTM

전처리 단계 후 적용된 비트코인 데이터를 Savitky-Golay, 단순 지수 평활법, 그리고 가중치 이동평균 등 평활 기법을 적용한다. 더불어 적용된 평활 기법이 시계열 데이터 분석에 미치는 정도를 평가하도록 한다. 비트코인의 경우 외부적 요인에 의한 가격 변동이 큰 만큼 급격한 변화를 제거하는 것이 중요하다. 평활화된 데이터는 LSTM 알고리즘을 통해서 학습되고 검증됨으로써 미래 가격 예측 성능을 높이는데 얼마나 기여하는지 그 효과성을 확인하는데 있다. Fig. 6은 '종가'에 따른 의미 있는 비트코인 데이터를 연도별 '가격'으로 표시한 그래프이다.

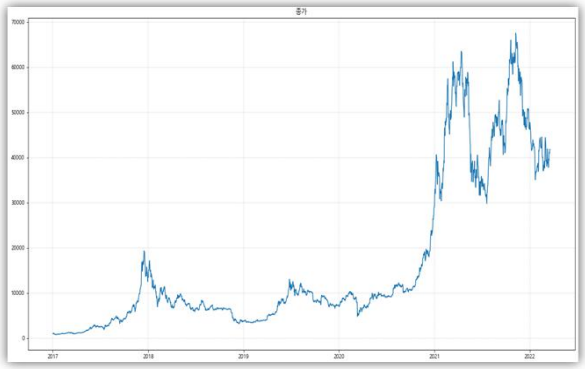


Fig. 6. Bitcoin Data Based on Annual Price(2017~2022)

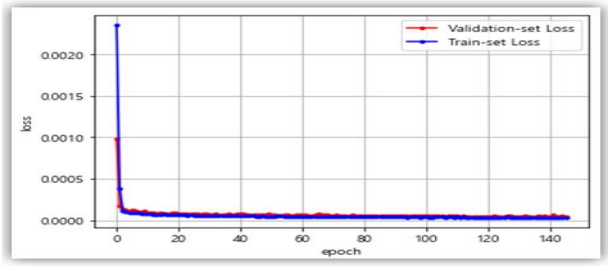
3.1 Comparison of Prediction Model Algorithms: GRU and LSTM

GRU와 LSTM 알고리즘은 비트코인과 같은 시계열 데이터 분석 차원에서 적용되는 기울기 소실 문제를 해결하기 위해서 개발된 모델이다. 주어진 모델의 효과성 확인을 위해 케라스(Keras)에서 신경망 생성과 밀집층(Dense Layer), 모델 객체 생성 후 학습 훈련 전의 사용될 손실 여부와 측정 지표 등을 지정하였다. 추가해서 밀집층이 많은 신경망일수록 그 효과를 높이는데 유용한 렐루(Relu)함수를 적용하였다. 모델의 컴파일과 학습 훈련을 위해서 적용된 요소는 Table 1과 같으며, 훈련 손실과 검증 손실 그래프는 Fig. 7과 같다.

Table 1. Application and Measurement Elements of Neural Networks Based on Smoothing Techniques

Classification	Measurement elements	Measurement element (optional)
Experimental Data : Data Set	Train / Valid	923 / 231
	Training set	test_size=730 window_size=20
	Validation set	0.2
Hyper parameters : Dense	Number of epochs	200
	Batch Size	16
	Dense : Neurons	1
	Activation Function	Relu
	Callbacks	EarlyStopping : patience=10
Compiling	Optimizer	adam
	Loss	mean_squared_error
Model (Save/Restore)	File Save : Format	HDF5

(a) Application of LSTM Learning Model



(b) Application of GRU Learning Model

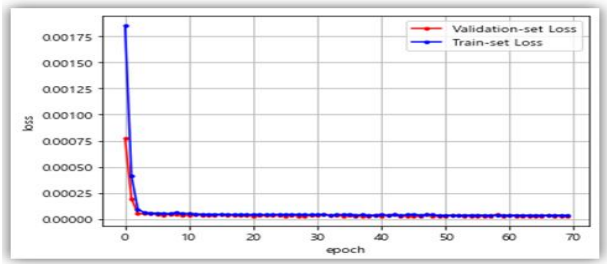


Fig. 7. Loss of Learning Training and Validation

학습 훈련 손실과 검증 손실 그래프에 따른 결과를 살펴보면 두 모델에서 큰 차이를 보이지 않는 것으로 보이나 본 실험 데이터 셋을 통해서 적용된 LSTM 학습 모델이 훈련과정에서 좀 더 효율적으로 수렴되고 있음을 알 수 있다.

3.2 Performance Assessment Based on Filter Application

시계열 데이터가 가지는 특성과 비트코인이 가지는 정량적, 정성적 평가 변수 및 큰 폭의 거래량과 가격 변동에 따른 가중치 적용 등을 고려하면서 예측한다는 것이 무의미할 수도 있다. 따라서 신경망 학습 전 적용된 평활 기법의 선택에 따라 실제값에 수렴해 가는 정도를 달리할 수 있으며, 수렴해가는 정도 차이를 분석하는 것만으로도 유의미한 결과를 보일 수 있다는 점에서 아래 실험 단계에 맞추어 그 결과를 나타내고자 한다. 다음은 LSTM 예측 알고리즘 적용 시 유의미한 결과 도출이 가능할 수 있는 필터를 선정하였으며, 그 성능 평가를 비교 검증할 수 있도록 유용한 평활 기법을 나열하였다.

【단계:1】 평활 기법의 미적용에 따른 LSTM 학습 알고리즘 및 예측 성능을 분석한다.
【단계:2】 Savitky-Golay 평활 기법의 적용과 LSTM 학습 시 예측 성능을 분석한다.
【단계:3】 지수평활 기법의 적용과 LSTM 학습 시 예측 성능을 분석한다.
【단계:4】 가중치 이동 기법의 적용과 LSTM 학습 시 예측 성능을 분석한다.

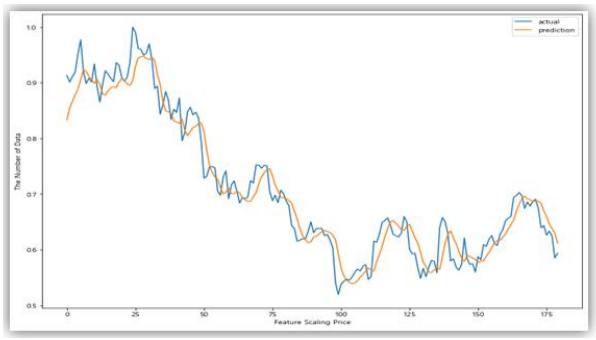


Fig. 8. Learning Algorithm of LSTM Model Without Smoothing Techniques

평활 기법이 적용되지 않은 비트코인 데이터에서의 LSTM 모델은 주어진 시계열 데이터가 시간의 흐름에 따라 점차적으로 실제값과 예측값이 수렴되고 있음을 알 수 있다(Fig. 8).

(a) LSTM Learning Model with Savitky-Golay Smoothing Technique

(b) LSTM Learning Model with Exponential Smoothing Technique

(c) LSTM Learning Model with Weight Moving Technique

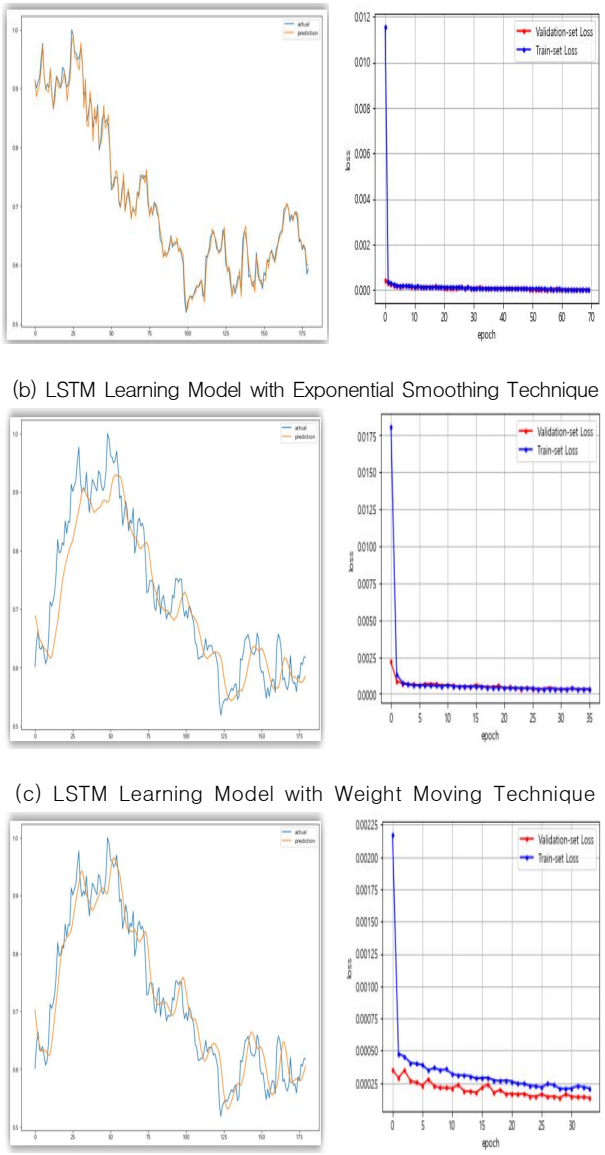


Fig. 9. Evaluation of Predictive Performance Based on the Application of the Smoothing Technique (Left:Predictive Performance Evaluation, Right:Loss of Learning Training and Validation due to Epoch)

지수평활 기법이나 주식 시장과 파생 상품 시장을 기술적으로 분석할 때 사용하는 가중치 이동 기법의 경우 시간의 흐름에 따라 반영되는 시계열 데이터에 가중치를 조정함으로써 미래 가격을 예측한다는 특징을 가지고 있으며, 방대한 데이터를 가지면서 비트코인 가격 변동이 적은 패턴을 보일 경우에서 좋은 성능을 예측할 수 있다. 그러나 비트코인 데이터의 급격한 변화에서는 가중치 비율을 조정하는 것이 쉽지 않은 만큼 예측의 신뢰성에도 한계가 있다. Savitky-Golay 필터는 예측값과 실제값이 아주 유사했다. 디지털 데이터 포인트에 적용하는 필터임에도 불구

하고 상당히 좋은 성능을 보여주었다. Table. 2와 Table. 3에서 보듯이 시계열 데이터 분석과 더불어 평활 기법의 적용만으로도 신경망 학습에 좋은 효과가 있음을 확인할 수 있었다.

Table 2. Comparison of Learning Loss and Validation Loss According to Application of LSTM and Smoothing Technique

Model&Filter	Train Loss	Validation Loss
LSTM	2.085	0.00026
Savitzky-Golay : LSTM	1.906	0.00012
Exponential Smoothing : LSTM	8.464	0.00063
Weighted Moving Average : LSTM	5.477	0.00054

Table 3. Comparison of learning loss and validation loss according to application of LSTM/GRU and smoothing technique(2020 times) : 365 days

Model&Filter	Train Loss	Validation Loss
LSTM	3.0556	0.00005
GRU	3.4426	0.00003
Savitzky-Golay LSTM	1.4659	0.00002
Savitzky-Golay GRU	4.7761	0.00003
Exponential Smoothing : LSTM	7.7591	0.00009
Exponential Smoothing : GRU	8.8429	0.00007
Weighted Moving Average : LSTM	7.4146	0.00006
Weighted Moving Average : GRU	6.0544	0.00006

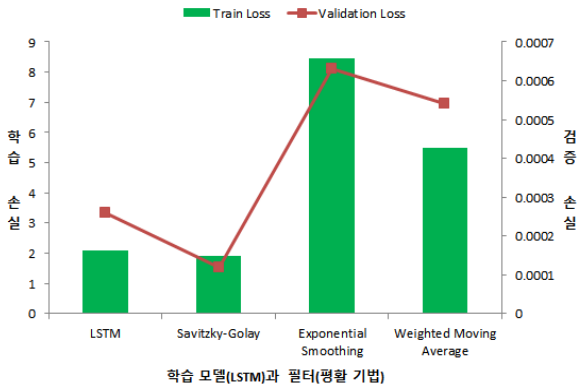
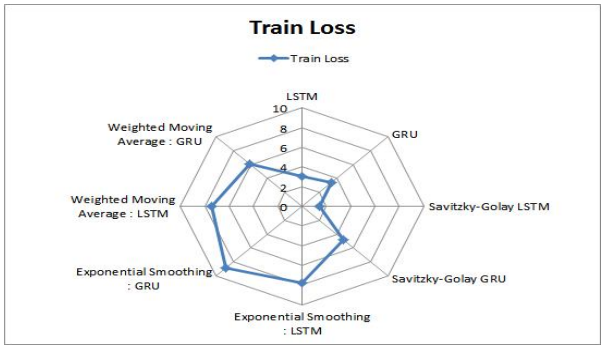


Fig. 10. Comparison of Neural Network Learning Models with LSTM and Smoothing Techniques

(a) Comparison of Train Losses in LSTM and GRU Models



(b) Comparison of Validation Losses in LSTM and GRU Models

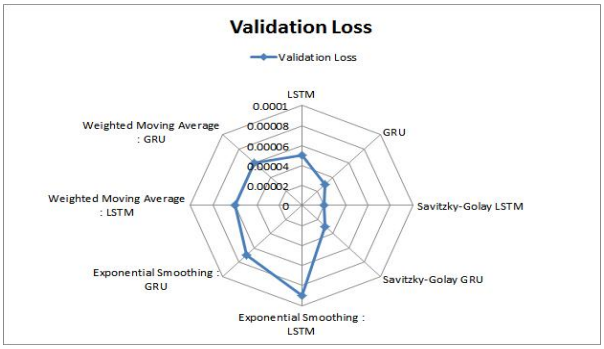


Fig. 11. Comparison of Train Loss and Validation Loss of LSTM&GRU Model According to the Application of Smoothing Technique:20 iterations

Fig. 10, Fig. 11에서 보듯이 Savitky-Golay 필터가 다른 평활 기법에 비해서 학습 및 검증 손실을 통한 성능 평가에도 좋은 결과를 보이고 있다. 더불어 필터가 적용되지 않은 신경망 학습 알고리즘(LSTM)과 비교해도 필터 적용 후 신경망 학습에 미치는 영향 또한 유의미한 결과를 보이고 있음을 알 수 있다.

IV. Conclusions

본 연구를 위해서 적용된 실험 분석 기간은 비트코인의 첫 거래가 기록되어있는 시점부터 최근 연구 시점까지의 데이터(4310×5)를 활용했으며, 실험 분석 기간은 비트코인의 첫 거래일 2010년 07월 18일부터 2022년 05월 05일까지 적용되었으며, 분석 기간의 시작 비트코인 종가/시가는 121.45(원)/0(원)이고, 최근 연구 시점에서의 종가/시가는 50,595(천원)/47,575(천원)을 나타내고 있다. 본 연구 논문에서는 신경망 학습 모델의 결과를 반영하기 전 전처리 과정에서 수행된 필터 기법들의 적용이 미래 가격 예측에 얼마나 영향을 줄 수 있는지, 적용된 필터 중 가장 의

미 있는 효과를 보인 평활 기법은 무엇인지 실험을 통해서 비교 분석하는데 중점을 두었다. 결과적으로 신경망 학습 전 적용된 평활 기법의 선택에 따라 실제값에 수렴해 가는 정도에도 차이가 있음을 실험을 통해 알 수 있었다. 연구 실험에서 적용된 신경망 알고리즘은 학습에 따른 미래 가격 예측 효율성을 높일 수 있도록 LSTM 모델을 선정하였다. 본 실험 결과를 통해서 비트코인 데이터에 LSTM모델 적용 시 보여준 결과 값보다 전처리 과정에서 적용된 Savitky-Golay 평활 기법이 급격한 기울기에 따른 문제를 별도의 조정 없이도 충분한 개선효과를 주었으며, 본 알고리즘(Saviztky-Golay)이 적용되었을 때 가장 높은 예측 값을 유추할 수 있었다. 학습 손실율과 검증 손실율을 비교해보면 LSTM의 경우 (2.085, 0.00026), Savitzky-Golay LSTM은 (1.906, 0.00012) 값으로 분석되었으며, 20회 평균값으로 실험한 결과 (3.0556, 0.00005), (1.4659, 0.00002)의 값을 얻을 수 있었다. 이러한 결과를 통해 시계열 데이터의 전처리 과정만으로도 비트코인 예측률을 높이는데 유용한 의미를 가질 수 있으며, 평활 기법이 시계열 데이터 분석에도 유용하게 적용되었다. 향후 연구에서는 좀 더 다양한 평활 기법들의 적용과 동시에 시간, 분 단위의 빠른 변화에 대응하면서 예측 성능 평가를 높일 수 있도록 비정형 쿼리의 특성을 최대한 활용한다. 시계열 특성을 가지는 비트코인 데이터는 미래 가격 예측을 하는데 변동성이 매우 크므로 외적 요인에 따른 급격한 변화와 가중치를 고려해서 데이터 셋을 관리하는 것이 효율적일 수 있다. 즉 비정형 쿼리 특성을 가지는 MongoDB의 컬렉션(Collection)을 생성, 삽입함으로써 데이터 셋을 효율적으로 운영할 수 있을 것이고 컬렉션 설정을 달리 하는 것만으로도 기대 효과를 높일 수 있다. 더불어 MongoDB의 인덱싱 기능을 적극적으로 활용한다면 학습 데이터 및 검증 데이터 관리에 따른 계산 능력 향상과 비트코인 데이터 거래에 따른 내역을 저장하거나 분석함에 있어서도 가치 있는 결과를 반영할 수 있다.

REFERENCES

[1] Satoshi Nakamoto, "Bitcoin : A Peer-to-Peer Electronic Cash System,", March. 2009, www.bitcoin.org

[2] S.Y. Hong, S.R. Cho, S.H. Kim, "Blockchain Beyond Bitcoin," Electronics and Telecommunications Research Institute, Vol. 32, No. 1, pp. 72-81, Feb. 2017.

[3] Euiseok Kim, "A Study for the Innovativeness of Blockchain," The Journal of Society for e-Business Studies, Vol. 23, No. 3, pp. 173-187, August 2018.

[4] Jun-Ho Kim, Hanul Sung, "A Study on the Hyper-parameter Optimization of Bitcoin Price Prediction LSTM Model," Journal of The Korea Convergence Society, Vol. 13. No. 4, pp. 17-24, April 2022. DOI: https://doi.org/10.15207/JKCS.2022.13.04.017

[5] S. W. Kim, "Performance Analysis of Bitcoin Investment Strategy using Deep Learning,", Journal of the Korea Convergence Society, Vol. 12, No. 4, pp. 249-258, April 2021.

[6] Investing.com [Internet], https://kr.investing.com/

[7] S Hochreiter, J Schmidhuber, "Long Short-Term Memory," Neural Computation, Vol. 9, No. 8, pp. 1735-1780, November 1997.

[8] Colah, "Understanding LSTM Networks,", August 2015. https://colah.github.io/posts/2015-08-Understanding-LSTMs/,

[9] Ji-Soo Hwan, "Bitcoin time series data prediction using machine learning," KANGWON national univ., pp. 1-31, 2019.

[10] Schafer, R. W., "What is a Savitzky-Golay filter?,", [lecture notes] IEEE Signal Processing Magazine, Vol. 28, No. 4, pp. 111-117, July 2011. DOI: 10.1109/MSP.2011.941097

[11] R. W. Schafer, "On the frequency-domain properties of Savitzky-Golay filter,", in Proc. 2011 DSP/SPE Workshop, Sedona, AZ, pp. 54–59, Jan. 2011.

Authors

Tae-Jin Park was received the B.S., in 1988, and then received M.S. and Ph.D. degrees in an Computer Science from Pukyong national university, Pusan, Korea, in 1995 and 2008, respectively.

He was served as a visiting professor in department of shipbuilding information technology(before, Computer Science) in Geoje university, Geoje city, Korea, 2003 from 2000. And that He was worked part-time instructor and as a visiting professor at College of liberal-arts-course, Silla University, Pusan, Korea, 2000 to 2015. He is currently lecturing at Busan National University, Pukyong National University's College of Liberal Arts, and Department of Computer Science at Gyeongsang National University. He is interested in control system of big data-based, Artificial Intelligence, classification and extraction of learning content.

Gab-Sig Sim received the B.S., M.S. and Ph.D. degrees in Computer Science and Statistics from Chonnam National University, Gwangju, Korea, in 1985, 1987 and 1993, respectively.

Prof. Sim joined the faculty of the Department of Liberal Arts at Jinju National University, Jinju, Korea, in 1993. He was a visiting professor in the Dept. of Computer Science at San Jose State University in San Jose, California, from March 2004 to February 2005. He is currently a professor in the Department of Human Health Care at Gyeongsang National University, Jinju, Korea. He is interested in Digital Health Care, Artificial Intelligence, and Neural Network.