Journal of The Korea Convergence Society
Vol. 12. No. 11, pp. 99-107, 2021

I-SSN 2233-4890 / e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2021.12.11.099

주성분 분석 기법을 활용한 시계열 데이터 분석 및 예측 시스템

진영훈[1], 지세현[1], 한군희[2*]
[1]백석대학교 스마트IT공학부 교수, [2]백석대학교 컴퓨터공학부 교수

Time Series Data Analysis and Prediction System Using PCA

Young-Hoon Jin[1], Se-Hyun Ji[1], Kun-Hee Han[2*]
[1]Professor, Division of Smart IT, Baekseok University
[2]Professor, Division of Computer Engineering, Baekseok University

요 약 우리는 무수히 많은 데이터 속에서 살고 있다. 다양한 데이터는 우리가 활동하는 모든 상황 속에서 만들어지는데 빅데이터 기술을 통해 데이터의 유의미를 발굴한다. 유의미한 데이터를 발굴하기 위해 많은 노력이 진행 중이다. 본 논문은 주성분 분석(Principal component analysis) 기법으로 시계열 데이터의 추이 및 예측을 통해 인간이 더 나은 선택을 가능케 하는 분석 기법을 소개한다. 주성분 분석은 입력된 데이터를 통해 공분산을 구성하고, 데이터의 방향성을 추론할 수 있는 고유벡터와 고윳값을 제시한다. 제안하는 방법은 비슷한 방향성을 갖는 시계열 데이터 집합에서 기준 축을 구성하고, 데이터 집합을 이루는 각 시계열 데이터들의 방향성이 기준 축과 이루는 사잇각을 통해 다음 구간에 존재하게 될 데이터의 방향성을 예측한다. 본 논문에서는 가상화폐의 추이를 통해 제시한 알고리즘의 정확도를 LSTM(Long Short-Term Memory)과 비교 검증한다. 비교/검증 결과 제안된 방법은 변동성이 큰 데이터에서 LSTM에 비해 상대적으로 적은 트랜잭션과 높은 수익(112%)을 기록하였다. 이는 상대적으로 정확하게 신호를 분석하여 예측했다는 의미로 볼 수 있으며, 보다 정확한 임계치 설정을 통해 더 나은 결과를 도출할 수 있을 것으로 기대된다.

주제어 : 주성분 분석, 예측, 분석, 빅데이터, 암호화폐, 핀테크, 시계열 데이터

Abstract We live in a myriad of data. Various data are created in all situations in which we work, and we discover the meaning of data through big data technology. Many efforts are underway to find meaningful data. This paper introduces an analysis technique that enables humans to make better choices through the trend and prediction of time series data as a principal component analysis technique. Principal component analysis constructs covariance through the input data and presents eigenvectors and eigenvalues that can infer the direction of the data. The proposed method computes a reference axis in a time series data set having a similar directionality. It predicts the directionality of data in the next section through the angle between the directionality of each time series data constituting the data set and the reference axis. In this paper, we compare and verify the accuracy of the proposed algorithm with LSTM (Long Short-Term Memory) through cryptocurrency trends. As a result of comparative verification, the proposed method recorded relatively few transactions and high returns(112%) compared to LSTM in data with high volatility. It can mean that the signal was analyzed and predicted relatively accurately, and it is expected that better results can be derived through a more accurate threshold setting.

Key Words : PCA, Prediction, Analysis, BigData, CryptoCurrency, Fintech, Time Series

*This research was supported by 2021 Baekseok University Research Fund.

*Corresponding Author : Kun-Hee Han(hankh@bu.ac.kr)

Received October 3, 2021 Revised November 11, 2021
Accepted November 20, 2021 Published November 28, 2021

1. 서론

연속성을 갖는 데이터의 방향성은 외적 요인이 많을수록 예측하기 어렵다. 일반적으로 통계학에서는 회귀분석 방법을 통해 예측한다. 회귀분석은 연속적으로 측정되는 데이터들에 대해 각 변수 사이를 설명할 수 있는 적합한 모델을 계산하여 최적의 적합도를 갖는 모델을 선택하는 방식이며 선택된 모델은 측정될 데이터를 예측할 수 있는 근거가 된다. 하지만 변수가 많은 상황에서는 정확한 가설을 설정하기 어려운 단점을 갖는다. 고전역학이 물질에 작용하는 힘과 운동 관계를 파악하면 물질의 위치를 정확하게 예측할 수 있지만, 물질이 매우 빠르게 움직이는 계와 미시적인 계에서는 물질의 움직임을 예측하기 어려운 것과 같은 이치이다.

연속된 데이터의 정확도 높은 예측은 다양한 산업구조에서 필요하다. 예를 들어, 최근 기후변화는 전 세계에 많은 위험요소를 초래한다. 뿐만 아니라 국가는 경제변동을 예측하여 경제계획을 위한 예산을 책정하고, 국제정세에 대비한다. 의료, 에너지, 곡물, 주식, 유가, 센서 데이터 등 연속된 데이터의 분석 및 예측은 우리 생활과 밀접한 관계를 맺는다. 이렇듯 연속된 데이터에서 패턴을 찾고, 예측모델을 구성하여 미래를 예측하거나, 인간의 행동을 추정하기 위한 다양한 기법들이 연구되고 있다.

연속된 데이터의 패턴 분석은 딥러닝, 분류기, 군집화 알고리즘 등을 통해 분석하는데, 과적합(OverFitting), 가중치 설정, 설명 불가능한 출력 결과, 계산량 및 적절한 데이터 수집 등 다양한 난제가 존재한다. 이를 극복하고자 차원 축소, 빠른 계산 처리, 쉬운 구현, 적은 파라미터 설정 등의 장점이 있는 주성분 분석을 많이 활용한다.

본 연구는 주성분 분석의 기하학적 의미를 근간으로 시계열 데이터 분석/예측방법을 소개한다. 논문의 구성은 관련 연구와 주성분 분석 및 시계열 데이터 분석에 자주 쓰이는 LSTM을 2장에서 소개하고, 제안하는 방법을 3장에서, 실험을 통한 증명을 4장 평가 및 분석에서 다루고, 5장에서 결론을 맺는다.

2. 이론적 배경

시계열 분석은 주어진 데이터의 패턴을 통해 시간적 인과관계를 찾는 과정으로 많은 학계에서 연구된다.

이응환은 6가지 빅데이터를 통해 독립변수를 선정하고, IBM SPSS MODELER 18을 사용하여 모델링하였다[1]. 송상화는 효율적인 난방을 위해 지역에서 난방수요를 정확하게 예측하고, 난방 에너지 생산 계획 최적화를 위해 사용량 패턴을 분석하여 회귀분석 기반 난방수요 예측 모형을 제시하였다[2]. 김태훈은 KNN과 AdaBoost가 프로야구 승패 예측에 최적 모델임을 증명하였다[3]. 나광택은 SHAP Value를 이용하여 해석 가능한 증권사 금융 고객의 이탈예측 모델을 제시했다[4]. Pramod B S는 LSTM을 이용하여 주가를 예측하는데, 확률적 경사 하강법을 사용하여 각 데이터 포인트에 대해 프로세스 가중치를 수정하였다. 제시된 알고리즘은 다른 주가 예측 알고리즘보다 더 정확도가 높다고 주장한다[5-10]. Myoung-Jong Kim은 주가지수를 예측하기 위해 다중 분류기 접근 방식을 제안하였다. 다중 분류기의 조합은 추정 오차의 분산을 줄이고, 전반적인 분류 정확도를 향상할 수 있는 방법이다[11]. 김상호는 유전자 알고리즘을 활용하여 주가를 예측하였다. 관계도가 높은 주가지수를 활용하는데, 관계도 높은 지수를 찾기 위해 유전자 알고리즘을 사용하였다[12]. N.H.M. Radzi는 SVM(Support Vector Machine) 및 PCA(Principle Component Analysis)를 사용하여 교통사고에서 가장 주요한 사망 요인을 분석/예측하는 방법을 제안하였다[13]. Tsun-Kuo Lin은 SVM과 PCA를 통해 멀티 센서 패턴 인식 시스템을 제시하였다. 다양한 센서의 데이터를 PCA를 통해 저차원으로 구성하고, SVM을 이용하여 분류하는 방식으로 다른 패턴을 동시에 감지할 때 여러 감지 장치를 사용하지 못할 수 있기 때문에 수행하기 어려운 단점을 극복하였다[14]. Mohd. Mustaqeem은 소프트웨어 결점 탐지를 위해 과적합 방지, 다양한 커널 트릭 등 적은 데이터에서 강점을 갖는 SVM과 시간 복잡성 및 분석의 강인함을 더하기 위해 PCA를 결합한 하이브리드 모델을 제안하였다[15]. 백승훈은 기온 변화에 따른 판매량 예측모델을 제시하였다[16]. 이윤정은 주식 포트폴리오 추천을 위해 주식시장 네트워크를 분석하고, K-Means 및 Kruskal 최소신장 트리를 통해 추천알고리즘을 구성하였다[17].

일반적으로 딥러닝은 시계열 데이터의 정확한 예측을 위해 많은 데이터가 필요하며, 과적합 오류를 극복해야 한다. 또한, 정확한 예측을 위한 feature 설정이 어렵고, 학습에 많은 시간이 필요하다. SVM은 초평면의 마진을 최대로 하는 이진 분류기로 범주나 수치예측 및 노이즈에 강하다. 뿐만 아니라 과적합 확률이 낮고, 딥러닝에 비해

사용이 쉽다. 하지만 커널과 모델에 대해 많은 테스트가 필요하고, 입력데이터 셋이 많을수록 학습에 많은 시간이 필요하다. 신경망과 더불어 결과에 대한 해석이 불가능한 단점 역시 존재한다. 반면 PCA는 데이터의 차원을 축소하여 빠르게 처리할 수 있고, 비교적 노이즈에 강한 특성을 보인다. 즉 다차원 데이터에서 방향성을 찾아 데이터의 주된 흐름을 파악하여 특징을 분석하는 데 적합한 방식이다.

PCA를 활용하는 기존의 연구들은 데이터의 차원 축소를 통해 기계학습에서 과적합을 방지하고, 학습 속도를 높일 수 있는 특징점을 찾는 데 활용한다. 반면 본 논문이 제안하는 방법은 고윳값과 고유벡터의 기하학적 특성에 초점을 맞춰 데이터를 분석하고, 더 정확한 예측 가능성을 증명할 것이다

본 논문은 2차원 시계열 데이터 분석/예측을 위해 계산이 빠르고, 비정상성(non-stationary) 데이터에 강인한 특성을 갖는 PCA를 사용하여 시계열 분석에서 자주 활용되는 LSTM과 비교/ 분석을 통해 제안된 방법을 검증한다.

2.1 LSTM

LSTM은 RNN(Recurrent Neural Network)의 한 종류인데, 일반적으로 신경망은 과거의 사건은 별개로 라벨링된 데이터의 출력에 관심을 갖는다. 과거의 데이터가 현재의 데이터에 영향을 미치는 데이터 구조에는 적합하지 않다. 이를 해결하기 위해 RNN을 활용하게 된다. 하지만 RNN은 Fig. 1 (a)와 같이 긴 의존 기간에 어려움을 겪기 때문에 긴 기간에 적합한 LSTM을 이용한다[18].

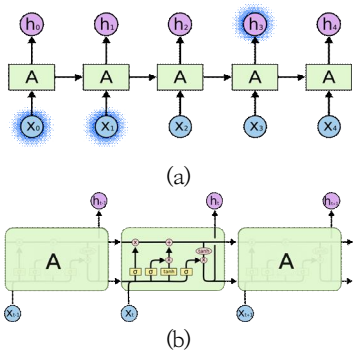


Fig. 1. RNN and LSTM Characteristics of RNN and LSTM[19]

(a) A strong RNN in the short term

(b) The repeat module of LSTM has four interactive layer structures

Fig. 1의 (b)를 보면 LSTM은 cell state에서 정보의 사용 여부를 정하거나 cell state에 정보 저장을 결정하는 레이어를 각각 두어 cell state의 갱신을 처리하고, 출력 데이터를 정한다[19]. 따라서 시계열 데이터를 예측하기에 적합하다.

2.2 주성분 분석

주성분 분석은 통계적 접근법으로 Fig. 2와 같이 데이터들의 분산이 가장 큰 방향 e1과 e1에 수직이며 e1 다음으로 분산이 큰 방향 e2, e3 등을 추정할 수 있는 기법이다[20,21].

주성분 분석은 식 (1), (2)와 같이 공분산 행렬(Covariance Matrix)을 구성하고, 공분산 행렬을 고윳값 분해(Eigen Decomposition)하여 방향벡터를 얻는다. 고윳값 분해는 $Ax=\lambda x$가 성립하는 0이 아닌 벡터 x와 고윳값 λ 계산이 가능하며 식(3)과 같다.

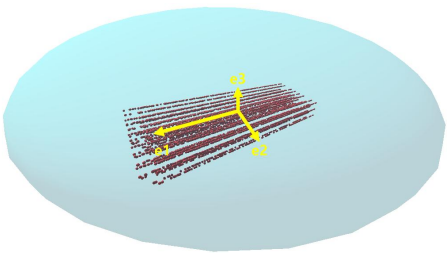


Fig. 2. Principal Component Analysis Eigenvector Estimation

$$cov(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)} \quad (1)$$

$$C_m=\begin{vmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{vmatrix} \quad (2)$$

$$\begin{aligned} &Ax=\lambda x \\ &(A-\lambda I)x=0 \\ &\det(A-\lambda I)=0 \end{aligned} \quad (3)$$

Where, A denotes n×n matrix, I denotes identity matrix, x denotes eigenvector, λ denotes eigenvalue

주성분 분석은 데이터가 이루는 주된 방향성과 신호의 강도를 파악할 수 있고, 데이터의 차원을 축소하여 많은 데이터를 빠르게 처리할 수 있으며, 비교적 노이즈에

강하다. 또한, 딥러닝과 다르게 분석 결과를 이해할 수 있는 장점이 있다.

2.3 특징점 설정

PCA를 통한 시계열 데이터 분석에 있어 의미있는 특징점을 추출하기 위해 주성분의 고윳값을 이용한다. Fig. 3은 고윳값을 연결한 그래프로 고윳값이 큰 순으로 정렬하여 구성되며, 각 고윳값에 대응하는 분산의 비율이 80~90%의 영역에 속하는 특성만을 이용한다. 즉 대부분의 정보를 포함하는 상위 특성들을 선택하면 더 나은 예측이 가능하게 된다.

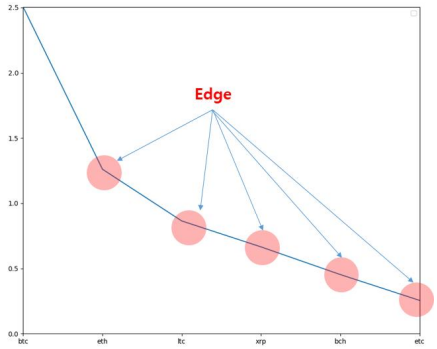


Fig. 3. Variance Chart for Feature Selection

3. 제안모델

시계열 데이터는 Fig. 4와 같이 일정한 시간 간격으로 나열되는 데이터들의 집합을 의미한다.

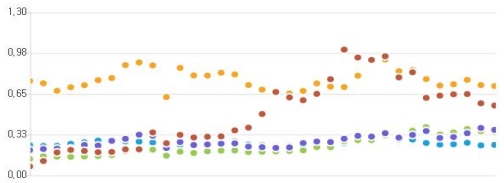


Fig. 4. Time Series Data

종합 주가지수, 유가 지수, 환율 등은 시계열 데이터로 볼 수 있고, 시계열 데이터 해석은 이후 도래하는 사건을 예측하는 도구로 활용된다.

시계열 데이터를 분석하기 위해 본 논문은 같은 의미가 있는 다양한 지표를 활용한다. 예를 들어 주가를 예측하기 위해 같은 추이를 갖는 데이터의 집합을 이용한다. 다시 말해, 같은 종목군의 데이터를 통해 각 종목의 추이를 분석한다. 주성분 분석은 많은 데이터에서 방향성을 찾아 데이터의 흐름을 분석한다. 이때 데이터의 수가 적으면 작은 변화에 민감하게 반응하므로 정확한 예측이 불가능하게 된다. 따라서 비슷한 특성을 갖는 시계열 데이터를 통해 기준이 되는 축을 구성한다. 기준 축이 구성되면 식 (4)와 같이 각 시계열 데이터의 방향성과 이루는 각을 통해 각 시계열이 앞으로 증가 혹은 감소할지 판별하게 된다. 식 (5)는 각 시계열 데이터의 축이 기준 축의 왼편 혹은 오른편에 위치하는지 판별하며, 결과값이 $\vec{v_1} \times \vec{v_2} > 0$ 이면 기준 축 대비 왼편에 놓이고, 이는 기준 축 대비 증가량이 크다는 것을 의미한다.

$$\theta = \vec{v_i} \cdot \vec{v_{std}} \tag{4}$$

$$det = \vec{v_i} \times \vec{v_{std}} \tag{5}$$

Where, $\vec{v_i}$ denotes each time series eigenvector, $\vec{v_{std}}$ denotes standard Axis

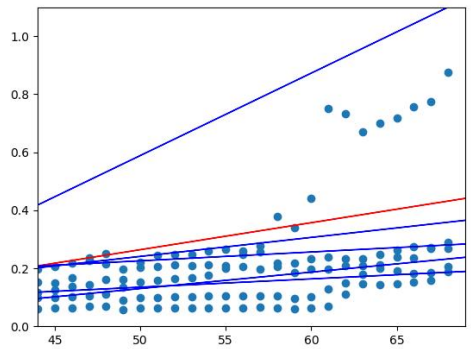


Fig. 5. Time Series Analysis with PCA

Fig. 5는 비슷한 방향성을 갖은 5개의 시계열 데이터를 통해 방향성을 계산한 모습이다. 빨간색 직선은 기준 축이고, 파란색 직선은 각 시계열 데이터의 방향성이다. 이 중 기울기가 기준 축에 비해 가파른 데이터는 상승 추세로 전환됨을 예측할 수 있게 된다.

기준 축을 생성하기 위해서 각 시계열 데이터는 식 (6)과 같이 정규화 과정이 필요하며, 식 (7)과 같이 기울기의 변화량을 통해 상승 강도를 예측할 수 있고, 기준 축의 상승은 각 시계열의 상승 가능성을 의미한다.

$$\frac{d_i}{d_{\max}} \tag{6}$$

$$\frac{\vec{v}_t - \vec{v}_{t-1}}{\triangle t} \tag{7}$$

Where, d_i denotes a element in each time series, $d_{\max}$ denotes max element in all time series

4. 평가 및 분석

실험은 Python 및 가상화폐 시계열 데이터를 이용하며 Fig. 6의 절차를 따른다. 먼저 시계열 데이터를 PCA를 통해 분석한다. 2.2절과 같이 분산을 통해 특징점을 선정하고, 선정된 데이터를 통해 기준 축이 될 고유벡터와 각 시계열의 고유벡터를 계산한다. 기준 축과 각 시계열 축들은 내적과 외적을 통해 추천 여부를 결정하게 된다. 외적의 결과 양수이면 기준 축 대비 더 상승 국면임을 의미하며, 내적의 결과가 클수록 상승 가능성은 높다고 볼 수 있다.

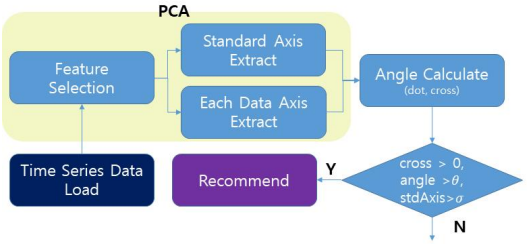


Fig. 6. Flow Chart of Proposed Method

제안된 방법의 효율성을 증명하기 위해 LSTM(Long Short-Term Memory) 알고리즘을 이용한 예측 시스템과 비교한다.

시계열 데이터는 bithumb에서 제공하는 데이터를 활용하며, 활발한 거래가 이루어지는 상위 5개 종목을 설정한다. 기간은 2017년 9월 8일 부터 2021년 8월 1일이다. 이렇게 구성된 프로그램은 Fig. 7과 같다.

Fig. 8은 2017/09/08부터 종가를 기준으로 30일 데이터를 분석한 결과이며, BCH는 0.282도를 이룬다. 하지만 기준 축의 상승도가 낮기 때문에 상승 여력은 미지수이다. 10일 후 BCH는 기준 축과 0.364도를 이루며 상승 추세에 있음을 알 수 있다. 기준 축 역시 기울기가 상승 추세임을 보인다. 이와 같이 각 축의 기울기를 통해 상승 혹은 하강 추세를 예측하여 데이터의 흐름을 분석할 수 있다. 이후 BCH는 0.272, 0.265도를 나타낸다.

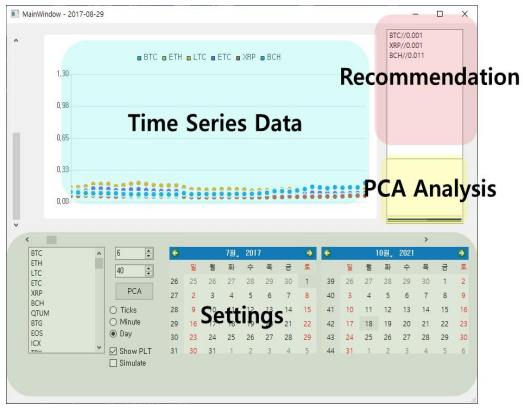


Fig. 7. MainWindow for Analysis with PCA

Fig. 8. Standard Axis and each Data Axis (10 days each)

(a) 2017-09-08 (b) 2017-09-18
(c) 2017-09-28 (d) 2017-10-08

Fig. 9는 Fig. 8의 PCA 분석을 확대한 결과이다. 검은색 축은 기준 축이 되고, 나머지 축은 각 시계열 데이터의 색과 일치한다. Fig. 9의 (a)에서 기준 축은 수평을 이루고, BCH 데이터는 증가추세를 나타낸다. (b)에서 BCH는 기울기가 증가하며 기준 축과 이격을 보인다. 따라서 확실한 증가 추세임을 알 수 있다. (c)는 이격이 더 증가하지만, 중심 축과 이루는 각이 (b)보다 감소함을 알 수 있고, (d)는 이격이 증가하지만 기울기가 하락함을 보이는데 이는 기준 축의 상승세가 있기 때문이다. 이와 같이 시계열 데이터 분석에 있어 중요한 각 축의 기울기와 이격을 통해 시계열 데이터 추이를 분석할 수 있다.

상승 조건을 분석하면 기준 축의 기울기가 수평보다 σ만큼 클 때의 각 시계열 데이터와 기준 축의 사잇각 θ가 일정 범위 이상일 때 증가이다. 반대로 사잇각 $\theta < 0$ 이고, 기준 축 기울기가 σ보다 작으면 하락 조건이다.

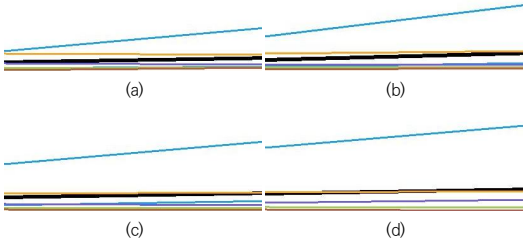


Fig. 9. PCA Analysis with Standard Axis(black) and each Data Axis(10 days each)

(a) 2017-09-08 (b) 2017-09-18
(c) 2017-09-28 (d) 2017-10-08

Fig. 9의 기간에 LSTM 알고리즘을 이용하여 BCH를 분석하면 Fig. 10과 같다. 2017-08-04부터 35개의 데이터로 윈도우를 구성하고, 전체 데이터의 90%를 학습하여 예측하면 거의 근사하게 예측하는 것을 볼 수 있다. Fig. 9와 같은 기간의 데이터를 예측해 보면 실제 데이터와 예측치의 차이가 크지 않음을 알 수 있다. 하지만 LSTM의 특성상 하루 이상의 예측은 가능치 않고, 그림에서와 같이 예측치가 실제 데이터를 후행하는 모습을 보인다. 다시 말해 예측이라기 보다 과거 데이터와 예측치의 오차를 줄여 최적화하기 위한 학습 결과로 풀이된다.

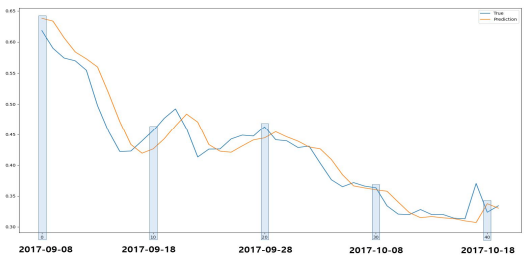
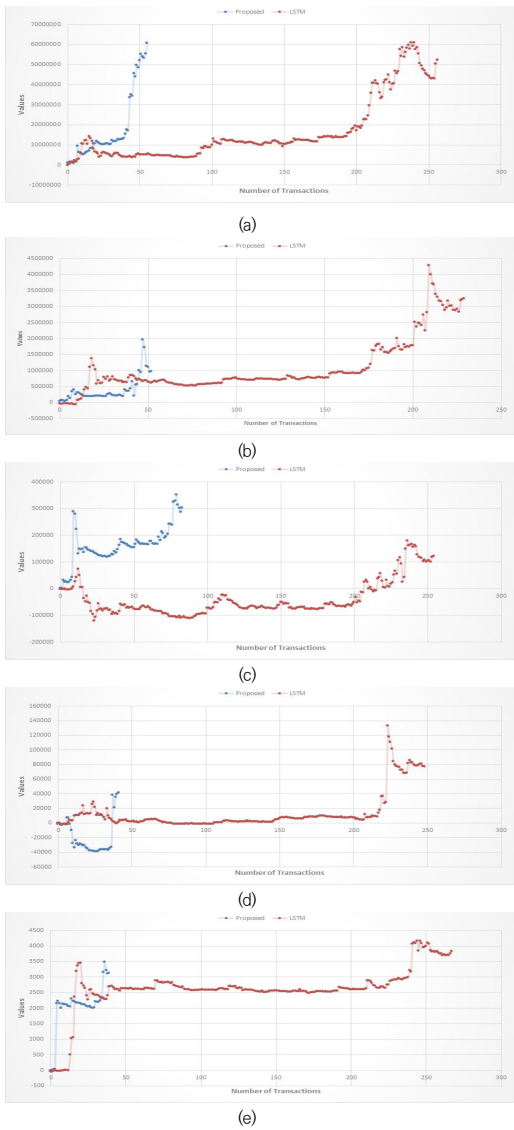
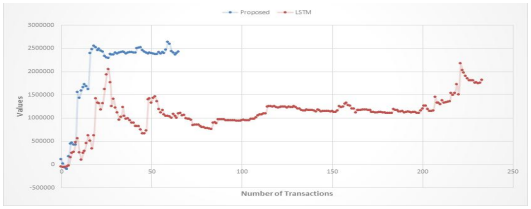


Fig. 10. Time Series Analysis with LSTM

Fig. 11은 두 알고리즘을 통한 누적 수익을 나타낸다.





(f)

Fig. 11. Comparison of revenue between Proposed Method and the LSTM (Axis X : Transactions; Axis Y : Values; Proposed Method : blue; LSTM : red)

(a) BTC (b) ETH (c) LTC (d) ETC (e) XRP (f) BCH

Fig. 11 (a)에서 제안된 방법(파란색)은 50여 회 남짓 거래한 모습이고, LSTM(빨간색)은 250여 회 거래한 모습이다. 최종 수익은 제안된 방법이 16% 정도 높음을 알 수 있다. (a), (c), (f)는 제안된 방법이, (b), (d), (e)는 LSTM이 좋은 결과를 나타낸다. 아쉽지만 (b), (d)는 눈에 띄게 LSTM이 좋은 결과를 보인다.

전체적으로 두 알고리즘을 바라보면 테이블 1과 같이 제안된 방법은 적은 횟수의 거래가 일어났음을 알 수 있고, 꾸준히 증가하는 모습을 보인다. 반면 LSTM은 거래 횟수가 많고, 특히 변동이 없는 구간에도 거래가 많다. 또한, 90%의 데이터를 학습했지만, 실험 구간이 100% 포함된 결과임에도 몇몇 데이터에서는 제안된 방법이 더 나은 결과를 보인다.

테이블 1에서 거래량이 적으며 수익이 높다는 의미는 비교적 정확하게 신호를 분석하여 예측했다는 의미로 볼 수 있다. 다시 말해 의미 없는 신호에 흔들리지 않았음을 보이는 지표이다.

Table 1. Cumulative Revenue for Each Data

ticker(P/L)	Proposed		LSTM	
	Transaction	Revenue	Transaction	Revenue
BTC(1.16)	55	60,670,000	256	52,284,000
ETH(0.29)	52	965,400	229	3,252,300
LTC(2.50)	83	303,590	254	121,360
ETC(0.53)	41	41,455	248	77,298
XRP(0.81)	39	3,129	267	3,830
BCH(1.33)	65	2,426,100	233	1,819,050
Total(1.12)	335	64,409,674	1,487	57,557,838

5. 결론

시계열 데이터의 흐름에서 의미 있는 패턴을 찾기 위한 다양한 연구가 진행되었지만, 복합적인 변수에 의해 예측이 어려운 측면이 존재한다. 기존의 신경망을 이용한 알고리즘에서 LSTM은 장기 의존성에 알맞아 많이 이용되지만, 딥러닝 특성상 어떤 패턴에 의해 출력되는지 알 수 없고, 비정상성 데이터의 분석에 대응하기 어려운 단점을 갖는다. 반면 제안된 방법은 비슷한 방향성을 갖는 데이터들을 중심으로 PCA를 통해 주된 방향성을 중심축으로 설정하고, 각 시계열 데이터의 방향성과 비교 분석하는 방식이다. 비정상성 데이터 집합에서 고유벡터를 중심으로 분포하는 데이터의 집합이 확률적으로 정상성에 가까울 것을 상정하여 방향성을 예측한다.

실험 결과를 통해 제안된 방법은 상대적으로 적은 트랜잭션과 비슷하거나 우위에 있는 누적 수익을 기록하였다. 또한, 변동성에 따라 윈도우의 크기를 조절하여 대응할 수 있고, 차원 축소 및 적은 계산량의 장점을 갖는다. LSTM과는 다르게 결과에 대한 추론 가능한 것 역시 장점으로 부각된다. 물론 데이터 흐름을 보며 임계치를 찾는 과정이 수반되어야 하는 단점이 존재한다.

제안된 방법은 연속적인 데이터의 예측이 중요하게 작용하는 날씨, 경제 추이, 산업 장비 센서의 분석, 계절성 수요 패턴, 요일 별 데이터 추이 등 다양한 산업 분양에 응용될 수 있을 것이다. 알맞은 임계값 설정을 통해 더 나은 결과를 도출할 수 있을 것이며, 향후 최적의 임계값 도출을 위해 딥러닝을 이용한 연구가 진행될 것이다.

REFERENCES

[1] E. H. Lee & J. P. Woo. (2019). A model for predicting the number of movie audiences and sales through big data analysis. *Journal of the Korean Big Data Society, 4(2),* 185-194. DOI : 10.36498/kbigdt.2019.4.2.185

[2] S. H. Song, K. S. Shin, J. H. Lee, Y. J. Jeong, J. S. Lee & S. M. Yoon. (2020). A study on the development of a short-term heat demand forecasting model using real-time calorimeter information. *Journal of the Korean Big Data Society, 5(2),* 17-27. DOI : 10.36498/kbigdt.2020.5.2.17

[3] T. H. Kim, S. W. Lim, J. K. Ko & J. H. Lee. (2020). A study on predictive analysis of win/loss of Korean professional baseball according to artificial intelligence model. *Journal of the Korean Society for Big Data, 5(2),* 77-84. DOI : 10.36498/kbigdt.2020.5.2.77

[4] K. T. Na, J. Y. Lee, E. C. Kim & H. C. Lee. (2020).

Predicting and inferring reasons for churn of securities financial instruments trading customers. *Journal of the Korean Society for Big Data, 5(2),* 215-229.
DOI : 10.36498/kbigdt.2020.5.2.215

[5] B. S. Pramod & M. S. PM. (2020). *Stock Price Prediction Using LSTM.*

[6] A. S. Saud & S. Shakya. (2020). Analysis of look back period for stock price prediction with RNN variants: A case study on banking sector of NEPSE. *Procedia Computer Science, 167,* 788-798.
DOI : 10.1016/j.procs.2020.03.419

[7] A. Moghar & M. Hamiche. (2020). Stock market prediction using LSTM recurrent neural network. *Procedia Computer Science, 170,* 1168-1173.
DOI : 10.1016/j.procs.2020.03.049

[8] Y. Li & H. Cao. (2018). Prediction for tourism flow based on LSTM neural network. *Procedia Computer Science, 129,* 277-283.
DOI : 10.1016/j.procs.2018.03.076

[9] J. Qiu, B. Wang & C. Zhou. (2020). Forecasting stock prices with long-short term memory neural network based on attention mechanism. *PloS one, 15(1),* e0227222.
DOI : 10.1371/journal.pone.0227222

[10] M. Roondiwala, H. Patel & S. Varma. (2017). Predicting stock prices using LSTM. International *Journal of Science and Research (IJSR), 6(4),* 1754-1756.
DOI : 10.21275/ART20172755

[11] M. J. Kim, S. H. Min & I. Han. (2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index. *Expert Systems with Applications, 31(2),* 241-247.
DOI : 10.1016/j.eswa.2005.09.020

[12] S. H. Kim, D. H. Kim, C. H. Han & W. I. Kim. (2008). Stock index relationship and stock prediction using genetic algorithm. *Journal of the Korean Institute of Intelligent Systems, 18(6).*
DOI : 10.5391/JKIIS.2008.18.6.781

[13] N. H. M. Radzi, I. S. B. Gwari, N. H. Mustaffa & R. Sallehuddin. (2019, August). Support Vector Machine with Principle Component Analysis for Road Traffic Crash Severity Classification. *In IOP Conference Series: Materials Science and Engineering, 551(1),* pp. 012068. IOP Publishing.
DOI : 10.1088/1757-899X/551/1/012068

[14] T. K. Lin. (2018). PCA/SVM-based method for pattern detection in a multisensor system. *Mathematical Problems in Engineering, 2018.*
DOI : 10.1155/2018/6486345

[15] M. Mustaqeem & M. Saqib. (2021). Principal component based support vector machine (PC-SVM): a hybrid technique for software defect detection. *Cluster Computing,* 1-15.
DOI : 10.1007/s10586-021-03282-8

[16] Se. H.Baek, J. Y. Oh, J. S. Lee, J. K. Hong & S. C. Hong. (2019). Sales forecast model for temperature change using big data analysis. *Journal of the Korean Big Data Society, 4(1),* 29-38.
DOI : 10.36498/kbigdt.2019.4.1.29

[17] Y. J. Lee & G. Woo. (2013). Stock market network analysis for stock portfolio recommendations. *Journal of the Korean Contents Association, 13(11),* 48-58.
DOI : 10.5392/JKCA.2013.13.11.048

[18] R. Pascanu, T. Mikolov & Y. Bengio. (2013, May). On the difficulty of training recurrent neural networks. *In International conference on machine learning* (pp. 1310-1318). PMLR.

[19] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[20] L. I. Smith. (Feb. 2002.). *A tutorial on principal components analysis.* Cornell University, USA

[21] https://en.wikipedia.org/wiki/Principal_component_analysis

진 영 훈(Young-Hoon Jin) [정회원]

· 2021년 3월 ~ 현재 : 백석대학교 스마트IT공학부 교수
· 관심분야 : AR, VR, 메타버스, 패턴매칭, 지능형 소프트웨어
· E-Mail : devjay@bu.ac.kr

지 세 현(Se-Hyun Ji) [정회원]

· 2021년 3월 ~ 현재 : 백석대학교 스마트IT공학부 교수
· 관심분야 : 금융공학, 핀테크, 트레이딩, 머신러닝
· E-Mail : sehyun@bu.ac.kr

한 군 희(Kun-Hee Han) [종신회원]

· 2001년 3월 ~ 현재 : 백석대학교 컴퓨터공학부 교수
· 관심분야 : 데이터베이스, 암호프로토톨, 네트워크 보안, 영상처리
· E-Mail : hankh@bu.ac.kr