비트코인 트랜잭션 수 예측을 위한 LSTM 학습데이터 선택기법

지세현, 구영훈, 백의준, 박지태, 윤성호*, 김명섭
고려대학교, *LG전자
{sxzer, gyh0808, pb1069, pjj5846, tmskim}@korea.ac.kr, *sungho.sky.yoon@lge.com

LSTM Learning Data Selection Technique for Number of Bitcoin Transactions Prediction

Se-Hyun Ji, Young-Hoon Goo, Ui-Jun Baek, Jee-Tae Park, Sung-Ho Yoon*,
Myung-Sup Kim
Korea Univ. *LG Electronics

요 약

블록체인 기술을 기반으로 만들어진 온라인 암호화폐인 비트코인은 오늘날 개인, 기업, 정부, 금융기관 등 모두의 관심을 끌고 있다. 지난 몇 년간 비트코인 트랜잭션 수가 증가함에 따라 비트코인 시장의 규모는 나날이 증가하고 있다. 비트코인 트랜잭션 수를 예측하는 것은 비트코인 네트워크에 있어 중요한 항목이다. 본 논문은 비트코인 트랜잭션 수를 예측하기 위해 기계학습 알고리즘 중 하나인 LSTM 모델의 학습데이터 선택기법을 제안한다. 본 연구팀이 수집한 비트코인 블록의 트랜잭션 통계데이터와 해당 블록에 담긴 트랜잭션 수의 상관분석 방법을 적용하여, LSTM 모델의 학습데이터로 선정한 뒤, LSTM 모델이 예측한 값과 실제 값의 차이를 비교하여 제안하는 기법의 타당성을 검증한다.

Ⅰ. 서론

사토시 나카모토에 의해 개발된 비트코인은 블록체인 기술을 기반으로 만들어진 온라인 암호화폐이다[1]. 현재 비트코인은 정부, 기업, 금융기관 등 모두의 관심을 끌고 있다. 지난 몇 년간 비트코인의 트랜잭션 수는 엄청나게 증가하고 있다. 2019년 4월 기준, 비트코인의 시가 총액은 약 100조 원에 달한다. 비트코인 트랜잭션 수의 증가에 따라 비트코인 네트워크는 급속도로 발전을 하고 있지만, 그에 따른 문제도 발생하고 있다. 예를 들어 트랜잭션 처리 비용은 증가했지만, 트랜잭션 확인 시간은 지연되고 있다. 이러한 이유로 미래의 비트코인 트랜잭션 수를 예측하는 것은 비트코인 네트워크의 성장 및 문제에 대응하는 것에 있어 중요하다[2]. 그러나 현재 비트코인 트랜잭션 수를 예측하기 위한 연구는 거의 없다.

기계학습 알고리즘 중 하나인 Long Short Term Memory (LSTM)은 순환신경망 구조로부터 파생된 시계열 데이터를 예측하는 데 있어 특화된 알고리즘이다[3]. 일정 시간 간격으로 생성되는 비트코인 블록데이터는 시계열 데이터로써 LSTM 모델의 학습데이터로 적합하다. 그러나 비트코인 트랜잭션의 수를 예측하는데 어떤 항목의 비트코인 블록데이터를 학습해야 하는지는 알 수 없을 뿐만 아니라 적합한 학습데이터를 찾기 위해 모든 경우의 수를 이용하여 데이터를 학습하는 것은 비효율적이다. 따라서 LSTM 알고리즘을 적용한 비트코인 트랜잭션 수를 예측하는데 적합한 학습데이터를 찾는 효율적인 방법이 필요하다.

본 논문은 서론에서 연구 배경과 목표를 서술하고, 본론에서 비트코인 트랜잭션 수 예측을 위해 기계학습 알고리즘 중 하나인 LSTM 모델의 학습데이터를 선정하기 위한 2가지의 상관관계 분석을 제안한다. 제안하는 방법은 실험을 통해 타당성을 검증한다.

※ 이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2018R1D1A1B07045742)과 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2018-0-00539-001,블록체인의 트랜잭션 모니터링 및 분석 기술개발)

Ⅱ. 본론

본 장에서는 비트코인 블록 통계데이터와 비트코인 트랜잭션 수와의 상관분석 방법 및 본 연구팀이 수집한 비트코인 블록 통계데이터와 상관분석 방법을 적용하여 선정된 실험데이터에 대해 언급한다.

2.1 상관분석

본 절에서는 상관분석 방법에 대해 언급한다. 상관분석은 두 변수 사이에 어떤 선형적 관계가 있는지를 분석하는 방법이다. 두 변수는 서로 독립적인 관계이거나 상관된 관계일 수 있으며 이때 두 변수 사이의 관계 강도를 상관관계라 한다.

표 1. 피어슨 상관계수 해석표

상관계수 r의 범위	해석
$-1.0 \le r \le -0.7$	강한 음의 선형적 관계
$-0.7 \le r \le -0.3$	뚜렷한 음의 선형적 관계
$-0.3 \le r \le -0.1$	약한 음의 선형적 관계
$-0.1 \le r \le +0.1$	무시 될 수 있는 선형적 관계
$+0.1 \le r \le +0.3$	약한 양의 선형적 관계
$+0.3 \le r \le +0.7$	뚜렷한 양의 선형적 관계
$+0.7 \le r \le +1.0$	강한 양의 선형적 관계

본 논문에서는 2가지의 상관분석 방법을 이용하였다. 첫 번째 상관분석 방법은 피어슨 상관계수(Pearson correlation coefficient)를 이용하는 것이다. 피어슨 상관계수는 두 변수 사이의 선형 상관관계를 계량화한 수치다. 피어슨 상관계수는 코시-슈바르츠 부등식에 의해 +1과 -1 사이의 값

을 가지며, +1은 완벽한 양의 선형적 상관관계, 0은 선형 상관관계가 없음, -1은 완벽한 음의 선형적 관계를 의미한다[4]. 피어슨 상관계수에 대한 해석은 표1과 같다. 두 번째 상관분석 방법은 스피어만 상관계수(Spearman correlation coefficient)를 이용하는 것이다. 스피어만 상관계수는 자료의 값 대신 순위를 이용하는 경우의 상관계수로써, 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용해 상관계수를 구한다. 스피어만 상관계수는 두 변수 사이의 연관 관계가 있는지 없는지를 밝혀준다. 스피어만 상관계수의 값이 -1과 +1 사이의 값을 가지는데 두 변수 안의 순위가 완전히 일치하면 +1이고, 두 변수의 순위가 완전히 반대이면 -1이 된다[5].

2.2 비트코인 블록 통계데이터

본 절에서는 비트코인 블록의 트랜잭션 통계데이터에 대해 언급한다. 비트코인 네트워크의 100,000 높이의 블록부터 200,000 높이의 블록에 담긴 트랜잭션 데이터를 수집하였고, 수집한 트랜잭션 데이터를 재정렬하여 80가지 항목의 트랜잭션 통계데이터를 추출하였다. 추출한 비트코인 트랜잭션 통계데이터 항목은 표 2와 같다. 비트코인 데이터를 블록, 트랜잭션 단위로 구분하였고 각 데이터 항목에 대한 합, 평균, 최댓값, 최솟값, 표준편차이다.

표 2. 비트코인 트랜잭션 통계데이터 항목

비트코인 트랜잭션 통계데이터 항목		
단위	항목	통계정보
Block	Tx Size	Sum
	Tx vSize	Mean
	Fee	Max
	Value	Min
	n_Vin	Stdvar
	n_Vout	
Transaction	Input	
	Output	

2.3 실험데이터

본 절에서는 비트코인 블록 통계데이터를 2가지 상관분석을 적용하여 선정된 실험데이터에 대해 언급한다. 80가지 항목의 비트코인 통계데이터와 비트코인 트랜잭션 수와의 상관분석을 하였다. 80가지의 비트코인 통계데이터 항목 중 상관분석을 적용하여 비트코인 트랜잭션 선정된 실험데이터 항목은 그림1, 2와 같이 상관계수 값의 크기에 대해 오름차순으로 나열하였다.

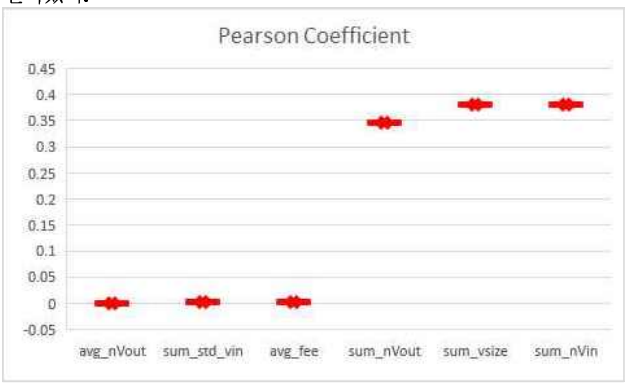

그림 2. 피어슨 상관계수 실험데이터 항목

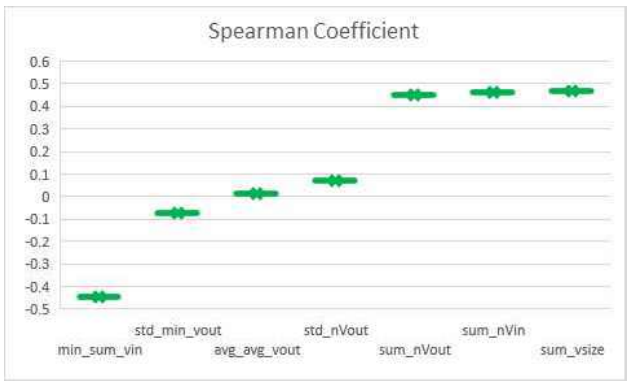

그림 1. 스피어만 상관계수 실험데이터 항목

피어슨 상관계수 분석을 적용한 경우 블록 당 비트코인 트랜잭션의 수와 양의 선형적 관계가 있는 데이터 3개와 성능 비교를 위해 선형적 관계가 없는 데이터 3개를 선정하였다. 음의 선형적 관계를 나타내는 데이터는 없었다. 스피어만 상관계수 분석을 적용한 경우 -1 또는 1에 가까운 4개의 데이터와 성능 비교를 위해 상관관계가 없는 데이터 3개를 선정하였다. sum_nVout, sum_nVin, sum_vsize는 2가지 상관분석 전부에 대해 높은 상관관계를 보이는 선정된 항목이다. 두 가지 상관분석을 적용하여 선정된 실험데이터는 모델의 성능 평가를 위해 학습데이터 80%, 검증데이터 10%, 실험데이터 10%의 비율로 구성하였다. 실험데이터 구성에 대한 정보는 표3과 같다.

표 3. 실험데이터 구성

블록의 높이(데이터 개수)	구분
100,000~180,000	학습데이터
180,001~190,000	검증데이터
190,001~200,000	실험데이터

Ⅲ. 실험 및 결과

본 장에서는 상관분석을 통해 선정된 LSTM 학습데이터를 이용하여 실험을 진행한다. LSTM 모델은 학습, 검증, 시험의 단계를 거친다. 학습단계는 같은 데이터를 여러 번 학습 하는 단계이다. 같은 데이터를 반복적으로 학습하게 되므로 해당 데이터에 맞는 예측 모델이 완성된다. 검증 단계는 학습에 사용되지 않은 데이터를 이용하여 모델의 성능을 검증하는 단계이다. 마지막 시험 단계는 학습 및 검증과정을 통해 완성된 모델의 성능을 평가하는 단계이다. 모델의 성능은 실제 값과 모델이 예측한 값의 평균 제곱 오차(Mean Square Error) 수치를 통해 평가한다. 평균 제곱 오차가 0에 가까울수록 모델의 성능이 좋다. 실험에 사용된 LSTM 모델의 정보는 표4와 같다.

표 4. 실험에 사용된 LSTM모델 정보

Hyper-Parameter	Method / Value
Data_Normalization	Min-Max Scaler
Sequence_Length	5
Number of Hidden_Unit	1
Loss_Function	MSE(Mean Square Error)
Optimizer	Adam

실험데이터 항목 간의 편차를 줄이기 위해 0과 1 사이의 값으로 정규화를 해주는 Min-Max Scaler 기법을 적용하였고, Sequence_Length는 5로 설정하였다. 실험데이터의 성능 비교를 극대화하기 위해 Hidden_Unit의

개수는 1로 설정하였다. Loss_Function은 모델의 성능을 객관적으로 비교하기 위해 평균 제곱 오차를 적용하였고, Optimizer는 기계학습에 보편적으로 쓰이는 Adam Optimizer를 적용하였다.

실험데이터를 학습하여 나온 모델의 실험 결과는 그림 3, 4와 같다. 피어슨 상관계수 분석을 적용하여 선정된 데이터를 학습한 LSTM 모델의 평균 제곱 오차는 피어슨 상관계수의 값이 클수록 작게 나왔다. 피어슨 상관계수가 가장 큰 데이터 항목인 sum_nVin은 LSTM 모델의 평균 제곱 오차가 가장 작게 나왔다. 따라서, 피어슨 상관계수 분석을 적용하여 나온 상관관계가 있는 데이터는 상관관계가 없는 데이터보다 LSTM 모델의 학습에 적합한 데이터라고 판단할 수 있다.

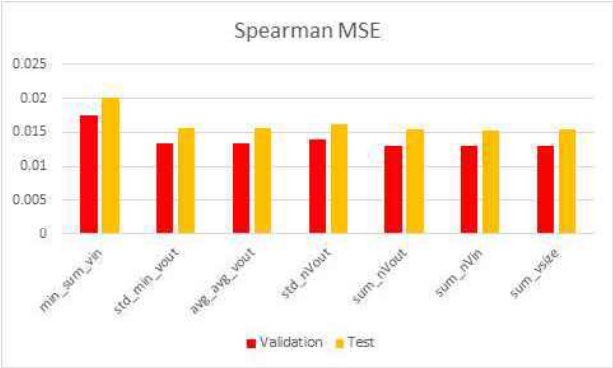

그림 3. 스피어만 상관계수 분석을 적용한 실험 결과

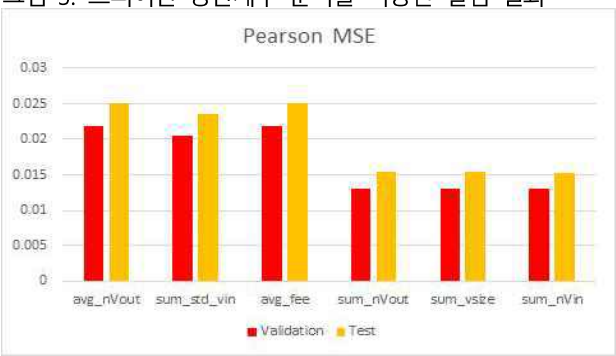

그림 4. 피어슨 상관계수 분석을 적용한 실험 결과

스피어만 상관계수 분석의 경우, 전반적으로 상관관계가 없는 데이터보다 상관관계가 있는 데이터를 학습한 LSTM 모델의 평균 제곱 오차가 낮게 나왔지만, 스피어만 상관계수 값이 가장 큰 데이터 항목인 sum_vsize는 sum_nVin 보다 평균 제곱 오차가 높게 나왔다. 피어슨 상관계수 분석과 비슷한 데이터 항목이 선택되었으나 스피어만 상관계수와 LSTM 모델의 평균 제곱 오차의 관계는 불규칙적이다. 2가지의 상관분석을 적용하여 실험한 결과 피어슨 상관계수 분석이 비트코인 트랜잭션 수를 예측하는 LSTM 모델의 학습데이터를 찾는 방법으로 조금 더 적합하다.

Ⅳ. 결론 및 향후 연구

본 논문은 비트코인 네트워크에 있어 중요한 항목인 비트코인 블록의 트랜잭션 수를 예측하기 위한 LSTM 모델의 학습데이터 선택기법을 제안하였다. 제안하는 방법은 실험을 통해서 타당성을 검증하였다. 향후 연구로는 본 논문에서 언급한 2가지 상관관계 분석 이외의 다양한 통계분석을 적용하여, 올바른 학습데이터를 선정한 뒤, 비트코인 블록의 트랜잭션 수를 정밀하게 예측하는 LSTM 모델을 설계할 계획이다.

참 고 문 헌

[1] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." (2008).

[2] Bianconi, Gabriel, and Mahesh Agrawal. "Predicting Bitcoin Transactions with Network Analysis." (2017).

[3] Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." (1999): 850-855.

[4] Benesty, Jacob, et al. "Pearson correlation coefficient." Noise reduction in speech processing. Springer, Berlin, Heidelberg, 2009. 1-4.

[5] Zar, Jerrold H. "Significance testing of the Spearman rank correlation coefficient." Journal of the American Statistical Association 67.339 (1972): 578-580.