통계분석을 적용한 비트코인 데이터 처리에 관한 연구

이승진, 지세현, 백의준, 김명섭

고려대학교

{nanfan00, sxzer, pb106, tmskim}@korea.ac.kr

A Study on Bitcoin Data Processing Using Statistical Analysis

Seung-Jin Lee, Se-Hyun Ji, Ui-Jun Baek, Myung-Sup Kim

Korea Univ

요 약

비트코인은 블록체인 기술을 기반으로 하는 온라인 암호 화폐이다. 지난 몇 년간 관심이 증가함에 따라 암호 화폐의 거래량 및 시장규모는 엄청난 속도로 발전하고 있다. 거래를 정리한 데이터 묶음의 크기는 날마다 커지고 있으며 하루 평균 거래량(트랜잭션 수)도 증가하고 있다. 그러나 그에 따라 데이터의 수도 많아지고 커지면서 데이터를 어떻게 가공해야 할지 중요한 이슈가 발생하고 있다. 본 논문에서는 빅데이터를 해석하고 활용하는데 적합한 통계분석을 기존의 기계학습 모델로 사용될 데이터 적용하여 더 나은 기계학습 모델을 만드는 방법을 제안한다. 제안하는 방법은 완성된 모델의 정확도 비교를 통해 적합성을 검증한다.

Ⅰ. 서 론

2009년 사토시 나카모토에 의해 등장한 비트코인은 암호화 기술과 블록체인, 네트워크 기술 등을 절묘하게 결합해 화폐 발행과 거래 내역에 대해peer-to-peer 기술을 적용한 최초의 암호 화폐이다. 화폐라는 형태를 벗어나 다수 관계자가 동시에 기록하고 관리하는 데이터베이스로써 활용하므로 더 안전하면서도 저렴한 신기술로 금융시장에 새로운 패러다임을 가져왔다. 일정 기간 동안 거래를 정리한 데이터 묶음 블록에는 이전 블록의 데이터를 짧게 요약한 해시와 이 블록 고유의 임의 값(nonce)이 포함되어 있다. 비트코인 초기 블록은 거래 데이터 36MB를 가질 수 있었다. 하지만 2010년 스팸이나 잠재적인 디도스 공격에 대비해 1MB로 줄었다. 이 제한은 비트코인 블록이 대량으로 증가하고 거래도 급증한 지금도 계속되고 있다. 트랜드블록(TrendBlock)에 따르면 2013년 이후 평균 블록 크기는 125KB에서 425KB로, 거래량은 일 2.5배 증가했다. 하루 평균 4번은 이 블록 크기 한계에 도달하고 있다고 한다. 2019년 10월 기준, 비트코인의 하루 평균 거래량은 약 34만에 달한다. 비트코인의 데이터는 점점 커지고 기하급수적으로 데이터의 양이 증가하여 트랜잭션의 처리시간이 지연되는 현상이 일어났다. 이렇게 많은 양의 데이터 확보가 가능해지고 그에 따라 이를 어떻게 분석하고 빠르게 처리해야 하는 문제가 대두되었다. 데이터들을 통한 기계학습 모델은 데이터의 차원이 커질수록 학습이 저하되는 문제를 가지고 있다. 많은 양의 데이터 확보가 가능해지고 이를 어떻게 분석할지는 매우 중요하다. 원 데이터의 완결성을 유지하고 그 속에 담겨있는 의미와 원리를 찾아낼 수 있어야 더 좋은 기계학습 모델을 만들 수 있다. 이러한 점에서 데이터에서 의미를 찾아내는 방법을 다루는 학문인 통계학은 빅데이터를 해석하고 활용하는 데 적합하다.

따라서, 본 논문에서는 통계분석방법을 적용시켜 데이터를 가공하는 방법으로 모수적 검정 중 5가지 검정을 파이썬의 Scipy 패키지에 내장된 stats 함수를 사용한다. 얻은 결괏값(검정통계량과 p-value)을 이용해 가설에 대한 검정 지표로 이용하여 더 좋은 기계학습 모델을 만드는 것을 제안한다. 제안하는 방법은 실험을 통해 결과를 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 통계분석방법과 데이터에 통계분석을 적용하는 방법을 소개하고 실험 데이터에 대해 언급한다.

2..1 통계분석

본 절에서는 통계분석의 종류와 의미, 주어진 데이터에 통계분석을 적용시켜 어떤 결과를 도출할 수 있는지를 언급한다. 통계분석은 세워진 가설이 참인지 거짓인지 판단하기 위해 가설을 검증하는 단계에서 사용된다. 과학적 의사결정은 p-value를 이용하여 진행하게 되고 이 값으로 구축한 가설을 통계적으로 판단할 수 있다. 이를 통해 여러 데이터를 수집하여 그 속에 담겨있는 의미와 원리를 찾아낼 수 있다. 사용된 통계분석의 종류는 다음과 같이 5가지(빈도분석, 평균분석, 분산분석, 상관분석, 회귀분석) 통계 분석방법을 사용한다.

첫 번째, 빈도분석(카이스퀘어 검증)은 측정하여 얻은 데이터가 사람 수, 횟수 등의 빈도인 경우, 사용이 가능하고 집단 간 빈도 차를 비교한다. 즉, 그룹 간의 차이가 있는지(독립인지)에 대한 검정으로 카이스퀘어 값을 통하여 두 집단이 동일 분포인지 확인할 수 있고 계산 방법은 (1)에 의해 구해진다. 표본의 관측값과 지정된 분포의 기댓값이 통계적으로 다른지를 확인하려면 p-value를 유의 수준과 비교하여 유의수준보다 작으면 동일하다는 결과를 얻을 수 있다.

$$\sum \frac{(O-E)^2}{E} \sim X^2(k), \text{(단 k는 자유도)} \quad (1)$$

두 번째, 평균분석은 집단 간의 평균과 분산 등을 통해 차이를 검정하는 방법으로 평균 간의 차이가 통계적으로 유의한지를 확인하려면 p-value와 유의 수준과 비교하여 유의 수준 값보다 크면 같은 평균을 가진다고 볼 수 있다.

세 번째, 분산분석은 둘 이상의 집단 간에 평균값의 차이가 있는지를 검증하는 분석방법으로 평균분석은 두 개의 집단이지만 분산분석은 셋 이상의 집단에서 평균분석을 할 때 사용한다.

네 번째, 상관분석은 두 변수 사이에 관계가 있는지를 검증하는 분석방법으로 상관계수를 통해 관계의 정도를 알 수 있다. 상관계수를 통해 선형관계를 갖는지를 알 수 있고, 그 값에 따라 크기와 방향을 알 수 있다. 상

관계수는 -1부터 +1 사이의 값을 가지고 (2)에 의해 값을 구한다. 상관계수가 ± 1에 가까울수록 두 변수는 큰 관계성을 갖고 있다. 비슷하게 상관계수가 ± 0.9이상 일 때 중복 데이터라 볼 수 있으며 이 방법으로 데이터를 축소 할 수 있다.

$$p = \frac{\sum_{i=1}^{n}(B_i - \overline{B})(T_i - \overline{T})}{(n-1)S_B S_T} \qquad (2)$$

다섯 번째, 회귀분석은 독립변인이 종속 변인에 영향을 미치는지를 알아보고자 할 때 실시하는 분석방법으로 상관분석과 달리 인과 관계를 적용할 수 있다. 회귀분석의 핵심은 독립변수와 종속변수로 구한 상관계수에 제곱한 값인 결정 계수(r^2)를 구하는 것이다. 결정 계수로 독립변수를 가지고 얼마만큼 의미 있게 종속변수를 예측할 수 있는지를 판별할 수 있다. 이 값은 0과 1 사이의 값으로 나타나는데 1에 가까울수록 추정된 회귀 식이 해당 자료를 잘 설명하고 본다. 단순 회귀분석의 경우 결정 계수는 독립변수 x와 종속변수 y의 상관계수 제곱과 같다. 이를 통해 종속 변인 데이터에 어떤 독립변수 데이터가 유효한 영향을 끼치는지 알 수 있다.

2.2 실험 데이터

본 절에서는 실험에 사용된 데이터에 대해 언급한다. 데이터는 본 연구팀이 수집한 비트코인 블록 데이터를 사용하였다. 수집한 비트코인 블록 및 통계 데이터로부터 통계정보를 얻기 위해 통계처리를 한다. 트랜잭션 단위에 대해서 통계처리를 적용하고, 통계처리의 종류는 합, 최대, 최소, 중앙값, 표준편차, 평균값, 사분위간 범위(IQR), 비대칭도, 첨도 Q1, Q2, Q3 을 구한다. 통계처리를 통해 수집한 데이터에 대한 자세한 설명은 표 1과 같다. 트랜잭션 단위의 데이터 6가지 항목은 1번의 통계처리를 하고, 2가지 항목은 2번의 통계처리를 한다. 총 312종류의 비트코인 블록 및 트랜잭션의 통계 데이터로 구성하고 유의 수준은 0.05로 기준을 하였다.

데이터 단위	데이터 항목	1^{st} 통계처리	2^{st} 통계처리	항목 수
블록	nTx(블록에 포함된 트랜잭션 수)			1
	Weight(블록의 Weight)			1
	Size(블록의 사이즈)			1
	Vsize(블록의 가상 크기)			1
트랜잭션	nVin(트랜잭션이 포함하고 있는 input 의 수)		sum Min Max Mean Stdv Avg IQR skewness kurtosis Q1 Q2 Q3	11
	nVout(트랜잭션이 포함하고 있는 Output 의 수)			11
	Value (트랜잭션의 거래 금액)			11
	Fee(트랜잭션 수수료)			11
	Tx.Size(트랜잭션의 크기)			11
	Tx.Vsize (트랜잭션의 가상 크기)			11
	Vin.value (트랜잭션의 입력 값)	sum Min Max Mean Stdv Avg IQR skewness kurtosis Q1 Q2 Q3		121
	vout.valu (트랜잭션의 출력 값)			121

표 1. 비트코인 블록의 통계 데이터

Ⅲ. 실험 및 결과

본 장에서는 앞선 비트코인 블록의 통계 데이터에 5가지의 통계분석을 하는 실험을 진행한다. 100,000개의 데이터에 빈도분석을 하여 같은 분포를 가지는 집단 데이터 추출할 수 있었다. 마찬가지로 평균분석과 분산분석을 통해 서로 같은 평균을 가진 집단을 추출할 수 있었다. 상관분석을 했을 때 상관계수가 ± 0.9이상인 데이터를 찾아내어 동일 데이터로 간주하여 제거하였다. 상관분석을 한 결과 312개의 특징을 101개의 특징으로 축소하였고, 검증으로는 모델을 트랜잭션의 유지, 증가, 감소를 예측값으로 설정한 뒤 학습시켜 정확도를 비교하여 성능을 확인한다. 모델 성능 평가 지표는 기존의 정확도와 유사하거나 높을수록 성능이 좋다. 검증 결과는 표2와 같다.

표 2. 상관분석을 적용한 기계학습 모델의 성능

기계학습 모델 학습 비율	데이터 특징의 수	정확도
학습 70% , 실험30%	312개 특징(상관분석 전)	57.45%
	101개 특징(상관분석 후)	57.45%
학습 60% , 실험40%	312개 특징(상관분석 전)	57.05%
	101개 특징(상관분석 후)	57.13%

회귀분석을 위해 트랜잭션의 유지, 증가, 감소를 종속변수로 설정하여 다항 로지스틱 회귀분석을 실행한 결과 기존 312개의 특징 중 종속변수에 영향을 주는 225개의 특징을 찾을 수 있었다. 추가로 더 좋은 효율을 내기 위해 회귀분석을 적용한 뒤 상관분석을 적용하여 정확도를 비교해보았다. 검증은 상관분석과 같이 학습을 시킨 뒤 정확도를 비교하였다. 결과는 아래의 표 3과 같다.

표 3. 회귀분석 및 상관분석을 적용한 기계학습 모델의 성능

기계학습모델 학습 비율	데이터 특징의 수	정확도
학습 70% 실험 30%	312개 특징(회귀분석 전)	57.45%
	225개의 특징(회귀분석 후)	99.35%
	76개 특징(회귀분석 후 상관분석)	99.03%
학습 60% 실험 40%	312개 특징(회귀분석 전)	57.05%
	225개의 특징(회귀분석 후)	99.2%
	76개 특징(회귀분석 후 상관분석)	99.16%

Ⅳ. 결론

본 논문은 비트코인 데이터를 통계분석을 통해 비트코인 데이터를 필요에 맞게 가공하는 방법을 제안하였다. 제안하는 방법은 실험을 통해서 적합성을 검증하였다. 상관분석의 결과 기존의 정확도와 유사하거나 더 높은 결과를 보였으며 데이터의 차원도 줄일 수 있었다. 회귀분석 결과 필요한 독립변수 데이터를 추출함으로써 정확도를 높이고 데이터를 축소시킬 수 있었다.

참 고 문 헌

[1] 지세현, 구영훈, 백의준, 신무곤, 윤성호, 김명섭, “비트코인 트랜잭션 수 예측을 위한 LSTM 학습데이터 선택기법”, KNOM Conference 2019, accepted May 14, 2019.

[2] 신무곤, 백의준, 구영훈, 지세현, 박준상, 김명섭, "주성분 분석을 적용한 클러스터링을 이용한 비트코인 네트워크 분석 방법", KNOM Conference 2019, May 30-31, 2019