주성분 분석을 적용한 클러스터링을 이용한 비트코인 네트워크 분석 방법

신무곤, 백의준, 구영훈, 지세현, *박준상 김명섭
고려대학교, *LG Electronics
{tm0309, pb1069, gyh0808, sxzer, tmskim}@korea.ac.kr, *junsang.park@lge.com

Bitcoin Network Analysis Method Using Clustering with Principal Component Analysis

Mu-Gon Shin, Ui-Jun Baek, Young-Hoon Goo, Se-Hyun Ji, *Jun-Sang Park, Myung-Sup Kim
Korea Univ., *LG Electronics

요 약

블록체인 기술을 기반으로 만들어진 온라인 암호화폐 비트코인은 개인, 기업, 정부 등 모두의 관심을 끌고 있다. 지난 몇 년간 블록체인 기술과 암호화폐에 대한 관심이 꾸준히 증가함에 따라 암호화폐의 거래량 및 시장규모는 놀라운 속도로 증가했다. 이에 따라 블록체인 네트워크와 블록, 트랜잭션에 대한 분석 및 모니터링 방안은 중요한 이슈가 되고 있다. 본 논문에서는 비트코인 네트워크 분석 방안으로 차원축소를 적용한 클러스터링 방법을 제안한다. 제안된 방법은 본 연구팀이 수집한 비트코인 내의 블록 데이터에 PCA를 적용한 K-means알고리즘을 이용한 분석 방법을 적용한다.

Ⅰ. 서 론

2008년 10월 사토시 나카모토가 개발한 비트코인(bitcoin)은 블록체인 기술을 기반으로 만들어진 온라인 암호화폐이다[1]. 지난 몇 년간 블록체인 기술과 암호화폐에 대한 관심이 꾸준히 증가함에 따라 암호화폐의 거래량 및 시장규모는 놀라운 속도로 증가했다. 2019년 4월 기준, 비트코인의 하루 평균 거래량(트랜잭션 수)는 약 38만건에 달한다. 비트코인의 거래량이 증가하고 블록체인에 대한 관심이 깊어지고 있지만, 블록체인에 대한 모니터링 및 분석에 대한 연구는 많지 않다. 비트코인을 포함한 암호화폐를 통한 불법적인 거래 등이 늘어남에 따라 암호화폐 블록과 트랜적션에 대한 모니터링하고 분석하는 것은 매우 중요하다.

클러스터링 알고리즘 중 하나인 K-Means 알고리즘은 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다[2]. 여러 개의 feature들로 구성된 비트코인의 블록 및 트랜잭션 데이터는 클러스터링을 통하여 여러 군집으로 묶일 수 있다. 또한 효과적인 클러스터링을 위하여 feature를 선택하는 것은 매우 중요하다. 그리고 클러스터링 된 데이터를 시각화 하는데 있어 차원축소도 중요한 이슈가 될 수 있다.

차원축소 알고리즘 중 하나인 PCA(Principal Component Analysis)는 고차원의 데이터를 저차원의 데이터로 환원시키는 기법이다[3]. 블록 데이터들이 여러 개의 고차원 데이터로 표현 되기 때문에 PCA를 활용하여 고차원의 블록 데이터들을 저차원의 데이터로 변환하여 이 정보들을 클러스터링에 이용하였다. 또한 저차원으로 변환된 데이터들은 2차원이나 3차원의 그래프로 시각화하여 데이터들의 분포를 명확히 알 수 있다.

본 논문은 서론에서 연구 배경과 목표를 서술하고, 본론에서 비트코인 블록데이터 분석을 위한 PCA를 적용한 K-Means 알고리즘을 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 PCA와 K-Means에 대해 소개하고 본 연구팀이 수집한 비트코인 블록 데이터와 PCA를 적용한 K-Means 클러스터링 방법에 대해 언급한다.

2.1 PCA(Principal Component Analysis)

본 절에서는 PCA에 대해 언급한다. PCA(Principal Component Analysis)는 데이터의 분산(variance)을 최대한 보존하면서 서로 직교하는 새 기저(축)를 찾아, 고차원 공간의 표본들을 선형 연관성이 없는 저차원 공간으로 변환하는 기법이다.

$$\arg\min_{\hat{x}} ||x - U\hat{x}||^2 \quad \text{[식 1]}$$

식 1에서 U는 역변환 행렬을 의미하고 χ는 원래의 벡터를 의미한다. 또한 $\hat{\chi}$은 원래 벡터와 가장 비슷해지는 차원축소 벡터를 의미한다. 이러한 과정을 통하여 PCA는 원래 입력 벡터와 가장 비슷한 축소된 새로운 차원의 벡터를 구한다.

PCA는 다음과 같은 단계로 이루어진다.

1. 학습 데이터셋에서 분산이 최대인 축(axis)을 찾는다.
2. 이렇게 찾은 첫번째 축과 직교(orthogonal)하면서 분산이 최대인 두 번째 축을 찾는다.
3. 첫 번째 축과 두 번째 축에 직교하고 분산을 최대한 보존하는 세 번째 축을 찾는다.
4. 1~3과 같은 방법으로 데이터셋의 차원(특성 수)만큼의 축을 찾는다.

2.2 K-Means Clustering

본 절에서는 K-Means에 대해 언급한다. 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다. 이 알고리즘은 자율 학습의 일종으로, 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아주는 역할을 수행한다.

n개의 d-차원 데이터 오브젝트 (x1, x2, ⋯, xn) 집합이 주어졌을 때, K-Means 알고리즘은 n개의 데이터 오브젝트들을 각 집합 내 오브젝트 간 응집도를 최대로 하는 k개의 집합 S = {S1, S2, ⋯, Sk} 으로 분할한다. 다시 말해, μi가 집합 Si의 중심점이라 할 때

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \quad \text{[식 2]}$$

각 집합별 중심점 ~ 집합 내 오브젝트간 거리의 제곱합을 최소로하는 집합 S를 찾는 것이 목표이다.

K-Means는 다음의 두 단계를 반복한다.

1. 클러스터 설정: 각 데이터로부터 각 클러스터들의 중심까지의 유클리드 거리를 계산하여, 해당 데이터에서 가장 가까운 클러스터를 찾아 데이터를 배당한다.

2. 클러스터 중심 재조정: 클러스터 중심을 각 클러스터에 있는 데이터들의 무게중심 값으로 재설정해준다.

K-means 알고리즘은 클러스터 개수 k값을 파라미터로 지정해 주어야 한다. 클러스터 개수에 따라 결과값이 완전히 달라지기 때문에 k값의 설정은 매우 중요하다. 따라서 본 논문에서는 k값 설정을 위하여 elbow 기법을 사용하였다.

Elbow 기법이란 K-means 알고리즘의 적정 클러스터 수를 찾아주는 기법으로 오차제곱 합의 값이 최소가 되도록 중심을 결정하는 과정에서 클러스터 개수를 하나씩 늘려 오차제곱의 값이 현저히 작아질 때의 k값을 구하는 방식이다.

2.3 실험 데이터

본 절에서는 실험에 사용된 데이터에 대해 언급한다. 본 논문에서는 본 연구팀이 수집한 비트코인 블록 데이터를 사용하였다. 블록 데이터 중에서 해쉬 값을 제외한 정수데이터들을 중심으로 실험 데이터를 선정하였다. 또한 블록의 높이 200,000부터 560,000 까지의 블록 데이터를 사용하였다.

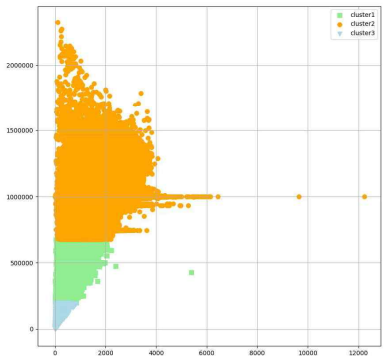
비트코인 블록 데이터	타입
height	Int32
nTx	Int64
size	Int32
nonce	Int32
weight	Int32
Difficuly	Int32
Confirmations	Int32
Strippedsize	Int32

[표 1. 실험데이터]

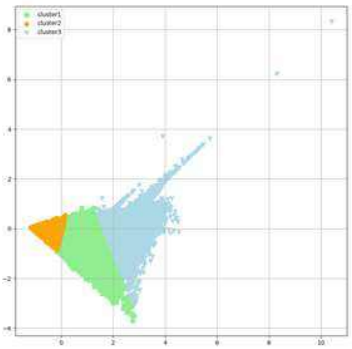
Ⅲ. 실험 및 결과

본 장에서는 PCA를 적용한 K-means 클러스터링 방법을 통해 수집한 비트코인 블록에 대해 실험을 진행한다. PCA를 적용하고 PCA를 통해 축소된 데이터를 활용하여 클러스터링을 진행하였다.

PCA는 2개의 항목(size, nTx) 그리고 36만개의 데이터에 대해 적용하였다. PCA를 적용한 데이터에 대해 K-Means 클러스터링을 적용하였다.



[그림 1. PCA 적용 전 클러스터링 결과]



[그림 2. PCA 적용 2차원 -> 2차원]

2개의 feature를 가지고 실험한 결과 결과들이 조금 더 가독성 있게 나타난다는 것을 발견하였다. 여러 개의 featrue를 가지고 실험을 진행한다면 블록데이터들의 특징을 잘 찾아낼 수 있을 것으로 예상한다.

Ⅳ. 결론 및 향후 연구

본 논문은 비트코인 네트워크에 있어 중요한 항목인 비트코인 블록 데이터의 특징을 분석하기 위한 PCA적용 K-means 클러스터링 방법을 제안하였다. 제안된 방법은 실험을 통하여 타당성을 검증하였다.

향후 연구로는 본 논문에서 실험에 사용한 2가지 feature 이외에 다른 데이터들에 클러스터링 기법을 적용하여 각 클러스터에 속하는 블록들의 특징을 연구 할 계획이다. 또한 같은 방법을 트랜잭션 데이터에도 적용하여 트랜잭션 데이터들의 특징을 분석할 계획이다.

참 고 문 헌

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

[2] J.A. Hartigan,. “Clustering algorithms” (1975).

[3] Jolliffe I.T. “Principal Component Analysis“ (2002)