블록체인 에코시스템 분석을 위한 클러스터링 알고리즘 성능 분석

박수철, 신무곤, 박지태, 김명섭
고려대학교
{such7185, tm0309, pjj5846, tmskim}@korea.ac.kr

Analysis in Clustering Algorithm for Blockchain Echo-system

Su-Cheol Park, Mu-Gon Shin, Jee-Tae Park, Myung-Sup Kim
Korea Univ.

요 약

블록체인 기술이 개발되면서 블록체인 기술을 활용한 다양한 사례가 등장하고 있다. 또한, 블록체인 기술의 취약점을 이용한 범죄들이 발생하고 있다. 블록체인 기술의 취약점 개선과 범죄 예방을 위해 클러스터링을 사용하는 많은 연구가 진행되고 있다. 연구에 사용되는 클러스터링은 많은 데이터를 사용해서 모델을 생성해야 하고, 적합한 클러스터 개수를 선정하기 위해 반복적인 클러스터링이 필요하다. 따라서 많은 연구시간이 소요되고, 속도 및 정확도 관점에서 최적의 클러스터링 알고리즘 선정을 위한 분석이 필요하다. 이에 본 논문에서는 블록체인 시스템 분석 연구에 적합한 클러스터링 알고리즘을 제시한다. 본 실험에서는 Spark machine learning library의 K-Means, Bisecting K-means, Gaussian Mixture Model 알고리즘의 성분 비교를 통해 블록체인 트랜잭션데이터에 대한 클러스터링 속도 성능과 특징을 서술한다.

Ⅰ. 서 론

블록체인은 사토시 나카모토에 의해 개발된 탈중앙화 분산 데이터 관리 기술이다 [1]. 블록체인 기술이 적용된 첫 암호화폐인 비트코인이 개발되고 난 후 블록체인 기술이 적용된 많은 암호화폐가 개발되고 있으며, 블록체인 기술을 활용한 다양한 사례가 등장하고 있다 [2]. 이러한 블록체인 기술의 발전에 따라 블록체인 기술의 취약점과 익명성을 이용하는 악성 행위들이 증가하고 있으며, 블록체인 기술의 취약점 개선과 악성 행위 예방을 위한 많은 연구가 진행되고 있다 [3]. 이러한 연구 중 클러스터링 알고리즘을 사용하는 연구는 클러스터링 과정에서 모델을 생성을 생성하기 위해 많은 트랜잭션데이터를 사용하게 되고, 적합한 클러스터 개수를 찾기 위해 반복적인 클러스터링이 필요하다. 클러스터링 과정에서 많은 실험 시간이 소요되기 때문에 클러스터링 알고리즘들이 가지는 특징을 기반으로 연구환경에 맞는 알고리즘 분석이 필요하다.

본 논문에서는 특징이 다른 세 가지 알고리즘에 대해 분석한다. 데이터의 거리 유사도를 통해 군집화하는 K-means 알고리즘과 계층적 군집화 알고리즘인 Bisecting K-means, 혼합 정규분포를 이용한 모수적 접근법인 Gaussian Mixture Model 알고리즘을 분석하여, 알고리즘 특징에 따른 클러스터링 속도 성능을 확인한다. 또한, 세 가지 알고리즘은 클러스터 개수 혹은 정규분포의 개수 k를 사용자가 직접 지정해야 하므로 적절한 k를 찾는 과정에서 반복적인 클러스터링이 필요하고, k의 개수 별 속도 성능에도 차이가 있다. 따라서 연구환경에 맞는 적합한 클러스터링 알고리즘을 분석하기 위해 세 가지 알고리즘의 클러스터 개수 별 속도 성능을 비교하여, 알고리즘을 선정할[1] 수 있는 기준을 제안한다.

Ⅱ. 본론

이 장에서는 K-means, Bisecting K-means, Gaussian Mixture Model 알고리즘의 특징과 특징에 따른 속도 성능에 대해 언급한다.

K-means 알고리즘은 주어진 데이터에서 임의의 K개의 지점을 설정하고, 해당 지점에서 거리 혹은 유사성을 기준으로 군집화한 후에 무게중심을 기준으로 군집의 중심을 옮기는 과정을 반복하는 알고리즘이다 [4]. K-means 알고리즘은 군집 k의 개수를 사용자가 직접 지정해야 한다는 단점을 가지기 때문에 적정 클러스터 개수 k를 찾는 것이 중요하다.

K-means 알고리즘의 k의 개수를 찾기 위해 elbow 기법을 사용해서 가장 중심점 오차값이 적은 k개의 클러스터를 찾아야 한다. K-means 알고리즘의 경우 무게중심을 찾는 동안 반복 과정이 생기지만 클러스터링 자체는 한 번만 진행된다. 따라서 클러스터 개수에 따라 속도 성능의 큰 차이를 주지 않는다.

Bisecting K-means 알고리즘은 주어진 데이터를 K-means 알고리즘을 통해 두 개의 군집으로 분류하고, 분류된 군집들을 다시 각각 두 개의 군집으로 분류해서 k개의 군집이 만들어질 때까지 클러스터링을 반복하는 계층적 클러스터링 알고리즘이다 [5]. K-means와 마찬가지로 k개수를 직접 지정해야 하므로 같은 과정을 통해 k개의 클러스터를 찾는다. Bisecting K-means는 계층적으로 클러스터링을 반복하는 알고리즘이기 때문에 한 번의 클러스터링 속에 클러스터 개수 k에 따라 $\log_2 K$ 횟수만큼 클러스터링이 진행된다. 따라서 Bisecting K-means의 경우 클러스터 개수가 속도 성능에 K-means보다 상대적으로 큰 영향을 준다.

Gaussian Mixture Model 알고리즘은 k개의 정규분포를 혼합해서 만들어진 혼합 정규분포를 통해 주어진 데이터의 분포특성을 알고, 해당 데이터가 각 군집에 속할 확률을 확인하기 위해 사용하는 알고리즘이다 [6]. Gaussian Mixture Model 역시 정규분포 k의 개수를 사용자가 직접 지정해야 하므로 같은 과정을 통해 성능 분석을 진행한다. Gaussian Mixture Model의 경우, 정규분포를 통해 클러스터링하는 알고리즘이기 때문에 정규분포 개수가 시간에 비례적으로 영향을 준다.

Ⅲ. 실험 및 결과

3.1 실험환경 및 데이터

본 연구실에서 블록체인 마이닝 풀을 통해 수집한 블록체인 트랜잭션

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2018R1D1A1B07045742)와 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2018-0-00539-002,블록체인의 트랜잭션 모니터링 및 분석 기술개발)

데이터 약 50만 개를 알고리즘별로 실험한다. 해당 데이터는 fee, vinValue, voutValue, txSize, txVsize 와 같은 value를 가지고, value들의 summation, minimum, maximum, average, standard deviation으로 Raw data가 구성된다. 아래 표 1은 데이터 형식을 보여준다. 본 실험은 Mongo Database에 저장된 데이터를 읽어와서 리눅스 서버의 docker 환경에서 클러스터링 실험을 진행한다. Spark machine learning library를 이용하기 위해서 JSON 형태의 데이터를 RDD 데이터로 변환한 후 사용한다.

표 1. Data format of Transaction

	Data	Extraction
Transaction	fee	sum,min,max,avg,stdv
	vinValue	
	voutValue	
	txSize	
	txVsize	

3.2 K-means

K-means 알고리즘은 50개의 클러스터까지 클러스터링을 진행한 결과 100~200초 사이에서 시간이 조금씩 증가한다. 알고리즘에서 클러스터 개수가 증가함에 따라 비교해야 할 중심점의 개수가 증가하므로 다음과 같이 클러스터 개수가 증가함에 따라 시간도 소폭 증가하는 성능을 나타낸다.

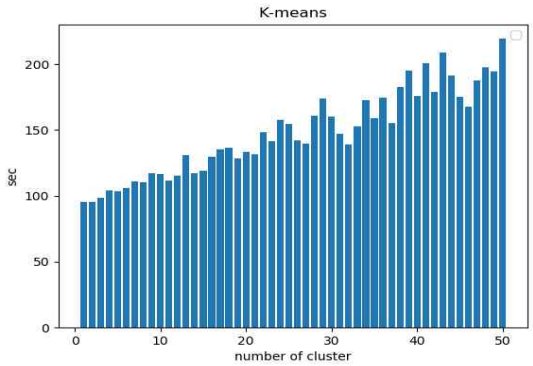

그림 1. K-means 성능 그래프

3.3 Bisecting K-means

Bisecting K-means 알고리즘은 2개의 클러스터까지 한 번의 계층까지만 클러스터링이 진행되기 때문에 K-means 알고리즘과 비슷한 성능을 보이지만, 3개의 클러스터부터 2의 제곱 개수의 클러스터를 기준으로 다시 계층적으로 클러스터링을 반복한다. 따라서 같은 계층의 범위 안에 속해있는 클러스터 개수는 비슷한 성능을 보이지만 계층의 범위를 벗어나서 새로운 계층이 만들어지면 $\log_2 K$번째 계층까지 비례적으로 증가하는 성능을 보인다.

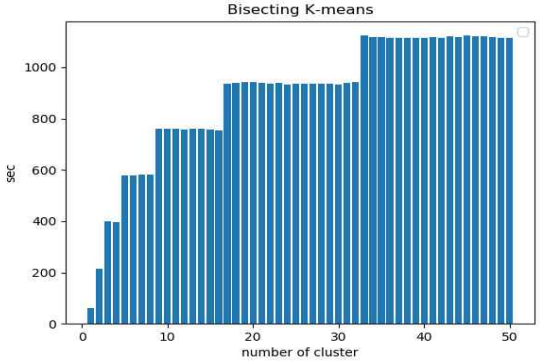

그림 2. Bisecting K-means 성능 그래프

3.4 Gaussian Mixture Model

Gaussian Mixture Model 알고리즘은 사용자가 직접 지정하는 k개의 정규분포 개수에 따라 시간이 비례적으로 증가하는 모습을 보인다. 정규분포의 개수가 증가함에 따라 클러스터링 시간도 배로 증가하기 때문에 클러스터 개수가 많은 경우에는 앞서 실험한 알고리즘들보다 많은 시간을 소비한다.

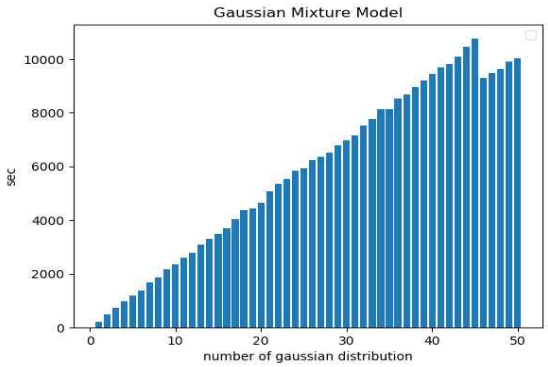

그림 3. Gaussian Mixture Model 성능 그래프

Ⅲ. 결론 및 향후 연구

본 논문 블록체인 시스템과 관련된 많은 연구를 위한 클러스터링 알고리즘 속도 성능을 실험하였다. 블록체인 시스템의 보안과 관련된 연구와 같이 주기적으로 클러스터링을 하는 경우 최소한의 시간 내에 적정 클러스터를 찾아 클러스터링하고 결과를 분석해야 하며, 새로운 데이터가 추가될 때 적정 클러스터의 개수가 달라질 수 있다. 따라서 주기적인 블록체인 데이터 연구에는 클러스터 개수에 영향을 적게 받고, 속도 성능이 가장 좋았던 K-means 알고리즘이 적합하다.

향후 연구로는 클러스터링 알고리즘들의 정확도 및 데이터가 분포된 클러스터의 특징을 파악하기 위해 Gaussian Mixture Model을 활용해서 데이터가 분포된 결과를 다른 알고리즘들과 비교하는 실험을 계획하고 있으며, 정상적인 블록체인 트랜잭션이 아닌 특징 있는 데이터 집단을 클러스터링함으로써 해당 집단이 속해있는 클러스터가 가지는 특징을 확인하는 연구를 진행할 계획이다.

참 고 문 헌

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

[2] Matthias Mettler, “Blockchain technology in healthcare: The revolution starts here”, 2016.

[3] Butian Huang, Zhenguang Liu, “Behavior pattern clustering in blockchain networks”, 2017.

[4] Jain, A.K "Data clustering: 50 years beyond K-means.", Pattern recognition letters, 2010.

[5] Savaresi, Sergio M.a | Boley, Daniel L.b "A comparative analysis on the bisecting K-means and the PDDP clustering algorithms", 2004.

[6] Carl Edward Rasmussen "The Infinite Gaussian Mixture Model", 2000.