유전알고리즘을 이용한 암호화폐 거래정보의 군집화 분석 및 분류

박준형* · 정석현 · 박은식 · 김경섭** · 원유재

충남대학교 컴퓨터공학과

Clustering analysis and classification of cryptocurrency transaction using genetic algorithm

Junhyung Park* · Seokhyeon Jeong · Eunsik Park · Kyungsup Kim** · Yoojae Won

Dept of Computer Engineering, Chungnam National University

E-mail : schwazodiac@gmail.com / arcane1028@gmail.com / kongse92@gmail.com / sclkim@cnu.ac.kr / yjwon@cnu.ac.kr

요 약

본 논문은 암호화폐 거래정보의 유사성과 거래패턴을 파악해서 군집화를 하고 학습을 통해서 다른 거래정보를 자동으로 분류해내는 모델을 제시한다. 유전알고리즘의 특성을 이용하여 군집화 과정에서 불필요한 요소를 최대한 제거하여 더 좋은 군집화 성능을 보여준다. 군집화 값이 포함된 거래정보를 훈련 데이터로 정하고 분류 알고리즘을 통해 거래정보의 예측이 가능해진다. 이는 암호화폐의 다양한 거래정보들로부터 자동으로 비정상 거래를 검출하는데 활용될 수 있다.

ABSTRACT

In this paper, we propose a model that classifies different transaction information by clustering and learning through similarity and transaction pattern of cryptocurrency transaction information. By using characteristics of genetic algorithms, we can get better clustering performance by eliminating unnecessary elements in clustering process. The transaction information including the clustering value is set as the training data, and the transaction information can be predicted through the classification algorithm. This can be used to automatically detect abnormal transactions from various transaction information of the cryptocurrency.

키워드

Blockchain, Bitcoin, Clustering, Genetic Algorithm, Classification

Ⅰ. 서 론

암호화폐의 급성장하면서 블록체인을 포함한 암호화폐와 관련된 많은 기술들이 주목을 받기 시작했고, 4차 산업혁명 돌풍에 있어서 하나의 핵심 키워드로 자리 잡기 시작했다.

2009년에 사토시 나카모토에 의해 처음 공개된 비트코인[1]을 시작으로 블록체인을 이용한 다양한 암호화폐가 생겨났다. 이 암호화폐의 특징은 기존의 화폐 시스템과는 달리 거래를 감독하는 중앙기관 없기 때문에 사용자가 원하는 거래를 Peer-to-Peer 네트워크를 통해 연결된다는 것이다. 또한 전 세계의 모든 네트워크 참여자들은 시스템의 전체 거래 데이터를 가지고 있어야 하며, 이를 통해 새로운 거래를 검증 및 인증하게 되는 구조이다.

하지만 암호화폐는 사용자에게 익명성을 제공하는 한편 모든 거래를 공개하는 특징을 가지고 있다. 이러한 특징으로 암호화폐를 이용한 거래는 개

* speaker

** corresponding author

인 사이의 거래뿐만 아니라 기부, 채굴거래, 랜섬웨어, 사기거래 등 다양한 거래방식이 존재한다. 이러한 사기거래의 방지, 탐지를 위해 기존 기업들이 비정상거래 방지를 위해 도입한 사기거래탐지(FDS), 자금세탁방지(AML) 등의 여러 가지 솔루션은 현금이나 카드거래를 기준으로 고안되었으며, 암호화폐의 거래는 적용할 수 없다는 문제가 있다. 이에 따라 암호화폐 상의 무수한 정보들의 연관성을 찾아서 거래상의 특징을 해석하려는 연구가 활발히 이루어지고 있다[2][3]. 특히, 암호화폐는 서로의 거래를 나타내는 트랜잭션에 거의 모든 정보가 포함되어 있으며, 이 정보는 다른 거래와 함께 블록들에 묶여있어서 블록체인의 특성과 거래에 대한 정보들을 내포하고 있다.

이처럼 본 논문에서는 암호화폐의 트랜잭션으로 군집을 형성하며, 해당 정보를 학습하여 거래를 분류하는 모델을 제시한다. 군집화 방법에는 기존의 K-mean, PAM, Clara 알고리즘을 이용하였다. 특히 군집화 과정에서 유전 알고리즘이라는 최적화 기법을 이용해서 지역최적화 문제에서 벗어나 연산속도와 정확성이라는 측면에서 좋은 성능을 낼 것으로 본다. 이를 실제 비정상 데이터에 적용하며, 기존의 분류모델과 비교를 통하여 분석한다.

논문의 구성은 2장에서 관련 연구에 대해 설명을 하고 3장에서 연구에서 사용된 모델에 대한 설명과 데이터의 처리 과정, 유전알고리즘을 통한 군집화 과정을 진행하며, 분류 알고리즘을 적용해 거래정보를 예측한다. 이후, 4장에서 군집화와 분류 결과에 대한 평가가 이루어지며 마지막 5장에서 결론을 맺는다.

II. 관련 연구

2.1 유전알고리즘(Genetic Algorithm)[4]

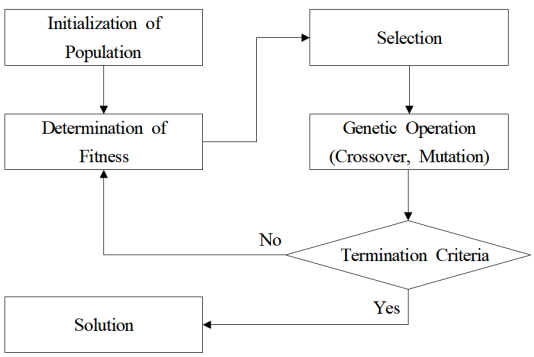

그림 1. 유전 알고리즘 흐름도

유전알고리즘은 자연세계의 진화과정에 기초한 계산모델로 최적화 문제를 해결하는 기법이다. 이 알고리즘은 그림 1과 같이 진행되며, 주요 연산 시스템으로는 초기화(Initialization), 선택(Selection), 유전 연산(Genetic operation), 종료(Termination)로 구분할 수 있다.

초기화(Initialization) 단계에서는 최적화 문제 해결을 위해 불규칙적으로 유전자 집단을 생성한다. 선택(Selection) 단계는 초기 유전자 집단에 대한 적합도 계산을 하고 문제 해결능력이 뛰어난 유전자들을 선택하게 된다. 유전 연산(Genetic operation) 단계에서는 선택된 유전자를 교배시키고, 일정 확률로 돌연변이가 나타나서 다음 세대를 위한 새로운 유전자 집단을 생성한다. 마지막으로 종료(Termination) 단계는 종료 조건에 도달했는지 검증하고 조건을 충족하지 못하면 세대교체를 하며 위의 단계를 반복한다.

2.2 암호화폐 블록 구조

블록체인 구조를 이루고 있는 블록들은 거래정보를 포함한 입출금에 관한 여러 정보를 가지고 있으며, 블록 헤더와 트랜잭션으로 구성된다[5]. 블록헤더는 해당 소프트웨어의 버전, 블록의 생성시간 등의 정보가 포함되어 있으며, 이 정보들을 이용해서 블록의 식별을 위한 블록 해쉬값이 결정된다. 생성된 블록 해쉬는 다음 블록의 이전 블록 해쉬에 연결이 되고, 처음 블록부터 이어져 하나의 체인 구조를 띄게 된다.

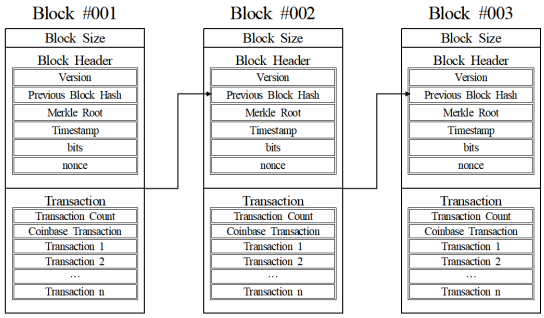

그림 2. 블록체인 구조

블록체인 구조를 기반으로 하는 암호화폐들은 거래가 이루어지는 행위들이 모두 트랜잭션이라는 단위로 각각의 블록에 기록이 된다. 트랜잭션은 이전의 트랜잭션들과 연결되어 있으며, 모든 정보들은 암호화 되어있지 않고 확인이 가능하다. 이러한 특징으로 하나의 트랜잭션에서도 블록정보나 입출력정보 등 다양한 값들을 나타낼 수 있다.

2.3 실루엣(Silhouette) 군집화 평가

군집화는 구분하려는 데이터들에 대한 지식 없이 유사한 패턴을 띄는 데이터를 무리지어 구별하는 과정이다. 이 과정에서 더 나은 군집화를 위해 현재 결과의 평가가 이루어져야 한다. 이에따라 평가의 기준으로 실루엣(Silhouette) 기법을 사용한다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

실루엣 $s(i)$는 각각의 데이터 i에 대해 같은 군집 내의 요소들 사이의 평균거리를 나타내는 $a(i)$와 다른 군집들 사이의 거리를 나타내는 $b(i)$를 비교한 값이다[6]. 이는 $a(i)$가 작을수록 군집 내부의 결집성이 높고, $b(i)$가 클수록 군집들 사이의 거리가 먼 것을 의미하므로 $s(i)$는 1에 가까울수록 올바른 군집화로, -1에 가까울수록 잘못된 군집화 알고리즘으로 평가한다. 이를 유전 알고리즘의 적합도 연산 척도로 사용하여, 동일한 군집화 알고리즘 내에서 가장 좋은 모델을 찾는다.

2.4 분류 알고리즘(Classification Algorithm)

분류 알고리즘은 이미 알려진 몇 개의 그룹에 속하는 데이터들로부터 각각의 그룹이 어떠한 특징을 가지는지 분류 모델을 만든 후, 새 관측치가 어떤 그룹에 분류될지를 결정하는 것이다. 이 방법을 통해 기존의 트랜잭션으로 분류 학습을 진행한 뒤, 새로운 트랜잭션을 정상 또는 비정상으로 분류한다.

III. 시스템 설계

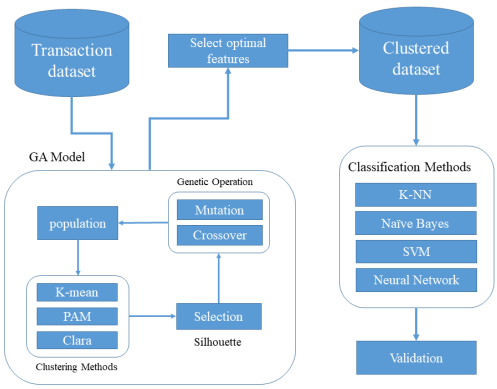

그림 3. 시스템 구조

전체적인 분석모델은 그림 3과 같이 구성하였다. 트랜잭션 데이터들을 유전 알고리즘을 통해 최적의 군집화 요소를 찾는다. 이 후, 가장 좋은 군집화 정보를 이용해 군집화 데이터셋을 만들고, 여러 분류 알고리즘을 통해 학습을 진행한다. 또한, 기존의 방법과 비교를 위해 데이터 군집화 과정에서부터 유전 알고리즘을 적용시키지 않는 모델을 만들며, 5개의 실제 비정상 address들의 트랜잭션을 통하여 군집화와 분류학습 모델들의 실제적인 성능비교를 한다.

3.1 데이터 수집 및 전처리

데이터 수집에 있어서 트랜잭션이 최근 들어 큰 폭으로 증가했다는 점과 사기 및 범죄 행위들을 포함한 비정상 거래 역시 최근 트랜잭션에 포함되어있다는 점[7]을 토대로 2017년 12월에 1105개의 Address에서 생성된 97,816개의 트랜잭션을 수집했다. 수집 데이터는 수정된 Bitcoin Core Client를 이용하였으며[8], 트랜잭션에 포함되지 않은 속성들은 block header의 정보를 통하여 추가하였다. 이 방법으로 표 1과 같은 12개의 속성을 가지는 트랜잭션 데이터를 구성하였다. 이 과정에서 주관적으로 불필요해 보이는 특정 속성을 제거하지 않고 모두 입력 데이터로 사용한다.

또한 트랜잭션의 크기나 거래량처럼 단위의 차이가 심하기 때문에 효율적인 학습을 위해 표준화(Standardization)를 적용시켰다.

표 1. 트랜잭션 데이터 속성

Input Transaction Features Definition

Number	features	Number	features
1	addr_ID	7	n_in
2	block_ID	8	n_out
3	tx_ID	9	input_seq
4	timestamp	10	prev_txID
5	n_txs	11	sum
6	output_seq	12	balance

Output Features

Number	features
1	clusters

3.2 데이터 군집화

전처리가 진행된 트랜잭션 데이터에는 여러 속성이 존재하며, 그 사이에는 블록의 번호처럼 군집화에 있어서 효율을 떨어트릴만한 부정적인 요소들이 존재한다. 하지만 이러한 의심스러운 요소들을 전처리를 통해 제거해 나가서 군집화를 하다보면 지역 최적화 문제에 빠질 가능성이 크므로[9] 이를 해결하기 위해 유전 알고리즘을 사용한다.

우선, 유전알고리즘은 트랜잭션 데이터의 속성여부가 이진수로 표기된 염색체 30개를 임의로 선별한다. 각각의 염색체들은 그림 3에서 제시된 K-mean, PAM, Clara알고리즘으로 군집화가 이루어지며, Silhouette 값을 통해 평가가 매겨진다. 높은 점수들의 염색체를 선별해서 80%의 교차연산과 10%의 변이연산이 이루어지고 다음 세대를 위한 새로운 염색체들이 구성된다.

이런 방식으로 총 10세대의 진화를 거쳐서 군집화에 있어서 최적의 트랜잭션 속성들을 찾게 되며, 이를 통해 평가가 가장 높은 모델을 이용해서 최종적으로 군집화를 한다.

3.3 데이터 분류 학습

최적의 성능을 보이는 군집화 모델을 이용해서 분류 학습에 필요한 데이터셋을 생성한다. 데이터 분류에 있어서 사용될 모델은 K-Nearest Neighbor, Naive Bayes Classifier, Support Vector Machine, Neural Network 총 4가지 모델을 사용하며 각각의 성능에 대한 비교가 이루어진다. 각각의 모델들은 동일하게 70%의 훈련데이터와 30%의 검증데이터로 구성하였으며, 각각의 학습을 30번씩 진행하여 정확도의 평균을 측정하였다.

Ⅳ. 결과 및 분석

4.1 군집화 결과 분석

K-mean, Clara, PAM에 대한 세 가지의 GA모델을 사용하여 최적화를 진행하였고, 이 과정을 10회 반복한 평균은 아래 그림 4와 같은 결과를 보인다. K-mean 알고리즘을 적용한 모델에서는 세대를 거듭할수록 Silhouette 값을 나타내는 유전적합도 값이 증가하는 추세를 보인다. Clara나 PAM의 경우는 소폭으로 평균값과 중앙값이 상승하지만 적합도의 최대값이 증가폭이 작아 최적화에 있어서 부적합한 모습을 보였다.

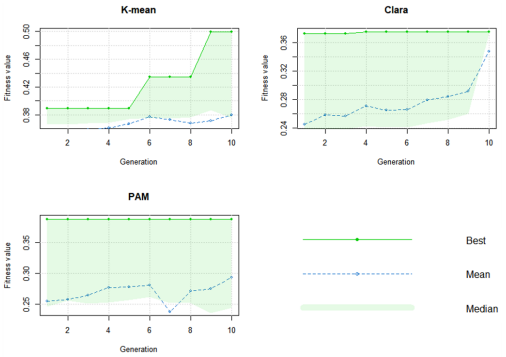

그림 4. 알고리즘 별 유전 적합도

다만 표 2에서처럼 모든 알고리즘들은 전체 속성을 모두 포함시킨 기존의 알고리즘과 유전 알고리즘모델을 비교했을 때, 값이 상승을 한 것을 볼 수 있으며, 이는 유전알고리즘으로 군집화의 효율을 높인다고 볼 수 있다.

표 2. 군집화 모델별 평균 Silhouette

	K-mean	Clara	PAM
Original Dataset	0.378816	0.226324	0.215183
GA Selected Dataset	0.445173	0.399716	0.398551

그림 5에서는 주성분분석을 통해 각각의 군집들의 상태를 시각적으로 확인할 수 있다. PAM과 Clara를 적용한 모델에서는 군집화 방식에서 기존과 큰 차이를 보이고 있지 않았다. 하지만 그림 5의 K-mean에 대한 군집의 분포를 보면, 매우 밀집된 부분과 일반 데이터들과 큰 편차를 보이는 부분으로 나타나는데 이는 블록체인 거래에 있어서 트랜잭션의 90%이상이 소수의 거래로 이루어져있는 형태[10]와 매우 유사하다.

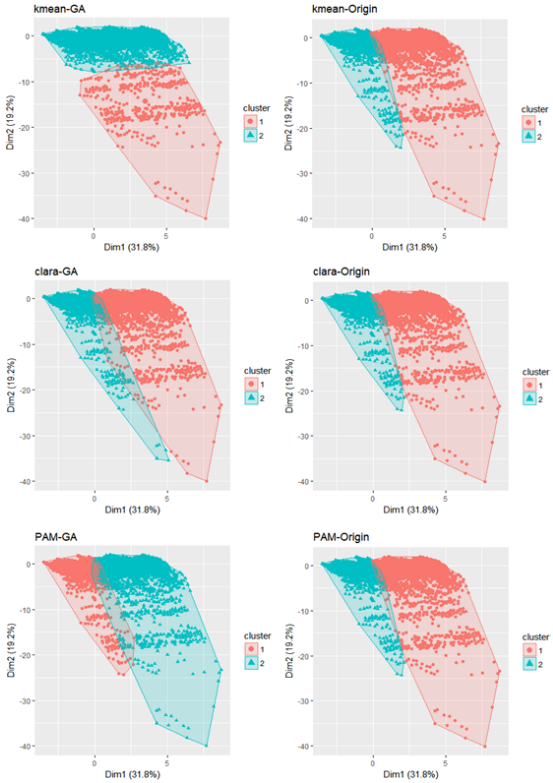

그림 5. 주성분분석을 이용한 군집화 결과

표 3에서는 K-mean을 이용하여 실제 비정상거래로 판명된 5가지 address와 거래량 상위 10%와 하위 90% address들에서 발생한 트랜잭션이 얼마나 군집화가 되었는지를 보여준다.

표 3. K-mean 군집화 비교

	Original K-mean Clustering		GA Selected K-mean Clustering	
Cluster	A	B	A	B
Donation A	21	0	19	2
Donation B	229	0	220	9
Donation C	62	0	62	0
Fraud A	291	0	286	5
Fraud B	83	0	58	25
10% Address	42236	52494	93008	1722
90% Address	3086	0	2922	164
Total	45322	52494	95930	1886

기존의 군집화에서는 두 개의 군집을 거의 동일한 비율로 나눈 것을 볼 수 있다. 이는 그림 5의 오른쪽 그래프와 같으며, 실제 비정상거래의 경우 하나의 군집에만 속한 것으로 나온다.

반면에 유전 알고리즘을 이용한 군집화에서는 두 군집의 분포 비율이 전반적으로 일정한 모습을 띄었다. Fraud B 사기 거래의 경우는 전체 데이터의 43%나 차지하는 수치를 보여주며 해당 값이 정

상적이지 않다는 것을 보여준다. 이처럼 실제 비정상 데이터들을 군집화 하는데 있어서 기존의 방법보다 향상된 결과를 보여준다.

4.2 분류 결과 분석

기존 군집화와 유전 알고리즘을 이용한 군집화를 이용해서 분류를 한 결과는 표 4와 같이 나타났다. 두 모델에서는 군집화 알고리즘으로 학습데이터들이 생성되어서 분류 알고리즘과 유사성으로 인해 높은 정확성을 보인다.

또한 유전 알고리즘을 이용한 분류 결과를 살펴보면, Naive Bayes를 제외한 다른 모델에서 1% 내외로 모두 향상된 모습을 보이고 있다.

표 4. 분류 모델별 평균 정확도 비교

	K-NN	Naive Bayes	SVM	NN
Original Accuracy	0.99583	0.95928	0.99804	0.99701
GA Accuracy	0.99930	0.95877	0.99959	0.99725
rate of change	+0.34%	-0.05%	+0.19%	+0.02%

Ⅴ. 결 론

본 논문에서는 유전 알고리즘을 통한 군집화방법과 분류학습에 적용시켜 기존의 모델과의 비교를 하였다. 군집화 과정에 있어서 유전 알고리즘이 K-mean과 같은 특정 모델에 성능이 향상되는 결과를 보였으며, 실제 비정상 데이터에도 유의미한 결과를 나타내는 군집화 성능을 보였다. 또한 분류학습에는 다수의 모델에서 소폭의 정확도 향상을 보였으나 기존과 큰 차이를 보이는 결과는 나타나지 않았다.

제시된 모델은 실제 비정상 트랜잭션을 포함한 사용자를 탐지하는 데 있어서 기존의 모델에 비해 향상되었음을 알 수 있다. 즉, 암호화폐에서 발생하는 트랜잭션들의 비정상 여부에 대한 평가가 자동으로 이루어 질 수 있다. 또한, 위의 모델을 활용하여 암호화폐 거래에 대한 사기거래탐지모델 구축이 가능하다.

향후 전체 블록체인 네트워크에 대한 유전 알고리즘을 적용할 수 있는 시스템을 구축을 통해, 향상된 암호화폐 거래 탐지 모델을 설계하는 연구가 필요하다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2018-2016-0-00304)

References

[1] Bitcoin: A peer-to-peer electronic cash system [Internet]. Available: https://bitcoin.org/bitcoin.pdf

[2] A Bayesian Approach to Identify Bitcoin Users [Internet]. Available: https://arxiv.org/abs/1612.06747.

[3] Data mining for detecting Bitcoin Ponzi schemes [Internet]. Available: https://arxiv.org/abs/1803.00646.

[4] K.F. Man, K.S. Tang, S. Kwong, “Genetic Algorithms: Concepts and Applications”, IEEE Transactions on Industrial Electronics, Vol. 43, No. 5, pp. 519-534, Oct. 1996.

[5] Bitcoin Developer Reference [Internet].Available: https://bitcoin.org/en/developer-reference.

[6] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”, Journal of Computational and Applied Mathematics, Vol. 20, pp. 53-65, Nov. 1987.

[7] Bitcoin hits new record high as warnings grow louder [Internet]. Available: https://www.reuters.com/article/us-global-markets-bitcoin/bitcoin-hits-new-record-high-as-warnings-grow-louder-idUSKBN1E919T.

[8] D. Kondor, “Do the rich get richer? An empirical analysis of the BitCoin transaction network”, PLOS ONE, Vol. 9, No. 5, Feb. 2014

[9] D. Steinley, “Local optima in K-means clustering: what you don't know may hurt you.”, Psychol Methods, Vol. 8, No. 3, pp. 294-304, Sep. 2003.

[10] D. Ron, A. Shamir, “Quantitative Analysis of the Full Bitcoin Transaction Graph”,in Financial Cryptography and Data Security, Berlin, pp. 6-24, 2013.