Block Analysis in Bitcoin System Using Clustering with Dimension Reduction

Mu-Gon Shin
Computer and Information Science
Korea University
Sejong, Korea
tm0309@korea.ac.kr

Ui-Jun Baek
Computer and Information Science
Korea University
Sejong, Korea
pb1069@korea.ac.kr

Kyu-Seok Shim
Computer and Information Science
Korea University
Sejong, Korea
kusuk007@korea.ac.kr

Jee-Tae Park
Computer and Information Science
Korea University
Sejong, Korea
pjj5846@korea.ac.kr

Sung-Ho Yoon
A&B Center
LG Electronics
Seoul, Korea
sungho.sky.yoon@lge.com

Myung-Sup Kim
Computer and Information Science
Korea University
Sejong, Korea
tmskim@korea.ac.kr

***Abstract*—The online cryptocurrency bitcoin, created based in blockchain technology, is attracting the attention of individuals, businesses and the government as well. As interest in blockchain technology and cryptocurrency has steadily increased over the past few years, trading volume and market size of cryptocurrency have increased at an astonishing speed. As a result, analysis and monitoring measures for blockchain networks, blocks, and transactions have become an important issue. In this paper, the method of clustering applied dimension reduction as a method of bitcoin network analysis is proposed. The proposed method applies the analysis way using K-means algorithm with PCA to block data in bitcoin collected by this research team.**

Keywords—blockchain, clustering, PCA, dimension reduction, analysis, block, transaction

I. INTRODUCTION

Bitcoin, developed by Satoshi Nakamoto in October 2008, is an online cryptocurrency created based on blockchain technology[1]. As interest in blockchain technology and cryptocurrency has steadily increased over the past few years, transaction amount and market size of cryptocurrency have increased at an alarming rate. As of April 2019, Bitcoin's average daily transaction volume(transaction count) was about 380,000 cases. Although transaction amount of bitcoin is increasing and interest in blockchain is deepening, there is not much research in monitoring and analysis about blockchain network, blocks, transactions and so on. Monitoring and analyzing cryptocurrency blocks and transactions is critical as more and more illegal transactions are made through cryptocurrency, including bitcoin.

The K-Means algorithm, one of the clustering algorithms, is an algorithm that binds a given data into k's clusters and works in a way that minimizes the variance of each cluster and the distance difference[2]. Block and transaction data from bitcoin, consisting of several features, can be grouped together through clustering. Analyzing the characteristics of each cluster by the configures cluster, and analyzing the feature of the data that is far from the clusters, called outlier, is also important task. The outlier might be bad transactions or blocks. Finding the number of k, the number of clusters, that can bets characterize data in K-means clustering is very big issue. We use the elbow technique to find an effective number of k. It is also very important to select features for effective clustering. We will use the And dimensionally shrinking can be an important issue in visualizing clustered data.

One of the dimensional reduction algorithms, PCA (Principal Component Analysis), is a technique that returns high-level data to low-dimensional data[3]. Because block data is represented by multiple high-dimensional data, we converted high-level block data into low-level data using PCA to use this information for clustering. In addition, data converted to low dimensions can be visualized in two-dimensional or three-dimensional graphs to clearly identify the distribution of the data.

We consider block data and transaction data analysis of bitcoin as a clustering problem. Clustering is an important class of unsupervised learning problems[4], which focuses on splitting data into groups and has a variety of approaches to its solution [5], [6]. However, because bitcoin data is high-level data, it is difficult to visualize clustering results. Therefore, this paper propose the clustering technique applying dimension reduction to bitcoin data.

Following the introduction of Section I, Section II lists and classifies related works. Section III describes the key algorithms, and Section IV explain the collecting data. Section V describes the experimental procedure including data preprocessing and all preprocessing. Section VI describes the conclusions, including limitations and future works.

II. RELATED WORK

A. Bitcoin Address Clustering

In general, the analysis of bitcoin has a few research on clustering to the address of bitcoin wallet, mining pool, or exchange[7]. Address clustering tries to construct the one-to-many mapping from entities to addresses in the Bitcoin system[8]. These studies use addresses to find a mining pool, to find an exchange or to find similar wallet. But there is few clustering study of bitcoin blocks or transactions.

B. Bitcoin Price Clustering

As cryptocurrency is expected to become the next-generation trading currency, Bitcoin's price is also a subject of high interest. Therefore, there is a lot of analysis on the market trends and prices of bitcoin[9][10]. As Katsiampa (2017)

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2018R1D1A1B07045742) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2018-0-00539, Development of Blockchain Transaction Monitoring and Analysis Technology)

notes, Bitcoin is the most popular cryptocurrency with 41% of the estimated cryptocurrency capitalisation in Bitcoin. However little is known about the behavior Bitcoin prices. Sure, these study is also important to understand Bitcoin, but analyzing blocks and transactions of Bitcoin is more important.

III. Key Method

This chapter introduces PCA and K-means algorithms that are key algorithms in this paper. Also mentions the elbow technique to find the right number of clusters in K-means clustering.

A. Principal Component Analysis(PCA)

This section refers to PCA, which is a technique for finding new baselines(axes) that are orthogonal to each other while preserving the greatest variances of data, and converting samples of high-dimensional space into low-dimensional space with no linear association[3].

$$\arg\min_{\hat{x}} \|x - U\hat{x}\|^2$$

In above Equation, U means a matrix of reversals, and x means the original vector. $\hat{x}$ also means a dimensionally reduced vector that is most similar to the original vector. Through this process, PCA obtains a new reduced level vector that is most similar to the original input vector.

In general, PCA consists of the following steps:

1) Locate the axis with the largest variance in the learning data set
2) Locate the second axis with the largest variance, orthogonal to the first axis found.
3) Locate the third axis, orthogonal to the first and second axis and preserving the variance as much as possible.
4) Locate the axis by the dimension(number of features) of the dataset in the same way as 1 to 3.

In this paper, two methods are used: K-means applied after applying PCA to learning dataset and PCA applied after K-means applied.

B. K-means Clustering

This section refers to K-means clustering. An algorithm that binds a given data into k clusters and operates in a way that minimizes the variance of each cluster and the distance difference. This algorithm is a type of unsupervised learning, and it acts as a label for input data that is not labeled.

When n d-dimensional data object(x1, x2, … , xn) set is given, the K-means algorithm maximizes the cohesion of n data objects within each set, with the maximum of k-set S = {S1,S2, … , Sk}. In other words, when μi is the center of the set Si, the goal is to find a set S at the center point of each group, which minimizes the sum of squares between the objects in the group.

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

K-Means repeat the following two steps:

1) *Cluster settings :* Calculate the Cuclidean distance from each data to the center of each cluster, locate the nearest cluster from that data and allocate the data.

$$S_i^{(t)} = \{x_p : \left|x_p - \mu_i^{(t)}\right|^2 \le \left|x_p - \mu_j^{(t)}\right|^2 \forall j, 1 \le j \le k\}$$

2) *Cluster-centric rebalancing :* Resets the center of the cluster to the center of weight of the data in each cluster.

1) $\mu_j^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x \in S_i^{(t)}} x_j$

K-means clustering has some limitations. The K-means algorithm should specify the number of clusters k as a parameter. The setting of k value is very important because the resultant value depends entirely on the number of clusters. Therefore, in this paper, elbow method was used to overcome this limitation and set appropriate k values.

C. Elow Method

Elbow method is a technique that finds the appropriate number of clusters in the K-means algorithm. In the process of determining the center so that the agreed error-squared values are minimal, the number of clusters is increased one by one to obtain the k value of the error-squared when the value of error-squared becomes significantly smaller.

The results are monitored by increasing the number of clusters sequentially. If adding one cluster does not produce much better results than before, set the number of previous clusters to be obtained.

IV. Data

The experiments in this paper use bitcoin block data, transaction data, and transaction statistics data collected by this research team. This section introduces the data used in the experiment.

A. Block Data

Bitcoin block is consist of block header and data part. The block data can be collected through the *getblock* command of Bitcoin client program, and transaction data contained in the block was also collected by setting *verbosity* option to 2. Information from block data except hash values is used. The block level data collected are as follow:

- **Height** : The height(number)of block.
- **nTx:** The number of transactions.
- **Size:** The size of block.
- **Nonce:** The number of block mining calculations.
- **Weight:** The weight of block.
- **Difficulty:** The difficulty of block.
- **Confirmations:** The number of confirmations.
- **Strippedsize:** The strippedsize of block.

We use this 8 features for block data clustering. Hash value such as merkleroot, previousblockhash, hash, nextblockhash and so on is excepted because they cannot be clustered.

B. Transaction Data

- **nVin:** The number of inputs the transaction contains.
- **nVout:** The number of outputs the transaction contains.

- **Value:** The amount of value of transactions.
- **Fee:** The fee of the transaction.
- **Tx_Size:** The size of each transaction.
- **Tx_vSize:** The virtual size of each transaction.

Finally, input and output level data collected is as follows:

- **Vin_Value:** The transaction input's value**.**
- **Vout_Value:** The transaction output's value.

C. *Transaction Statistic Data*

We extract statistical data using the collected block data and transaction data of Bitcoin. The statistical data consist of five total values: summation, minimum, maximum, average, and standard variation. Extraction data is based on blocks and transactions. The extracted statistical data are described in table I. We will use this data for analysis of clustered data, means outlier.

TABLE I. Figure 1. Result of clustering before PCA applied
Transaction statistic data

Data	Raw data	Extraction	Number of data
Transaction	nVin	Sum Max Min Avg Stdv	5
	nVout		5
	Value		5
	Fee		5
	Tx_vSize		5
	Tx_Size		5
	Vout_value		25
	Vin_value		25
Total number of data			80

V. Experiment and Result

A. *Result of K-means after Applying PCA*

We show the experimental results on various parameters in 2d-graph and analyzed them through this graph. We use block data mentioned section IV. Table II describe the result of the K-means with data applied PCA and the PCA after K-means applied. The result of PCA after K-means applied is not mentioned this paper. But the results are described for comparison of the data. Also elbow method was used to find the optimal number of cluster k.

The data used in the experiment (figure 2) are block height 0 to 218,593 blocks. Not all block statistical information has yet been obtained because it takes a long time to calculate the block statistical information to be used for the analysis of the results. Once all the statistics are collected, clustering experiments will be carried out on all blocks created to date.

In figure 1, We can see that the data is clustered in three clusters. Each cluster has data that is detached from the center, which we define as an outlier. The analysis will be performed using statistical information from data (block) that is far from the center and is less cohesive.

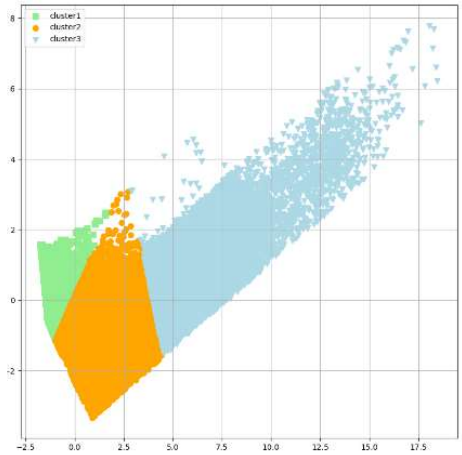

Figure 1. K-means result with data applied PCA (0-218,593)

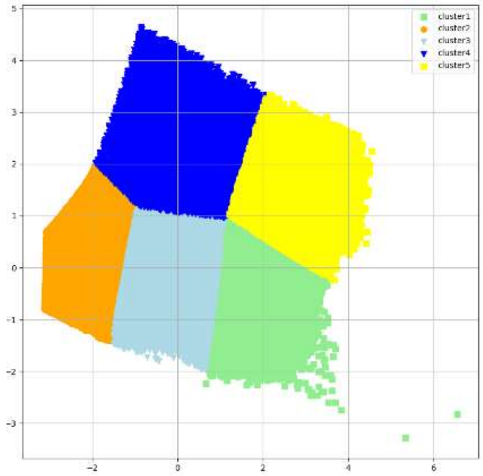

Figure 2. K-means result with data applied PCA (200,000-565,691)

In figure 2, Five clusters have been created, which could mean that the less noise in block information, the more detailed features there are: we use block height 200,000-ㄴ 565,691 that is the information except for block data with low variation. This shows that the clusters of blocks become tighter if data has no noise. The graph also shows that the number of outliers has decreased a lot.

B. *Cluster Analysis*

This section conducts an analysis of the clusters derived from the experiments introduced earlier. First, we mention the number of members in each cluster and then introduces the characteristics of each cluster. In addition, the characteristics of each cluster were analyzed by means and by standard variance. The number of each cluster is described in table II. The mean values and variance values for each cluster are described in table III through VI. And each category's number is used in table III to VI.

TABLE II. Cluster information about Experiments

Category	Number of cluster	Number of items
1. K-means with data applied PCA (0 to 218593)	3	121,871
		83,456
		13,267
2. K-means with data applied PCA (200,000 to 565691)	5	108,141
		149,685
		50,735

		17,751
		39,380
3. PCA after K-means applied (0 to 218593)	3	129,301
		76,646
		12,647
4. PCA after K-means applied (200,000 to 565691)	2	169,749
		195,943

TABLE III. Cluster Characteristic of category 1

Average						
Cluster #	nTx	size	nonce	weight	difficulty	strippedsize
1	3.77	1275.85	1.17E+09	5103.42	40175.77	1275.85
2	62.23	27098.18	2.39E+09	108392.73	1603560.85	27098.18
3	469.57	213564.6	2.12E+09	854258.21	2669305.58	213564.55
Standard Variance						
Cluster #	nTx	size	nonce	weight	difficulty	strippedsize
1	9.54	3504.41	1.24E+09	14017.67	176520.83	3504.41
2	60.64	25950.78	1.25E+09	103803.14	843786.64	25950.78
3	190.66	71906.38	1.24E+09	287625.54	674524.99	71906.38

TABLE IV. Cluster Characteristic of category 2

Average						
Cluster #	nTx	size	nonce	weight	difficulty	strippedsize
1	290.50	146651.9	2.15E+09	586607.6	1996143268	146651.9
2	31.874	14355.37	1.5E+09	57421.49	24177883.45	14355.37
3	110.80	51301.43	2.09E+09	205205.7	192064023.4	51301.43
4	570.58	327272.3	2.14E+09	1309089	36648268585	327272.3
5	624.59	358044.2	2.15E+09	1432177	38357867903	358044.2
Standard Variance						
Cluster #	nTx	size	nonce	weight	difficulty	strippedsize
1	264.68	133542.3	1.24E+09	534169.1	4316600383	133542.3
2	100.02	46834.56	1.38E+09	187338.3	393810586.1	46834.56
3	180.84	84298.44	1.27E+09	337193.7	1642273628	84298.44
4	499.76	259909.8	1.24E+09	1039639	10753741124	259909.8
5	555.26	277790.4	1.24E+09	1111162	12222124072	277790.4

TABLE V. Cluster Characteristic of category 3

Average						
Cluster #	nTx	size	nonce	weight	difficulty	strippedsize
1	4.3318	1478.507	1.36E+09	5914.027	27730.67	1478.50
2	68.496	29997.92	2.18E+09	119991.7	1785889.45	29997.92
3	480.27	218224.6	2.16E+09	872898.2	2662301.41	218224.6
Standard Variance						
Cluster #	nTx	size	nonce	weight	difficulty	strippedsize
1	10.934	3886.43	1.36E+09	15545.74	89462.13	3886.43
2	63.196	27200.31	1.25E+09	108801.2	699747.67	27200.31
3	188.57	70389.38	1.24E+09	281557.5	678416.60	70389.38

TABLE VI. Cluster characteristic of category 4

Average						
Cluster #	nTx	size	nonce	weight	difficulty	strippedsize
1	74.743	37843.28	1.64E+09	151373.1	2115824893	37843.28
2	364.58	195021.5	2.14E+09	780086.1	11639087951	195021.5
Standard Variance						
Cluster #	nTx	size	nonce	weight	difficulty	strippedsize
1	204.93	108079.8	1.38E+09	432319.2	8754739931	108079.8
2	394.84	206940.8	1.24E+09	827763.2	17742199069	206940.8

VI. Conclusion

In this paper, we proposed a clustering method for bitcoin block and transaction data analysis. The proposed method defines the data that can be collected from the Bitcoin network and the statistical data from the blocks that can be extracted from the collected data. We conducted a clustering experiment by applying PCA to the extracted data, and we also tested how to apply PCA to the clustered data. By experimenting with various clustering, the results were derived as to which method was more effective. In conclusion, the results of the clusters produced by each experiment were analyzed. Although analysis of cluster results was conducted only on block data, analysis using statistical data introduced in this paper will be carried out in future studies. The analysis will also be conducted by setting the criteria for the outliers of each cluster.

References

[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

[4] Ghahramani, Zoubin. "Unsupervised learning." Advanced lectures onmachine learning. Springer Berlin Heidelberg, 72-112. 2004.

[5] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 264-323. 1999

[6] S. Fortunato. "Community detection in graphs." Physics reports 486.3.Pp. 75–174. 2010.

[7] Dmitry Ermilov. "Automatic Bitcoin Address Clustering" 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017

[8] Martin Harrigan, "The Unreasonable Effectiveness of Address Clustering", 2016

[9] Andrew Urquhart , "Price clustering in Bitcoin", Economics Letters. pp 145-148. 2017

[10] P Ciaian. "The economics of BitCoin price formation", Applied Economics. pp1799-1815. 2015