

Analysis of Refined Big Data Clustering Algorithm using Power Method on Block Chain Network

D. Jayalatchumy

Assistant Professor/ Dept. of Computer Science and Engineering
Perunthalaivar Kamarajar Institute of Engineering and Technology, Karaikal. India
djpkiet@gmail.com

Abstract: The need to redesign and convert the traditional clustering technique has paved a way to analyze and refine a new clustering algorithm that address the challenges like efficiency and scalability when dealing with Big Data. Clustering deanonymizes block chain data. It links various blocks that belongs to the same entity. It ensures secure transaction and reduces the cost. Also, it speeds up data processing. This paper aims in refining the clustering technique for block chain data that can process huge volumes of data using the Power Method. In addition, the block chain system helps to secure and interpret the information.

Keywords— Power Method, Clustering, Big Data, Block Chain, Inflation.

1. INTRODUCTION (Heading 1)

Clustering is an unsupervised Learning Algorithm that splits data into several subsets. It makes a group of abstract objects into classes of similar object. cluster analysis works as a means to achieve understanding into the distribution of data to detect characteristics of each cluster. There exist many clustering algorithms [4,7] that can handle outliers easily and can handle high dimensional datasets, but most of them are not suitable for larger datasets [5,11]. Though clustering is an efficient tool to analyze big data there come an issue on applying clustering techniques to Big Data [1,6,8,9,16,17]. They have high computational cost and consumes much time. Hence, existing algorithms has to be refined to handle Big Data.

Block chain is a distributed database that shares and processes data over a network in a secure manner. There is no any central authority over the network. The main disadvantage of using block chain is massive energy consumption. Each block chain contains a timestamp, a hash value and a link to the previous block. The ledger records the transactions and manages the assets. A block chain network is the

infrastructure through which the applications can access the ledger and the smart contracts. Block chain network [21,22] are classified into four types as shown in Table 1.

Table 1. Types of Block Chain Network

Public	Private	Hybrid	Consortium
Anyone in the world can join this network and can become a node of the network	People who have access and can join the network only can become a node in private	It is a combination of the Public and the private. Controlled by a single organization	The consortium block chain is governed by the group of organizations.
There is no central authority	It is Controlled by one authority	Controlled by one authority with some permission less process	It is Controlled by group.

Block chain helps to secure the information and also interpret them correctly. It uses the consensus algorithm that verifies all transactions hence assuring correctness and integrity. Block chain [23] can be used in various applications like communication, healthcare, finance, banking etc.

In this paper we have designed and clustering algorithm that uses the power method [2,18] and it is refined in such a way that it can handle huge dataset with low computational cost. The algorithm designed is efficient and implemented using the block chain SHA 256 algorithm [19] since it adds to secure and privacy of the information.

2. POWER METHOD

Power method is a simple iterative method to find the dominant eigen value and eigen vector for a matrix A where λ is the largest of eigen value and v^0 is the largest of eigen vector. The power method is found using [2]

$$v^{t+1} = cWv^t$$

where v^t is the vector at iteration t.

W is the Square matrix

c is the normalizing constant to keep v^t from getting too large or too small.

Since the power method for approximating the eigen value is iterative the initial approximation x_0 of the dominant eigen vector is chosen.

2.1 DRAWBACKS OF THE POWER METHOD

- If the initial column vector $x^{(0)}$ is an eigenvector of A other than corresponding to λ_1 the dominant eigen value, then it fails or converges to a wrong value [2].
- The convergence of the power method depends on the magnitude of the dominant eigenvalue λ_1 and the magnitude of next largest eigenvalue. If the ratio is small then convergence become slow or can even fail if $|\lambda_1| = |\bar{\lambda}_1|$
- It gives only one pseudo eigenvalue and its difficult when it happens to be more than one dimension.
- It is complicated in the case of many eigen values.
- The power method needs an initial guess.

3. POWER ITERATION CLUSTERING (PIC)

Spectral Clustering is a method based on eigen value decomposition of a matrix [12]. It solves problems of different cluster shapes. Compared to many other clustering algorithms, spectral clustering has better performance. The most important disadvantage of this technique is its time and space complexity. The time complexity of spectral clustering is $O(n^3)$ because of eigen value decomposition of a $N \times N$ matrix. Power Iteration Clustering is a family of spectral clustering, designed by Lin and Cohen [13] to address the drawbacks of *spectral* clustering. PIC uses the power method to compute the largest eigen vector. Power Iteration is an iterative method that starts with an arbitrary vector and continuously performs its update. Given, a set of data points $x_1, x_2, \dots, x_n$ where x is a d dimensional vector, the similarity function [13] is given by

$$S(x_i, x_j) = \exp\left(\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$$

where σ is the parameter that controls the width.

The affinity matrix is built using $a_{ij} = S(x_i, x_j)$. The diagonal matrix is given by $D_{ii} = \sum_j a_{ij}$. The Laplacian matrix L is defined as $L = D - A$, here Δ is the identity matrix. The normalized affinity matrix is given by $W = D^{-1} A$. Instead of finding n eigen vectors PIC finds only one pseudo eigen vector that is a combination of eigen vectors.

A pseudo eigen vector is actually not the eigen vector but it is created linearly from it. The method of computing the eigen vectors of a matrix is as follows. First, the iterative vector is set equal to the random initial vector v^0 . This iteration vector is multiplied by matrix W. Then it is normalized to prevent the vector from becoming too small or too large. The power method is given as [2]

$$v^t = \gamma W v^{t-1}$$

where W is a square matrix, v^t is the vector at iteration t, and γ is the normalizing constant to prevent the vectors from becoming too small or too large. The vector is made to converge if the difference between the iterations has no change. At this point the local convergence is achieved but there is no global convergence or it may be slow. The procedure for PIC [13] is shown in Algorithm 1

Algorithm 1 Power Iteration Clustering

- Step1. Input the dataset and similarity function
 Step 2. Construct the similarity matrix $A \in R^{n \times n}$
 Step 3. Normalize the matrix $W=D^{-1}A$
 Step 4. Generate the initial vector using its norm function
 Step 5. Find its new vector and velocity by setting
- $$v^{t+1} = Wv^t \text{ and } \delta^{t+1} \leftarrow |v^{t+1} - v^t|$$
- Step 6. Repeat the process until the convergence criteria are met.
 Step 7. The convergence is decided when the difference between the two iterations is reduced.
 Step 8. Obtain the vectors and cluster the resulting vectors using K- means algorithm
 Step 9. Output the clustered results.

The computational complexity of PIC is $O(n^2)$ since, it requires matrix vector multiplication. PIC does not distribute the original data distribution and hence there is no loss in information. It is computationally fast and effective. It can be used for solving larger datasets. But, it becomes infeasible for matrices where its dimension is very large. Hence, new algorithms like IPIC and HPIC are proposed and are implemented on a distributed framework like Hadoop.

4. INFLATED POWER ITERATION CLUSTERING ALGORITHM (IPIC)

Spectral Clustering [12] consumes more time as it computes many eigen values and vectors. PIC uses the power method to find the largest eigen vector. But, it suffers from achieving global minima. i.e. it has slow global convergence hence Lagrangian constraints are introduced. This constraint induces an inflation [3] to obtain an optimal solution. Lagrangian multiplier is used to regulate the inflation where the lowest eigen vector grows exponentially fast relative to their neighbors. The Lagrangian approach is given by [3,14]

$$L = \sum_i x_i^2 - \sum_{ij} x_i A_{i,j} x_j + \lambda \left(\sum_i x_i^2 - 1 \right)$$

where, λ is the Lagrangian multiplier that enforces normalization. The corresponding Euler equation is

$$\ddot{x}_i = - \sum_j A_{i,j} x_j + \lambda(t) x_i, \quad i=1, 2, \dots, n.$$

Integrating with time δt gives,

$$x_i(t + \delta t) = x_i(t) + p_i(t + \delta t) \delta t \quad \text{where,}$$

$$p_i(t + \delta t) = p_i(t) - \sum_j [A_{i,j} - \lambda(t) \delta_{ij}] x_j(t) \delta t$$

Assume $\lambda(t)$ be a constant for small time step we have,

$$\xi_i(t) = a_i \cos(w_i t + \delta_i),$$

If $e_i > \lambda$, where δ_i is a real phase shift & $w_i = \sqrt{\lambda - e_i}$ for low eigen value $e_i < \lambda$, they coordinate as

$$\xi_i(t) = (a'_i e^{w'_i t} + b'_i e^{-w'_i t})$$

The eigen modes with eigen values $e_i < \lambda(t)$ gets inflated exponentially at a time t . $\lambda(t)$ standardizes inflation by handling the inflated and the non-inflated states. The need for eigen pair and accuracy can be adjusted based on the convergence constraints. Algorithm shows the steps involved in Inflated PIC. Based on the convergence factor, in this paper the computational time and accuracy is improved. The Algorithm 2 given below describes the steps for Inflated PIC [14].

Algorithm 2. Inflated Power Iteration Clustering

Input: Converted Normalized similarity matrix W

- Step 1. Generate the initial vector v^0 using its row sum.
- Step 2. Find the new vector v^t from w^t using the power method $v^t = \gamma W v^{t-1}$ where

$$\gamma = \frac{1}{||W v^{t+1}||}$$

- Step 3. Inflate the vector and perform update through iteration and normalize it from getting too large for the time δt using

$$v^t = \sqrt{e_1 - e_0} t$$

- Step 4. Calculate the new acceleration using

$$\delta t \sim \sqrt{\frac{e_1 - e_0}{e_{max} - e_0}}$$

- Step 5. Check the acceleration $|\delta^t - \delta^{t-1}|$ and increment the value of t.
- Step 6. On incrementing the value of t repeat steps 2 to 5 on incrementing the value of t and calculate the rate of convergence of the inflated algorithm using the constraint

$$(|\delta^t - \delta^{t-1}| \simeq 0 \ \&\& \ \frac{v_{max} - v_0}{2} \neq 0)$$

- Step 7. Cluster the points v^t on a k-dimensional subspace using K- Means algorithm.

Output: Clusters $c_1, c_2, \dots, c_k$.

4.1 ADVANTAGES OF IPIC

- IPIC uses the Lagrangian constraint that makes the algorithm converge a faster rate.
- IPIC is computationally better than other spectral algorithms.
- IPIC uses both the objective function and the Lagrangian constraints to find its convergence.
- The number of iteration in IPIC is lesser than DPIC and PIC.
- IPIC is highly scalable and suitable for handling larger datasets.
- IPIC is fast and easy for implementation.
- Though IPIC do not increases the accuracy it maintains the accuracy for most of the datasets.

4.2 IMPLEMENTATION OF IPIC

To evaluate the performance of IPIC the execution time is compared with the existing traditional algorithms. Experiments were carried out using an Intel core i3 processor, with 2.5 GHZ of processor speed, 4 GB RAM and 320 GB HDD. We compare the execution time of proposed IPIC with K- means [10,15] and PIC. The dataset includes IRIS, WINE, BREAST, HOUSE, BIRCH, and MNIST taken from the UCI machine learning repository database. The execution time is measured in milliseconds. Fig. 1 shows the graph comparing the execution time of various datasets. The datasets are given along the X-axis and the execution time measured in milliseconds (ms) are given along the Y-axis.

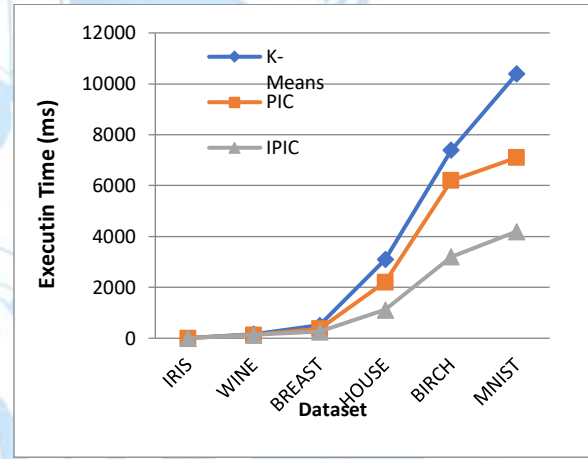


Fig 1. Execution time for IPIC

5. DATA CLUSTERING USING SHA 256

The refined algorithm is to be implemented using the SHA 256 algorithm. It is a popular hash algorithm. The hash or the digest is a unique string of characters that generates the hash value. The output of the hash will always be unique.

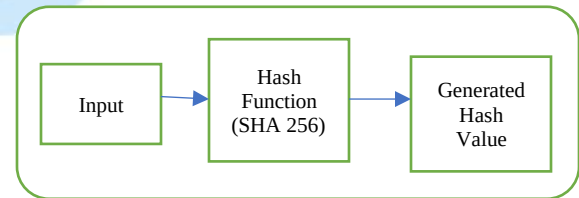


Fig 2. Hash Function

Hash function -The input can be either numbers, videos, string etc. the variable length is first transformed to a fixed length depending on the hash

function [20] that is being used. The procedure is shown in Fig 2. The use of hash functions ensures integrity and security. They can be used in digital signature the authentication the block chain transaction. Apart from it, hash function can be used as Merkle Tree to ensure integrity, chain to avoid modification.

6. BLOCK CHAIN AND BIG DATA CLUSTERING IN HEALTH CARE

Block chain can be used in healthcare with respect to drug traceability, clinical trials, supply chain management, medication monitoring, fraud prevention and data management. The proposed Big Data clustering algorithm [14] the IPIC was tested with a health record dataset that is collected from the Primary Health Centre nearby. IPIC algorithm is highly efficient when compared to the existing clustering algorithm. It has high execution speed and maintains its accuracy. The algorithm is to be implemented using the block chain technology using SHA256 algorithm to prove its effectiveness and privacy of the data. Using block chain the users can control access and can authorize sharing of information. Records can be stored and permissioned block chain can be used by Individuals or even third party access. Information can be stored on a block chain and can be allowed to accessed using trusted environment.

7. CONCLUSION AND FUTURE WORK

In this paper a simple Big Data Clustering algorithm based on the power method has been discussed. The power iteration clustering is refined to handle huge datasets using the inflation technique. Finally, the clusters are grouped using the K means algorithm. IPIC algorithm has higher and groups data quickly when compared to existing clustering algorithm. This algorithm has been implemented on health care dataset as well as UCI machine learning datasets to check its performance. IPIC seems to be highly efficient and fast. In future work has to be done to improve the accuracy of the algorithm. Also, several case studies are to be done to prove its applicability for Big Data.

REFERENCES

- [1] Ali Seyed Shirkhorshidi et al.2014. Big Data Clustering: A Review ICCSA 2014, Part V, LNCS 8583, (2014) pp.707–720.
- [2] Dartmouth, Numerical methods. Chapter 10.3 Power method for approximating eigen values, ENGS 9 Dartmouth College Hanover. New Hampshire.
- [3] Eric J.Heller, Lev Kaplan, Frank Polimann, Inflationary dynamics for matrix eigen value problems, PNAS, 105 22(June 2008), 7631-7635 doi/10.1073/ pnas. 0801 04 7105.
- [4] T. Sanjana et al. 2016.A Survey on Clustering Techniques for Big Data Mining. Indian Journal of Science and Technology.9, 3 (January 2016) DOI:10.17485/ijst /2016 /v9i3/75971.
- [5] Chanchal Yadav et al. 2013. Algorithm and approaches to Handle Large Data- A Survey. IJCSN International Journal of Computer Science and Network. 2, 3 (2013) ISSN 2277-5420
- [6] A. Sherin, S. Uma, K.Saranya and M. Saranya Vani. Survey on Big Data Mining Platforms, Algorithms and Challenges. International Journal of Computer Science & Engineering Technology. 5 (2014).
- [7] A. Fahad, N. Alshatri, Z. Tari, A. ALAmri, A. Y. Zomaya, I. Khalil, F. Sebt and A. Bouras.2014. A Survey of Clustering Algorithms for Big Data:Taxonomy & Empirical Analysis, IEEE transactions on emerging topics in computing (2014).
- [8] S.Arora, I.Chana. 2014. A survey of clustering Techniques for Big Data Analysis. Confluence the Next Generation Information Technology Summit Confluence, 5th International Conference-. IEEE (2014), pp. 59-65.
- [9] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan. 2014. Big Data Clustering: A Review. Computational Science and Its Applications–ICCSA (2014), pp. 707-720
- [10] Beste Eren 2015. A K-Means Algorithm Application Big Data in Proceedings of the World Congress on Engineering and Computer Science. WCECS 2015, (October. 2015), San Francisco, USA.
- [11] J. Dean et al. 2008. MapReduce: Simplified data processing on large clusters. Communications of the ACM. 51,1 (2008): pp.107–113.
- [12] Chen,W.Y , Song, Y, Bai,H. and Lin. C. Parallel spectral clustering in distributed systems, IEEE Transactions on Pattern Analysis and Machine Intelligence 33,3(2011) pp. 568–586.
- [13] Frank Lin frank, William W. 2010 Power Iteration Clustering, International Conference on Machine Learning, Haifa, Israel. (2010).
- [14] Dhanapal Jayalatchumy, Perumal Thambidurai. 2016. Inflated Power Iteration Clustering

- Algorithm to Optimize Convergence Using Lagrangian Constraint. Advances in Intelligent system and computing, Proceedings of 5th Computer Science Online Conference (CSOC 16) (2) pp. 227-238
- [15] Velade, O. J. et al.2010. Application of k-Means Clustering algorithm for prediction of Students' Academic Performance (IJCSIS) International Journal of Computer Science and Information Security.7, 1(2010). ISSN 1947-5500
- [16] Dhanapal Jayalatchumy, Perumal Thambidurai MIKE 2015: Proceedings of the Third International Conference on Mining Intelligence and Knowledge Exploration - Volume 9468, December 2015 Pages 370–381https://doi.org/10.1007/978-3-319-268323_35
- [17] D Jayalatchumy, P Thambidurai, Clustering algorithm based on Hadoop for big data, Advances in Natural and Applied Sciences 9 (6 SE), 92-97
- [18] Loughborough et al., “Workbook 22: Eigenvalues and eigenvectors,” Helping Engineers Learn Mathematics, 2008. http://ask.fxplus.ac.uk/tools/HELM/pages/workbooks150_jan_2008/Workbook22/22_4_numrcl_detrmntn_eignvl_eignvc.pdf
- [19] <https://komodoplatform.com/en/academy/sha-256-algorithm/>
- [20] <https://www.javatpoint.com/blockchain-hash-function>
- [21] <https://www.analyticssteps.com/blogs/4-types-blockchain-networks>
- [22] <https://cointelegraph.com/blockchain-for-beginners/a-beginners-guide-to-the-different-types-of-blockchain-networks>
- [23] Sudarshan S. Chawathe, Clustering Blockchain Data: Techniques, Toolboxes and Applications, In book: Clustering Methods for Big Data Analytics (pp.43-72), 2019DOI:10.1007/978-3-319-97864-2_3

