IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Volume-5, Issue-12 E-ISSN: 2347-2693

Survey on Partition based Clustering Algorithms in Big Data

E. Mahima Jane[1] and E. George Dharma Prakash Raj[2*]

[1]Department of Computer Application , Madras Christian College, Tambaram, India
[2]*Department of Computer Science and Engineering, Bharathidasan University, Trichy, India

[*]*Corresponding Author: georgeprakashraj@yahoo.com*

Available online at: www.ijcseonline.org

Received: 17/Nov/2017, Revised: 29/Nov/2017, Accepted: 15/Dec/2017, Published: 31/Dec/2017

***Abstract*-** Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. As Big Data is referring to terabytes and petabytes of data and clustering algorithms are come with high computational costs, the question is how to cope with this problem and how to deploy clustering techniques to big data and get the results in a reasonable time. This paper focuses on the traditional partition based clustering algorithms such as KMeans, K Medoids, PAM, CLARA and CLARANS and its advantages and disadvantages.

Keywords: KMeans, PAM, CLARA, CLARANS

I. INTRODUCTION

Big data analytics is the process of examining large and varied data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions [1].Clustering algorithms have developed as a powerful meta learning tool which can precisely analyze the volume of data produced by modern applications.Partitioning-based algorithms, all clusters are determined promptly. Initial groups are specified and reallocated towards a union. In other words, the partitioning algorithms divide data objects into a number of partitions, where each partition represents a cluster. These clusters should fulfill the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group. In the K-means algorithm, for instance, a center is the average of all points and coordinates representing the arithmetic mean. In the K-medoids algorithm, objects which are near the center represent the clusters. There are many other partitioning algorithms such as K-modes, PAM, CLARA, CLARANS[2].In this paper section II discusses the four partition algorithms. In Section III the advantages and disadvantages of the four algorithms and finally the conclusion in section IV.

II. PARTITION BASED CLUSTERING ALGORITHMS

Partition based clustering create k partition of data set with n data object. It is an iterative relocation technique is used to improve the clustering by moving up the object from one group to another. Partition based clustering is represent by centroid or medoid. [3]

K-Means Algorithm

The k-means algorithm is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as the centre of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.[4]

K-Means Algorithm:

```
    Step 1: Randomly select k data objects from data set
    D as initial centers.
    Step 2: Repeat;
    Step 3: Calculate the distance between each data
    object di (1 <= i<=n) and all k clusters C j(1 <=
    j<=k) and assign data object di to the nearest cluster.
```

Step 4: For each cluster j (1 <= j<=k), recalculate the cluster center.
Step 5: Until no change in the center of clusters.

PAM (Partitioning Around Medoids)

It was proposed in 1987 by Kaufman and Rousseau. It starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resultant clustering. It selects k representative medoid data items arbitrarily. For each pair of non-medoid data item x and selected medoid m, the total swapping cost S is calculated. If S< 0, m is replaced by x. Thereafter each remaining data item is assigned to cluster based on the most similar representative medoid. This process is repeated until there is no change in medoids. [5][6]

PAM Algorithm:

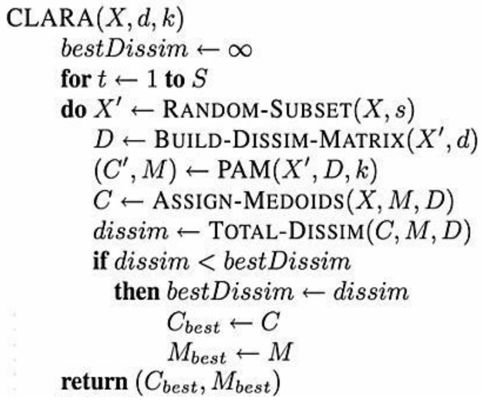
1. Use the real data items in the data set to represent the clusters.
2. Select k representative objects as medoids arbitrarily.
3. For each pair of non-medoid item xi and selected medoidmk, calculate the total swapping cost S(ximk). For each pair of xi and mk If S < 0, mk is replaced by xi Assign each data item to the cluster with most similar representative item i.e. medoid.
4. Repeat steps 2-3 until there is no change in the medoids.

Use real object to represent the cluster

- Select *k* representative objects arbitrarily
- For each pair of non-selected object *h* and selected object *i*, calculate the total swapping cost TC_{ih}
- For each pair of *i* and *h*,
  - If TC_{ih}< 0, *i* is replaced by *h*
  - Then assign each non-selected object to the most similar representative object
- repeat steps 2-3 until there is no change

CLARA (CLusteringLARge Applications)

CLARA was also developed by Kaufmann &Rousseeuw in 1990. It draws multiple samples of the data set and then applies PAM on each sample giving a better resultant clustering. It is able to deal more efficiently with larger data sets than PAM method. CLARA applies sampling approach to handle large data sets. Rather than finding medoids for the entire data set D, CLARA first draws a small sample from the data set and then applies the PAM algorithm to generate an optimal set of medoids for the sample. The quality of resulting medoids is measured by the average dissimilarity between every item in the entire data space D and the medoid of its cluster. The cost function is defined as follows: Cost(md,D) = ∑ n i-1 d(xi , rpst(md, xi) / n where, md is a set of selected medoids, d(a, b) is the dissimilarity between items a and b and rpst(md, xi) returns a medoid in md which is closest to xi . The sampling and clustering processes are repeated a pre-defined number of times. The clustering that yields the set of medoids with the minimal cost is selected. [7][8]



```
CLARA($X, d, k$)
    $bestDissim \leftarrow \infty$
    for $t \leftarrow 1$ to $S$
    do $X' \leftarrow$ RANDOM-SUBSET($X, s$)
        $D \leftarrow$ BUILD-DISSIM-MATRIX($X', d$)
        $(C', M) \leftarrow$ PAM($X', D, k$)
        $C \leftarrow$ ASSIGN-MEDOIDS($X, M, D$)
        $dissim \leftarrow$ TOTAL-DISSIM($C, M, D$)
        if $dissim < bestDissim$
          then $bestDissim \leftarrow dissim$
               $C_{best} \leftarrow C$
               $M_{best} \leftarrow M$
    return ($C_{best}, M_{best}$)
```

Figure.1. CLARA ALGORITHM [6,7]

CLARANS Algorithm

CLARANS draws sample of neighbours dynamically. This clustering technique mimics the graph search problem wherein every node is a potential solution, here, a set of k medoids. If the local optimum is found, search for a new local optimum is done with new randomly selected node. It is more efficient and scalable than both PAM and CLARA. [6]

```
CLARANS Algorithm
Set mincost to MAXIMUM;
For i=1 to h do  // find h local optimum
        Randomly select a node as the current node
        C in the graph;
        J = 1;  // counter of neighbors
        Repeat
        Randomly select a neighbor N of C;
        If Cost(N,D)<Cost(C,D)
        Assign N as the current node C;
        J = 1;
    Else  J++;
    Endif;
        Until J > m
        Update mincost with Cost(C,D) if
        applicableEnd for;
End For
Return bestnode;
```

III. ANALYSIS OF PARTITION BASED CLUSTERING ALGORITHMS

Table 1. Advantages and Disadvantages of Algorithms

S.No	Name of the Algorithm	Advantages	Disadvantages

1	K-Means	1. K-Means algorithm is simple and less expensive when compared to other clustering algorithms. 2. If the variables are large, then K-Means most of the time computationally faster than hierarchical clustering methods. 3. The results are easily interpretable and are often quite descriptive for real data sets. 4. The clusters are non-hierarchical and they do not overlap.	1. It is difficult to predict the K Value. 2. More difficulty in comparing quality of cluster.
2	PAM	PAM is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean	PAM works efficiently for small data sets but does not scale well for large data sets.
3	CLARA	Handles larger data than PAM	Depends on the Sample size
4	CLARANS	CLARANS is more efficient than PAM and CLARA in terms of Execution Time and Number of Iterations	1.It doesn't guarantee to give search to a localized area. 2. It uses randomize samples for neighbors.

The above Table1 describes the advantages and disadvantages of various partition based clustering algorithms.

IV. CONCLUSION

This paper we present the various partition based clustering algorithms. The size of the data generated every day is huge and the variety of data is also expanding day by day. This paper focuses on the partition based algorithms. Finally, an analysis of the four algorithms with their advantages and disadvantages is also given. Based on the disadvantages given in this paper, research on this topic can be done with respect to the partition based clustering algorithms.

REFERENCES

[1]. T Saha, K Dhas " *Inregration and Interelation of Bigdata With Cloud Computing: A Review* " International Journal of Computer Sciences and Engineering Vol.5(11), Nov 2017, E-ISSN: 2347-2693

[2]. Prateeksha Tomar, Amit Kumar Manjhvar, "*Clustering Classification for Diabetic Patients using K-Means and M-Tree prediction model*", International Journal of Scientific Research in Multidisciplinary Studies , Vol.3, Issue.6, pp.48-53, 2017.

[3]. Shalini S Singh, N C Chauhan," *K-means v/s Kmedoids: A Comparative Study*", National Conference on Recent Trends in Engineering & Technology, May 2011.

[4]. C. Zhang, and Z. Fang, *An improved k-means clustering algorithm*, Journal of Information & Computational Science, 10(1), 2013, 193-199.

[5]. 5.Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil A. Zomaya, S. Foufou, and A. Bouras, *A Survey of Clustering Algorithms for Big Data:Taxonomy& Empirical Analysis*, Accepted for IEEE transaction on emerging topics in computing 2014.

[6]. Gopi Gandhi, RohitSrivastava ,"*Review Paper: A Comparative Study on Partitioning Techniques of Clustering Algorithms* "-International Journal of Computer Applications (0975 – 8887) Volume 87 – No.9, February 2014

[7]. AzharRauf, Sheeba, SaeedMahfooz, Shah Khusro and HumaJaved" "*Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity* "Middle-East Journal of Scientific Research 12 (7): 959-963, 2012

[8]. Ali SeyedShirkhorshidi, SaeedAghabozorgi, Teh Ying Wah and TututHerawan, "*Big Data Clustering: A Review*", Research Gate, Jun, (2014)