Volume 5, Issue 4, 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Review of Various Enhancement for Clustering Algorithms in Big Data Mining

Rishikesh Suryawanshi
M-Tech Computer, MPSTME
SVKM'S NMIMS University,
Mumbai, India

Shubha Puthran
Assistant. Professor,
MPSTME, SVKM'S NMIMS University,
Mumbai, India

Abstract— The enlarging volumes of information emerging by the progress of technology, makes clustering of big data a challenging task. The K-means clustering algorithm is most commonly used algorithms for clustering analysis. The existing K-means algorithm is, inefficient while working on big data and improving the algorithm remains a problem. K-means algorithm is computationally expensive. The quality of the resulting clusters heavily depends on the selection of initial centroids. The existing k-means algorithm, does not guarantee optimality. The conventional database querying methods are inadequate to extract useful information from the large data sets. This paper discusses about the various methods for enhancing k-means clustering algorithm based on different research papers referred. These methods are Refined initial cluster centers method, A parallel K-means algorithm, A parallel k-means clustering algorithm based on Map Reduce technique, Determine the initial centroids of the clusters and Assign each data point to the appropriate clusters, An efficient enhanced k-means clustering algorithm, Variation in K-means algorithm and proposed parallel K-means clustering algorithm, A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm, Dynamic Clustering of Data with Modified K-Means Algorithm. In this review paper the limitations of clustering technique are mapped with the all the above mentioned modified k-means approaches.

Keywords— clustering, parallel kmeans, improvement kmeans, kmeans, Big Data

I. INTRODUCTION

Large volume of data processed by many applications will routinely cross the big-scale threshold, which would in turn increase the computational requirements. From the large amount of data, it is very difficult to access the useful information and provide the information to which it is needed within time limit. So data mining is the tool for extracting the information from big database and present it in the form in which it is needed for the specific task. The use of data mining is very vast. Cluster analysis of data is a principal task in data mining. Cluster analysis aims to group data on the basis of similarities and dissimilarities among the data elements.

Clustering method is the process of partitioning a given set of objects into dissimilar clusters. In this grouping data into the clusters so that objects in the same cluster have high similarity in comparison to each other, but are very dissimilar to objects in other clusters. Various Efficient methods is resolve the problem of large data clustering. Parallel clustering algorithms and implementation are the key to meeting the scalability and performance requirements in such scientific data analyses. By using Cluster analysis technique's it is easy to organize and represent complex data sets. K-means is a widely used partitional clustering method. The k-means algorithm is efficient in producing clusters for many applications [2] [7] [8].

In the Figure 1 shows basic Cluster formation while applying the K-means Clustering algorithm on the dataset. In the First part of the figure clusters the data object in the dataset according to the randomly selected initial centroid. In the next part of the figure the cluster is reformed by recalculating the centroid in the first iteration. In this stage figure shows that some of the data object is moved from one cluster to the other cluster. In the third part of the figure the centroid is not changed which means the convergence is occurred. All the data object is clustered to the respective cluster centroid. This cluster formation for the data object depends on the initial centroid selection.

The rest of the paper is organized as follows. Section II discusses the related work of the existing k means algorithms. The III section discusses the study of all the modified k-means algorithms. Section IV describes Conclusion.

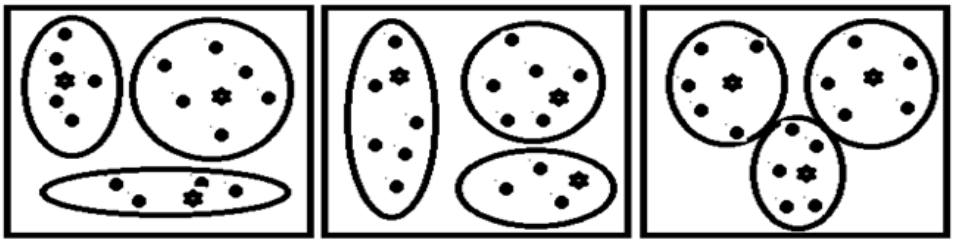

Fig. 1. Block Diagram of Cluster Formation

II. RELATED WORK

The "k-means" clustering algorithm was first used by James MacQueen in 1967 [3]. The basic original algorithm was first proposed by Stuart Lloyd in 1957 as a technique for signal processing, though it wasn't published until 1982. K-means algorithm is a widely used partitional clustering algorithm in the various application. The partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$ (data objects) [6]. It clusters the data into k (no. of cluster) groups, which together fulfil the following requirements:

i) Each group must contain at least one object, and
ii) Each object must belong to exactly one group [8].

K-means Algorithm [3]:

Input:
D = {d1, d2,.....,dn} //D contains data objects.
k // user defined number of cluster
Output:
A set of k clusters.
Steps:
1. Randomly choose k data-items from D as initial centroids;
2. Repeat the loop
Assign each item di to the cluster which has the closest centroid;
Calculate new mean for each cluster;
Until convergence criteria is met.

As shown in above Algorithm, the original k-means algorithm make up of two phases: In the first phase determining the initial centroids and the other for assigning data object to the nearest clusters and then recomputed the cluster centroids. The second phase is carried out continuously until the clusters get stable, i.e., data objects stop moving over cluster boundaries [3].

The k-means algorithm is effective in producing good clustering results for many applications [4]. The reasons for the popularity of *k*-means are ease and simplicity of implementation, scalability, speed of convergence and adaptability to sparse data [4]. K-means is simple and can be easily used for clustering of data practice and the time complexity is O(nkt), n is the number of objects, k is the number of clusters, t is the number of iterations, so it is generally regarded as very fast. Original k-means algorithm is computationally expensive and the quality of the resulting clusters heavily depends on the selection of initial centroids. K-means clustering is a partitioning clustering technique in which clusters are formed with the help of centroids. On the basis of these centroids, clusters can vary from one another in different iterations. Also, data elements can vary from one cluster to another, as clusters are based on the random numbers known as centroids [6].

The K-Means algorithm has some drawbacks. These are as follows

- Lack of knowledge how to treat with inappropriate and clatter attributes.
- Lack of universal method how to choose the initial location of cluster centroids.
- The algorithm stuck in local minima value [7].
- The K-means algorithm takes number of clusters (K) as input from the user. But in the practical scenario, it is very difficult to fix the number of clusters in advance [8].
- Lack of universal method how to choose the initial location of cluster centroids because the result of K-means algorithm depends on initial clustering centres; different centroids can cause different clusters, and can even lead to no resolution.

Various Application of clustering analysis is used in the rising areas like bioinformatics. Real life areas like speech recognition, genome data analysis and ecosystem data analysis also analysis of geographical information systems [3] [1]. Data clustering is used regularly in many applications such as data mining, vector quantization, pattern recognition, and fault detection & speaker recognition [5].

To analyse and identifying the cancerous Data: Clustering algorithm can be used in identifying the cancerous data set. In this take known samples of cancerous and non-cancerous data set [12]. Clustering Algorithm in Search Engines [13]: Clustering algorithm is used search engines. In this clustering is done on the basis of keywords or phrases of similarity also search engine is group the similar object and dissimilar object in well separated clusters based on data. Clustering Algorithm in Academics [14]: In this based on the student academic performance cluster is formed. Clustering algorithm monitor the student performance. Based on the students score similarities they are grouped into different clusters. By knowing students present per cluster we can measure the whole class performance. Application of clustering analysis in wireless sensor networks [15]: Wireless Networks used clustering algorithm efficiently to analyse the network data usage. One application where it can be used is in Landmine detection. In this clustering algorithm plays the role of finding the cluster centre which collects all the data in its respective cluster.

III. STUDY OF THE VARIOUS APPROACHES OF MODIFIED K-MEANS ALGORITHMS

All the algorithms reviewed in this paper define the same common problems of the k means algorithm like clustering large dataset, number of iteration in the algorithm, defining no of cluster, selection of initial cluster centre. So a comparison of all the algorithms can be made based on all these problems.

Approach 1:
Refinement initial cluster centre [1]

- This technique's result gives better effects and less iterative time than the existing k-means algorithm.
- This approach adds nearly no burden to the system.
- This method will decrease the iterative time of the k means algorithm, making the clustering analysis more efficient.

Approach 2:
Parallel k means implementation [1]
The computational complexity of the parallel k-means algorithm can be given as

$$T_{m \approx T_m^{comp}} \approx \frac{(3nkd).\Im.T^{flop}}{m} \text{ [1]}$$

- It can be seen from the above formula that k-mean algorithm has lessening computational complexity close to 1/m of that of the original k-means algorithm when groping large data.
- From this it enhances the cluster analysis efficiency from perspectives of both time and space.
- These improvements can greatly enhance k-means algorithm, i.e., allow the grouping of a large number of data sets more accurately and more quickly.

Approach 3:
Parallel implementation by using Map Reduce [2]
In this algorithm performance is evaluated with respect to speedup, scale up and size up.
PK-Means (Parallel K-Means) has a very good speedup performance. Specifically, as the size of the dataset increases, the speedup performs better. Therefore, PK-Means algorithm can treat large datasets efficiently.
The PK-Means algorithm handles larger datasets, in this performed scale up experiments where increased the size of the datasets in direct proportion to the number of computers in the system. The PK-Means algorithm scales very well. In this algorithm performed scale up experiments where increased the size of the datasets in direct proportion to the number of computers in the system. The datasets size of 1GB, 2GB, 3GB and 4GB are executed on 1, 2, 3 and 4 computers respectively. It clearly shows that, the PK-Means algorithm scales very well.
Size up analysis holds the number of computers in the system constant, and grows the size of the datasets by the factor m. Size up measures how much longer it takes on a given system, when the dataset size is m-times larger than the original dataset. In this algorithm PK-Means has a very good size up performance. In this fixed the number of computers to 1, 2, 3, and 4 respectively. PK-Means shows the size up results on different computers has a very good size up performance.

Approach 4:
Determine the initial centroids of the clusters and Assign each data point to the appropriate clusters [3]
In this enhanced algorithm, the data object and the value of k are the only inputs required since the initial centroids are computed automatically by using the algorithm. A systematic method for finding initial centroids and an efficient way for assigning data object to clusters. This method ensures the entire process of clustering in O (n2) time without sacrificing the accuracy of clusters. Thus the overall time complexity of the enhanced algorithm (Algorithm 2) becomes O (n2), since k is much less than n. A limitation of the proposed algorithm is that the value of k, the number of desired clusters, is still required to be given as an input, regardless of the distribution of the data points.

Approach 5:
An efficient enhanced k-means clustering algorithm [4]
In this approach from every iteration some heuristic value is kept for less calculation in next iteration from data object to the centroid. i.e. in each iteration the centroid closer to some data objects and far apart from the other data objects, the points that become closer to the centroid will stay in that cluster, so there is no need to find its distances to other cluster centroids. The points far apart from the center may change the cluster, so only for these data object their distances to other cluster centers will be calculated, and assigned to the nearest center. This is simple and efficient clustering algorithm based on the k-means algorithm. This algorithm is easy to implement, requiring a simple data structure to keep some information in each iteration to be used in the next iteration. Improve the computational speed of the k-means algorithm by the magnitude in the total number of distance calculations and the overall time of computation.

Approach 6:
Variation in K-means algorithm and proposed parallel K-means clustering algorithm [6]
In this paper algorithm is used to design data level parallelism. The algorithm is work as divide the given data objects into N number of partitions by Master Processor. After then each partition will be assigned to every processor. In next step Master processor calculates K centroids and broadcast to every processor. After that each processor calculates new centroids and broadcast to Master processor. Master processor recalculates global centroids and broadcast to every processor. Repeat these steps until unique cluster is found. In this algorithm number of clusters are fixed to be three and the initial centroids are initialized to minimum value, maximum value and the N/2th value of data point of the total data object. The proposed algorithm produces the more accurate unique clustering results.

Approach 7:

A New Initialization Method to Originate Initial Cluster Centres for K-Means Algorithm [7]

In this paper, a new algorithm is proposed for the problem of selecting initial cluster centers for the cluster in K-Means algorithm based on binary search technique. Search technique Binary search is a popular searching method that is used to find an item in given list of array. The algorithm is designed in such a way that the initial cluster centers have obtained using binary search property and after that assignment of data object in K-Means algorithm is applied to gain optimal cluster centers in dataset. This proposed method obtains high accuracy rate. The proposed method also takes less time as compare to most of algorithms for execution. It is concluded that proposed approach provides good results among all other initialization methods with simple K-Means

Approach 8:

Dynamic Clustering of Data with Modified K-Means Algorithm [8]

In the above paper a dynamic clustering method is presented with the intension of producing better quality of clusters and to generate the optimal number of cluster. In the former case it works same as K-means algorithm. In the latter case the algorithm calculates the new cluster centroids by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality. In this algorithm modified k-means algorithm will increase the quality of cluster compared to the original K-means algorithm. It assigns the data object to their suitable cluster or class more effectively. The practical implementation results of the algorithm show that the proposed method outperforms K-means algorithm in quality and optimality for the unknown data set with better results than K-means clustering. The new algorithm works efficiently for fixed number of clusters as well as unknown number of clusters. The main disadvantage of the proposed approach is that it takes more computational time than the K-means for larger data sets.

Review of limitations on the above approaches based on the following listed five limitation of the k-means algorithm is shows in table 1

1. Selection of initial centroid
2. Defining number of cluster beforehand
3. Speed up/Scale up/Size up
4. Assignment of data object appropriately
5. Computational Complexity

Other limitations of the k-means algorithm is as follow:

1. Handling empty clusters:
2. Handling inappropriate data or outliers or noisy data
3. Reducing the SSE with post processing
4. Applicable only when mean is defined i.e. fails for categorical data
5. Algorithm fails for non-linear dataset

1. Handling empty clusters [17]:

This is the limitation of the k-means algorithm it may produce empty clusters depending on initial centroids. For static execution, this problem is considered trivial and can be solved by executing the algorithm for a number of times.

2. Handling inappropriate data or outliers or noisy data:

Before applying the k-means algorithm pre-processing should be done on the dataset to remove noisy data, filling missing value, data cleaning and selection should lessen the above limitation.

3. Reducing the SSE with post processing [18]:

In the k-means algorithm to get better clustering we have to reduce the SSE. This is most difficult task. There are various types of clustering methods available which reduces the SSE. Two strategies that decrease the total SSE by increasing the number of clusters are the following: Split a cluster and introduce a new cluster centroid

4. Applicable only when mean is defined:

The main limitation of the k-means algorithm is that it should work only when the mean is defined for number of cluster.

5. Algorithm fails for non-linear dataset:

In this the data objects is distributed in non-linear manner and clustering this data objects using k-means will not give optimal clusters.

Table 1 Review of limitations on the above approaches

K-Means	Approach1	Approach2	Approach3	Approach4	Approach5	Approach6	Approach7	Approach8
Limitation1	The result for the small data set is not very notable	-	-	-	-	In this approach number of clusters are fixed to be three	-	-
Limitation2	-	The limitation of the proposed	The limitation of	The limitation	The limitation	The limitation	Again the limitation	-

		algorithm is value of k.	the proposed algorithm is value of k.	of the proposed algorithm is value of k.	of the proposed algorithm is value of k.	of the proposed algorithm is value of k.	of the proposed algorithm is value of k	
Limitation3	-	-	Specifically, as the size of the dataset increases, the speedup performs better.	-	-	-	-	-
Limitation4	-	-	-	-	Data structure is required to hold some information in each iteration.	-	-	-
Limitation5	-	Computational complexity is combined computational complexity of computing and communication.	-	-	-	-	-	It takes more computational time than the K-means for larger data sets.

The table 1 shows that every approach tried to solve the limitation or disadvantage of the existing k-means algorithm. But it solved the problem with some constraints on the algorithm. The approaches concentrated only on the specific limitation. The approaches fails to improve the overall efficiency and effectiveness of the existing k-means algorithm problems. Such as clustering large dataset with optimal number of clusters without specific predefined number of clusters, effective initial centroid selection, execute the algorithm in less number of iteration and minimum time. It also takes into consideration for scale up, speed up and size up for the dataset, resources used and faster execution of the algorithm. The dash line indicates no limitation for this approach.

IV. CONCLUSIONS

In day-to-day life the large volume of data is generated in many applications. Clustering is the powerful and popular approaches in knowledge discovery and data mining. Cluster analysis is plays very important role in data mining and tools for big data analysis. To analyse the large volume of data effectively is very crucial for making decision on the field. In this paper discusses various limitations of existing k-means algorithm. And those limitations are selection of initial centroid, assigning data objects to the clusters, defining number of cluster beforehand and computational complexity of the algorithm, number of iteration. Also discusses and reviewed various approaches of modified k-means algorithm. This paper is also discusses the various algorithms which are used for clustering of big data using modified k-means algorithm. It also discusses the solution for finding best initial cluster centre, reduces number of iteration in existing k means algorithm, big data clustering, assigning data to the appropriate cluster, defining number of cluster efficiently. Parallel implementation by using Map Reduce techniques and k-means algorithm gives the better result in terms of speed up, size up, and scale up for big data as per the review. It also discusses the limitations of all the approaches in the paper.

ACKNOWLEDGMENT

This research was supported by my mentor Shubha Puthran and faculty guide Dr. Dhirendra Mishra. I am grateful to them for sharing their pearls of wisdom with me during the course of this research.

REFERENCES

[1] Tian, Jinlan, Lin Zhu, Suqin Zhang, and Lu Liu. "Improvement and parallelism of k-means clustering algorithm." Tsinghua Science & Technology 10, no. 3 (2005): 277-281.

[2] Zhao, Weizhong, Huifang Ma, and Qing He. "Parallel k-means clustering based on mapreduce." In Cloud Computing, pp. 674-679. Springer Berlin Heidelberg, 2009.

[3] Nazeer, KA Abdul, and M. P. Sebastian. "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm." In Proceedings of the World Congress on Engineering, vol. 1, pp. 1-3. 2009.

[4] Fahim, A. M., A. M. Salem, F. A. Torkey, and M. A. Ramadan. "An efficient enhanced k-means clustering algorithm." Journal of Zhejiang University SCIENCE A 7, no. 10 (2006): 1626-1633.

[5] Rasmussen, Edie M., and PETER WILLETT. "Efficiency of hierarchic agglomerative clustering using the ICL distributed array processor." Journal of Documentation 45, no. 1 (1989): 1-24.

[6] Dr.Urmila R. Pol, "Enhancing K-means Clustering Algorithm and Proposed Parallel K-means clustering for Large Data Sets." International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[7] Yugal Kumar, Yugal Kumar, and G. Sahoo G. Sahoo. "A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm." International Journal of Advanced Science and Technology 62 (2014): 43-54.

[8] Shafeeq, Ahamed, and K. S. Hareesha. "Dynamic clustering of data with modified k-means algorithm." In Proceedings of the 2012 conference on information and computer networks, pp. 221-225. 2012.

[9] Ben-Dor, Amir, Ron Shamir, and Zohar Yakhini. "Clustering gene expression patterns." Journal of computational biology 6, no. 3-4 (1999): 281-297.

[10] Steinley, Douglas. "Local optima in K-means clustering: what you don't know may hurt you." Psychological methods 8, no. 3 (2003): 294.

[11] Aloise, Daniel, Amit Deshpande, Pierre Hansen, and Preyas Popat. "NP-hardness of Euclidean sum-of-squares clustering." Machine Learning 75, no. 2 (2009): 245-248.

[12] Wang, Haizhou, and Mingzhou Song. "Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming." The R Journal 3, no. 2 (2011): 29-33.

[13] Al-Daoud, Moth'D. Belal. "A new algorithm for cluster initialization." In WEC'05: The Second World Enformatika Conference. 2005.

[14] Wang, X. Y., and Jon M. Garibaldi. "A comparison of fuzzy and non-fuzzy clustering techniques in cancer diagnosis." In Proceedings of the 2nd International Conference in Computational Intelligence in Medicine and Healthcare, BIOPATTERN Conference, Costa da Caparica, Lisbon, Portugal, p. 28. 2005.

[15] Liu, Ting, Charles Rosenberg, and Henry A. Rowley. "Clustering billions of images with large scale nearest neighbor search." In Applications of Computer Vision, 2007. WACV'07. IEEE Workshop on, pp. 28-28. IEEE, 2007.

[16] Oyelade, O. J., O. O. Oladipupo, and I. C. Obagbuwa. "Application of k Means Clustering algorithm for prediction of Students Academic Performance." arXiv preprint arXiv: 1002.2425 (2010).

[17] Akkaya, Kemal, Fatih Senel, and Brian McLaughlan. "Clustering of wireless sensor and actor networks based on sensor distribution and connectivity." Journal of Parallel and Distributed Computing 69, no. 6 (2009): 573-587.

[18] https://sites.google.com/site/dataclusteringalgorithms/clustering-algorithm-applications

[19] Pakhira, Malay K. "A modified k-means algorithm to avoid empty clusters." International Journal of Recent Trends in Engineering 1, no. 1 (2009).

[20] Singh, Kehar, Dimple Malik, and Naveen Sharma. "Evolving limitations in K-means algorithm in data mining and their removal." International Journal of Computational Engineering & Management 12 (2011): 105-109.

