Imperial Journal of Interdisciplinary Research (IJIR)
Vol.2, Issue-1 , 2016
ISSN : 2454-1362 , www.onlinejournal.in

Efficient clustering of data using improved K-means algorithm: A Review

Miss. Vrinda khairnar
Student Computer science and engineering, G.H. Raisoni College Of Engineering And Management, Jalgaon, India

Miss. Sonal patil
Assistant Prof. and head of information technology, G.H. Raisoni College Of Engineering And Management, Jalgaon, India

Abstract

The k-means clustering algorithm is one of the most widely used algorithm for clustering analysis of big data. The traditional K-means algorithm is inefficient while working on large numbers of data sets. Thus improving the algorithm efficiency remains a problem. The traditional k-means algorithm is computationally expensive in terms of time complexity. The quality of the resulting clusters mostly depends on the selection of initial centroids and the existing k-means algorithm, does not guarantee optimality. Several methods have been proposed in the literature for improving the performance of the k-means clustering algorithm. This paper discusses about the different techniques and improvements of K-means clustering algorithm based on different research papers referred. These methods includes Refined initial cluster center's 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 matching 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 Centre's for K-Means Algorithm, Dynamic Clustering of Data with Modified K-Means Algorithm. This review paper discusses about the limitations of these clustering technique and also compare each technique with other techniques.

1. Introduction

Data mining is a tool extracting the information from large datasets because it is very difficult to get important information and provide it within time limit. In data mining clustering that is cluster analysis of data performing principal task. Clustering is a task to group data on basis of their similarities and dissimilarities from data elements, mainly it is difficult at the time of big dataset.

Clustering method convert that information into various clusters where object in that cluster having similar proper-ties as compare to other but not same to other clusters properties.

There are various efficient techniques used to solve the problem for large data clustering. Clustering techniques and implementation used for getting scalability and performance in such data analysis. By using cluster analysis techniques it is very easy to handle complex data sets and K-means is widely used for producing clusters in many application. It is also used for automatically organized data, compression form and finding some hidden structure. [4][5][10]

In the figure 1 shows basic cluster formation from given dataset when we applying K-means algorithm. In first stage by selecting random initial centroid and clusters the data object in the dataset. Now in stage second recalculating centroid from first iteration due to this as figure shows some of the data object move from one cluster to another. In third stage of the figure centroid remain constant which means convergence is found. In this way all data object is clustered as respective cluster hence selection of initial centroid is main task in cluster formation.

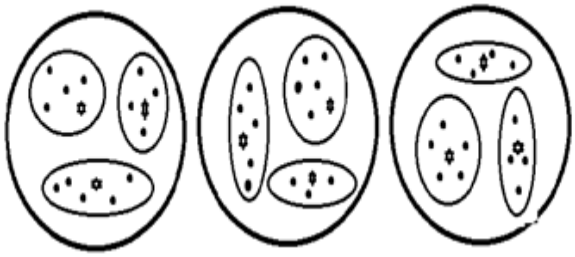

Fig 1: Block Diagram of cluster formation

In rest of the paper is organized as Section II discuss related work of existing K-means algorithms. The section III Discuss study of the various techniques of modified K- means algorithm.

2. Related work

First used of K-means clustering algorithm by James MacQueen in 1967[1]. In 1957 Stuart Lloyed first proposed the basic original algorithm as a technique for signal processing, through it wasn't publish until 1982. K-means algorithm mostly used partitioned clustering algorithm in many application. In this method partition depends the k partition data where k represent the cluster and $k \leq n$(data object). The following condition will be fulfil at the time of clusters the data into k that is number of cluster groups:

I) each group contain at least on object and
II) Each object belongs to exactly one group.[11]

K-means algorithm: [1]
Input: D= {d1, d2, d3…….,dn} // D contains data objects.
Number of cluster=k //User defined
Output:
A set of k clusters.
Steps:
Phase I: Randomly select initial cluster centroid from data items.
Phase II: Assign each data point to the cluster which has the closest (most similar) centroid.
Phase III: Recalculate the new mean for each cluster
Phase IV: Repeat step II and III until convergence is met.

The k-means algorithm mostly used for data mining purpose. The input for this is 'n' data items. And user gives number of clusters that is 'k'. Basically k-means algorithm is divided into two phase. In first randomly select initial centroid with better accuracy and assign each data point to which it is the most similar cluster. And in phase second. The new mean for each cluster is then calculated. This process iterates until the criterion function converges. [1]

For producing better accurate result many application used k-means algorithm which having efficient, easy and simplicity of implementation, adaptability to sparse data and scalability [5]. K-means can be easily used for clustering of data. The time complexity for k-means algorithm is O(nkt), where n is the number of object is number of iteration and k is number of cluster hence it is very fast algorithm. The traditional k-means is expensive and its result is mostly depends on selection of initial centroid. Clusters form in k-means with the help of initial centroid and these initial centroid calculated randomly so clusters are vary from one an-other in different iteration. Also data element vary from one cluster to another [5].

The drawback of k-means algorithms is given below:

- How to choose the initial location of centroid that is lack of universal methods.
- K-means algorithm has input k that is number of cluster, it is very difficult to fix the number of cluster in advance.
- Result of k-means depends upon the selection of initial centroid because different centroid can cause different cluster that affect result accuracy so how to select initial location for centroid [10].
- The algorithm stuck in local minima value [11].
- Lack of knowledge how to treat with inappropriate and clatter attribute.

Clustering analysis used in various application like analysis of geographical information system, Real life area like speech recognition, rising area like bioinformatics, genome data analysis and ecosystem data analysis. Similarly data mining, pattern recognition, vector quantization and fault detection, speak recognition [3] this application also regularly used data clustering.

There are many application of k-means clustering algorithm for data analysis some of these are given as follow: [2] Wireless sensor network: To analysis the network data usage efficiently so wireless network used cluster algorithm. Landmine detection also used clustering algorithms [9]. Cancerous data analysis and identifying [7]: For identifying cancerous data set used cluster algorithm as this take known cancerous and non-cancerous data set. Academics used cluster algorithm: For checking student performance used cluster analysis based on similarities they are grouped into different clusters hence we say clustering monitor the student performance. In search Engine: Here clustering is done by keywords or phrases of similarity also is decided well separated clusters based on data, in this way cluster algorithms used in search engine. In this way clustering algorithms efficiently used analyses and finding cluster center where it's collect data in its respective clusters.

3. Study of various techniques of modified k-means algorithms

K-means clustering algorithms have some common problems which are reviewed in this paper like number of iteration in the algorithm, selection of initial cluster Centre, clustering large data set, and defining number of cluster. By discuss this problem we compare this all clustering algorithm can be made based on this problem.

Technique I: Implementation parallel of K-means [2]

Parallel k-means algorithm computational complexity is given as:

$$T_{m \approx T_m^{comp}} \approx \frac{(3nkd).\vartheta.T^{flop}}{m}$$

- K-means algorithm has Lessing complexity when grouping large data is near to 1/m of the traditional k-means algorithm it can be seen by above formula.
- From this with perspective both time and space we enhances cluster analysis efficiency.
- K-means algorithm greatly enhance due to these improvements that is grouping of large data is more quickly and accurately.

Technique II: Refinement of initial cluster centroid [2]

- The iterative time of k-means algorithm is decreases due to this method for making the clustering analysis more efficient.
- This approach gives better effects result and less iterative time than the existing k-means algorithm
- This techniques adds nearly no burden to the sys-tem.

Technique III: Variation between k-means algorithm and proposed k-means clustering algorithm [12]

In this algorithm given data objects divide into N numbers of partitions by master processor, then each partition assign to everyone processor. In next stage master processor calculate initial centroid and pass it to all processor. Again centroid recalculation done by master processor and broadcast it to other processor. This step is done continue until unique cluster is found. But in this algorithm number of cluster is fixed to three and initial centroid declared to minimum value and N/2th value of data point of the total data object. For design data level parallelism algorithm used from this pa-per and proposed algorithm gives the more accurate unique clustering results.

Techniques IV: Parallel implementation using Map Reduce [4]

In this paper performance is calculated by considering speedup, scale up and size up. Whenever size of dataset in-creases parallel k=means gives better performance in terms of speed hence parallel k-means algorithm treat large data sets efficiently. In this paper algorithm performed scale up experiment where increase the size of the datasets in directly proportional to the number of computers in the system hence we say parallel k-means handles larger data sets having size is 1GB,2GB,3GB and 4GB executed sequentially like 1, 2 , 3, 4 computers respectively. Due to this it's clearly shown that parallel k-means algorithm scales very well.

Number of computer in the given system is constant which holds size up analysis, size of datasets grows by the factor m. Similarly size up measure how much longer it takes on system, where the data set size is m-times larger than the traditional dataset. So parallel k-means very good size up results performance on different and same computer also.

Technique V: Modified K-means algorithm with dynamic clustering of data [11]

For better quality of clusters and also generate the optimal number of cluster this paper provide intestinally such techniques dynamic clustering method. This algorithm calculate the new cluster centroid by increasing the cluster counter by one in each iteration until it satisfying the validity of cluster quality, otherwise its work same as a k-means algorithm. Modified k-means definitely increase the quality of cluster as compared with original k-means. Also assign the data object according to their cluster or class is efficiently. Same as its optimality and performance good for unknown data set as compared to k-means. The modified k-means work efficiently where we know number of cluster or not. Main disadvantage for this algorithm is, it's take more computational time for large data set as compare to k-means algorithm.

Technique VI: An efficient improved k-means clustering algorithm [5]

In this techniques every iteration containing some heuristic value for decreases the calculation of centroid in next iteration because data object kept in that centroid where its distance less and far away from the other centroid. Due to heuristics value data object is closer to centroid in each iteration so no need to find its distance from other cluster center and its assign to nearest center. This algorithm is easy to implement and required less computational time but need a simple data structure to maintain the information required for next iteration.

Technique VII: Originate initial cluster Center's for k-means algorithm: A new initialization method [1]

The main problem in k-means is selection of initial cluster centroid so here implement new algorithm which is based on binary search techniques. Whether we find out the item from list of array we used mostly binary search. Using same technique algorithm design in such way initial centroid find by binary search property. This proposed algorithm required less time as compare to other algorithm for execution. This conclude that the algorithm provide good result as other initialization methods with simple k-means.

Technique VIII: Calculate the initial centroid of the cluster and Assign each data point to the appropriate cluster [1]

Input required for this is data object and number of cluster that is k and initial centroid calculation is done automatically using this algorithm. Hence finding initial Centre systematically and assign data object to cluster.

This method over-all need time complexity of the improved algorithm is , Since k must less than n. Main drawback for this method is that value of k, that is number of cluster is still required as an input.

Review of the limitation on the above approaches based on the following five limitation of the k-means shown in table

Basic limitation of k-means is:

- Computational complexity
- Speed up/Size up/Scale up
- Selection of initial centroid
- Assignment of data object
- Declared number of cluster beforehand.

Other limitation of K-means is below:

I) Empty cluster handle:

Depending upon initial centroid it produce the empty cluster but for static execution this problem take a trivial and can be solve by executing algorithm with number of times.

II) Algorithm fails for irregular data set:

Distributed non-linearly mannered data object and using k-means clustering of this data object will not give optimal clusters.

III) Calculating mean is necessary in clustering algorithms:

K-means applicable only when the mean is defined for number of clusters.

IV) Handling of noisy data:

Pre-processing is required to remove noisy data before applying k-means algori8thm.

V) Reducing the SSE with post processing: [13]

To better clustering result for k-means we have to reduce the SSE. There are various method used for reducing SSE but it is difficult task.

From the Table 1 dash line indicate no limitation for that technique. And other gives limitation or disadvantage for the techniques for clustering. This technique concentrate on some specific limitation but this technique fail to improve the efficient and effective result of the k-means algorithms. It also takes into consideration for scale up, speed up and size up for the dataset, resources used and faster execution of the algorithm.

Table 1: Review of limitation on the above techniques

K-Means	Technique I	Technique II	Technique III	Technique IV	Technique V	Technique VI	Technique VII	Technique VIII
Limitation I	-	The result for small data set is not notable.	Number of clusters is fixed that is three.	-	-	-	-	-
Limitation II	Computational complexity combine by computational complexity computing and communication	-	-	-	For larger data set it take more computational time as compared to k-means	-	-	-
Limitation III	-	-	-	-	-	For each iteration to hold some information we required data structure.	-	-
Limitation IV	Value of k is limitation for this.	-	The value of k is limitation of proposed algorithm	Value of K is limitation for this.	-	Value of k is limitation for this proposed algorithm	Value of k is limitation for this proposed algorithm	
Limitation V	-	-	-	Speedup performs increases as the size of the dataset increases	-	-	-	-

4. Conclusion

In today's era, very large volume of digital data is generated by many applications. Clustering is widely used technique for knowledge discovery and data mining tool. Clustering plays very crucial role in data mining and used by tools for big data analysis. The analysis of these large volume of data is very important steps for making decisions on the field. This paper we discusses various limitation and disadvantages of various clustering algorithms techniques. And those limitation are Selection of initial centroid, Assignment of data object, declared number of cluster beforehand, number of iterations and computational complexity. Also discuss and reviewed various techniques of modified k-means clustering algorithm. This paper is also discusses the various clustering algorithms techniques which are used for clustering of large data using modified and improved k-means algorithm. It also discusses the solution for finding best initial cluster, efficient improved k-means clustering algorithm, Calculate the initial centroid of the cluster and Assign each data point to the appropriate cluster, Modified K-means algorithm with dynamic clustering of data, Implementation parallel of K-means And k-means clustering algorithm gives best result in terms of speed up, scale up, size up for a large data from this review paper. Similarly we discusses limitation for this technique in this review paper.

5. Acknowledgement

This research was supported by my guide Miss. Sonal patil and faculty HOD Prof.P.P.Rewagad. I am thankful to them for sharing their pearls of wisdom and knowledge with me during the course of this research

6. References

[1] 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.

[2] 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..

[3] 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. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[4] Zhao, Weizhong, Huifang Ma, and Qing He. "Parallel k-means clustering based on mapreduce." In Cloud Computing, pp. 674-679. Springer Berlin Heidelberg, 2009. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/ contrib. /supported/ IEEEtran/

[5] 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.

[6] 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): 29-33A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[7] Ben-Dor, Amir, Ron Shamir, and Zohar Yakhini. "Clustering gene expression patterns." Journal of computational biology 6, no. 3-4 (1999): 281-297. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.

[8] 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).

[9] sLu, 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.