A Novel Approach for Data Clustering using Improved K-means Algorithm

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

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

ABSTRACT
In statistic and data mining, k-means is well known for its efficiency in clustering large data sets. The aim is to group data points into clusters such that similar items are lumped together in the same cluster. The K-means clustering algorithm is most commonly used algorithms for clustering analysis. The existing K-means algorithm is, inefficient while working on large data and improving the algorithm remains a problem. However, there exist some flaws in classical K-means clustering algorithm. According to the method, the algorithm is sensitive to selecting initial Centroid. The quality of the resulting clusters heavily depends on the selection of initial centroids. K-means clustering is a method of cluster analysis which aims to partition 'n' observations into k clusters in which each observation belongs to the cluster with the nearest mean. In the proposed project performing data clustering efficiently by decreasing the time of generating cluster. In this project, our aim is to improve the performance using normalization and initial centroid selection techniques in already existing algorithm. The experimental result shows that, the proposed algorithm can overcome shortcomings of the K-means algorithm.

Keywords
Data Analysis, Clustering, k-means Algorithm, Improved k-means Algorithm

1. 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. [26] 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] [26]. K-means algorithm results in different types of clusters depending on the random choice of initial centroids. Several attempts were made by researchers for improving the k-means clustering algorithm performance. This paper deals with a method for improving the accuracy of the k-means algorithm.

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.

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].

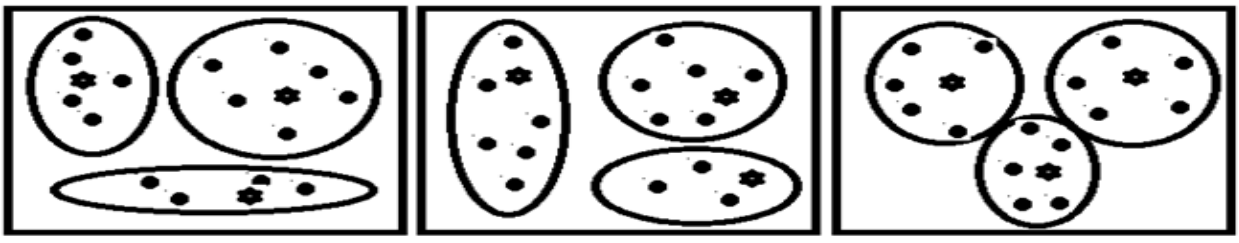

Figure. 1. Block Diagram of Cluster Formation

2. 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 ≤ 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 [1]:

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 is the most extensively studied clustering algorithm and is generally effective in producing good results. The k-means algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations.

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

Several attempts were made by researchers to improve the effectiveness and efficiency of the k-means algorithm [1, 3, 4, 6, 7, 8]. 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 center. So a comparison of all the algorithms can be made based on all these problems.

Tian et al. [1] proposed a systematic method for finding the initial centroids. 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. The result for the small data set is not very notable when the refinement algorithm operates over a small subset of a quite large data set.

Abdul Nazeer et al. [3] proposed a systematic method for finding the initial centroids. 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. 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.

FAHIM A et al. [4] proposed a systematic method for finding the initial centroids. 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.

Dr. Urmila R. et al. [6] proposed a systematic method for clustering large dataset. 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.

Yugal Kumar et al. [7] proposed a systematic method for finding the initial centroids. 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.

Shafeeq et al. [8] proposed a systematic method for dynamic clustering of data. 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 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.

4. DATASETS AND PROPOSED ALGORITHM

In the improved clustering method discussed in this paper, the original k-means algorithm are modified to improve the accuracy and reduce execution time. The improved method is outlined in the following steps.

The improved method

Steps

1. Input: In this step take input from the user the dataset and pass it to the algorithm.
2. Apply the normalization technique to the given dataset
3. Apply the sorting technique to the given dataset
4. Apply the algorithm to find initial centroid from the dataset
5. Assign data object to the centroids (repeat until convergence occur)
6. recalculate centroid
7. Check for the convergence

In the modified algorithm first initial cluster size is calculated by using total attributes divide by number of cluster. Then normalization technique is used to normalize the dataset to scale up the values is the range. In the next step the sorting technique is used on the dataset because processing the sorted array is faster than unsorted array. Then calculate the initial centroid by mean of the cluster. In the next step assign the data object to the initial centroid by calculating ecludien distance. Check for the convergence criteria. Repeat the steps until no more changes in the last centroids and updated centroid. Because of the initial centroid is generated by calculation the number of iterations is fixed, the initial centroids are determined systematically so as to produce clusters with better accuracy.

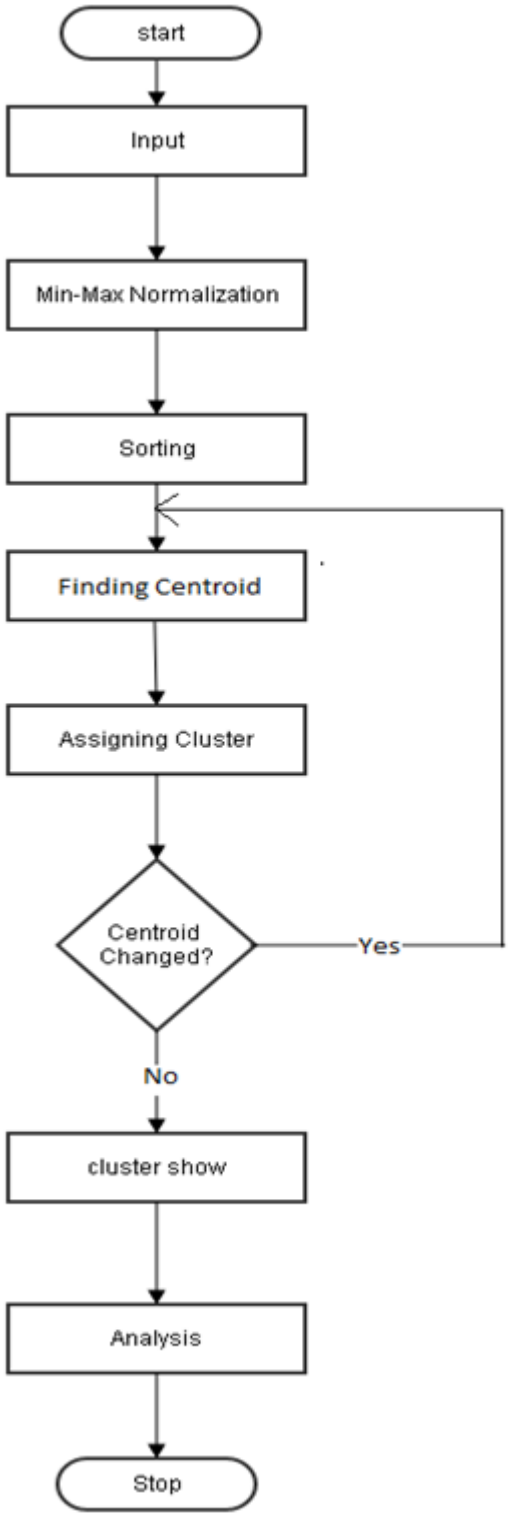

Figure 2: Block Diagram for proposed algorithm

The data sets used to tests the algorithms from the Machine Learning repository. Selection of the datasets further depended on their size, larger data sets generally means higher confidence. In this choose different kinds of data sets, because in this want to test if the performance of an algorithm depended on the kind of set that is used.

Table 1: Datasets properties

Datasets	Instances	Attributes
Transfusion	500	5
Wavesurge	1000	3
Iqitems	1500	17
Ds1.10	2000	5
bfi	2500	29

5. EXPERIMENTAL RESULTS

In this project, decided to work on bfi and different datasets which is used in several papers in order to do experiments and give results also used for testing the accuracy and efficiency of the improved algorithm. The same data set is given as input to the standard k-means algorithm and the improved algorithm. The value of k, the number of clusters, is taken as 4. The results of the experiments are tabulated in Table II. The standard k-means algorithm select the values of the initial centroids randomly as input, apart from the input data values and the value of k. The experiment is conducted for standard k-means and proposed algorithm. The accuracy of clustering is determined by comparing the clusters obtained by the experiments with the clusters finding in the proposed algorithm. For the improved algorithm, the data values and the value of k are the only inputs required since the initial centroids are computed automatically by the program. The accuracy, number of iteration, and execution time taken in the case of this algorithm are also computed and tabulated.

- Comparison of K-means and Proposed Algorithm K=4 (Iterations, Accuracy, Execution Time) for iqitems dataset

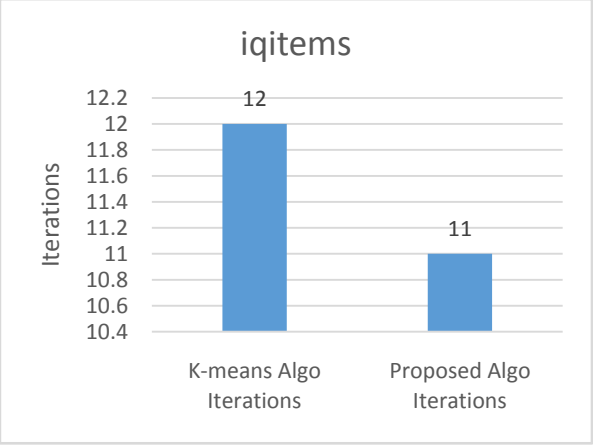

Figure 3: Graph for Iterations

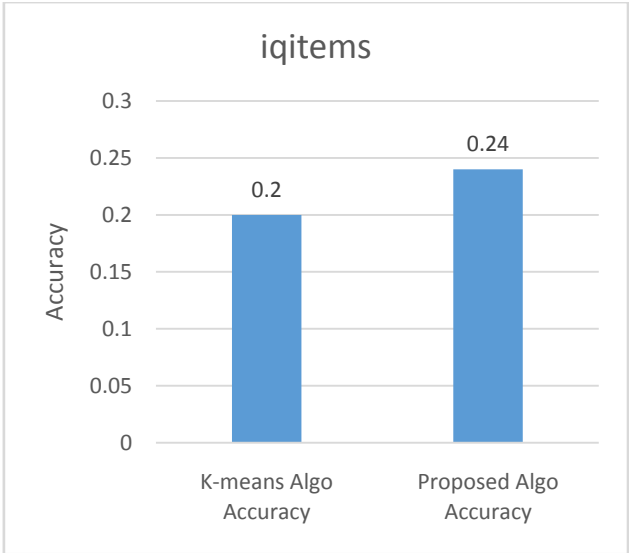

Figure 4: Graph for Accuracy

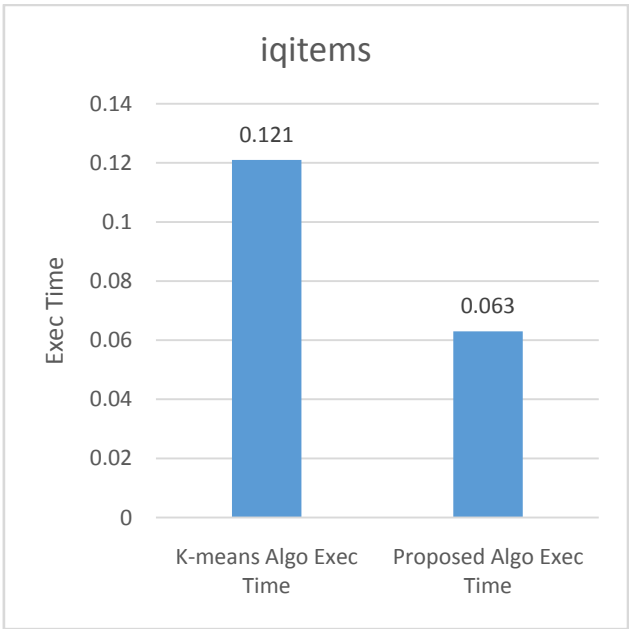

Figure 5: Graph for Execution time

Table 2. Performance Comparison

Dataset	Algorithm	Iterations	Accuracy	Execution time
Transfusion	K-means	9	0.65	0.089
	Proposed Algorithm	7	0.81	0.039
wavesurge	K-means	15	0.022	0.035
	Proposed Algorithm	13	0.027	0.03
iqitems	K-means	26	0.20	0.108
	Proposed Algorithm	13	0.45	0.076

ds1.10	K-means	20	0.0034	0.101
	Proposed Algorithm	9	0.0051	0.064
Bfi	K-means	30	0.084	0.38
	Proposed Algorithm	18	0.16	0.31

Table II depicts the performances of the standard k-means algorithm and the improved algorithm in terms of the accuracy, number of iteration and efficiency. It can be seen from the above experiments that the improved algorithm significantly outperforms the original k-means algorithm in terms of accuracy, number of iterations and efficiency.

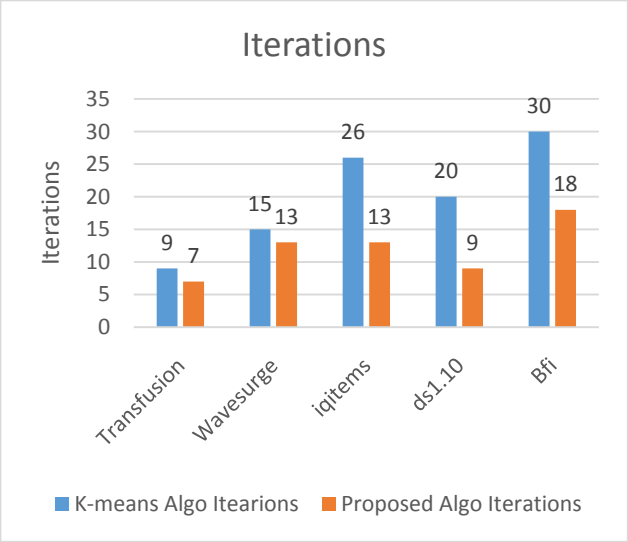

Figure 6: comparison of Iterations on different datasets

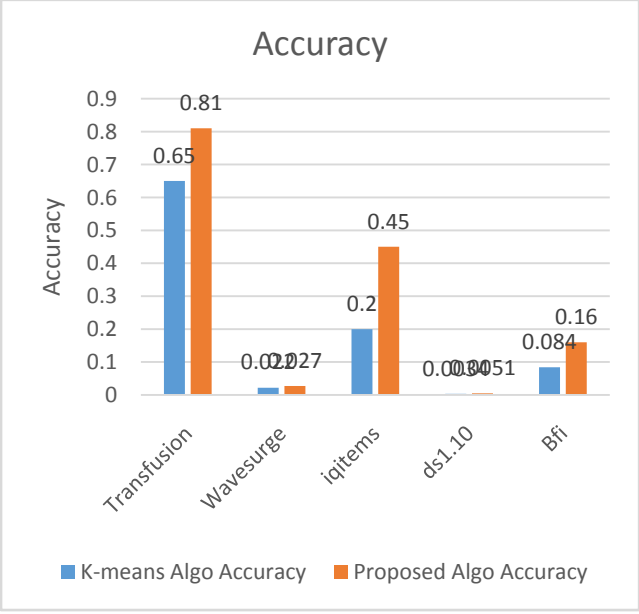

Figure 7: comparison of Accuracy on different datasets

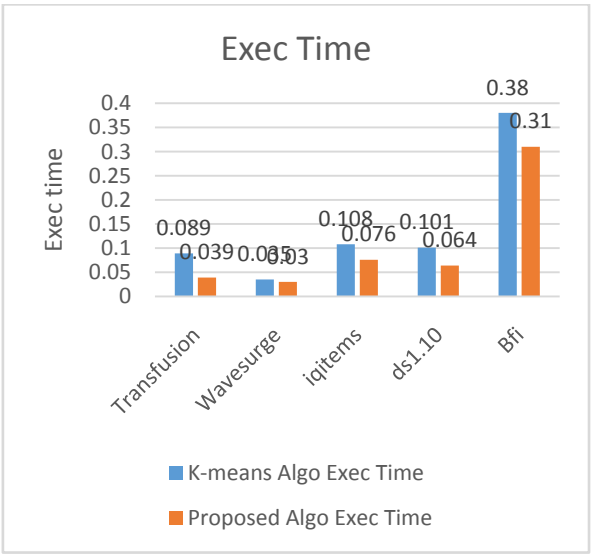

Figure 8: comparison of execution time on different datasets

Hike in accuracies for all datasets for k=5 is 0.13, .013, 0.22, .0018, .055 respectively. Average of all this comes to 8.396%. Hence the accuracy of proposed algorithm is 8% better than that of K-means algorithm.

6. CONCLUSION

This study proposes a new modified implementation of the k-means clustering algorithm. This modified k-means algorithm clusters the large datasets effectively with less execution time. This study provides a new method of clustering datasets with normalization technique and sorting technique used for faster accessing the data object in the datasets and because of this the overall execution time of the algorithm is reduced. Also the proposed algorithm is effectively work on the different datasets. From this experiments on different datasets it is also concluded that the modified approaches have better accuracy for datasets. The Average accuracy of proposed algorithm is 8% greater than k-means algorithm. . The Average Execution time of proposed algorithm is 0.0602 sec lesser than k-means algorithm. In this used purity to compare existing k-means and proposed k-means clustering algorithm is producing accurate clusters.

7. 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.

8. REFERENCES

[1]. Farajian, Mohammad Ali, and Shahriar Mohammadi. "Mining the banking customer behavior using clustering and association rules methods."International Journal of Industrial Engineering 21, no. 4 (2010).

[2]. Bhatia, M. P. S., and Deepika Khurana. "Experimental study of Data clustering using k-Means and modified algorithms." International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol 3 (2013).

[3]. Jain, Sapna, M. Afshar Aalam, and M. N. Doja. "K-means clustering using weka interface." In Proceedings of the 4th National Conference. 2010.

[4]. Kumar, M. Varun, M. Vishnu Chaitanya, and M. Madhavan. "Segmenting the Banking Market Strategy by Clustering." International Journal of Computer Applications 45 (2012).

[5]. Namvar, Morteza, Mohammad R. Gholamian, and Sahand KhakAbi. "A two phase clustering method for intelligent customer Segmentation." In Intelligent Systems, Modelling and Simulation (ISMS), 2010 International Conference on, pp. 215-219. IEEE, 2010.

[6]. 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.

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

[8]. 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.

[9]. 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.

[10]. 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.

[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]. 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.

[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.

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

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

[16]. 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.

[17]. 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.

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

[19]. 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.

[20]. 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.

[21]. 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).

[22]. 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.

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

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

[25]. 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.

[26]. Rishikesh Suryawanshi, Shubha Puthran,"Review of Various Enhancement for Clustering Algorithms in Big Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering(2016)

[27]. http://nlp.stanford.edu/IR-book/html/htmledition/k-means-1.html#sec:kmeans

[28]. https://archive.ics.uci.edu/ml/datasets.html

[29]. http://stats.stackexchange.com/questions/70801/how-to-normalize-data-to-0-1-range

[30]. http://stackoverflow.com/questions/11227809/why-is-processing-a-sorted-array-faster-than-an-unsorted-array