Improved Dynamic Parallel K-Means Algorithm using Dunn's Index Method

Prashant Yadav
M.Tech Scholar, KITE
prashantyadav23jan@gmail.com

Abstract:- K-Means is popular and widely used clustering technique in present scenario. Many research has been done in same area for the improvement of K-Means clustering algorithm, but further investigation is always required to reveal the answers of the important questions such as 'is it possible to find optimal number of clusters dynamically while ignoring the empty clusters' or 'does the parallel execution of any clustering algorithm really improves it performance in terms of speedup'. This research presents an improved K-Means algorithm which is capable to calculate the number of clusters dynamically using Dunn's index approach and further executes the algorithm in parallel using the capabilities of Microsoft's Task Parallel Libraries. The original K-Means and Improved parallel modified K-Means algorithm performed for the two dimensional raw data consisting different numbers of records. From the results it is clear that the Improved K-Means is better in all the scenarios either increase the numbers of clusters or change the number of records in raw data. For the same number of input clusters and different data sets in original K-Means and Improved K-Means, the performance of Modified parallel K-Means is 20 to 50 percent better than the original K-Means in terms of Execution time and Speedup.

*****

1. Introduction

Clustering is an approach that classifies the raw data logically and searches the hidden patterns that may be present in datasets [1]. It is procedure of collection data items into disjointed clusters so that the data's in the same cluster are similar. The demand for organizing the sharp increasing data's and taking valuable data from information which makes clustering procedure are broadly connected in numerous application, region for example pattern recognition artificial intelligence, marketing biology, data compression, data mining, customer relationship management, retrieval of information, image processing, psychology, medicine, machine learning statistics and so on [2].A definition of clustering could be "the procedure of organizing items into sets whose associates are similar in somehow". A cluster is a grouping of data objects which are "similar" among the same cluster objects and are "dissimilar" to the objects which belongs to other clusters [4].

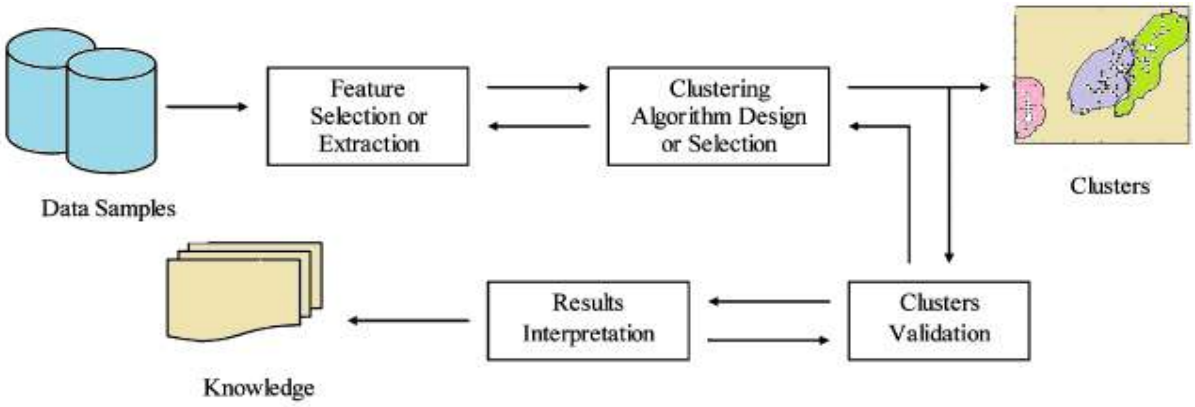

Figure 1.1: Clustering Procedure Steps

2. Literature Survey

The classification of Clustering methods can be done under following .

- Partitioning Method
- Hierarchical Method
- Density-Based Method
- Grid-Based Method
- Model-Based Method
- Constraint-Based Method

Suppose a database Contain 'n' objects and the partitioning method constructs 'k' partition of data. A cluster is represented by each partition and $k \leq n$. Which simply means that it will categorize the data objects into k groups that satisfy the requirements given as follows

- At least one date object is there in each group.

- It is must for each data object to belong to exactly one group.
- In case when number of partitions (suppose k) are given, initial partitioning will be created by the partitioning method.
- Finally to improve the partitioning quality by moving the date objects between different groups, the iterative relocation technique is used.

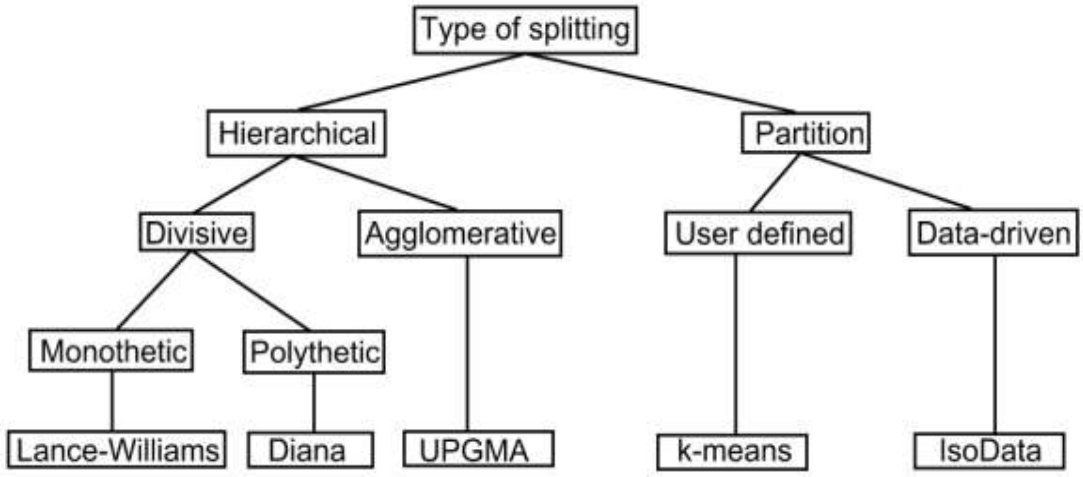

Figure 2.1: Types of Clustering Method

In this section the related works that propose the problems of K-Means clustering algorithm are discussed. A number of papers have been published regarding the improvement of quality of K-Means clustering, here are few of them that are found most appropriate with this research work.

JeyhunKarimov and Murat Ozbayoglu: in year 2015 in their research titled "Clustering Quality Improvement of K-Means using a Hybrid Evolutionary Model" presented an approach for choosing good candidates for the initial centroid selection process for compact clustering algorithms, such as K-Means, is essential for clustering quality and performance.

In their research researchers proposed a novel hybrid evolutionary model for K-Means clustering (HE-kmeans). Their model uses meta-heuristic methods to identify the "good candidates" for initial centroid selection in K-Means clustering method. The results indicate that the clustering quality is improved by approximately 30% compared to the standard random selection of initial centroids.

Improvement of the clustering quality was done but did not solve the problem of fixed numbers of static cluster as input. Any approach was also not targeted with parallel computation of clusters.

GrigoriosTzortzis and AristidisLikas: in year 2014 under their research titled "The MinMax k-Means clustering algorithm" proposed the Min Max k-Means algorithm, a method that assigns weights to the clusters relative to their variance and optimizes a weighted version of the k-Means objective.

In their approach Weights are learned together with the cluster assignments, through an iterative procedure. The proposed weighting scheme limits the emergence of large variance clusters and allows high quality solutions to be systematically uncovered, irrespective of the initialization.

Researchers performed some Experiments to verify the effectiveness of their approach and its robustness over bad initializations, and compared it with both k-Means and other methods from the literature that consider the k-Means initialization problem.

Researchers targeted the K-Means initialization problem but did not considered the parallel execution and computation of clusters.

Ahamed Shafeeq B M and Hareesha K S: in 2012 under their research titled "Dynamic Clustering of Data with Modified K-Means Algorithm" proposed an approach for modified Kmeans algorithm with the intension of improving cluster quality and to fix the optimal number of cluster. With their approach user has the flexibility either to fix the number of clusters or input the minimum number of clusters required.

Finally it was showed that how the modified k-mean algorithm will increase the quality of clusters compared to the K-Means algorithm.

3. Proposed Technique

 ISSN: 2321-8169
342 – 348

The standard K-Means algorithm need to calculate the distance between every data object and the centers of k clusters when it executes the iteration every time; it takes up more execution time mainly for large datasets as it executes in serial manner.

Proposed approach overcomes this problem as its uses Dunn's index approach for the calculation of finding the total number of clusters from given data objects.
This approach also speed ups the execution time as it performs the execution in parallel manner with the help of Microsoft's task parallel libraries.

4. Result Analysis

All the programs are written in Microsoft Visual studio.net(C#). Different machine architectures can produce differ results for the total runtime in case of same algorithms. Here the runtime means the period between input given and output is ready to collect or the execution time, rather than the CPU time calculated in the experiments under some literature. To illustrate the numerical behavior of the modified k-mean algorithm and to compare it with the standard k-mean algorithm of randomly choosing initial starting points, first solve a problem in detail by standard and then modified k-mean algorithm with the same data set. In this research, the most representative algorithms K-Means and proposed algorithm, modified K-Means were examined and analyzed based on their basic approach for different raw data consisting multiple numbers of rows. The best algorithm in each category will found out based on their performance. Comparison between K-Means and modified K-Mean algorithm with numbers of records and execution time (in milliseconds) is shown in the following tables and figures.

4.1 Time for Serial v/s Parallel K-Means for Different Clusters

The Following results were found for serial and parallel K-Means while having different cluster size is on different size of datasets.

4.1.1 When Number of Clusters K=6

TABLE 4.1

EXECUTION TIME OF SERIAL V/S PARALLEL K-MEANS FOR 6 CLUSTERS

Data	K-Means	Parallel K-Means	Speed Up
400	1097	803	36
1200	1420	1150	23
2300	2551	1760	44

Figure 4.1: Running Time of Serial v/s Parallel K-Means for 6 Clusters

In Figure 4.1 on increasing the number of data records the execution time of Serial K-Means is very high in comparison with Parallel K-Means. On increasing the number of data records the execution time of serial K-Meansis always high in comparison with the Parallel K-Means Average Speed up in Parallel K-Means is almost 35% in comparison with Serial K-Means.

ISSN: 2321-8169
342 – 348

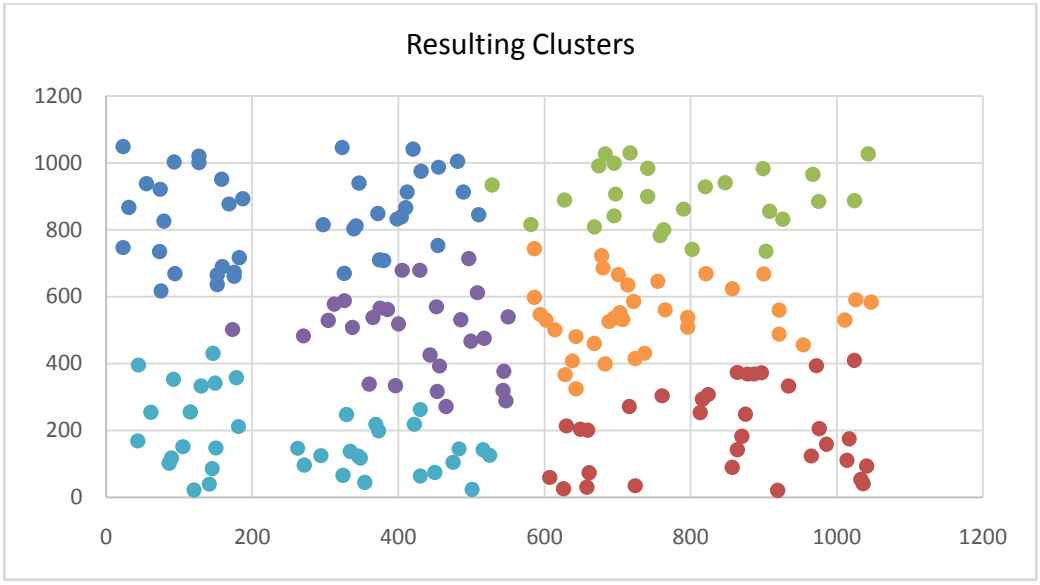

Figure 4.2: Resulting 6 Clusters for K=200

Figure 4.2 shows the different clusters for data set consisting 200 rows and number of clusters are 6. Different colors display the different group or cluster of different item sets. Items consisting same colors are within the same cluster. These are the results for the K-Means Algorithm that makes the clusters according their similarity and having minimum distance.

4.1.2 When Number of Clusters K=9

Following results were found for Serial and parallel K means while having cluster size is 9, on different size of datasets.

TABLE 4.2

EXECUTION TIME OF SERIAL V/S PARALLEL K-MEANS FOR 9 CLUSTERS

Data	K-Means	Parallel K-Means	Speed UP
400	985	865	13
1200	1854	1358	36
2300	2640	1743	51

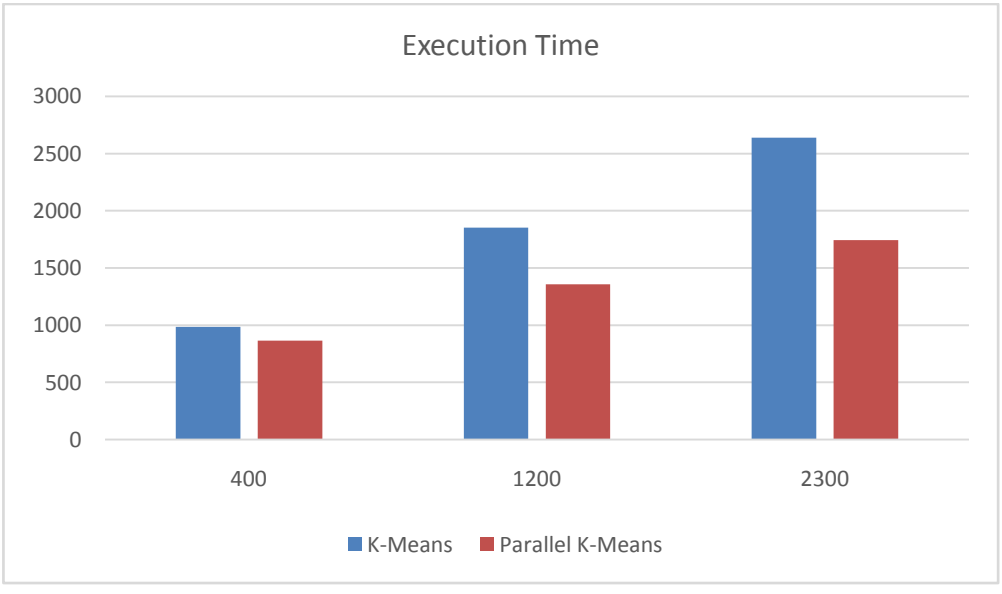

Figure 4.3: Running Time of Serial v/s Parallel K-Means for 9 Clusters

 ISSN: 2321-8169
342 – 348

In Figure 4.3 on increasing the number of data records the execution time of Serial K-Means is much increasing respectively in comparison with Parallel K-Means. On increasing the number of data records the execution time of serial K-Meansis always high in comparison with the Parallel K-Means. The performance of Parallel K-Means is significantly increasing in terms of Speed up. Average Speed up in Parallel K-Means is around 34% in comparison with Serial K-Means.

4.1.3 When Number of Clusters K=18

Following results were found for Serial and parallel K means while having cluster size is 18, on different size of datasets.

TABLE 4.3 EXECUTION TIME OF SERIAL V/S PARALLEL K-MEANS FOR 18 CLUSTERS

Data	K-Means	Parallel K-Means	Speed UP
400	1450	1126	28
1200	1766	1296	36
2300	2274	1688	34

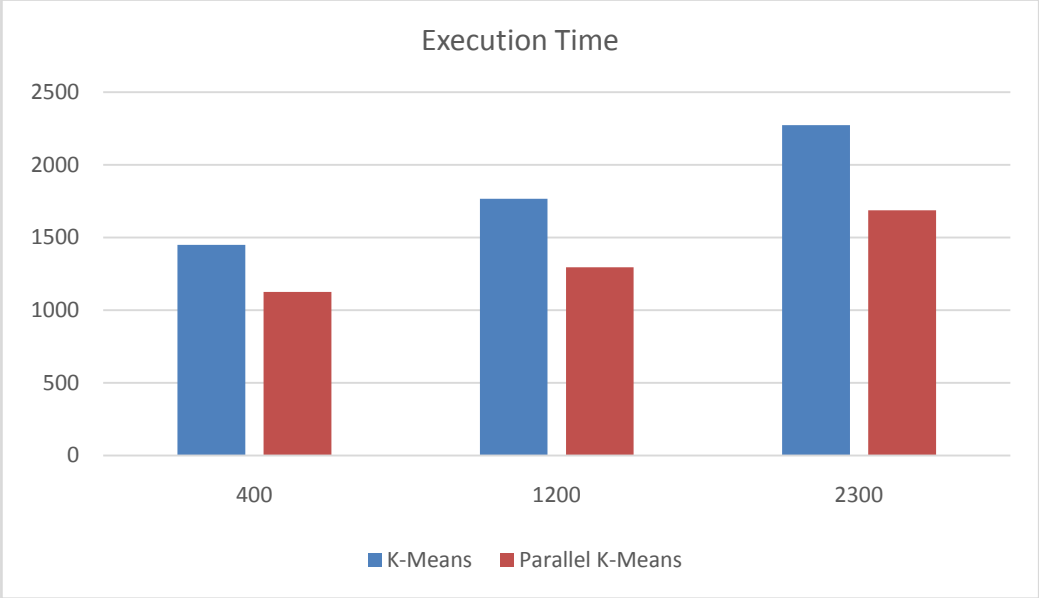

Figure 4.4: Running Time of Serial v/s Parallel K-Means for 18 Clusters

In Figure 4.4 on increasing the number of data records the execution time of Serial K-Means is very high in comparison with Parallel K-Means. As the number of data recordsare increased the execution time of serial K-Meansis always high in comparison with the Parallel K-Means Average Speed up in Parallel K-Means is almost 33% in comparison with Serial K-Means.

With different number of Clusters the performance of parallel K-Means is always better for any number of data records in comparison with serial K-Means execution. From the analysis it is clear that performance of Parallel K-Means is almost 35% better than serial K-Means in terms of Speed up of execution time.

4.2 Time for Serial v/s Parallel K-Means for Different Raw Data

These results came out from the analysis between the Serial and Parallel K-Means with Dunn's index. Here operations are performed on different size of datasets and in Dynamic K-Means clusters size is calculated by using Dunn's Index.

TABLE 4.4

EXECUTION TIME OF SERIAL V/S PARALLEL K-MEANS FOR DIFFERENT RAW DATA

Data	K-Means	Parallel K-Means	SpeedUP
100	390	312	25
400	1097	803	36
1200	1420	1150	23
1800	1710	1340	27
2300	2551	1760	44

 ISSN: 2321-8169
342 – 348

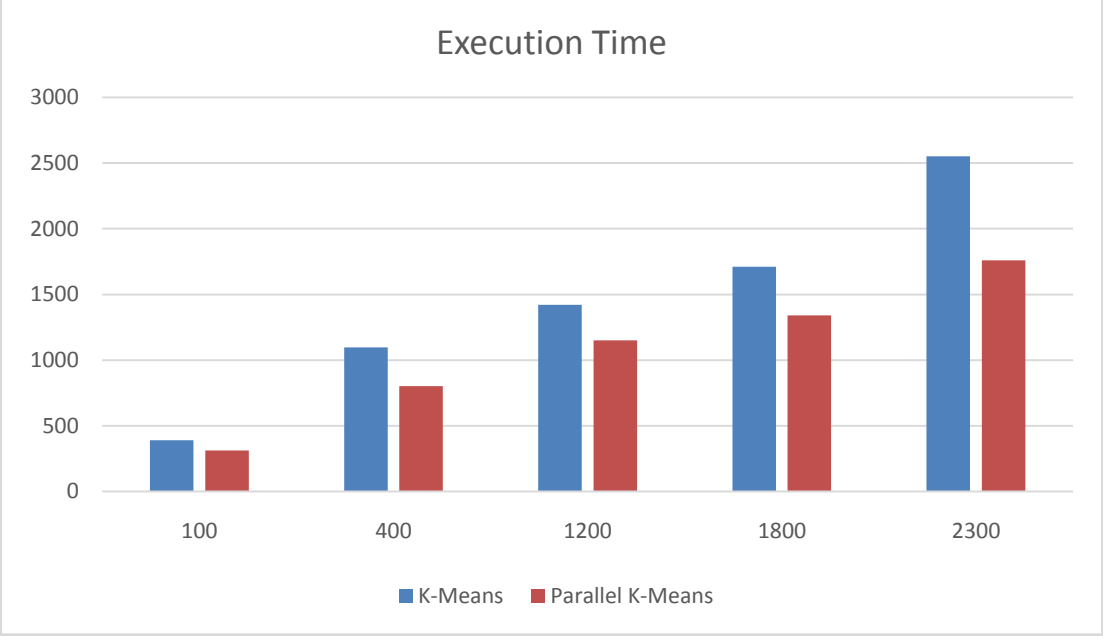

Figure 4.5: Running Time of Serial v/s Parallel K-Means by Dunn's Index

Figure 4.5 shows the execution time for different data sets consisting dynamic number of clusters for Parallel K-Means. On increasing the number of data records the execution time of serial K-Meansis always high in comparison with the Parallel K-Means Average Speed up in Parallel K-Means is almost 30% in comparison with Serial K-Means for a large range of data sets.

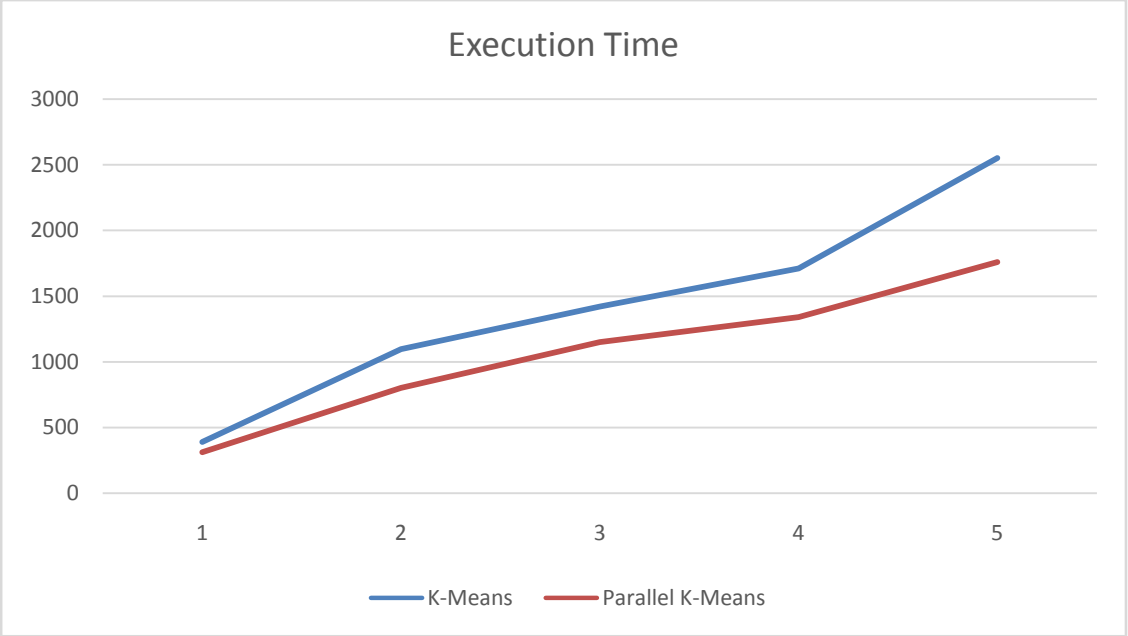

Figure 4.6: Speedup Ratio between K-Means & Parallel K-Means

Figure 4.6 gives the analysis between the Serial and Parallel K-Means with Dunn's index. Here operations are performed on different size of datasets and in Dynamic K-Means clusters size is calculated by using Dunn's Index. This figure shows the speed up of parallel operation over serial operation for table 4.4

5. Conclusion and Future work

The algorithm works fine for the unknown data set with improved results than traditional K-Means clustering. K-Means algorithm is widely known for its simplicity and the alteration is done in the proposed method with maintenance of simplicity. The traditional K-Means algorithm obtains number of clusters (K) as input from user. The major problem in traditional K-Means algorithm is that it fixed the

total number of clusters to be generated in advance. The results shows that the proposed approach has overcome the problem by calculating the feasible number of clusters with the help of Dunn's index, and use of parallel libraries enables the algorithm to perform better than conventional K-Means algorithm in all scenarios with small or large datasets. Future work can be done on how to minimize the time complexity with no compromising in quality of cluster and its optimality. More experiments can be performed with natural datasets using different features. One can also use some more powerful parallel programming models like Intel's Cilkplus and OpenMP to obtain reduced execution time.

REFERENCES

[1] Michelin andJ.Hanmorgan Kauffman "Data mining concepts and techniques", 2006.

[2] U.Maulik, and S.Bandyopadhyay, "Genetic Algorithm-Based Clustering Technique" Pattern Recognition 33, 1999.

[3] M.Murty&K.Krishna,"Genetic k-means Algorithm,"IEEE Transactions on System, Vol.29, No.3, 1999.

[4] Manish Verma, Mauly Srivastava, NehaChack, Atul Kumar Diswar andNidhi Gupta, " A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Reserch and Applications (IJERA),Vol. 2, Issue 3, pp.1379-1384, 2012

[5] Mukul Sharma and Pradeep Soni, "Comparative Study of Parallel Programming Models to Compute Complex Algorithm",International Journal of Computer Applications (0975 – 8887) Volume 96– No.19, June 2014

[6] TIAN Jinlan ZHU Lin ZHANG Suqin, "Improvement and Parallelism of K-Means Clustering Algorithm", TSINGHUA SCIENCE AND TECHNOLOGY, ISSN 1007-0214 Pages 277-281, Volume 10, Number 3, June 2005

[7] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh et. al, "Top 10 algorithms in data mining", Knowledge and InformationSystems, Volume 14, Issue 1, Pages 1-37, January 2008

[8] E. Kijsipongse and S. U-ruekolan, "Dynamic load balancing on GPU clusters for large-scale K-Means clustering", IEEE International Joint Conferenceon Computer Science and Software Engineering (JCSSE), Pages 346-350,2012

[9] P.Bradley, and U.Fayyad, "Refining Initial Points for K-means Clustering,", In Proceeding of 15th International Conference on Machine Learning, 1998.

[10] K.A Abdul Nazeer and M.P Sebastian, "Improving the Accurancy and Effiency of the K-means Clustering Algorithm", WCE London, 2009

[11] O. A. Abbas,"comparisons between data clustering algorithms", the international Arab journal of information technology, vol. 5, no. 3, July 2008.

[12] Monika Sharma and Jyoti Yadav, "A Review of K-mean Algorithm",International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013.

[13] GrigoriosTzortzis and AristidisLikas, "The MinMax k-Means clustering algorithm", Pattern Recognition 47(2014)2505–2516, Pages 2505-2516, Elsevier Ltd, February 2014.

[14] JeyhunKarimov and Murat Ozbayoglu, "Clustering Quality Improvement of k-means Using a Hybrid Evolutionary Model", Procedia Computer Science, Volume 61, Pages 38–45, Complex Adaptive System, Elsevier Ltd, 2015.

[15] AhamedShafeeq B M and Hareesha K S, "Dynamic Clustering of Data with Modified K-Means Algorithm", International Conference on Information and Computer Networks IACSIT Press, Singapore, Vol 27, 2012.

[16] Amandeep Kaur Mann, NavneetKaur**,**"Survey Paper on Clustering Techniques ",Volume 2, Issue 4,ISSN: 2278 – 7798 International Journal of Science, Engineering and Technology Research (IJSETR), April 2013

[17] RomanaRiyaz andMohdArifwani, **"**Review and Comparative Study of Cluster ValidityTechniques Using K-Means Algorithm",International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE), Volume 1, Issue 3, August 2014.