K-Meansusing OpenMP: An Approach

Prateek Swamy*[1]
Student M.tech 2nd year,
Dept of Computer Science & Engineering,
Rajiv Gandhi College of Engineering & Research, Nagpur, India.
swamy.prateek11@gmail.com

Dr. M. M Raghuwanshi[2]
Professor,
Dept of Computer Technology,
Yeshwantrao Chavan College of Engineering
Nagpur, India
m_raghuwanshi@rediffmail.com

Ashish Gholghate[3]
Asst. Professor,
Dept of Computer Science & Engineering,
Rajiv Gandhi College of Engineering & Research, Nagpur, India
ashishgolghate@gmail.com

Abstract–Serial execution of K-means algorithm on large dataset takes more execution time and does not give accurate results. Parallel processing is one of the ways to improve the performance of K-Means algorithm. But the execution time and accuracy is largely dependent on selection of initial cluster centers. In this paper, parallel processing of K-Means is proposed using an initialization method to originate initial cluster centers, which not only reduces the execution time but also gives accurate results.

Keywords—*Serial execution,large dataset,Parallel processing, K-Means, execution time, accuracy, initial cluster centers.*

*****

I. INTRODUCTION

Clustering of data is a key technique for analysis of data, which is used to find the similarity or dissimilarity between groups of item in a dataset such that item in one group are more similar than other group and vice versa[1].Because of modern methods for scientific data collection, size of database is increasing day by day. As a result, data mining is getting practically difficult by using conventional techniques [2].An efficient algorithm for mining of data is the need of the hour so that useful information from large databases can be extracted.

Large number of algorithms has been developed for data clustering task. K-Means is the oldest and probably the most popular algorithm proposed fordata clustering task [3].It is easy, efficient, fast and sensitive. But, K-Means algorithm has some issues [2, 4, 5, 6, 8]. These are

- Initialization of initial cluster center is random, which affects the accuracy.
- Serial execution takes more time
- No information about number of clusters in the dataset.
- Stuck in local optima

In this paper, parallel processing of K-Means is proposed using an initialization method to originate initial cluster centers, which not only reduces the execution time but also gives accurate results. Hence, a novel methodis proposed which tries to improve K-Means by using an initialization method to address the first issue [7] and parallel processing using OpenMP [8] to address the second issue.

A. Exisiting Algorithm: K-Means clustering algorithm

Pseudocode for the k-means algorithm [9]is given as follows:

Input:D= {d1, d2,…..,dn } //set of n data items

K //Number of desired clusters

Output: A set of K clusters

Steps:

1. Arbitrarily choose K data items from D as initial centroids
2. **Repeat**
   - 2.1 Assign each data item di to the cluster which has the closest centroids.
   - 2.2 Calculate the new mean of each cluster,

   Until convergence criterion is met

It can be observed from the above algorithm that initial centroids are chosen randomly. This affects the accuracy of algorithm [2]. In case if the dataset is large, serial execution of k-means takes more time [8]. These are the two important drawbacks of k-means clustering algorithm.

To improve the accuracy of k-means, the initial centroids are originated by using binary search technique [7].To speed up the execution time of k-means, parallel processing can be used UsingOpenMP [8]. A given task is broken down into discrete parts and parallel execution is done with the help of child process. Execution of child process is simultaneous on different CPUs [10, 11].

The OpenMPApplication Programming Interface is one of the best emerging standards for parallel programming on shared-memorymultiprocessors.It extends existing languages such as FORTRAN and C/C++ with a set of directives. Touse parallelism with the code in OpenMP, thecompiler directives are used. [12].

Rest of the paper is structured as related work in section II; Section III describes data clustering using K-Means; proposed approach in section IV and conclusion of paper in section V.

II. RELATED WORK

A lot of efforts have been made by researchers to improve the accuracy and efficiency of k-means algorithm.K.A Abdul Nazeer *et al*. [2] have proposed a heuristic technique to find better initial centroids but the time complexity of the algorithm is not enhanced in this method. Moreover, if the number of attribute is more, then there is a chance that efficiency of the algorithm can be affected.

 ISSN: 2321-8169
319– 322

Yuan *et al.*[13] have proposed a technique to find out the initial centroids. The centroids obtained by this method produce clusters, which are more accurate than that of original k-means algorithm. But, efficiency of k-means is not improved using Yuan's method.

Fahim A M *et al.*[14] proposed a better approach for assigning data-points to clusters. The original k-means algorithm is computationally very expensive because there is a need to calculate the Euclidean distance between data points and all preliminary centroids. In Fahim's approach, two distance function are used, one similar to k-means algorithm and another one based on prediction is used to minimize the number of distance computations. But, in this method, centroids are selected randomly, which affects the accuracy of final clusters.

DS Bhupal Naik *et al.*[8] have proposed a technique in which parallel processing of Enhanced K-means using OpenMP is done. Using this approach, the time taken for parallel processing of Enhanced K-means is optimized as compared to the serial approach. But, approach used in the selection of initial centroids is a heuristic one and hence it is not that accurate.

Yugal kumar *et al.* [7] has proposed a new initialization method to originate initial cluster centers for k-means algorithm based on binary search technique. The accuracy obtained in this method is impressive as compared to other method. But the time taken is more because the method is implemented in sequential manner.

III. DATA CLUSTERING USING K-MEANS ALGORITHM

As discussed in section I, clustering of data is a key technique for analysis of data. K-means algorithm is widely used for clustering of data [3]. K-means algorithm is implemented in serial manner using multivariate dataset with known clustering available at UCI repository of machine learning.

a. Dataset Information

The data set used is the Twenty Newsgroups Data Set [15].20 News Groups is quite data set used for text clustering and classification. It has a collection about 20,000 documents across 20 different newsgroups from Usenet. Each newsgroup is stored in a subdirectory, with each article stored as a separate file. Clustering time is recorded for different quantity of documents. Number of iterations is also changed.

TABLE I
PERFORMANCEOF K-MEANS ALGORITHM WHEN IMPLEMENTED IN SERIAL MANNER

Size(Mega Bytes)	Number of documents	Number of iterations	Time taken(seconds)
21	6000	50	4895
17	4000	30	1766
8	2000	20	525

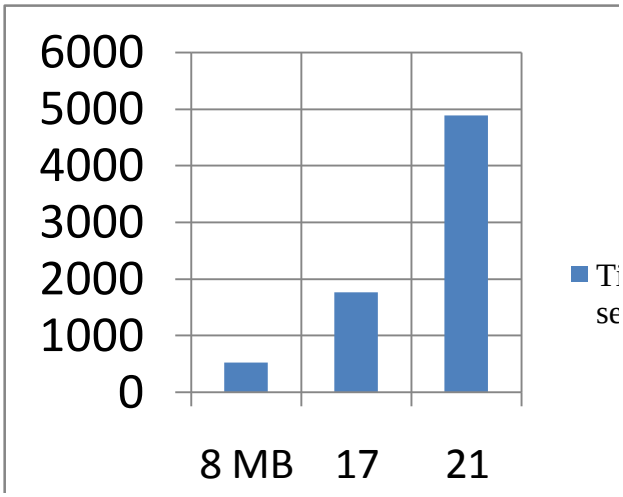

Fig.1. K-Means execution time Vs size of documents

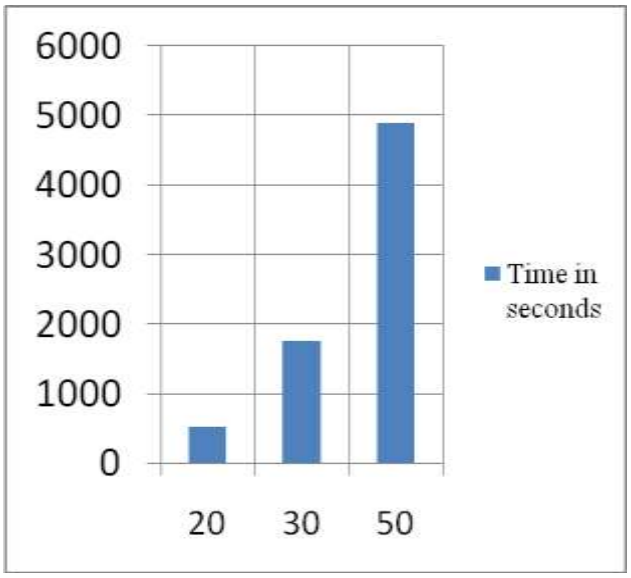

Fig.2. K-Means execution time Vs Number of iterations

From figure 1, it is clear that the time taken for clustering is in direct proportion with the size of data. From figure 2, it can be observed that as the number of iterations increases, the clustering time also increases. Serial execution of k-means algorithm slows down the process of clustering. Clustering of large dataset using original k-means is a tedious task. So, there is a need of an efficient method, which not only reduces the execution time for clustering of data, but also gives accurate results.

IV. PROPOSED APPROACH

In this section, parallel processing of k-means algorithm using OpenMP is introduced [8] with an efficient method to find initial centroid [7].We believe that combination of these two approaches is new technique in the field of clustering of data. Using this approach, clustering time can reduced to a great extent with great efficiency.

Number of cluster **K** and dataset is given as an input. Then, data is partitioned and each part is assigned to each process. For every partition, initialization of OpenMP process is done to

perform parallel processing in multi-core systems. Now, in each partition, value of initial centroids is calculated and**K** local clusters are formed till convergence criterion is met.

Synchronization of result obtained at each **K** local cluster formed at P process is done to get **K** global clusters. This process is continued till convergence is met.

In every partition, values of initial centroids arecalculated using an initialization method, which is based on the unique property of binary search algorithm [7, 16]. In binary search algorithm, the value of middle item in list is calculated as follows:

$$A[mid] = A[beg] + A[end]/2. \quad (1)$$

The above property is modified to find initial cluster points for K-means algorithm [7].

- A [beg] is replaced by A [max]
- A [end] is replaced by A [min]
- 2 is replaced by K , numbers of clusters
- A [mid] is replaced by any variable such as M
- Plus symbol is replaced by minus symbol

Now, the equation (1) is formulated in another equation as given below:

$$M = A[max] - A[min] /2. \quad (2).$$

The generalization form of the equation (2) can be written as:

$$M_i = \max (A_i) - \min (A_i)/K \quad (3).$$

The equation (3) is used to calculate the value of the variable M that specifies the range of initial cluster centers but not give the cluster centers.

The cluster centers for K-Means algorithm are generated using given equation.

$$C_k = \min (A_i) + (K-1) * M \quad (4).$$

Consider an example dataset D that is given in Table II. The given dataset is applied with proposed method to get the initial cluster points. This dataset is consist total number of instances (N) = {9}, no. of attributes (i) = {2} and number of Clusters (K) = {3}. The working of proposed method is given below:

Table II: Example dataset to generate the initial cluster center

Objects	X1	X2	X3	X4	X5	X6	X7	X8	X9
A	1.1	1.3	1.2	3.2	2.8	2.9	2	1.9	2.2
B	4.3	3.9	3.8	4.8	3.9	3.7	3.6	3.3	3.2

- Calculate the maximum and the minimum values of each attribute in the dataset.

  Maximum = (3.2, 4.8) and Minimum = (1.1, 3.2)

- Calculate the value of M as

  M= {(3.2 – 1.1)/3, (4.8 – 3.2)/3}

  M= {0.70, 0.53}

- Generate the initial cluster centers for initialization as

  C_1 = (1.1 + ((1-1) * 0.70), 3.2 + ((1-1) * 0.53))

  = (1.1, 3.2)

  C_2= (1.1 + ((2-1) * 0.70), 3.2 + ((2-1) * 0.53))

  = (1.8, 3.73)

  C_3= (1.1 + ((3-1) * 0.70), 3.2 + ((3-1) * 0.53))

  = (2.5, 4.26)

The newly generated cluster centers (1.1, 3.2), (1.8, 3.73) and (2.5, 4.26) are used as initial cluster centers for K-Means algorithm.

A. Algorithm for Proposed approach

Input:*D*= {D1, D2, D3, D4,…Dn}
K= No. of cluster
Output: *K* Clusters
Procedure:

1. Initialize the OpenMP Processes= {P1, P2,.. Pn}.
2. Partition the dataset and assign each part of dataset to each process.
3. **For each process repeat steps a-e;**

a. Take the value of K as given by user.
b. Generate the range of the initial centroids using following:

$$\mathbf{M_i = \max (D_j) - \min (Dj)/K}$$

Where j = 1, 2, 3 …….n.

c. Obtain the initial cluster centers C_k using the following equation:

$$\mathbf{C_k = \min (D_j) + (K-1) * M}$$

d. Calculate the Euclidean distance as similarity measure of each attribute Dj and assigned to cluster center Ck using following equation:

$$\mathbf{Dist. = \min (|| Dj – Ck ||2)1/2.}$$

e. Continue the process to form final K local cluster.

Until convergence criterion is met.

1. Now, synchronize the results from individual process
2. Then obtain the Global *K* clusters from all local K clusters.
3. Continue the procedure till convergence is met.

The above approach can be used to improve the efficiency of k-means algorithm.

V. CONCLUSION AND FUTURE WORK

In this paper we have discussed clustering of data with original K-means algorithm. There is no doubt that K-means algorithm is still one of the most widely used techniques for data

clustering. But, as the size of dataset and the number of iterations increase, the execution time taken by K-means algorithm increases exponentially. There has been lot of improvement suggested for K-means algorithm as discussed in section II. Our approach is another step towards improving the efficiency and accuracy of K-means algorithm. We have proposed a novel approach by using Parallel Processing and an efficient an initialization method to originate initial cluster centers. In future, we will test our approach with the standard dataset to show its practical efficiency and accuracy.

REFERRENCES

[1] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: A review", *ACM Computing. Surveys*, vol. 31, pp. 264-323, 1999.

[2] K A Abdul Nazeer et.al., "Enhancing the k-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroids", IEEE Second International Conference on Emerging Applications of Information Technology,pp-261-264, 2011.

[3] J Macqueen, "Some methods for classification and analysis of multivariate observations". Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, California, 1967.

[4] S. Z. Selim and M. A. Ismail, "K-means type algorithms: A generalized convergence theorem and characterization of local optimality", IEEE Transactions on Pattern Analysis andMachine Intelligence., vol. 6, pp. 81-87,(1984).

[5] A. K. Jain, "Data clustering: 50 years beyond K-means", PatternRecognition Letters archive. Volume 31 Issue 8, June, 2010.

[6] Y. T. Kao, E. Zahara, and I. W. Kao, "A hybridized approach to data clustering, ", Expert Systems with Applications, vol. 34, no. 3, pp. 1754–1762, 2008.

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

[8] DS.Bhupal Naik, S. Deva Kumar, S.V Ramakrishna, "Parallel Processing Of Enhanced K-Means Using OpenMP", IEEE International Conference on Computational Intelligence and Computing Research, 2013.

[9] Margaret H. Dunham, "Data Mining- Introductory and Advanced Concepts", Pearson Education, 2006.

[10] Sanjay Goil, Harsha Nagesh, Alok Choudhary, "MAFIA: Efficient and Scalable Subspace Clustering for Very Large Data Sets" , 1999.

[11] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy: "Advances in Knowledge Discovery and Data Mining",American Association for Artificial Intelligence Press, 1996.

[12] Mitsuhisa Sato, "OpenMP: Parallel programming API for shared memory multiprocessors and on-chip multiprocessors", International Symposium on System Synthesis (ISSS) 2002, Kyoto, Japan, 2002.

[13] FANG Yuan, Zeng-Hui Meng, , Hong-Xia Zhanhz, Chun-Ru Dong , "A New Algorithm To Get the Initial Centroids", Third International Conference on Machine Laming and Cybernetics, Shanghai, 2004.

[14] Fahim A.M. Salem A.M. Torkey F.A. Ramadan M.A.,"An efficient enhanced k-means clustering algorithm", Journal of Zhejiang University, 10(7):1626-1633,2006.

[15] Tom Mitchell, UCI Repository of Machine LearningDatabasesAvailable:http://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

[16] Abdolreza Hatamlou, "In search of optimal centroids on data clustering using a binary search algorithm", Pattern Recognition. Letter archive.vol. 33, pp. 1756-1760, 2012.

[17] Fahad. A, Ashtari.N, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", *IEEE Transactions on Emerging Topics in Computing*, 2014.