International Journal of Engineering Research and Development
e-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com
Volume 5, Issue 12 (February 2013), PP. 24-28

Automated Clustering in K-Means Using Double Link Cluster Tree (DLCT)

R.Ranga Raj[1], Dr.M.Punithavalli[2],
[1]Head of the Department Computer Science, Hindusthan College of Arts and Science Coimbatore
[2]Director, Department of Computer Applications, Sri Ramakrishna Engineering College Coimbatore

Abstract:- Clustering is the process of grouping related documents from the large collection of database. The mining of such related documents from the enormous database which are unlabelled is a challenging one. To overcome this problem, clustering is used to filter the unlabelled documents from the large collection of database. Clustering can be achieved by various algorithms that differ significantly in their notion and how to efficiently find them. The Standard K-Means algorithm is a well known data mining algorithm which can effectively cluster data in the database. K-mean is a simple algorithm that has been adapted to many problem domains. Hence by using k-mean, the initializations of number of clusters can be done through manually. In this research paper, a new technique DLCT (Double Link Cluster Tree) is merged with the enhanced K-Mean algorithm which helps to makes clustering in an efficient manner by without initializing of number of clusters and optimal clusters. The result of k-mean with DLCT, which allows automatic determination of number of clusters on any type of data such as documents, images etc.
General Terms Effective Clustering Using DLCT

Keywords:- Clustering, K-Means Enhanced Approach Algorithm, Double Link Cluster Tree (DLCT), unsupervised clustering.

I. INTRODUCTION

Clustering algorithm can be categorized based on several cluster method. A cluster is a set of points such that a point in a cluster is closer to one or more other points in the cluster than to any point not in the cluster. A good clustering method will produce high quality cluster in which the intra cluster similarity is high and inter class similarity is low.

For example, in an image related datasets it is difficult to identify how many clusters are available. Image clustering which is an important technology for processing image that has been actively researched for a long period of time. Recently the growth of interest in supervised method makes to improve the way of representing image sets. Image clustering is the high level description of image content. Nowadays the grayscale images are very important for analyzing image contents which has the application of satellite images to medical images. Such analysis becomes very complex. By using certain mathematical approach [1] the clustered grayscale images are determined with the optimal cluster number. The clustering process which separates the data into number of segments those are in the form of n-dimensional space. These segmented data uses a specific function which helps to model the data distribution [8]. Based on intra cluster and inter cluster distance measure in the mathematical approach which allows the number of clusters to be determined automatically.

Cluster will be grow depend on the size of database. In other hand some existing subjects concentrates on reducing iteration in K-means method [2] during the clustering process so to obtain an optimized cluster output. These algorithm uses methods such as Genetic algorithm , PSO, Ant Colony Optimization (ACO) Using Genetics algorithm(GA) and PSO these are the optimization technique, to reduce the no of iteration. The new unsupervised k-means clustering algorithm [2] can be applied for the any type of datasets such as images, documents etc. This clustering process depends on the size of the database and concentrates on reducing iterations in K-Means method [1] to obtain an optimized cluster output.

II. EXISTING WORK

Clustering can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. If the cluster analysis is done on image clustering, the growth of interest in unsupervised method makes to improve the way of representing image sets. There are

many clustering algorithms that can be performed on the image but by using k-means algorithm on image will obtain several segments. The k-mean clustering algorithm [2] will not provide the optimal cluster number and also makes initialization of number of clusters on the given data set. One way to overcome this problem is the mathematical approach [1], which is used to specify optimal cluster number with intra and inter cluster distance measure and finally allows automatic determination of number of clusters.

III. K-MEANS ALGORITHM

Now, using K-Means clustering algorithm we can cluster an image to obtain segments. To run this algorithm, we need to provide the value of K which is nothing but the number of cluster centers.

3.1 Enhanced Approach

The algorithm for Enhanced K-Means is as Follows

Input:

D={d1,d2,d3,……..dn}//Set of n data points

C={c1,c2,…….ck}//Set of k clusters

Output:

A set of k clusters

Steps:

1. Compute the distance of each data points di(1<=i<=n) to all the centroids cj(1<=j<=k) as d(di,cj).
2. For each data point di, find the closest centroid cj and assign di to cluster j.
3. Set cluster id[i]=j;//j:id of the closest cluster.
4. Set Nearest_Dist[i]=d(di,cj);
5. For each cluster j(1<=j<=k),recalculate the centroid;
6. Repeat;
7. for each data point di;
   - 7.1 Compute its distance from the centroid of the present nearest cluster;
   - 7.2 If this distance is less than or equal to the present nearest distance, the data point stay in this cluster; Else
   - 7.2.1 For every centroid cj (1<=j<=k) Compute the distance d(di,cj);
   - End for;
   - 7.2.2 Assign the data point di to the cluster with the nearest centroid cj;
   - 7.2.3 Set Cluster id[i]=j;
   - 7.2.4 Set Nearest _Dist[i]=d(di,cj);
   - Endfor;
   - 8. For each cluster j(1<=j<=k);
   - Recalculate the centroids;
   - Untill the convergence criteria is met.

IV. DLCT

DLCT is a Double Link Cluster tree which is used as the enhancement of Enhanced K-means algorithm. In certain dataset there will need for algorithm which can cluster the data without any initialization (i.e.) NO_OF_CLUSTERS.

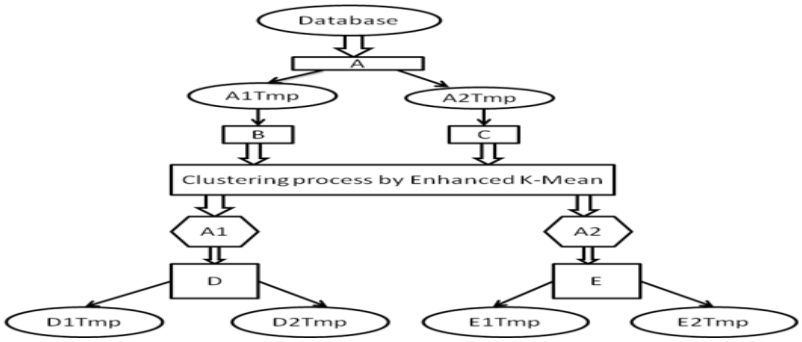

Fig 1: The Diagrammatical representation of the DLCT

DLCT is an algorithm, which works as a looping frame for Enhanced K-Means algorithm and makes Enhanced k-means algorithm to cluster the dataset several times.

The Algorithm for this method is shown below
Step 1: Get the Database as input for the purpose of processing Let it be **DB**
Step 2: Initialize the Level to 1 i.e., Level=1;
Step 3: Find mean for the **DB** let it is **A**
Step 4: By means of **A,** separation the **DB** Dataset is processed in to two groups let the First group be **A1tmp** which contains
the value should be minimum to **A** i.e. less than **A.** Let the Second group be **A2tmp** should contain the value, less than
A.
Step 5: Now have two clusters **A1tmp** and **A2tmp.**
Step 6: Like Step3 find mean for the **A1tmp** cluster and **A2tmp** Cluster Let it be **B** and **C**
Step 7: Now initialize K-Means Algorithm with Initial Cluster center as **B** and **C** and No of Cluster would be 2
Step 8: Begin to cluster the Database **DB.** When K-means algorithm finished clustering the Database after certain Iterations the
Outcome will be two clusters they are **A1** and **A2**
Step 9: Increment the Level to 1 (Now Level =2;)
(So at the level one, acquire 2 power 1 i.e 2 clusters at Level 2 and get 4 clusters vise versa)
Step 10: Now at Level 2 the same steps or procedure were handled for cluster **A1** and **A2** (Steps from 4 to 9)
Level inside our algorithm an algorithm
Level 0➔ 1 cluster (probably Original Dataset)
Level 1➔ 2 clusters
Level 2➔ 4 clusters
Level 3➔ 8 clusters and so on;

V. PROPOSED WORK

In the traditional K-Means algorithm it has limitations of getting no of cluster centers [4] by means of its user. This makes K-Means difficult to use, where there is a unpredictable datasets are available. So to solve this problem we proposed a method named DLCT (Double Link Cluster Tree). In certain dataset there will need for algorithm which can cluster the data without any initialization.

During the clustering processes are,

1. Sometimes the cluster centers of the tree node would be same. The algorithm will merge the two clusters which has unique cluster center
2. Due to data insufficiency some cluster will does not have any cluster element. So that cluster space was terminated.

In the standard algorithm, the usage of K-Means algorithm [2] is allowed for clustering the datasets only one time, hence the numbers of clusters are given manually. The concept K-Means Enhanced Approach Algorithm with Double Link Cluster Tree (DLCT) focuses on clustering of documents and images in an efficient way by without initializing the number of clusters. In DLCT, the K-Means algorithm is used as library to process the clustering among all the Database types, images always contains the unpredictable amount of cluster in the Fig 2 and 3. It shows input image before clustering and output image after clustering.

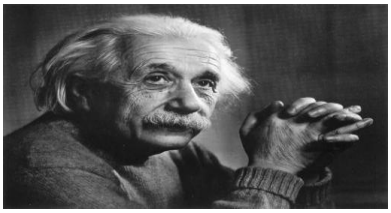

Fig 2: Input image before clustering

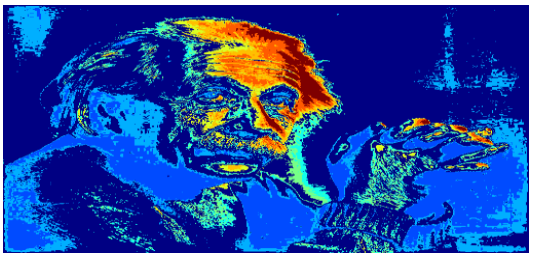

Fig 3: Output Image after Clustering

Database Contains➔ 200*200 Elements i.e. 40,000 items. Among them the cluster using DLCT is made as,
Cluster1=30307 Elements
Cluster2= 1300 Elements
Cluster3 =865 Elements
Cluster4=1997 Elements
Cluster5= 3167 Elements
Cluster6= 1621 Elements
Cluster7=743 Elements
For example, the Representation of Number of Clusters in the Character Database is shown in Fig 4 and Fig 5.

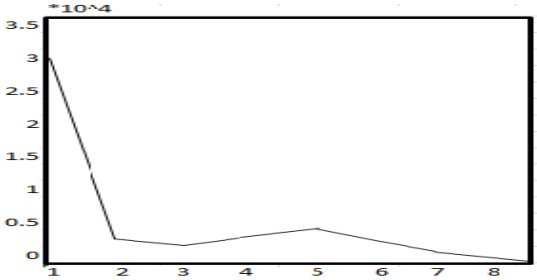

Fig 4: Number of Character Database

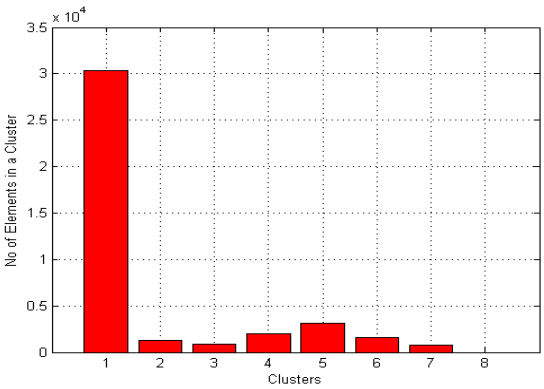

Fig 5: Number of Character Database

VI. CONCLUSION

The clustering of datasets with Enhanced K-Mean algorithm [2] and DLCT helps to make clustering in an efficient way, by without initializing the number of clusters in an unpredictable database. The designing of Double Link Cluster Tree (DLCT) algorithm is done in such a way is solves the problems in previous unsupervised methods implemented. When comparing the proposed algorithm with various existing clustering algorithm, the analysis of clustering process results in the cluster centers. The proposed algorithm is also an automatic optimization process since the separation of the clusters is done each time for every level of the process. In which every cluster centers should have a difference of about 60% else those clusters that have less difference below 60% will be merged and considered as single cluster. This process will be handles to avoid minimum distanced cluster to be a separate clusters.

REFERENCES

[1]. A Novel Approach for Determination of Optimal Number of Cluster", Debashis Ganguly.Computer Science and Engineering,Department,Heritage Institute of Technology,Anandapur Kolkata – 700107, India

[2]. Improving the accuracy and efficiency of K-Mean Clustering Algorithm", by K.A. Abdul Nazeer, M.P. Sebastian. Proceeding of the world congress on Engineering 2009 vol I WCE 2009, July 1-3, 2009, London, U.K.

[3]. Clustering Algorithms Based on Volume Criteria" Raghu Krishnapuram and Jongwoo Kim IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 8, NO. 2, APRIL 2000.

[4]. An Efficient k-Means Clustering Algorithm: Analysis and Implementation" Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE VOL. 24, NO. 7, JULY 2002

[5]. A Comparison of Document Clustering Techniques", Michael Steinbach,George Karypis. Department of Computer Science University of Minnesota Technical Report #00-034 steinbac, karypis, kumar@cs.umn.edu Vipin Kumar.

[6]. Document Image Segmentation Using Wavelet Scale–Space Features" Mausumi Acharyya and Malay K. Kundu, Senior Member, IEEE VOL. 12, NO. 12, DECEMBER 2002

[7]. Document Clustering in Correlation Similarity Measure Space" Taiping Zhang; Yuan Yan Tang; Bin Fang; Yong Xiang Knowledge and Data Engineering, IEEE Transactions on Volume: 24 ,,2012

[8]. A review on image segmentation techniques,Pattern Recognition" , N.R. Pal and S.K. Pal, vol. 26, pp. 1277-1294, 1993.

[9]. A Modified K-Means Algorithm for Circular Invariant Clustering", Dimitrios Charalampidis, Member, IEEE VOL. 27, NO. 12, DECEMBER 2005

[10]. Randomized Clustering Forests for Image Classification" Frank Moosmann, Student Member, IEEE, Eric Nowak, Student Member, IEEE, and Frederic Jurie, Member, IEEE Computer Society VOL. 30, NO. 9, SEPTEMBER 2008

[11]. Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and *k*-Means Clustering", Turgay Celik, IEEE GEOSCIENCE VOL. 6, NO. 4, OCTOBER 2009