SIMULATION OF ENHANCED KMEAN CLUSTERING MECHENISM ON MATLAB

Nancy, nancymalik339@gmail.com

ISSN 2454-308X
9 770024 543081

ABSTRACT: K-means clustering is very Fast, robust & easily understandable. If data set is separated from one other data set, then it gives best results. Clusters do not having overlapping character & are also non-hierarchical within nature. Some challenges are related to visualization & querying of data. Scientist has faced several challenges in e-Science such as meteorology, complicated physics simulation & environmental researches. Lot of challenges has been faced due to big data in case of biology & genomics. Problems with existing system were search, sharing, storage, transfer, and visualization, querying-updating. These problems can be reduced by using proposed algorithm. In this paper we have explain clustering & proposed algorithm is discussed. We have simulated the enhanced K-Mean clustering using MATLAB.

Keywords: Clustering, K-Mean, Data Mining, MATLAB

1. INTRODUCTION

Clustering is[1] a process of data into a group of meaningful sub-classes is called clustering. Used either as a stand-alone tool to get insight into data distribution or as a pre processing step for other algorithms.

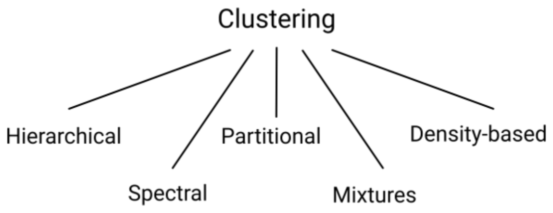

Fig 1 Clustering

By examining one or more attributes or classes, you may group individual pieces of data value together to form a structure opinion. At a simple stage, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is valuable to identify dissimilar info since this correlates with other examples so you may see where similarities & ranges agree[2]. Clustering may work both ways. You may assume that there is a cluster at a certain point & then use our credentials criteria to see if you are correct.

2. REQUIREMENTS OF CLUSTERING

The following points throw light on why clustering is required in data mining[3]

Scalability

We need highly scalable clustering algorithms to deal with large databases.

Capacity to contract within different type of attributes

The K-Mean Algorithms would be able to be implement on some kind of information like as interval based data binary data & categorical,.

Discovery of clusters with attribute shape:- clustering algorithm should be capable of detecting clusters of arbitrary shape[4]. We have not to be bounded to distance measures that care of to find round cluster of small sizes.

High dimensionality:- clustering algorithm should not only be able to handle low-dimensional data but also high dimensional space.

Capability to contract within noisy data

The Databases hold noisy absent or erroneous data. Some algorithms are responsive to such data & poor quality clusters.

Interpretability:-The clustering results should be interpretable, comprehensible, & usable.

3. PROPOSED WORK

Before specifying proposed work, K-Mean algorithm [5] is discussed. Proposed work is based on it algorithm. K-means clustering is known as partitioning method. In it objects are classified as belonging to one of K-groups. In each cluster there might be a centroid or a cluster presentative. In case where we think real valued data, mathematics mean of attribute vectors for all objects[6] within a cluster given an appropriate representative; alternative types of centroid might be required within other cases.

Suppose we had following data set

2
5
6
8
12
15
18
28
30

Suppose K=3

Cluster1=2

Cluster 2=12

Cluster 3=30

Cluster 1	2
	5
	6
	8
Cluster 2	12
	15
	18
	28
Cluster 3	30

The distance is calculated for each data point from centroid & data point having minimum distance from centriod of a cluster is assigned particular cluster.

So cluster according to distance are as follow

12-5>5-2

So cluster for data point 5 is Cluster 1

6-2>12-6

So cluster for data point 6 is Cluster 1

In same way cluster would be assigned

Cluster 1	2
Cluster 1	5
Cluster 1	6
Cluster 1	8
Cluster 2	12
Cluster 2	15
Cluster 2	18
Cluster 3	28
Cluster 3	30

Data member of Cluster 1 are 2,5,6

Data Member for Cluster 2 are 8,12,15,18

Data Member for Cluster 3 are 28,30

Clusters generated previously, centriod is again repeatly calculated means recalculation of centriod.

So mean of cluster C1 is (2+5+6)/3=4.3

So mean of cluster C2 is (8+12+15+18)/4=13.25

So mean of cluster C3 is (28+30)/2=29

Now distance would be recalculated within new mean & cluster[5] of data point would be changed according to new distance

Cluster 1	2
Cluster 1	5
Cluster 1	6
Cluster 1	8
Cluster 2	12
Cluster 2	15
Cluster 2	18
Cluster 3	28
Cluster 3	30

For example take 8 from C2 cluster

The issue within traditional system were search analysis, sharing, transfer storage, visualization & querying updating. One more problems within K-means clustering [14] is that empty clusters are generated during execution, if within no data points are allocated of cluster under consideration during assignment phase. proposed algorithm[5] overcome these problem. K-mean clustering proposed algorithm as follow.

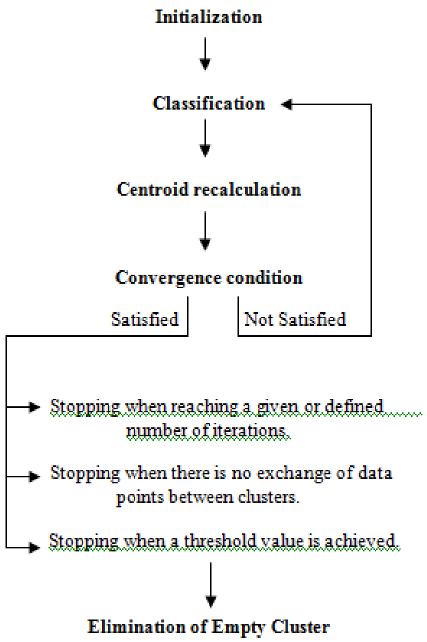
Initialization
Classification
Centroid recalculation
Convergence condition
Satisfied
Not Satisfied
Stopping when reaching a given or defined number of iterations.
Stopping when there is no exchange of data points between clusters.
Stopping when a threshold value is achieved.
Elimination of Empty Cluster

Fig 2 Proposed algorithm

STEPS OF K-MEANS CLUSTERING ALGORITHM

This algorithm is an thought within that there is require to categorize given data set into K clusters; value of K is defined by user which is fixed. The first centroid of each cluster is select for K-Mean clustering & then according to select centroid, data points having minimum distance from given cluster, is assigned to that particular cluster.

1. **Initialization:** In this first step data set, number of clusters & centroid that we defined for each cluster.

2. **Classification:** Distance is intended for every data system from centroid & data point having least space from centroid of a cluster is assigned to that particular cluster.

3. **Centroid Recalculation:** Clusters generated previously, centroid is again repeat calculated means recalculation of centroid.

4. **Convergence Condition:** Some convergence conditions are given as below:

4.1 Ending when reaching a provide or explain no. of monotony.

4.2 Ending when there is no replace of data system between clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of above conditions are not satisfied, then go to step 2 & whole process repeat again, until given conditions are not satisfied.

4. IMPLEMENTATION OF CLUSTER REMOVAL

K-Means algorithm converges to local minimum. Before *k*-means converges, centroid computed number of times, & all points are assigned to their nearest centroid, i.e., complete redistribution of points according to new centroid, this takes $O(nkl)$, where n is number of points, k is number of clusters & l is number of iterations. In existing enhanced *k*-means algorithm, to obtain initial clusters, this process requires $O(nk)$. In our research cluster generated previously is rechecked clusters where no data points are allocated to a cluster under consideration during assignment phase are eliminated.

Comparative analysis of result between Existing & Proposed K-MEAN

Number of record	Traditional (K-Mean)	Proposed Algorithm
1000	2	1
2000	3	2
3000	4	3
4000	6	4
5000	8	5
6000	8	6
7000	9	7
8000	13	8
9000	15	10
10000	17	12

Table 1 Comparative analysis of result between Existing & Proposed K-MEAN

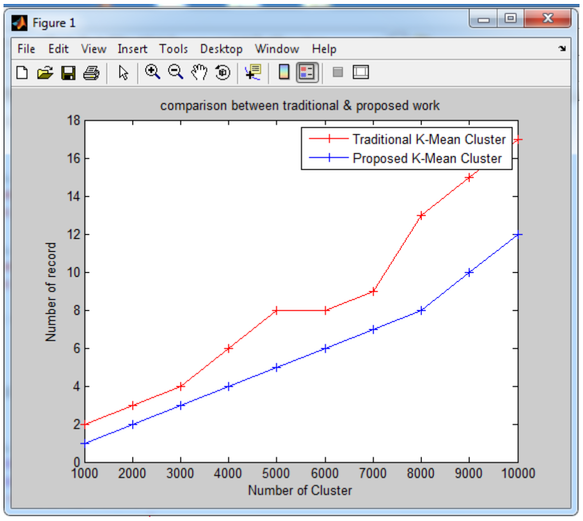

Fig 3 Analysis of Existing & Proposed cluster

Existing Total Size & Proposed Total Size

Number of record	Old Total Size	New Total Size
1000	1220	1123
2000	1843	1750
3000	2490	2276
4000	4945	4760
5000	6734	6593
6000	7554	7345
7000	8454	8322
8000	12344	12222

9000	13454	12954
10000	15667	14322

Table 2 Existing Total Size & Proposed Total Size

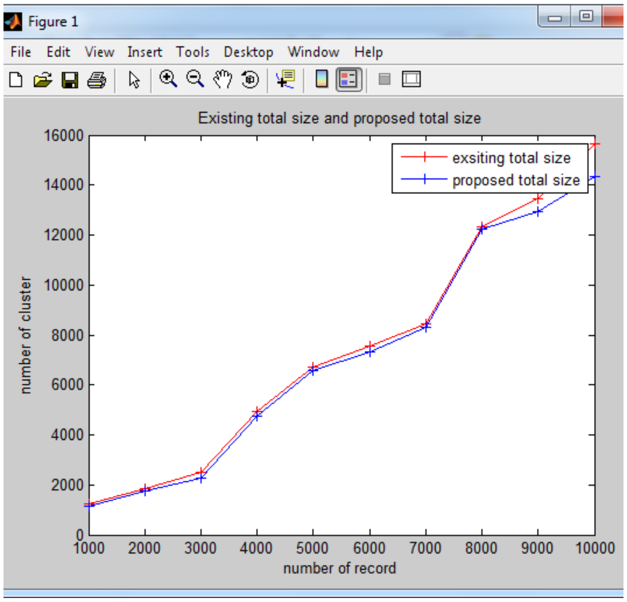

Fig 4 comparative analysis Existing Total Size & Proposed Total Size

Comparative analysis of result between old & enhanced K-MEAN

Number of record	Old K-Mean Algorithm	Enhanced Algorithm
1000	2	1
2000	3	2
3000	5	3
4000	6	4
5000	8	5
6000	8	6
7000	9	7

8000	12	8
9000	14	9
10000	14	10

Table 3 Comparative analysis of result between old & enhanced K-MEAN

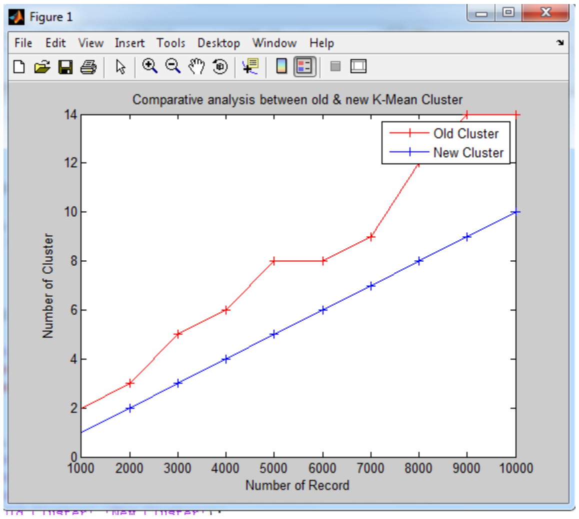

Fig 5 Analysis of old & new cluster

Above figure represent comparative analysis of number of clusters formed in case of old K mean clustering & enhanced K mean clustering. Number of vacant clusters has been removed in case of enhanced clustering algorithm so number of clusters get reduced in case of enhanced algorithm.

5. CONCLUSION

Clustering is process of grouping objects that belongs to same class. Similar objects are grouped in one cluster & dissimilar objects are grouped in another cluster. We have explain comparative analysis of number of clusters formed in case of existing K mean clustering & proposed K mean clustering. Number of vacant clusters had been removed in case of proposed clustering algorithm so number of clusters get reduced in case of proposed algorithm.

REFERENCES

1. Hong Liu 1,&Xiaohong Yu (2009), “Application Research of Clustering Algorithm in Image Retrieval System”.International Journal of Database Theory & Application.
2. Dr. Yashpal singh, 2alok singh chauhan in (2009), “neural networks in data mining” International Journal of Research, Mar (2009).
3. Jiawei Han & Jing on (2011) “Research Challenges for Data Mining in Science & Engineering” International Journal of Scientific&Research Publications ,March (2011).
4. Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur (2012), “Efficient Clustering Algorithm Using Ranking Method In Data Mining” ” International Arab Journal of Information Technology, July (2012).
5. Manjot Kaur * Navjot Kaur (2012), “Web Document Clustering Approaches Using Algorithm” International Journal Of Engineering & Science Issn.
6. Prabin Lama (2013), “Clustering system based on text mining using density algorithm” International Journal of Computer application, May-(2012).
7. Farhat Roohi in (2013), “Artificial Neural Network Approach to Clustering” IADIS international Journal of computer science & IT.
8. Waldemar Wójcik & Konrad Gromaszek in (2014)“Data Mining Industrial Applications” (Lublin University of Technology, Poland)
9. S.V.S. Ganga Devi(2014), “A Survey On Distributing Data Mining & Its Trends” International Journal of Research, Mar (2014)
10. Srimathi, M. Subaji, AnuSoosan Babyand Deepu Raveendran wrote a review A Study On Distributed Data Mining Frame Work Sarpn Journal of Engineering & Applied Sciences, december (2014)
11. Josenildo C. da Silva, Chris Giannella, Ruchita Bhargavan, Hillol Kargupta, & Matthias Klusch wrote a review on “Distributed Data Mining & Agents” Journal of Engineering & Applied Sciences, DECEMBER (2014)
12. Vuda Sreenivasa Rao, S Vidyavathi&G, Ramaswamy (2010) Distributed Data Mining& Agentmining Intreaction & Intrgration: A Novel approach by IJRRAS September (2010)
13. Vuda Sreenivasa Rao(2015), “Multi Agent-Based Distributed Data Mining: An Over view” International Journal of Reviews in Computing ,March (2015)
14. TrilokNathPandey, Niranjan Pandaand Pravat Kumar Sahu wrote a review on Improving performance of distributed data mining (DDM)with multi-agent system by International Journal of Computer Science Issues, March (2012)