International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

A Study of K-Means and Cure Clustering Algorithms

Sandra Sagaya Mary. D. A
Department of Computer Science
Dr.SNS Rajalakshmi College of Arts & Science
Coimbatore, India.

Tamil Selvi. R
Department of Computer Science
Dr.SNS Rajalakshmi College of Arts & Science
Coimbatore, India.

***Abstract*---Clustering is an important research topic and widely used unsupervised classification application in data mining. Clustering is technique which divides a data into meaningful groups. It aims to minimize intra-class similarity while to maximize interclass dissimilarity. Data clustering is under vigorous development. We can also say that, clustering is the process of organizing objects into groups whose members are similar in some way and the dissimilar objects are grouped into other clusters. The main goal of clustering is to reduce the amount of input data by grouping similar data items together. There are many clustering methods available and each of them may give different grouping of data sets. The widely used methods are partitioning clustering and hierarchical clustering. In this paper, the popular K-Means and CURE clustering algorithms are discussed with their merits and demerits.**

Keywords*---*K-means algorithm, Cure algorithm, Merits, Demerits and Comparison.

I. INTRODUCTION

A cluster is an ordered list of objects, which have some common characteristics. Clustering is a main task in exploring data mining and a common technique for statistical data analysis extensively used in many fields including artificial intelligence, libraries, insurance, city planning, earthquake studies, psychiatry, information retrieval, biology and marketing. Data clustering is under vigorous development. Clustering is an iterative task in knowledge discovery and interactive muti-objective process that involves trial and failure. Dissimilarities and similarities are done based on the attribute values describing the objects. Many clustering algorithms have been developed such as partition based, hierarchical based, grid based, density based, model based, constraint based and methods for high dimensional data. Among them, partition based clustering and hierarchical based clustering are the traditional and most widely used methods. We have taken the popular k-means in partition based method and CURE in hierarchical based method. **[1]** k-means is one of the simplest unsupervised clustering algorithms that is used fort well known clustering problem. **[2]** CURE (Clustering Using REpresentatives) is an efficient data clustering algorithm for large databases that is more robust to outliers and identifies clusters having non-spherical shapes, size and densities.

II. K-MEANS CLUSTERING ALGORITHM

The k-mean is a numerical, unsupervised, iterative and evolutionary algorithm that had its name from the operation method. The k-means clustering aims to find the positions of the clusters that minimize the distance from the data points to the cluster. This algorithm partition the n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation. A more efficient version was proposed and published in Fortran by Hartigan and Wong in 1975/1979. This clustering algorithm also referred to as **Lloyd's algorithm**. K-means algorithm is easy to implement and apply even on large data sets. K-means clustering algorithm is simply described as follows:

Step 1: First create k initial clusters by choosing k number of data randomly.

Step 2: Calculate the arithmetic mean for each cluster formed.

Step 3: Each record is assigned to nearest cluster by finding the similarity between the points using the formula,

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} (x_{i1} - m_{j1})}$$

Step 4: Reassign each record to most similar cluster and recalculate the arithmetic mean of all clusters in the dataset.

Step 5: The process continues from step 3 until no data point was reassigned and k-means procedure is complete.

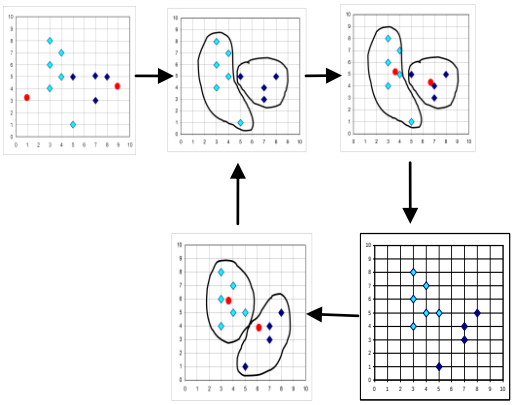


Fig 1. K-Means Process

A. *Advantages*

- K-means is computationally faster than hierarchical clustering, if we keep k smalls.
- Time complexity of k–means is O(tKn),

  where n- number of data points

  k- number of clusters

  t- number of iterations.
- This algorithm is suitable of very large data sets.

B. *Disadvantages*

- It does not work well with clusters of Different size and Different density.
- This algorithm does not provide the same result with each run.
- Euclidean distance measures can unequally weight underlying factors.
- This algorithm fails for categorical data and non-linear data set.
- Difficult to handle noisy data and outliers.

C. *Limitations of k-means and their overcomes*

- One of the problem with k-means is that empty clusters that is obtained if no points are allocated to cluster during assignment. This can be overcome by choosing the replacement centroid from the cluster that has highest SSE.
- K-means is very sensitive to outliers. This can be overcome by discovering outliers and eliminating them beforehand.
- The other limitation is, increased SSE makes algorithm difficult. To overcome this, we should reduce SSE either by splitting clusters or by introducing a new cluster centroid.

III. CURE CLUSTERING ALGORITHM

CURE is a hierarchical clustering algorithm for large datasets proposed by Guha, Rastogi and Shim in 1998. This algorithm is agglomerative hierarchical approach. Cure combines centroid and single linkage approaches by choosing more than one representative points from each cluster. At the end of each step, the clusters with the closest representative points are clustered together. Cure represents each cluster by a fixed number of points that are generated by selecting well scattered points from the cluster, then shrink them toward the center of the cluster by a specified faction. This enables CURE to correctly identify the clusters and makes it less sensitive to outliers. We cannot apply this algorithm directly to large datasets, instead we have to apply random sampling, partitioning for speedup – the advantage of partitioning the inputs is to reduce the execution time, labeling on disk.

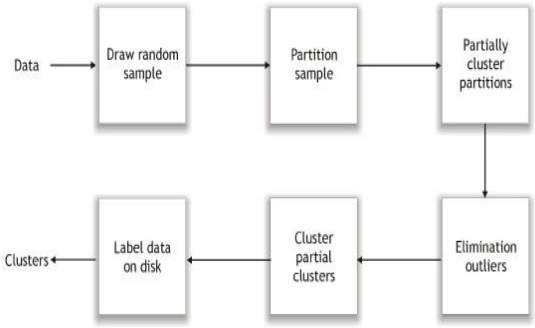


Fig 2. Architecture of Cure

A. *Cure Procedure*

Step 1: Draw a random sample s.

When all dataset is considered as input of algorithm, execution time could be high due to cost. For that, random sampling is done to lower the execution time.

Step 2: Partition sample to p partitions with size s/p

Step 3: Partially cluster partitions into s/pq clusters

Step 4: Cluster partial clusters, shrinking representatives towards centroid

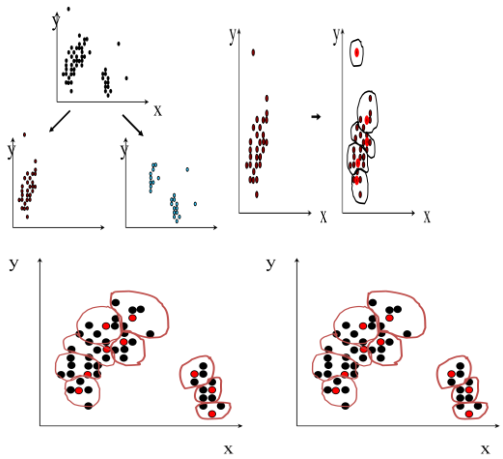
Random sampling is useless when the clusters are less dense. So, partition data points into p partitions. Partially cluster each partition until the final number of cluster created reduces to s/pq with q>1.

Step 5: Label data on disk.

In initial clustering, many data points are excluded. The data points that are not included in the first phase are added to the clusters whose representative point is closer.

For Example, Let us take,

s=50; p=2; s/p=25; s/pq=5



A. *Advantages*

- CURE can identify non-spherical shaped clusters and wide variances in size with the help of well scattered representative points and centroid shrinking.
- CURE can handle large databases by combining random sampling and partitioning method.
- CURE is robust to outliers
- The time complexity is O (n^2 log n) and space complexity is O (n).

B. *Disadvantage*

- CURE fails to explain the inter-connectivity of objects in clusters.

IV. COMPARISON

Properties	K-Means	CURE
Model	Static	Static
Sensitivity to outliers	More sensitive	Handles outliers effectively
Time Complexity	O(tKn)	O(n^2log n)
Shapes	Supports spherical shapes	Supports non-spherical shapes, densities.

Fig 3. Comparative Study

CONCLUSION

From the comparative study of k-means and cure algorithm, k-means deals with the data points of spherical datasets in 2D data sets, whereas the cure algorithm with random sampling and partitioning deals with outliers of non-spherical shapes and with large data sets. But by overcoming the limitations of k-means, this algorithm would be the fast and easy algorithm.

REFERENCES

[1] Bruce Moxon "Defining Data Mining, The Hows and Whys of Data Mining, and How It Differs From Other Analytical Techniques" Online Addition of DBMS Data Warehouse Supplement, August 1996.

[2] Guha, R. Rastogi, and K. Shim, CURE: An Efficient Clustering Algorithm for Large Databases, ACM 0-89791~996.5/98/006, 1998.

[3] G.G. Ma and J. Wu, Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Applied Probability, 2007.

[4] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi 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.