# A SURVEY OF CLUSTERING TECHNIQUES

**Sagar K.Naik[1], Gajanan Gawde[2]**

[1]*Computer Engineering Department, Goa College of Engineering, Farmagudi-Ponda, Goa, India*
[2]*Computer Engineering Department, Goa College of Engineering, Farmagudi-Ponda, Goa, India*

## Abstract

*Data mining involves working with very huge amount of data. This data always contains some natural groups of data points wherein the data points within these groups (clusters) are identical to each other. Finding these set of clusters is an unsupervised approach and various clustering methods exists to accomplish this task. Density based, partitioning based, hierarchical and grid based are some of the common clustering methods which groups the similar data in the clusters with common goal of increasing the intra cluster similarity and decreasing the inter cluster similarity. Each of these methods contains several algorithms which are briefly discussed in this review paper.*

***Keywords:*** *hierarchical,partitional,grid,density based clustering,spatial data;*

----------***----------

## 1. INTRODUCTION

Analyzing the large data and deriving useful patterns from it forms an integral goal of data mining. Classification, clustering, predictive data analysis etc. forms important aspects of data mining. Clustering aims at grouping the similar data into groups (clusters) such that inter cluster similarity is low and intra cluster similarity is high i.e. the data points within a cluster are more similar to each other as compared to points from other clusters. In order to find the similarity between the data objects, several similarity measures [1], [2] exists which can find the similarity/distance between the data points.

Euclidian, Mahalanobis, Manhattan/city block etc. are some of the distance measures that can be used in case of numerical/spatial data. In order to get the similarity between document data objects, Cosine, Sine, Jaccard's similarity measures are generally used. Given the data to be clustered, one does not know the total number of natural clusters in the data. Hence, the number of classes present in the data is unknown and therefore clustering the data becomes unsupervised approach of data mining. If we consider the machine learning perspective, then we can say that patterns that are hidden in the data correspond to clusters. Searching those hidden patterns forms unsupervised learning leading to data concept representation by the resulting system. So, we can conclude that clustering is an unsupervised learning of hidden data concept. There exist some of the important requirements that a clustering algorithm need to satisfy. The algorithm should be able to detect outliers/noise present in the data. Time complexity should not be on the higher side. The algorithm should be scalable and data order independent. Varying sized and shaped clusters must be detected by the algorithm along with detection of varied density clusters. The number of input parameters to the algorithm should be as low as possible since the user has very less information regarding the data under consideration. Each of the existing clustering algorithms satisfies only some of the above mentioned requirements and designing an algorithm which satisfies all the requirements is little difficult task. This paper gives the brief idea of most of the well-known clustering algorithms along with their comparison with respect to several issues of clustering.

## 2. CLUSTERING METHODS

Wide variety of clustering algorithms [1], [2] exists which are broadly classified into following groups.

They are hierarchical, partitioning, grid based, density based methods.

### 2.1 Hierarchical Clustering

Hierarchical algorithms build the clusters hierarchy in the form of tree also known as dendrogram. Every cluster may have a child cluster, sibling cluster which partitions the data as per the algorithm. Agglomerative (bottom-up) and divisive (top-down) are two forms of hierarchical clustering. Again, these algorithms can be single, complete or average link based on distance calculation between the clusters. CURE, DIANA, AGNES etc. are the hierarchical clustering algorithms.

### 2.2 Partitioning Based Clustering

These algorithms partition the data space and relocate the data points from one cluster to other in order to improve the cluster quality through iterative optimizations. After creating the initial partitions, these algorithms iteratively refines the quality of clusters to maximize intra-cluster similarity and minimize inter cluster similarity. K-means, K-medoids, CLARA, CLARANS etc. algorithms falls under this category.

### 2.3 Density Based Clustering

Density based algorithms detects the regions of high density points as the clusters which are surrounded by regions of

low density points. These algorithms mark points as core points, border points or outlier points. DBSCAN and OPTICS are some examples. These algorithms works suitably well only if the data points are uniformly distributed.

### 2.4 Grid Based Clustering

The data space is divided into the finite number of cells to form a grid like structure. Then, clustering is performed on these cells instead of entire data set. CLIQUE and STING are well known grid based clustering algorithms.

## 3. CLUSTERING ALGORITHMS

### 3.1 K-means Algorithm

K-means algorithm [2], [3] is a partitioning method which partitions the data into K disjoint sets/clusters by taking number of clusters (K) as input parameter from user. K number of initial centroids is selected and data points are iteratively relocated amongst these centroids. After each iteration, mean of each cluster forms its new centroid. Process continues until centroid values does not change from one iteration to next.

Algorithm can be stated as:
Input: K (number of clusters).
Output: Set of K clusters along with their centroids.
Method:
1. Randomly select K initial centroids.
2. Assign each data point to its closest centroid.
3. Recompute new centroid of each cluster by taking the mean value of that cluster.
4. Go to step 2 until centroid values do not change.

#### Advantages

1. Requires just single input parameter i.e. K, number of clusters.
2. The algorithm takes linear time i.e. O(nKt) where n is total number of points and t is number of iterations taken.
3. It is order independent.

#### Disadvantages

1. It cannot detect outliers and is very sensitive to outliers because mean values can be greatly affected by these outliers.
2. Not suitable to detect clusters with non-convex shapes.
3. Selection of initial centroids may affect number of iterations/convergence criteria.
4. Generally used only in case of numerical data and is not suitable for non-numerical data.

### 3.2 K-medoids Algorithm

Working procedure of K-medoids algorithm [4], [5] is almost very similar to K-means algorithm. But in case of K-medoids, each cluster is not represented by its mean. Each cluster is represented by one of the representative point from that cluster itself. It is called as medoid. After each iteration, medoid of cluster gets changed. Selection of new medoid is done based on swapping cost criteria [1], [2]. This algorithm is also known as PAM (Partitioning Around Medoids)

PAM Algorithm can be stated as:
Input: K (number of clusters).
Output: Set of K clusters along with their medoids.
Method:
1. Randomly select K initial medoids.
2. Assign each data point to its nearest medoid.
3. Recompute new medoid based on swapping cost criteria.
4. Go to step 2 until medoids do not change.

#### Advantages

1. Needs just single input parameter (K).
2. Less sensitive to outlier points as compared to K-means.

#### Disadvantages

1. Finding new medoids takes much computation.
2. Cannot detect outlier points.
3. Time complexity is on the higher side i.e. $O(tK(n-K)^2)$ where t is number of iterations and n is total number of data points.

### 3.3 CLARA Algorithm

To deal with large datasets, CLARA algorithm (Clustering LARge Applications) was proposed [6]. This algorithm takes random sample from the dataset and applies PAM algorithm on that sample by selecting the medoids from that sample only instead of selecting medoids from entire dataset. The sample selected should be random as much as possible. Multiple samples are tried and the best clustering is given as output.

Algorithm can be stated as:
Input: K (number of clusters).
Output: Set of K clusters.
Method:
1. For i=1 to number of iterations do,
2. Randomly select a sample of 40+2K points from the entire dataset. Apply PAM algorithm on that sample and get the K medoids.
3. For each data point P from the dataset, find its closest medoid.
4. Compute the average dissimilarity of clustering and if it is less than current minimum, then set it as current minimum. Medoids obtained are saved as best medoids obtained so far.
5. Go to step 1 to run next iteration.

#### Advantages

1. CLARA performs satisfactorily well for large datasets as compared to PAM.
2. Time efficient than PAM algorithm.

### Disadvantages

1. Best K medoids might not be selected during sampling process, and hence CLARA might not find best clustering.
2. It might not give good clustering if sampling is biased.

### 3.4 DBSCAN Algorithm

DBSCAN (Density Based Spatial Clustering of Applications with Noise) uses concept of density to detect the clusters [9], [10]. It involves use of two parameters Epsilon and MinPoints. Starting with arbitrary point P, within a distance of Epsilon, if there are at least 'MinPoints' number of points, then P forms a core point and cluster is formed with point P. Using the concepts of density connectivity,density reachable and directly density reachable [2], [9], various clusters present are detected along with outlier points.

Algorithm can be stated as:
Input: Epsilon,MinPoints.
Output: Set of clusters along with outliers.
Method:
1. Arbitrary select a point P.
2. If P is not a core point, and if no points are density-reachable from P then visit the next point.
3. If P is a core point, a cluster is formed with all points density-reachable from P with respect to Epsilon and MinPoints.
4. Continue the process until all of the points have been processed.

### Advantages

1. Can detect the outlier points.
2. Can detect varied shaped and sized clusters.
3. Algorithm is order independent.

### Disadvantages

1. Time complexity is $O(n^2)$ but it can be reduced to O(nlogn) using Kd tree.
2. It cannot detect clusters of varying densities.
3. Works well only if dataset is uniform.

### 3.5 CURE algorithm

CURE (Clustering Using REpresentatives) algorithm [1], [2], [10] is an agglomerative hierarchical method in which it starts with each point as a separate cluster and merges the closest pair of clusters at each successive iteration. Some C numbers of representative points from each cluster are identified. This is done by selecting C well scattered points from cluster and then shrinking them towards mean of the cluster. Distance between two clusters is the distance between closest representative points from both clusters.

The algorithm can be stated as:
Input: α (shrinking factor), C (number of representative points), s (sample points), p (partition number).
Output: Set of clusters.
Method:
1. Draw a random sample s.
2. Partition the sample to p partitions with size s/p.
3. Partially cluster the partitions into s/pq clusters.
4. Select C well scattered sample points from the cluster.
5. Shrunk the chosen scattered points toward the centroid.
6. These points are used as representative of clusters.
7. After each merging, C sample points will be selected from original representative of previous clusters to represent new cluster.
8. Cluster merging will be stopped until target K clusters are found.

### Advantages

1. It can detect arbitrary shaped cluster and can detect outlier points.
2. Computation time and memory loading can be reduced with random sampling and two pass clustering.

### Disadvantages

1. Time complexity is $O(n^2 \log n)$.
2. Involves too many parameters and is very sensitive to those parameters.

Table I provides brief summary [7], [11], [12], [13] of some clustering algorithms wherein the algorithms are compared with respect to some of the requirements of algorithm. Considering all the properties of the algorithms, we can conclude that DBSCAN algorithm can be considered as good algorithm for performing clustering task.

**Table 1:** Brief comparison of clustering algorithms

| **Algorithm** | **Input Parameters** | **Time & Space Complexity** | **Is Order Independent?** | **Outlier Detected?** | **Can tackle high dimensional data??** | **Shape of clusters detected** |
|---|---|---|---|---|---|---|
| K-means | K | O(nKt) time, O(n+K) space | Yes | No | No | Spherical |
| K-medoids | K | $O(tK(n-K)^2)$ time , O(n+K) space | Yes | No | No | Spherical |
| CLARA | K | $O(KS^2+K(n-K))$ time, O(n) space | Yes | No | No | Nearly Spherical |
| DBSCAN | Epsilon, MinPoints | O(nlogn) time, Matrix based & non matrix | Yes | Yes | No | Arbitrary |

| Algorithm | Input Parameters | Time & Space Complexity | Is Order Independent? | Outlier Detected? | Can tackle high dimensional data?? | Shape of clusters detected |
|---|---|---|---|---|---|---|
| | | based $O(n^2)$ & O(n) resp. space | | | | |
| CURE | α,C, s,p | $O(n^2 \log n)$ | Yes | Yes | Yes | Arbitrary |

## 4. CONCLUSION

Clustering task of data mining domain usually deals with finding the clusters present in the dataset. Various clustering methods and the algorithms exist in the literature some of which are briefly discussed in this review paper. Hierarchical, partitioning, grid and density based are the methods of clustering each of which involves many algorithms under that method. K-means, K-medoids (PAM), CLARA, DBSCAN and CURE algorithms are studied and briefly explained in this paper. Brief comparison of these algorithms is also provided with respect to some of the prime requirements/issues of a clustering algorithm.

## REFERENCES


[1] A.K.Jain, M.N.Murty , P.J.Flynn, "Data Clustering:A Review" , ACM Computing Surveys,Vol.3,September 1999.

[2] Pang-Ning Tan ,Michael Steinbach, Vipin Kumar "Introduction to Data Mining".

[3] Jyoti Yadav,Monika Sharma, "A Review of K-means Algorithm",International Journal of Engineering Trends & Technology (ITETT)-Volume 4 Issue 7-July 2013.

[4] Ji Wentian,Guo Qingju , Zhong Sheng, "Improved K-medoids Clustering Algorithm Under Semantic Web",Proceedings of The 2nd International Conference on Computer Science and Electronics Engineering(ICCSEE 2013).

[5] Shalini S.Singh, N C Chauhan, "K-means v/s K-medoids:A Comparative Study",National Conference on Recent Trends in Engineering and Technology.

[6] Raymond T.Ng,Jiawei Han, "CLARANS:A Method for Clustering Objects for Spatial Data Mining".

[7] Rui Xu,Donald Wunsch, "Survey of Clustering Algorithms",IEEE Transactions On Neural Networks,Vol.16,No.3,May 2005.

[8] K.Sasirekha,P.Baby, "Agglomerative Hierarchical Clustering Algorithm-A Review", International Journal of Scientific & Research Publications,Volume 3,Issue 3,March 2013.ISSN 2250-3153.

[9] Martin Ester, Hans-Peter Kriegal,Jorg Sander,Xiaowei Xu, "A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).

[10] Pooja Batra Nagpal,Priyanka Ahlawat Mann, "Comparative Study of Density based Clustering Algorithms",International Journal of Computer Applications(0975-8887),Volume 27 No.11,August 2011.

[11] Sudipto Guha,Rajeev Rastogi, Kyuseok Shim, "CURE:An Efficient Clustering Algorithm for Large Databases".

[12] Neha Soni,Amit Gonatra, "Categorization of Several Clustering Algorithms from Different Perspective:A Review",International Journal of Advanced Research in Computer Science & Software Engineering,Volume 2,Issue 8,August 2012,ISSN:2277 128X.

[13] Amandeep Kaur Mann,Navneet Kaur, "Review Paper on Clustering Techniques", Global Journal of Computer Science & Technology, Software & Data Engineering,Volume 13, Issue 5 Version 1.0 Year 2013,Print ISSN:0975-4350.

[14] P.IndiraPriya,Dr.D.K.Ghosh, "A Survey on Different Clustering Algorithms in Data Mining Technologies", International Journal of Modern Engineering Research(IJMER),Vol.3,Issue 1,Jan-Feb.2013 pp-267-274,ISSN 2249-6645.