A Review: Comparative Study of Image Clustering Algorithms

Jashanpreet Kaur[#1], Gaurav Deep[*2]

[#]*University College of Engineering, Punjabi University*
Patiala, Punjab, India
[1]jashan.31290@gmail.com

[*] *Assistant Professor, Department of Computer Engineering, Punjabi University*
Patiala, Punjab, India
[2]deepgaurav48@gmail.com

Abstract— **An image is a picture that can be captured and stored in computer. Clustering is a process of putting data into groups of similar objects. In an image, each group, known as a cluster, having similar objects within the cluster and dissimilar with the objects in other clusters. This paper reviews different data clustering algorithms based on different parameters. These algorithms are: k- means algorithm, hierarchical clustering and density based clustering algorithms.**

Keywords— ***K-Means Clustering, Hierarchical Clustering, DBSCAN Clustering, Vector Graphics.***

I. INTRODUCTION

An image is a visual representation of a person, animal , or thing , photographed , painted,or otherwise made visible. Thus an image is a picture that has been created or copied and stored in an electronic form. An image can be represented in terms of vector graphics. To represent an image in a computer, vector graphics uses geometrical primitives such as points, lines, curves and shapes. Vector graphics are based on vectors which lead through locations called control points. An image stored in graphic form is sometimes called a bitmap . An image is a section of random access memory that has been copied to another memory or storage location for processing[2].

II. CLUSTERING

Clustering is the process of organizing the objects in such a way that objects within the cluster are similar to each other and dissimilar to other objects. Clustering can also be viewed as a special type of classification. In classification, we have a set of predefined classes and want to know in which class a new object belongs to, whereas clustering tries to group a set of objects and find whether there is some relationship between the objects. In relation to machine learning, classification is supervised learning and clustering is unsupervised learning. The clusters formed as a result of clustering can be defined as a set of like objects. But the objects from different clusters are not alike[1][3].

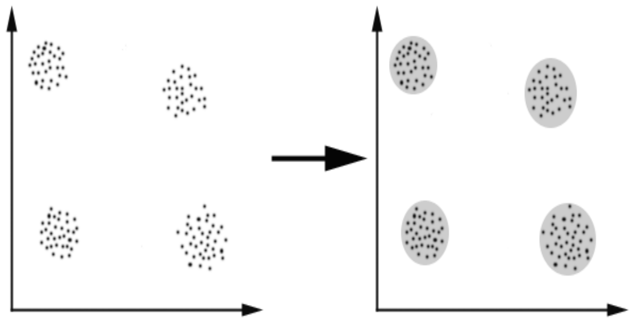

Fig. 1 Clustering on data objects

In fig 1, we easily identify the 4 clusters into which the data can be divided, the similarity criteria is distance.

A. *Role of Clustering*

Clustering is performed so as to achieve the following:

- Scalability;
- to deal with different attributes;
- to discover clusters having arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitive to order of input parameters;
- high dimensions;
- interpretability and usability.

III. CLUSTERING ALGORITHMS

Clustering is a technique of grouping data objects into multiple groups or clusters so that objects within the cluster have high similarity.Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc. Many clustering algorithms have been developed, each of which uses a different induction principle. Clustering algorithms are categorized from several aspects such as partitioning algorithms, hierarchical algorithms and density-based algorithms[6].

.

A. *Partitioning Algorithm*

The partitioning algorithm generally result in a set of N clusters, each object belongs to one cluster. Each cluster is expressed by a centroid or a cluster representative[7]. If there are large number of clusters, then centroids must be further clustered to produce hierarchy within a dataset.

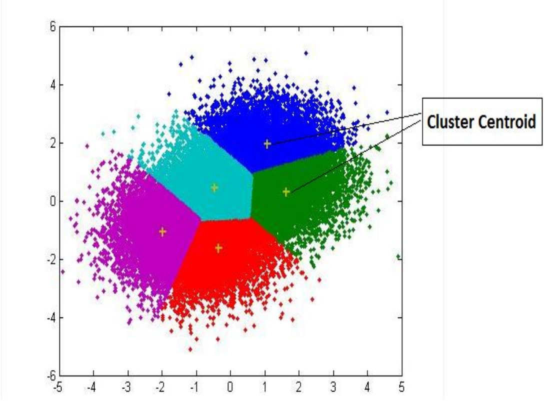

Fig. 2 Cluster representative of clusters

1) K-means clustering algorithm:

K-means is one of the simplest unsupervised learning algorithms in which each point is assigned to only one particular cluster. It generates a specific number of disjoint, flat(non-hierarchical) clusters. K-Means is an iterative clustering algorithm [8] [9] in which items are moved among set of clusters until the desired set is reached.

A K-means algorithm runs in the following steps:

Step 1: Initially, clusters are chosen at random. These represent the “temporary” means of the clusters.

Step 2: The Euclidean distance from each object to each cluster is computed, and each object is assigned to the closest cluster.

Step 3: For each cluster, the new centroid is computed – and each mean value is now replaced by the respective cluster centroid.

Step 4: The Euclidean distance from an object to each cluster is computed, and the object is assigned to the cluster with the smallest Euclidean distance.

Step 5: The centroids of clusters are recalculated based on the assignment of clusters.

Step 6: Steps 4 and 5 are repeated until no object moves from cluster to another.

B. *Hierarchical Algorithm*

A hierarchical algorithm creates a hierarchical decomposition of the given set of data objects. A tree of clusters called as dendrogram is built. The tree is not a single cluster, but rather a multilevel grouping, where clusters have sub-clusters, which in turn have sub-clusters. The hierarchical algorithms produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains[10][12].

A Hierarchica algorithm consists of the following steps:

Step 1: Initially assign each object into a cluster such that if we have N objects then we have N clusters.

Step 2: Find closest pair of clusters and merge them into single cluster.

Step 3: Compute distance between newly formed cluster and each of existing clusters.

Step 4: Repeat steps 2 and 3 until all objects are clustered into no. of clusters.

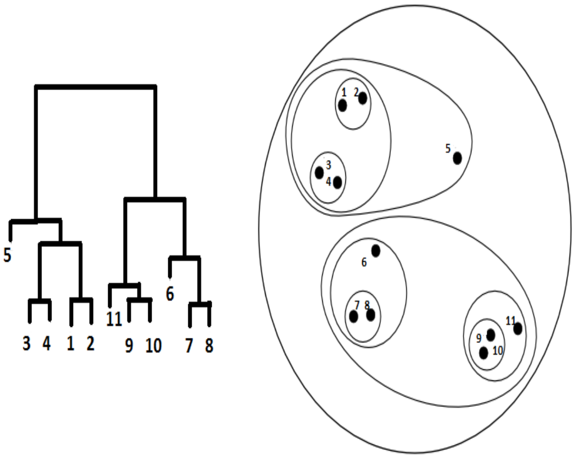

Fig. 3 Hierarchical Clustering

In fig 3. Each object is assigned into its own cluster. Now, the objects which are closest to each other are merged into a single cluster like 1,2 and 3,4 are merged and so on. A dendogram is built according to the distance between the clusters. The pairs of objects which are closest to each other (1,2) (3,4) (9,10) (7,8) are maintained at leaf nodes and further merging is done on the criteria of distance.

It can be subdivided into following:

1) Agglomerative hierarchical clustering:

It is a bottom-up clustering where clusters have sub-clusters, which in turn have sub-clusters, etc. It starts from an object in its own cluster and iteratively merges cluster into bigger clusters, until all the objects are in a distinct cluster or certain termination condition is satisfied[11]. For the merging, it finds the two clusters that are closest to each other on the basis of distance and combines the two to form one cluster.

2) Divisive hierarchical clustering:

A top-down clustering method and is less commonly used. It works in a similar manner like agglomerative clustering but in the opposite direction. This method starts with a one cluster containing all objects, and then divide the clusters in a sequence until only clusters of individual objects remain [13].

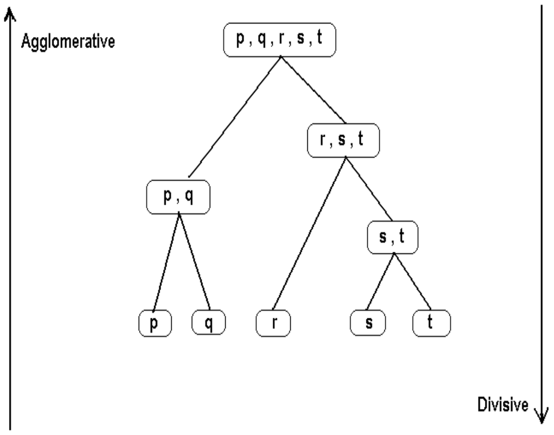

Fig 4.Agglomerative and Divisive Hierarchical Clustering

C. Density Based Algorithm

Density based clustering algorithm plays a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the idea of density reachability and density connectivity.

Density Reachability: A point 'p' is said to be density reachable from a point 'q' if point 'p' is within ε distance from point 'q' and 'q' has sufficient number of points in its neighbour which are within distance ε.

Density Connectivity: A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbors and both the points "p" and "q" are within the ε distance.

1) DBSCAN:

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Clusters found by this algorithm can be any arbitrary shape, as opposite to k-means which assumes that clusters are convex shaped. The central component of this algorithm is the core samples, which are samples in the areas of high density. A cluster is thus a set of core samples, adjacent to each other and a set of non-core samples that are adjacent to a core samples[4][5].

A cluster is thus a set of core samples, built recursively by taking a core sample, finding all of the neighbors of the core samples, and their neighbors that are core samples, and so on. A cluster therefore also consists of a set of non-core samples, which are neighbors of a core sample in cluster but they cannot be considered as core samples.

The algorithm is as follows:
Let X = $\{x_1, x_2, x_3, ..., x_n\}$ be set of data points. Density based algorithm requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

Step 1: Start with an arbitrary starting point that has not been visited.

Step 2: Extract the neighborhood of this point using ε.

Step 3: If there are sufficient neighbors around the point then the clustering process starts and point is considered as visited else the point is labeled as noise.

Step 4: If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is known.

Step 5: A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

Step 6: This process continues until all points are marked as visited.

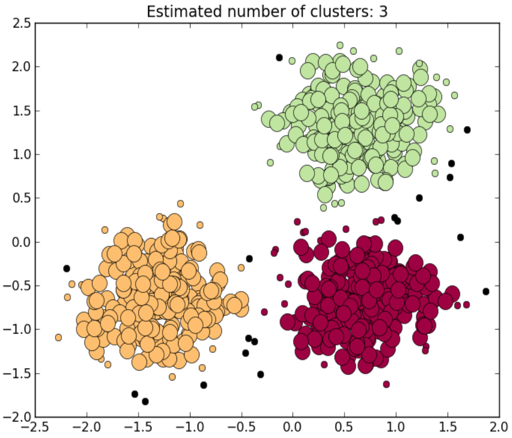

Fig. 5 Density based clustering

In fig 5, the color indicates cluster group, with large circles pointing core samples found by the algorithm. Small circles are non-core samples that are part of a cluster. Also, the outliers are indicated by black points.

D. Comparison of Clustering Algorithms

Parameters	**Hierarchical algorithm**	**Partioning algorithm**	**Density-based algorithm**
Prior Reqirements of algorithm	No prior information about the clusters is required	It requires the number of clusters in advance	It grows clusters on the basis of density of neighbours
Range of input numerical values	It is applicable to numeric values	Limited to numeric attributes	Applicable to numeric values
Applicable Time Complexity	Time complexity : O(n^2logn)	Linear time complexity : O(nktd)	Time complexity : O(n^2)
Efficiency	Less efficient; sometimes it is difficult to identify correct number of clusters by dendogram	More efficient; applicable only when mean is defined i.e fails for categorical data	Efficient for large spatial databases only
Condition of Termination	A termination condition has to be defined indicating when the process is to be terminated	In partiotining algorithm, no termination condition is specified	It continue growing the given cluster as long as the density in the neighborhood exceeds some threshold.
Noise sensitivity	Not sensitive to noisy data	Sensitive to noisy data,outliers and non-linear data	Robust towards noise detection
Modes of implementation	Easy to implement	Fast, robust and easy to understand	Fast for low dimensional data
Size of Dataset	Algorithm can never undo what was done	It gives best results on large datasets due to less computation time	Fails to identify clusters if density varies and if data set is too sparse
Cluster shape	Applicable to clusters of spherical shape	Applicable to clusters of spherical shape	Applicable to clusters having arbitrary shape
Verstality	Versatile	Not versatile	Less versatile
Number of Processors	Partitionable so that it can run on multiple processors	Partitionable into clusters	Not partitionable for multiprocessor system

E. Possible Applications

Clustering algorithms can be applied in many areas, for illustration:

- *Marketing:* For finding groups of customers having similar behavior from a large database of customer data containing their properties and past buying records;
- *Biology:* For assigning plants and animals according to their features;
- *Libraries:* For placement and ordering of books;
- *Insurance:* For identifying frauds and groups of motor insurance policy holders;
- *City-planning:* For identifying groups of houses according to their house type, cost and area;
- *WWW:* For document classification; clustering weblog data to discover groups of similar access patterns;
- *Image segmentation:* Clustering can be used to divide an image into different regions for border identification or object realization[3];
- *Software evolution:* Clustering is useful in software evolution as it helps to reduce legacy properties in code by reforming functionality that has become dispersed. It is a form of restructuring and hence is a way of directly preventative maintenance;

F. Issues in clustering

There are a number of problems with clustering. Some of them are given as:

- Current clustering techniques do not locate all the requirements in an appropriate manner;
- Due to high dimensionality and large number of datasets, there can be difficulty due to its time complexity;
- For distance based clustering, the efficiency of the method depends on the concept of distance
- There must be a definition of distance measure which is not common in multidimensional spaces;
- The output of the clustering algorithm can be explained in different ways.

IV. CONCLUSION

Among all the clustering algorithms defined above, k-means partitioning algorithm is popular due to:

- Linear time complexity: O(nktd) where n is objects in cluster, k is no of clusters, d is measurement of each object, and t is iterations.
- Its space complexity: O(k+m) It requires additional space to store the data objects.
- Performs best for large databases due to low computation time.
- Therefore Agglomerative and Divisive hierarchical algorithm was adopted for categorical data.

Hence performance of K- mean algorithm is better than Hierarchical Clustering Algorithm.

Density based methods such as DBSCAN are designed to find clusters of arbitrary shape. Density based algorithms doesnot require to specify the number of clusters in advance, as opposite to K-Means.

V. FUTURE SCOPE

This paper was intended to compare between clustering algorithms. As a future work, comparisons between the algorithms can be done based on other factors that are not included. Comparing between the results of algorithms using different data sets will give different results. Also, it will effect the performance of algorithm and enhance the results.

REFERENCES

[1] Chamkor singh, Gaurav Deep, “Cluster Based Image Steganography Using Pattern Matching”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 2, Issue 4, July – August 2013

[2] www.whatis.techtarget.com “Concept of Image”

[3] 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.

[4]www.pages.cs.wisc.cbir.pdf “Density based Algorithm”

[5] S. Thilagamani1 and N. Shanthi, ”A Survey on Image Segmentation Through Clustering”, International Journal of Research and Reviews in Information Sciences Vol. 1, No. 1, March 2011

[6] S. Revathi, Dr.T.Nalini, “ Performance Comparison of Various Clustering Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 2, Febraury 2013

[7] Periklis Andritsos,” Clustering Techniques”, March 11, 2002.

[8] Jain, A.K., Dubes, R.C., 1988. “Algorithms for Clustering Data”. Prentice-Hall Inc.

[9] Shi Na, L. Xumin, G. Yong, “Research on K-Means clustering alg orithm- An Improved K-Means Clustering Algorithm”, “IEEE Third International Symposium on Intelligent Information Technology and Security Informatics”, pp.63-67, Apr.2010.

[10] Aastha Joshi, Rajneet Kaur , “A Review: Comparative Study of Various Clustering Techniques in Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013

[11]Osama Abu Abbas, “Comparison between data clusterin algorithms”, The international Arab Journal of information technology, Vol.5, No. 3, July 2008

[12] Ying Zhao, George Karypis,” Comparison of Agglomerative and Partitional Document Clustering Algorithms”, University of Minnesota, Minneapolis, MN 55455

[13] N.F. Johnson and S. Jajodia, “Exploring Steganography: Seeing the Unseen,” *Computer,* vol. 31, no. 2, Feb. 1998, pp. 26-34.

.