CLUSTERING IN DATA MINING

By Anuradha Taluja

Clustering is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Clustering helps to splits data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for outlier detection.

For example, the data from customer base is divided into clusters; we can make an informed decision about who we think is best suited for this product. Suppose we are a market manager, and we have a new tempting product to sell. We are sure that the product would bring enormous profit, as long as it is sold to the right people. So, how can we tell who is best suited for the product from our company's huge customer base?

Fig

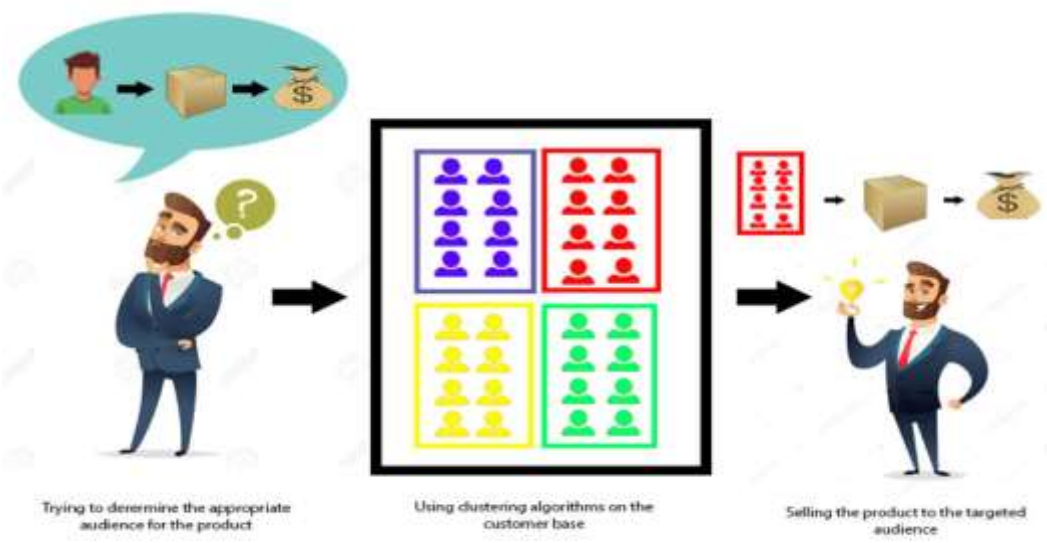

Figure 1: Application of Clustering Algorithm

- In machine learning, clustering is an example of unsupervised learning. Unlike classification, clustering and Unsupervised learning do not rely on predefined classes and class-labeled training examples. For this reason, clustering is a form of learning by observation, rather than learning by examples.

A Categorization of Major Clustering Methods

(1) Partitioning methods

It classifies the data into k groups, which together satisfy the following requirements:

- each group must contain at least one object, and
- each object must belong to exactly one group

- It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.
- The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects of different clusters are "far apart" or very different.
- Popular heuristic methods, such as

  (1) the **k-means algorithm**, where each cluster is represented by the mean value of the objects in the cluster, and

  (2) the **k-medoids algorithm**, where each cluster is represented by one of the objects located near the center of the cluster.

1 (a) Centroid-Based Technique: The k-Means Method

- The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low.
- Cluster similarity is measured in regard to the mean value of the objects in a cluster
- First, it randomly selects k of the objects, each of which initially represents a cluster mean or center.

  For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.

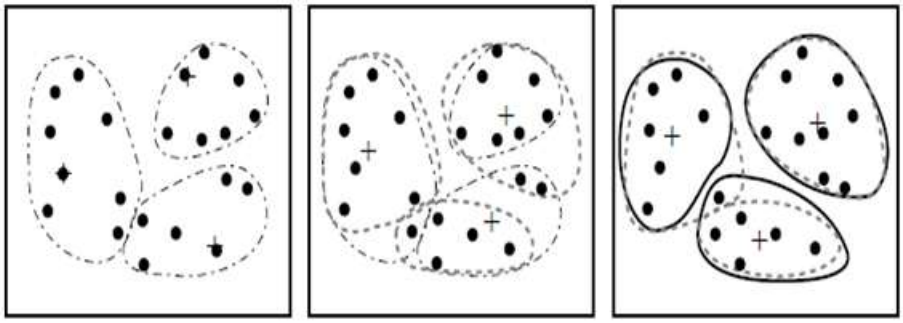

Figure 2: Clustering of set of objects based on k- means method

1(b) Representative Object-Based Technique: The k-Medoids Method

- The k-means algorithm is sensitive to outliers because an object with an extremely large value may substantially distort the distribution of data.
- Instead of taking the mean value of the objects in a cluster as a reference point, we can pick actual objects to represent the clusters, using one representative object per cluster.
- Each remaining object is clustered with the representative object to which it is the most similar.
- The partitioning method is then performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point .

- p is the point in space representing a given object in cluster Cj; and oj is the representative object of Cj. In general, the algorithm iterates until, eventually, each representative object is actually the metoid, or most centrally located object, of its cluster.
- Case 1: p currently belongs to representative object, oj. If oj is replaced by o random as a representative object and p is closest to one of the other representative objects, oi,
- i not equal j, then p is reassigned to oi.
- Case 2: p currently belongs to representative object, oj. If oj is replaced by o random as a representative object and p is closest to o random, then p is reassigned to o random.
- Case 3: p currently belongs to representative object, oi, i not equal j. If oj is replaced by o random as a representative object and p is still closest to oi, then the assignment does not change.
- Case 4: p currently belongs to representative object, oi, i not equal j. If oj is replaced by o random as a representative object and p is closest to o random, then p is reassigned to o random.

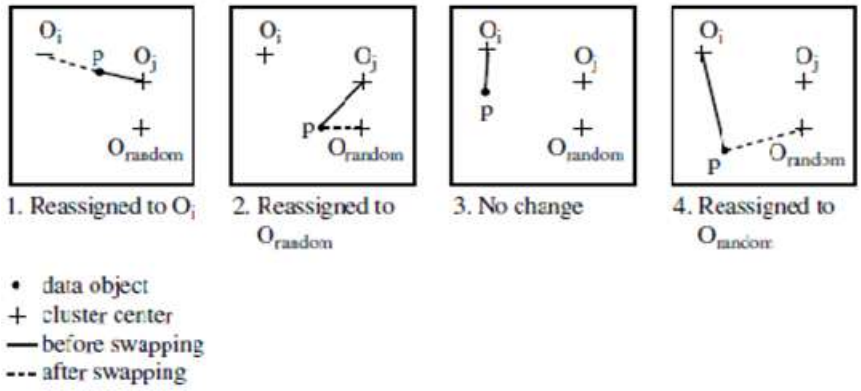

Figure 3: Four cases of the function k-medoids clustering

(2) Hierarchical methods

A hierarchical method creates a hierarchical decomposition of the given set of data objects.A hierarchical clustering method works by grouping data objects into a tree of clusters.

Classfication:

- Agglomerative & Divisive Hierarchical Clustering
- CURE
- Chameleon

There are two approaches to improving the quality of hierarchical clustering:

- perform careful analysis of object "linkages" at each hierarchical partitioning, such as in Chameleon, or
- integrate hierarchical agglomeration and other approaches by first using a hierarchical agglomerative algorithm to group objects into microclusters, and then performing macroclustering on the microclusters using another clustering method such as iterative relocation, as in BIRCH

(2a) **Agglomerative Hierarchical Clustering**

- A hierarchical method can be classified as being either agglomerative, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination conditions are satisfied.
- The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster.
- In each successive iteration, It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, each cluster is within a certain threshold.

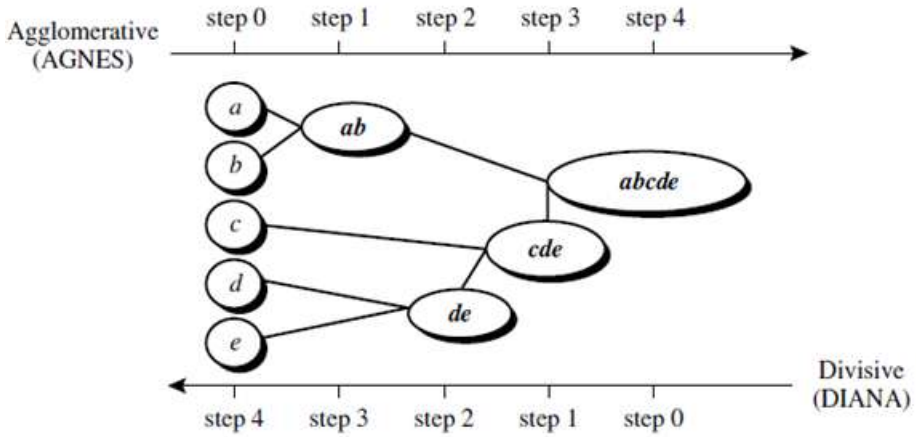

Figure 4: Agglomerative and Divisve Hierarchical clustering

EXAMPLE:

- The figure shows AGNES (AGglomerative NESting), an agglomerative hierarchical clustering method, and DIANA (DIvisive ANAlysis), a divisive hierarchical clustering method, to a data set of five objects,{**a, b, c, d, e }.**
- Initially, AGNES places each object into a cluster of its own.
- The clusters are then merged step-by-step according to some criterion.
- For example, clusters C1 and C2 may be merged if an object in C1 and an object in C2 form the minimum Euclidean distance between any two objects from different clusters.
- This is a single-linkage approach in that each cluster is represented by all of the objects in the cluster, and the similarity between two clusters is measured by the similarity of the closest pair of data points belonging to different clusters.
- The cluster merging process repeats until all of the objects are eventually merged to form one cluster.
- In DIANA, all of the objects are used to form one initial cluster.
- The cluster is split according to some principle, such as the maximum Euclidean distance between the closest neighboring objects in the cluster.
- The cluster splitting process repeats until, eventually, each new cluster contains only a single object

(2b) CURE

The cure algorithm assumes a Euclidean distance. It allows clusters to assume any shape.It uses collection of representative points to represent clusters

Figure 5: Clusters of different shapes

For example, the dataset of engineers and humanity people is been shown with their salary and age..

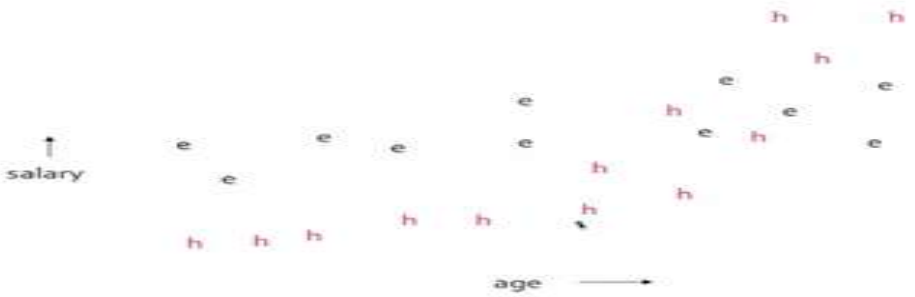

Figure 6: Dataset representation in terms of salary and age

We formed the two clusters of the dataset of engineers and humanities. The clusters formed are overlapping with each other which will not give the solution.

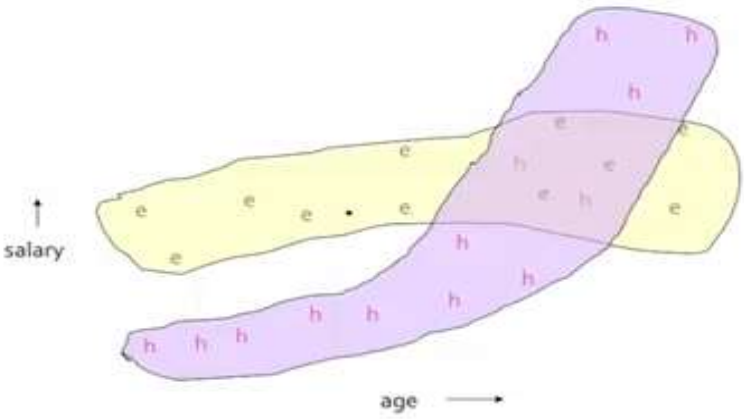

Figure 7: Two cluster formation

We tried to create three clusters for segregation by which results can be achieved. But after cluster formation still one cluster is formed having both the dataset values.

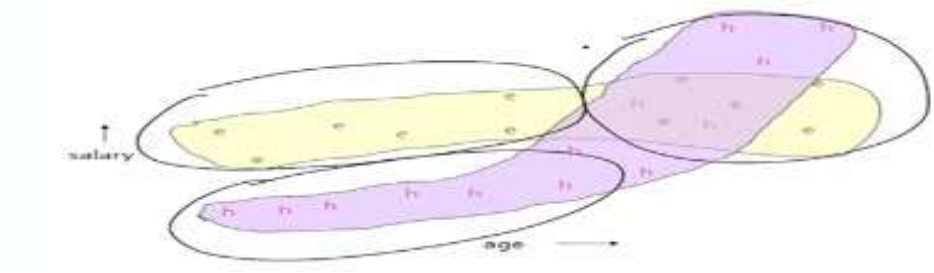

Figure 8: Three cluster formation

Algorithm for cure:

Pass 1 of 2:

- Pick a random sample of points that fit in main memory.
- Cluster sample points hierarchically to create the initial clusters.
- Pick representatives points:
    - For each cluster, pick k (eg.,4) representative points, as dispersed as possible
    - Move each representative point a fixed fraction (eg., 20%) toward the centroid of the cluster

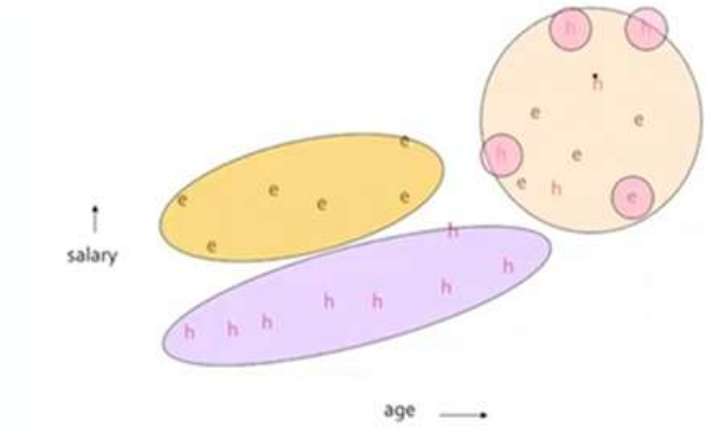

Figure 9: Representative points or remote points in cluster

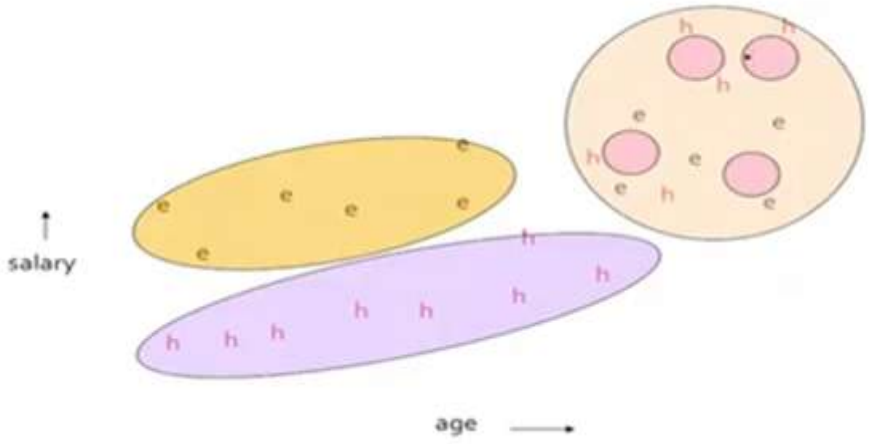

Figure 10: Remote points moving 20% toward centroid

Pass 2 of 2:

- Now, rescan the whole dataset and visit each point p in the data set.
- Place it in the "closest cluster"
    - Closest: that cluster with the closest (to p)among all the representative points of all the clusters.

(2C) Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling

- Chameleon is a hierarchical clustering algorithm that uses dynamic modeling to determine the similarity between pairs of clusters.
- It was derived based on the observed weaknesses of two hierarchical clustering algorithms: ROCK (ignores cluster nearness) and CURE (ignores cluster interconnectivity)

How does Chameleon work?

- Chameleon uses a k-nearest-neighbor graph approach to construct a sparse graph, where each vertex of the graph represents a data object, and there exists an edge between two vertices (objects) if one object is among the k-most-similar objects of the other.
- The edges are weighted to reflect the similarity between objects. Chameleon uses a graph partitioning algorithm to partition the k-nearest-neighbor graph into a large number of relatively small subclusters.

- It then uses an agglomerative hierarchical clustering algorithm that repeatedly merges subclusters based on their similarity.
- To determine the pairs of most similar subclusters, it takes into account both the interconnectivity as well as the closeness of the clusters

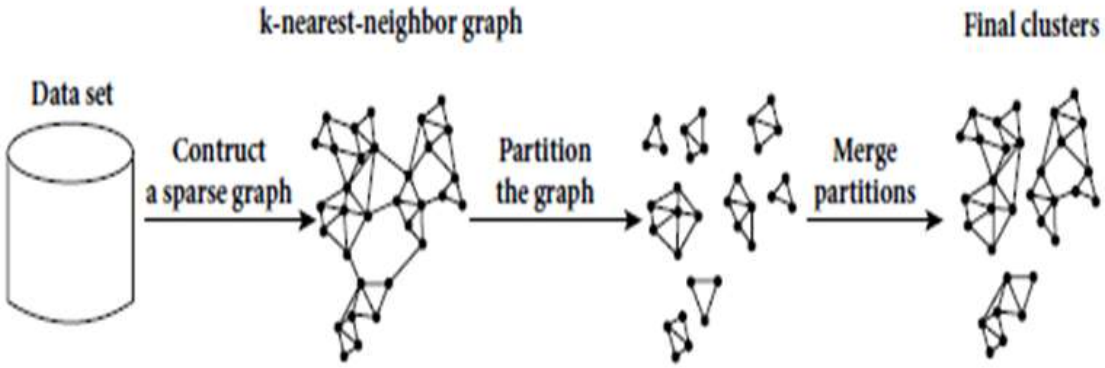

Figure 11: Chameleon – Hierarchical clustering based on k-nearest and dynamic modeling

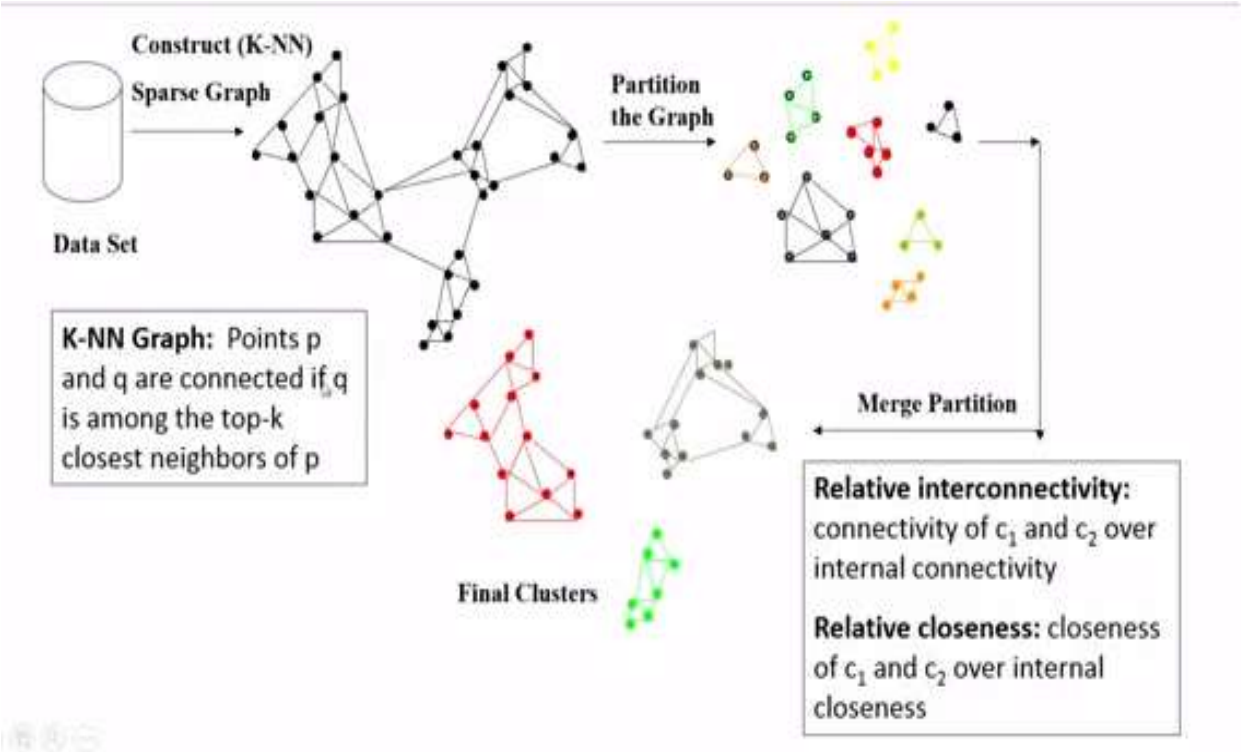

Figure 12: OVERALL FRAMEWORK OF CHAMELEON

(3) **Density-Based Methods**

To discover clusters with arbitrary shape, density-based clustering methods have been developed.

These typically regard clusters as dense regions of objects in the data space that are separated by regions of low density (representing noise).

DBSCAN grows clusters according to a density-based connectivity analysis.

OPTICS extends DBSCAN to produce a cluster ordering obtained from a wide range of parameter settings

(3a) DBSCAN: A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm.
- The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise.
- It defines a cluster as a maximal set of density-connected points.
- The basic ideas of density-based clustering involve a number of new definitions

➢ The neighborhood within a radius ε of a given object is called the **ε-neighborhood** of the object.

➢ If the ε-neighborhood of an object contains at least a minimum number, **MinPts**, of objects, then the object is called a **core object**

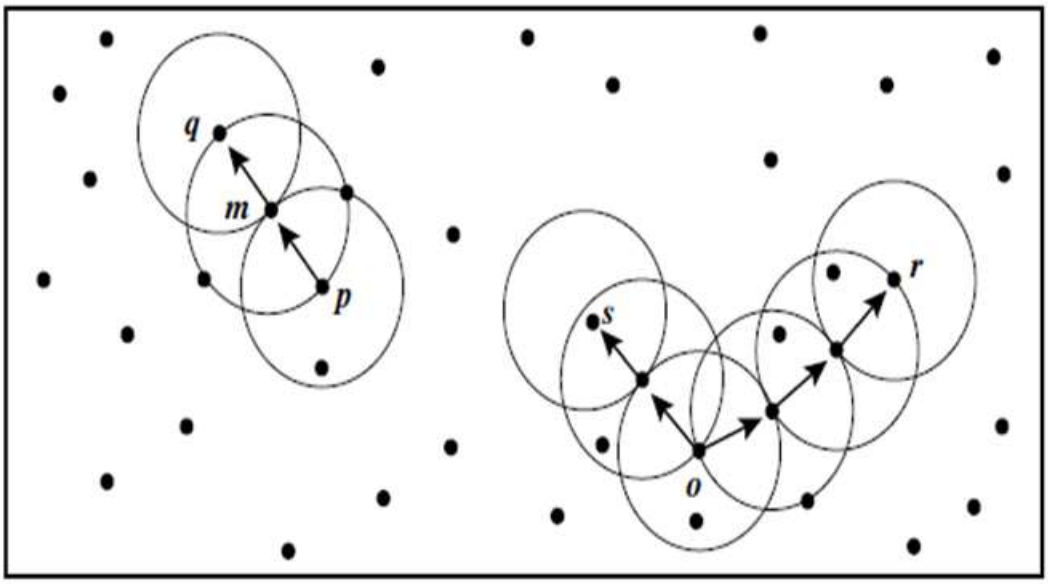

Figure 13: Density reachability and density connectivity in density based clustering

EXAMPLE:

Consider Figure for a given e represented by the radius of the circles, and, say, let

MinPts = 3.

- Of the labeled points ,m, p, o, and r are core objects because each is in an ε-neighborhood containing at least three points.
- q is directly density-reachable from m. m is directly density-reachable from p and vice versa.
- q is (indirectly) density-reachable from p because q is directly density-reachable from m and m is directly density-reachable from p.
- However, p is not density-reachable from q because q is not a core object. Similarly, r and s are density-reachable from o, and o is density-reachable from r.
- o, r, and s are all density-connected.

How does DBSCAN find clusters

- A density-based cluster is a set of density-connected objects that is maximal with respect to density-reachability.

- DBSCAN searches for clusters by checking the e-neighborhood of each point in the database.
- If the e-neighborhood of a point p contains more than MinPts, a new cluster with p as a core object is created.
- DBSCAN then iteratively collects directly density-reachable objects from these core objects, which may involve the merge of a few density-reachable clusters.
- The process terminates when no new point can be added to any cluster.

(3b) OPTICS: Ordering Points to Identify the Clustering Structure

- Rather than produce a data set clustering explicitly, OPTICS computes an augmented cluster ordering for automatic and interactive cluster analysis.
- This ordering represents the density-based clustering structure of the data.

  It contains information that is equivalent to density-based clustering obtained from a wide range of parameter settings
- To construct the different clusterings simultaneously, the objects should be processed in a specific order.
- This order selects an object that is density-reachable with respect to the lowest e value so that clusters with higher density (lower ε) will be finished first.

  Based on this idea, two values need to be stored for each object—core-distance and reachability-distance:
- The core-distance of an object p is the smallest ε ‘ value that makes {p} a core object. If p is not a core object, the core-distance of p is undefined.
- The reachability-distance of an object q with respect to another object p is the greater value of the core-distance of p and the Euclidean distance between p and q. If p is not a core object, the reachability-distance between p and q is undefined**.**

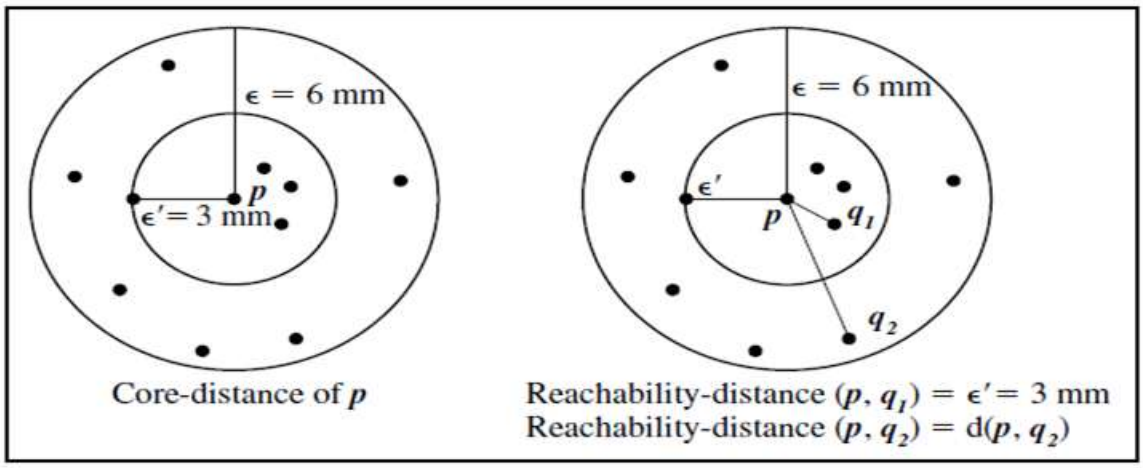

Figure 13 : OPTICS Terminology

Core-distance and reachability-distance

- Figure illustrates the concepts of core distance and reachability-distance.
- Suppose that ε =6 mm and MinPts=5. The core distance of p is the distance, ε', between p and the fourth closest data object.
- The reachability-distance of q1 with respect to p is the core-distance of p (i.e., ε' =3 mm) because this is greater than the Euclidean distance from p to q1.
- The reachability distance of q2 with respect to p is the Euclidean distance from p to q2 because this is greater than the core-distance of p.

4. Grid-Based Methods

- The grid-based clustering approach uses a multi resolution grid data structure.
- It quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed.
- The main advantage of the approach is its fast processing time, which is typically independent of the number of data objects, yet dependent on only the number of cells in each dimension in the quantized space.
- Some typical examples of the grid-based approach include STING, which explores statistical information stored in the grid cells; Wave Cluster, which clusters objects using a wavelet transform method; and CLIQUE, which represents a grid-and density-based approach for clustering in high-dimensional data space

4 a. STING: Statistical Information Grid

- STING is a grid-based multi resolution clustering technique in which the spatial area is divided into rectangular cells.
- There are usually several levels of such rectangular cells corresponding to different levels of resolution, and these cells form a hierarchical structure:
- each cell at a high level is partitioned to form a number of cells at the next lower level.
- Statistical information regarding the attributes in each grid cell (such as the mean, maximum, and minimum values) is precomputed and stored

4 b. CLIQUE: A Dimension-Growth Subspace Clustering Method

- CLIQUE (CLustering InQUEst) was the first algorithm proposed for dimension-growth subspace clustering in high-dimensional space.
- In dimension-growth subspace clustering, the clustering process starts at single-dimensional subspaces and grows upward to higher-dimensional ones.
- Because CLIQUE partitions each dimension like a grid structure and determines whether a cell is dense based on the number of points it contains, it can also be viewed as an integration of density-based and grid-based clustering methods

The ideas of the CLIQUE clustering algorithm are outlined as follows.

- Given a large set of multidimensional data points, the data space is usually not uniformly occupied by the data points.
- CLIQUE's clustering identifies the sparse and the "crowded" areas in space (or units), thereby discovering the overall distribution patterns of the data set.
- A unit is dense if the fraction of total data points contained in it exceeds an input model parameter

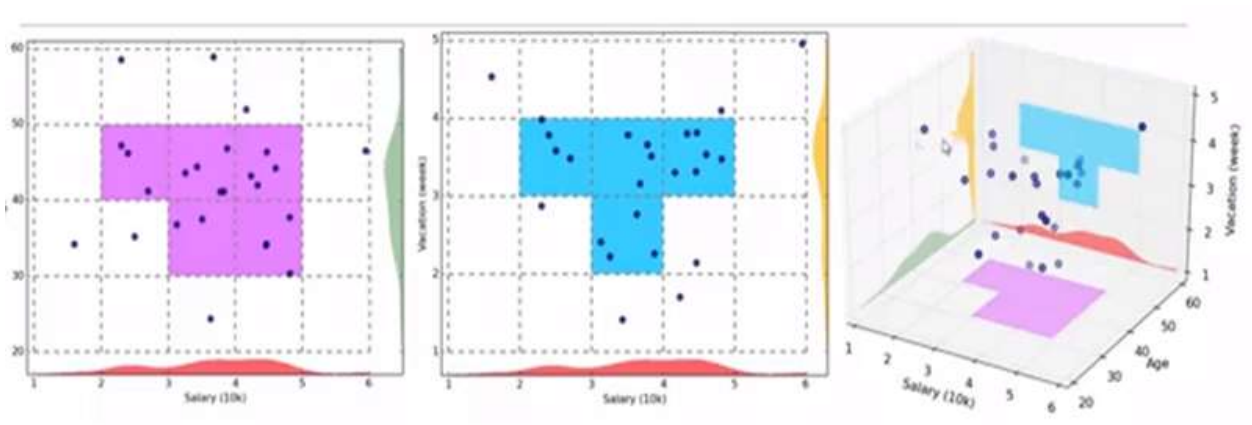

Figure 14: Density and Grid based clustering

Conclusion

So now we have learned many things about Data Clustering such as the methods of Data Clustering in Data mining. .