Clustering

- Cluster is a collection on data objects in which the objects are similar to one another within the same cluster and dissimilar to objects of another cluster.
- Given a database D = { t_1, t_2, t_3,…….., t_n}, a distance measure dist. (t_i, t_j,)defined between any two objects t_i and t_j and an integer value K (number of clusters), the clustering problem is to define a mapping f: D ->{ 1,2, ……., K) where each t_i is assigned ot on of cluster.
- Clustering is similar to classification where similar objects are placed together. Grouped are not predefined as in classification.
- Clustering is an example of unsupervised learning i.e. learning by observation.
- Clustering is also called data segmentation.
- Also used for outliers detection.

What is not Cluster Analysis:

- ***Supervised classification***: Have class label information
- ***Simple segmentation:*** Dividing students into different registration groups alphabetically, by last name.
- ***Results of a query:*** Groupings are a result of an external specification.
- ***Graph partitioning:*** Some mutual relevance and synergy, but areas are not identical

Distinctions between sets of clusters

Partitioning versus Hierarchical

- A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
- A set of nested clusters organized as a hierarchical tree.

Exclusive versus non-exclusive:

- In non-exclusive clustering, points may belong to multiple clusters.
- Can represent multiple classes or ‘border’ points.

Fuzzy versus non-fuzzy

- In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
- Weights must sum to 1
- Probabilistic clustering has similar characteristics

Partial versus complete

- In some cases, we only want to cluster some of the data

Heterogeneous versus homogeneous

- Cluster of widely different sizes, shapes, and densities

Types of Clusters

a. ***Well-separated clusters:*** Clusters are placed at different places.
b. ***Center-based clusters:*** The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster
c. ***Contiguous clusters:*** Nearest neighbor or Transitive
d. ***Density-based clusters:*** Used when the clusters are irregular or intertwined, and when noise and outliers are present.
e. ***Property or Conceptual:*** Finds clusters that share some common property or represent a particular concept
f. Described by an Objective Function:

- Finds clusters that minimize or maximize an objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function. (NP Hard).
- Can have global or local objectives.
    - Hierarchical clustering algorithms typically have local objectives
    - Partitioned algorithms typically have global objectives
- A variation of the global objective function approach is to fit the data to a parameterized model.
    - Parameters for the model are determined from the data.
    - Mixture models assume that the data is a 'mixture' of a number of statistical distributions.
- Map the clustering problem to a different domain and solve a related problem in that domain
    - Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points
    - Clustering is equivalent to breaking the graph into connected components, one for each cluster.
    - Want to minimize the edge weight between clusters and maximize the edge weight within clusters

Clustering techniques:

i. Partitioning Clustering

ii. Hierarchical Clustering
iii. Density-based Clustering
iv. Grid-based Clustering
v. Model-based clustering.

- Characteristics of the input data are important for the selection of the clustering technique.
  - Type of proximity or density measure.
  - Sparseness i.e. type of similarity
  - Attribute type
  - Type of Data
  - Dimensionality
  - Noise and Outliers
  - Type of Distribution

Partitioning Clustering (Iterative relocation method)

- A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
- Partition the database into predefined number of cluster.
- Attempt to determine K-partition that optimizes the certain function.
- Construct a partition of a database D of n-objects into a set of K clusters such that there is the minimum sum of square distance.
- Common partitioning algorithms are: K-means, K-mode, K-medioid.

K-means algorithm (Simple k-means)

- Choose number of cluster (K) to be determined.
- Choose k objects randomly from the data as the centers of k clusters.
- Assign each of the remaining objects to the cluster whose center it is most close to using Euclidean distance.
- Computer the new cluster centers of the clusters using mean points.
- Repeat step3 and 4 until no change in cluster centers or no object change in clusters.

Example: Refer class note

Advantages:

- Relatively simple.
- Simple Implementation.

Disadvantages (Weakness):

- Need to specify number of cluster (K) in advance.

- Unable to handle noisy data and outliers.
- Complexity increases with increase in size.
- Can't handle categorical data.
- Not efficient for highly non-uniform distributed data.
- Basic K-means algorithm can yield empty clusters.
- May generate empty cluster.

Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE). For each point, the error is the distance to the nearest cluster.
- To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(m_i, x)^2$$

- x is a data point in cluster Ci and mi is the representative point for cluster Ci i.e. mi corresponds to the center (mean) of the cluster.
- Given two clusters, we can choose the one with the smallest error.
- One easy way to reduce SSE is to increase K, the number of clusters.
- A good clustering with smaller K can have a lower SSE than a poor clustering with higher K.

Hierarchical Clustering

- A nested set of cluster is created with each level in the hierarchy. At each level it has separate set of clusters.
- At the lowest level, each item is in its own unique cluster.
- At the highest level, all items belong to the same cluster.
- Do not have to assume any particular number of clusters. Any desired number of clusters can be obtained by 'cutting' the dendogram at the proper level.

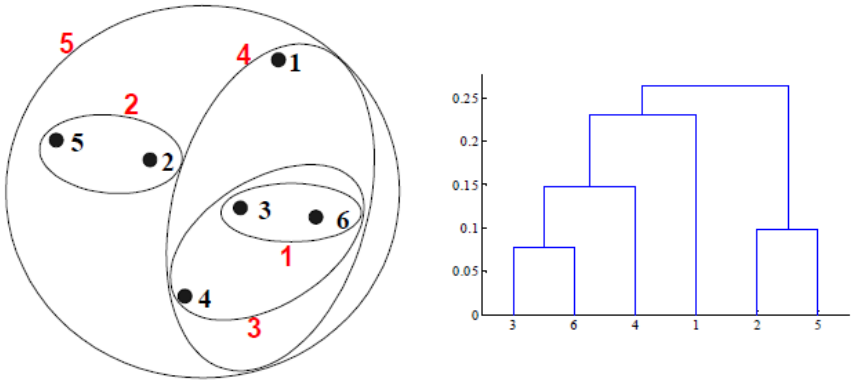
Types

i. Agglomerative (Bottom-Up)

- Start from clustering individual point only, with each cluster having only one record.
- Repeat merging the cluster until a certain number of clusters are left.
- The merging is done on the basis of pair nearest to each other.
- If the merging is continued, it terminates in the hierarchy of clusters which ends into a single cluster.
- Agglomerative is more powerful.

ii. Divisive

- Start from a cluster including all points.
- Repeat splitting the cluster until a certain number of clusters are left.
- The splitting is done on the basis of optimization function.
- If the splitting is continued, it terminates in the hierarchy of clusters which ends into a number clusters with having one data points in each cluster.



Nested Clusters **Dendrogram**

Bisecting K-means

- Variant of K-means that can produce a partitioned or a hierarchical clustering.
- Bisecting k-Means is like a combination of k-Means and hierarchical clustering.

Algorithm

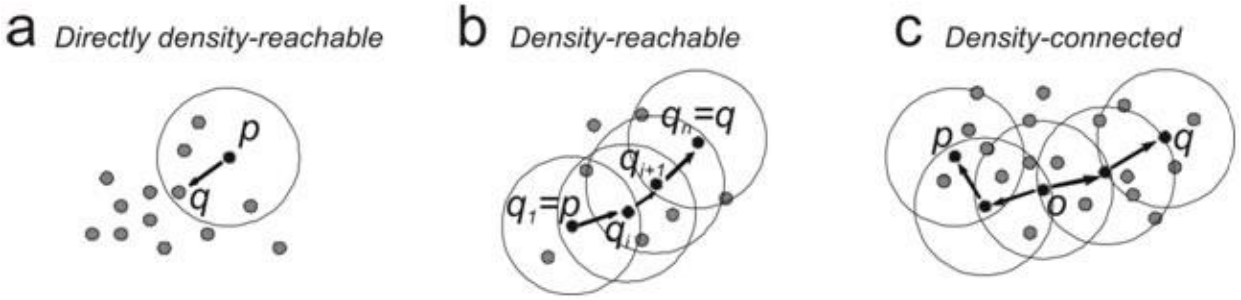
1. Initialize the list of clusters to contain the cluster containing all the points
2. Pick a cluster to split.
3. Find 2 sub-clusters using the basic k-Means algorithm (Bisecting step)
4. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
5. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

Problems and Limitations

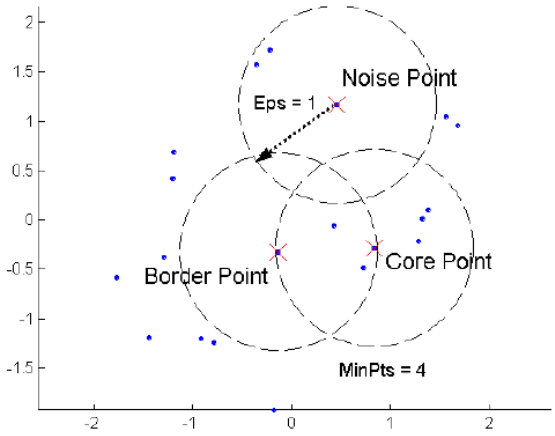
- Once a decision is made to combine two clusters, it cannot be undone.
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and convex shapes
  - Breaking large clusters

Density-Based Clustering

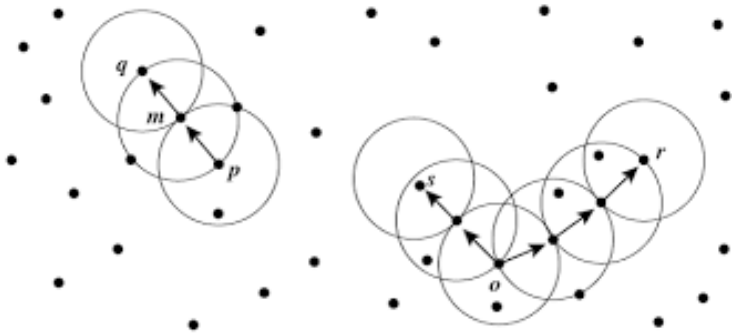
- Grows regions with sufficiently high density into clusters and discover clusters of arbitrary shape in spatial database with noise.
- The neighborhood within the radius 'ε' of a given object is called the ε-neighborhood of the object.
- If the ε-neighborhood of the object of an object contains at least a minimum number of points of object then the object is called ***core object***.
- Given a set of objects (D), an object 'p' is directly density reachable from object 'q' if 'p' is within the ε-neighborhood of q and q is a core object.
- An object 'p' is density reachable from object 'q' with respect to ε and minimum points in a set of objects (D), if there is a chain of objects p1,p2,……, pn such that p_{i+1} is density reachable from p_i with respect to ε and minimum points.
- An object 'p' is density reachable from object 'q' with respect to ε and minimum points in a set of objects (D), if there is an object O ϵ D such that both p and q are density reachable from O with respect to ε and minimum points.
- Density reachability is the transitive closure of direct density reachability and thus relationship is asymmetric.
- Only core objects are mutually density reachable and are symmetric relation.



- ***Border point*** has fewer than MinPts within ε-neighborhood but is in the neighborhood of a core point.
- A ***noise point*** is any point that is not a core point or a border point.



Eg:



- Above fig. shows two clusters with arbitrary shapes.
- m, p, o are core objects, each contain minimum point (4) in their ε-neighborhood.
- 'q' is directly density reachable from m.
- 'm' is directly density reachable from p and vice-versa.
- 'q' is 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' since 'q' is not a core object.
- Similarly 'o', 'r' and's' all are density connected.

Algorithm

- Search for cluster by checking the ε-neighborhood of each point in database.
- If the ε-neighborhood of any point contains more than minimum points, a new cluster with that point as core object is created.
- Iteratively collects directly density reachable objects from these core objects which may involve the merge of a few density reachable clusters.
- Terminates when no new points can be added to any cluster.

Issues:

Cluster Evaluation

1. Intrinsic

- Measure cluster quality based on how "tight" the clusters are.
- Do genes in a cluster appear more similar to each other than genes in other clusters?

Intrinsic Evaluation Methods

a. Cross-validation:

- Leave out *k* experiments (or genes) then perform clustering.
- Measure how well clusters group in left out experiment.

b. Rand Index

-The Rand index is a simple criterion used to compare an induced clustering structure (C_1) with a given clustering structure (C_2).

-Let a be the number of pairs of instances that are assigned to the same cluster in C_1 and in the same cluster in C_2;

- b be the number of pairs of instances that are in the same cluster in C_1, but not in the same cluster in C_2;

- c be the number of pairs of instances that are in the same cluster in C_2, but not in the same cluster in C_1;

- and d be the number of pairs of instances that are assigned to different clusters in C_1 and C_2.

-The Rand index is defined as:

$$RAND = (a + d)/ (a + b + c + d)$$

-The Rand index lies between 0 and 1. When the two partitions agree perfectly, the Rand index is 1.

c. **Sum of squares**: A good clustering yields clusters where genes have small within-cluster sum-of-squares (and high between-cluster sum-of-squares).

d. Silhouette: Good clusters are those where the genes are close to each other compared to their next closest cluster.

e. Gap statistic

Extrinsic:

Compare the results to some best standard labeled data.

Scalability

- Clustering techniques for large sets of data must be scalable, both in terms of speed and space.
- There may be millions of records, and thus, any clustering algorithm used should have linear or near linear time complexity to handle such large data sets. (Even algorithms that have complexity of $O(m_2)$ are not practical for large data sets.).
- Some clustering techniques use statistical sampling. Nonetheless, there are cases, e.g., situations where relatively rare points have a dramatic effect on the final clustering, where a sampling is insufficient.
- Clustering techniques for databases cannot assume that all the data will fit in main memory or that data elements can be randomly accessed.
- Accessing data points sequentially and not being dependent on having all the data in main memory at once are important characteristics for scalability.

Comparison

- The choice of clustering algorithm depends on type of data available and on the particular purpose of the application.
- Several algorithms can be applied on same data for descriptive or exploratory purpose cluster analysis.
- It is difficult to generalize the algorithm and techniques for clustering since some application may have clustering criteria that requires the integration of several techniques.
- Clustering techniques highly depends on dimensions and constraints.