Clustering : state of the art

1) Presentation of different clustering algorithms

General information on clustering :

- Clustering consists in partitioning a set while minimizing some cost function (usually a dissimilarity criterion).
- Clustering gives an overview of a set's distribution, it adds knowledge to exploratory data analysis.

Most popular clustering methods:

a) Iterative square-error partitional clustering
b) Hierarchical clustering
c) Grid-based clustering (Quantise the space into a finite nb of cells and apply operations on these cells and not the indivivuals)
d) Density-based clustering (Pb density fct using max-likelihood estimator)

Clustering can also be employed for feature extraction :

e) By using dissimilarity value as a criteria for feature extraction.
f) By representing data as a distance to cluster centroids.

A. Distance partitional clustering : KMEANS - KMENOIDS

Kmeans iteratively chooses the partition to minimize euclidean distance between individuals. Initialisation is done randomely for a specified number of centers.

How to select the optimal number of clusters with Kmeans :

- The Elbow method : plot the Total within sum of squares vs the number of clusters. Choose the first k before the decrease reduces.
- The Gap Statistic : the gap statstic compares the total within intra-cluster variation for different k with their expected values under null reference distribution of the data (random). Choose higher Gap
- The silhouette method : silhouette coefficient = for each obs calculate avg dissimilarity with all other points in the same cluster i belongs to(Di). than calculate dissimilatity with all other cluster points and get the lower dissimilarity (Ci). Si = (Ci-Di)/max(Di,Ci) => if >0 obs is well clustered.
- Sum of Squares : compare within sum of squares with between sum of squares.
- Consider how clusters vary depending on k (see Clustree library)

K-medoids : algorithm based on median points not to take into account outliers. At each iteration a point is taken as center unlike in k-means.
The medoid of a cluster is defined as the object in the cluster whose average dissimilarity to all the objects in the cluster is minimal, that is, it is a most centrally located point in the cluster. Goodness of clustering is assessed with silhouette.

B. Density Based Methods

Advantages of Density-based clustering are :

- No assumption on the number of clusters. The number of clusters is often unknown in advance. Furthermore, in an evolving data stream, the number of natural clusters is often changing.
- Discovery of clusters with arbitrary shape. This is very important for many data stream applications.
- Ability to handle outliers (resistant to noise).

Gaussian Mixture Modelling
DBSCAN algorithm

Gaussian Mixture Modelling :
https://en.proft.me/2017/02/1/model-based-clustering-r/#:~:text=Model%2Dbased%20clustering%20are%20iterative,to%20set%20of%20data%20points.

https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html#clustering

Method to detect complex patterns in data. Heavy computing.
Each cluster is modelled by a mixture gaussian distribution. The objective function to maximize is the Likelihood for the gaussian mixture model.

K-means can be expressed as a special case of the Gaussian mixture model. In general, the Gaussian mixture is more expressive because membership of a data item to a cluster is dependent on the shape of that cluster, not just its proximity.

As with k-means, training a Gaussian mixture model with EM can be quite sensitive to initial starting conditions. If we compare and contrast GMM to k-means, we'll find a few more initial starting conditions in the former than in the latter. In particular, not only must the initial centroids be specified, but the initial covariance matrices and mixture weights must be specified also. Among other strategies, one approach is to run k-means and use the resultant centroids to determine the initial starting conditions for the Gaussian mixture model.

Can initialise centers for EM algo with :

- hcPairs : matrix of merge pairs from hierarchical clustering.
- subset : subset of data to be used for initialisation.
- noise : initial guess as to which observations are noise in the data.

Model evaluation with : cv or AIC, BIC.

DBSCAN algorithm :
The goal is to identify dense regions, which can be measured by the number of objects close to a given point.

Two parameters are used : epsilon (eps) and minimum points (MinPts). The parameter eps defines the radius of neighbors around x within 'eps' radius. Any point x in the data with a

neighbor count greater than or equal to MinPts is marked as a core point. x is the border point if its number of its neigbors is less than MinPts. points outside eps are outliers.

- *Direct density reachable*: A point “A” is directly density reachable from another point “B” if “A” is in the ϵ-neighborhood of “B” and “B” is a core point.
- *Density reachable*: A point “A” is density reachable from “B” if there are a set of core points leading from “B” to “A.
- *Density connected*: Two points “A” and “B” are density connected if there are a core point “C”, such that both “A” and “B” are density reachable from “C”.

1. For each point *xi*, compute the distance between *xi* and the other points. Finds all neighbor points within distance *eps* of the starting point (*xi*). Each point, with a neighbor count greater than or equal to *MinPts*, is marked as *core point* or *visited*.

2. For each *core point*, if it’s not already assigned to a cluster, create a new cluster. Find recursively all its density connected points and assign them to the same cluster as the core point.

3. Iterate through the remaining unvisited points in the data set.

1. Unlike K-means, DBSCAN does not require the user to specify the number of clusters to be generated
2. DBSCAN can find any shape of clusters. The cluster doesn’t have to be circular.
3. DBSCAN can identify outliers
4. If there is variation in the density, noise points are not detected
5. Sensitive to parameters i.e. hard to determine the correct set of parameters.
6. The quality of DBSCAN depends on the distance measure.
7. DBSCAN cannot cluster data sets well with large differences in densities.

2) Combining multiple clustering models : Cooperative and collaborative clustering

Issues with clustering :

- There isn’t one method that proved to be efficient on all kind of datasets.
- Each method presents specifics weaknesses and strengths

- considering all data for each points
- need to specify number of clusters
- dependency on the random initialization of cluster centroid

K-means efficient when : clusters are compact, hyperspherical, well seperated in the feature space. Can improove K-means by using Mahalonobis distance (<=> PCA transformation) or Fuzzy C-means.

=> The goal is to find a method that selects the appropriate measure of similarity to define clusters AND specify the optimal nb of clusters.

Cooperative clustering (ensemble clustering) :

- Each clustering algorithm is ran on the dataset
- Find a way to aggregate the results in a single partition.

Collaborative clustering :

- At each iteration, algorithms interact with each-other to find a better ending result.

More complex method

Cooperative clustering (ensemble clustering)

Each clustering algorithm is run on the dataset => get for each clustering method a set of partitions.

Voting models

- naive approach : assign xi to the “majority cluster” => initial cluster can give different nb of clusters.
- Corresponding classes : compute intersection between clusters = similarity measure for 2 clusters (close to a cosine similarity measure) to create a corresponding function between clusters/partitions. Than choose for each xi its best class.

Correlation clustering

‘Master algorithm that combines local solutions’
Approaches to clustering according to some dissimilarity value.
Cost function takes into account total within-cluster similarity and between-cluster dissimilarity. Both to be maximized (Maximisation problem is Np-complete)
d() dissimilarity distance based on how often two patterns are placed in the same cluster among every partition.

Co-association matrix method
Vote between each partitions combination. Assumption that individuals from a cluster will often be in the same partition of another cluster.
Create co-association matrix S. Si,j =nij /M nij=nb of occurrences where xi and xj belong to the same cluster among the M partitions.

Si,j = frequency of associations between xi and xj

Then use agglomerative clustering on S (=Hierarchical agglo clustering). Each pattern is considered a singleton cluster. Each step aggregates 2 clusters.
Need to compute dissimilarity between clusters and a stop criterion.
At each iteration 2 columns and lines are merged into a cluster. stop when every point has been merged into a cluster. (costly with high values of n and p>>20)

Lowest dissimilarity (single linkage) = when 2 instances xi and xj are merged distance with outer instances xj is min(d (xi,xk)d (xj,xk)) for each instance.
Avg dissimilarity (avg linkage) = when 2 instances xi and xj are merged distance with outer instances xj is Avg(d (xi,xk)d (xj,xk)) for each instance.
Highest dissimilarity (complete linkage) =when 2 instances xi and xj are merged distance with outer instances xj is max(d (xi,xk)d (xj,xk)) for each instance.

Minimizing disagreement
This method is a rigorous formalisation of the co-assosiation matrix.
Consensus method to minimize disagreement between each partition.
d(C1, C2) =sum (dij) dij=1 if i and j belong to the same partition or both don't belong to the same partition. We then search a partition to minimize : D(C)=SUM(d(Ck,C)) for each k, which is the same as D(C) = total dissimilarity between patterns + total similarity between patterns. D(C*) <= M x D(C opt)

Collaborative clustering (ensemble clustering)

Method creates a communication network between algorithms which main objective is to reduce the variability of each algorithm.

Process is more abstract than cooperative clustering with wider techniques.

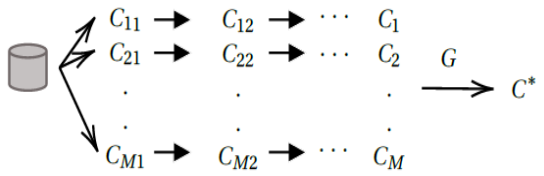


Figure 1: Cooperative Clustering

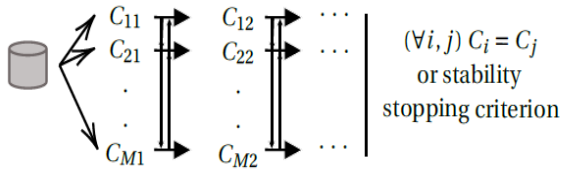
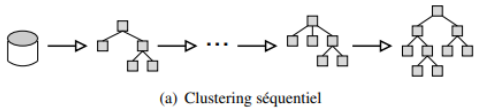
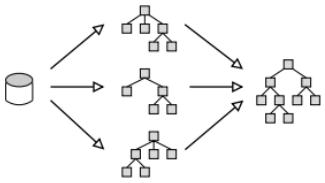


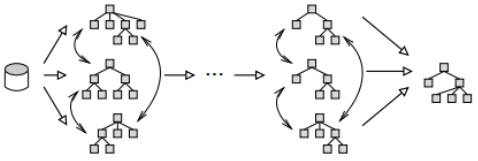
Figure 2: Collaborative clustering



(a) Clustering séquentiel



(b) Clustering coopératif



(c) Clustering collaboratif

At each iteration, each cluster method adapts its heuristic to the result of the other clusterings.

- *Single vs. multi strategy* : single strategy methods work with same algorithms, possibly running with different parameters, while multi strategy approaches can work with different algorithms.

- *Single objective vs. multiobjective* : single objective clustering works with algorithms that have similar cost functions, in contrary to *multiobjective* clustering.

- *Local vs. global clustering*: a global collaborative clustering will give to every algorithm the same data to work with. In local approaches, collaboration can be organized between algorithms that work with different subsets of D, or even with same data but using different features.