Data Clustering
Hierarchical Clustering, Density based clustering
Grid based clustering

Team 2

Prof. Anita Wasilewska
CSE 634 Data Mining
All Sources Used for the Presentation


Hierarchical Clustering. www.saedsayad.com/clustering_hierarchical.htm


https://cognitiveclass.ai/courses/machine-learning-with-python/


https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
Overview

- Hierarchical Clustering
- Density Based Clustering
- Grid Based Clustering
- Related Paper
Hierarchical Clustering
Hierarchical Clustering

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis) is a method of cluster analysis which seeks to build a hierarchy of clusters.
Two Types of Hierarchical Clustering

● **Agglomerative**(bottom-up)
  ○ Assign each observation to its own cluster
  ○ Compute the similarity (e.g. distance) between each of the clusters and join the two most similar clusters
  ○ Repeat until reach the termination condition

● **Divisive**(top-down)
  ○ Assign all the observations to a single cluster
  ○ Partition the cluster to two least similar clusters
  ○ Proceed recursively on each cluster until reach the termination condition
How to measure the distance?

- Single linkage -- the shortest distance between two points in each cluster
- Complete linkage -- the longest distance between two points in each cluster
- Average linkage -- the average distance between each point in one cluster to every point in the other cluster

\[
L(r, s) = \min(D(x_{ri}, x_{sj}))
\]

\[
L(r, s) = \max(D(x_{ri}, x_{sj}))
\]

\[
L(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})
\]
Example
Example

Hierarchical Clustering Dendrogram
Example
Example
Example
Example

Hierarchical Clustering Dendrogram

Sample Index
Example
Example
Example
Example
Example
Example
Termination Condition (Agglomerative)

- Method 1: pick a number n, the algorithm stops when n number of cluster are formed.
- Method 2: stops when the next merge would create a cluster with low “cohesion”
  - Example: set cohesion as some distance between two clusters. If distance between two clusters are larger than this cohesion, the algorithm stops.
- Method 3: let the algorithm run until all the nodes become a single cluster
Termination Condition (Divisive)

- Method 1: pick a depth $n$, the algorithm stops when $n$ number of split are performed.
- Method 2: like Agglomerative clustering, set a cohesion.
- Method 3: let the algorithm run until all the clusters all separated into leaves.
Complexity

- In the standard algorithm for hierarchical agglomerative clustering (HAC) has a time complexity of $O(n^3)$ and requires $O(n^2)$ for memory
  - because we exhaustively scan the $N \times N$ matrix for the largest similarity in each of $N - 1$ iterations.

- Divisive clustering with an exhaustive search is $O(2^n)$
  - but with the help of faster heuristics (such as k-means) to choose splits, the time complexity can be reduced massively.
DENSITY BASED CLUSTERING
Density Based Clustering

Minimum Points
Density Based Clustering

- Core Point
- Border Point
- Outlier

Min Points = 5
$\varepsilon = 1$ cm

https://cognitiveclass.ai/courses/machine-learning-with-python/
Density Reachability

- A point $y$ is said to be reachable from $x$ if there is a path $p_1,...,p_n$ with $p_1=x$ and $p_n=y$.
- Where each $p_{i+1}$ on the path must be a core point with the possible exception of $p_n$.

An object $y$ is directly density-reachable from object $x$, if $x$ is core object and $y$ is in $x$’s epsilon-neighborhood.

1. $a$ is directly density-reachable from $b$.
2. $b$ is directly density-reachable from $c$.
3. $a$ is indirectly density-reachable from $d$.
4. $d$ is not density-reachable from $a$, since $a$ is not a core point.
How does Density-based clustering works?

Min points = 3
ε = 1cm

Step 1: pick a random point that has not yet been assigned to a cluster or designed as an outlier. Determine if it is a core point. If not label it as outlier.

Step 2: once a core point has been found, add all directly reachable to its cluster. Then no neighbor jumps to each reachable point and add them to the cluster. If the outlier has been added, label it as a border point.

Step 3: repeat these steps until all point are assigned a cluster or label as outlier.

DBSCAN Smiley Face Clustering

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
Density based clustering vs K-means
Density based clustering vs K-means
Pros and Cons

Pros:

- discovers clusters of arbitrary shapes
- Handle Noise
- Needs density parameter as termination condition
Pros and Cons

Cons:

● Cannot handle varying densities
● Sensitive to parameters—hard to determine the correct set of parameters

(MinPts=4, Eps=9.92).

(MinPts=4, Eps=9.75)

Original Points

GRID BASED CLUSTERING
Grid based clustering

- Divide space into grid cells
- Evaluate density in each cell
- Form clusters of dense neighbour cells

WILL DISCUSS ABOUT CLIQUE ALGORITHM...
CLIQUE

- Divide space into cells
- Cell frequency:
  - Number of data inside cell
- Use Apriori Principle
- CELL = Itemset

ANY SUBSET OF A FREQUENT ITEMSET

A FREQUENT ITEMSET
PROS

✓ Efficient
✓ Reduce time complexity
✓ Clusters with arbitrary size and shape
CONS

- Non-uniformity
- Locality
- Dimensionality
consider the problem of determining the structure of clustered data, without prior knowledge of the number of clusters or any other information about their composition.

Cluster analysis means the partitioning of data into meaningful subgroups, when the number of subgroups and other information about their composition may be unknown.
Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Information

Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma and Ji-Rong Wen

Proceedings of the 12th annual ACM international conference on Multimedia Pages 952-959 2004
Introduction

Top 8 results of the query “Pluto” in Google and AltaVista’s search engine[2]
A web-page can be divided into semantic blocks. For each image, there is a smallest block which contains that image. We call it image block. The image block contains information that might be useful for describing the image. Text, links and visual information can be used to process image.
Features

Three kinds of representations:

- Visual feature based representation
- Textual feature based representation
- Link graph based representation
Visual feature based representation

Color correlogram
Color moments
Color histogram
Texture features
Textual Feature Based Representation

File name of the image file
URL of the image
Image ALT (alternate text) in web page source
The title of the web page
Link graph based representation

Which URL is connected to which URLs
Clustering

Combining the features, will result in a feature space and by using hierarchical clustering the images of a web-page can be put in different clusters.
Results
Conclusion

- A method to organize WWW image search results.
- First proposed three representations
- Then Spectral techniques were applied to cluster the search results into different semantic categories
- The reorganization of each cluster based on visual features makes the clusters more comfortable to the users
We are Team 2
How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis

CHRIS FRALEY AND ADRIAN E. RAFTERY

Department of Statistics, University of Washington, USA


Hierarchical methods proceed by stages producing a sequence of partitions, each corresponding to a different number of clusters. They can be either ‘agglomerative’, meaning that groups are merged, or ‘divisive’, in which one or more groups are split at each stage. Hierarchical procedures that use subdivision are not practical unless the number of possible splittings can somehow be restricted.
In agglomerative hierarchical clustering, however, the number of stages is bounded by the number of groups in the initial partition. It is common practice to begin with each observation in a cluster by itself, although the procedure could be initialized from a coarser partition if some groupings are known. A drawback of agglomerative methods is that those that are practical in terms of time efficiency require memory usage proportional to the square of the number of groups in the initial partition.

At each stage of hierarchical clustering, the splitting or merging is chosen so as to optimize some criterion.
In model-based clustering, it is assumed that the data are generated by a mixture of underlying probability distributions in which each component represents a different group or cluster.
the basis for a more general model-based strategy for clustering:

Determine a maximum number of clusters to consider (M) and a set of candidate parametrizations of the Gaussian model to consider. In general M should be as small as possible.

Do agglomerative hierarchical clustering for the unconstrained Gaussian model, and obtain the corresponding classifications for up to M groups.

Do EM for each parametrization and each number of clusters 2; : : : ; M, starting with the classification from hierarchical clustering.

Compute the BIC for the one-cluster model for each parametrization and for the mixture likelihood with the optimal parameters from EM for 2; : : : ; M clusters. This gives a matrix of BIC values corresponding to each possible combination of parametrization and number of clusters.

Plot the BIC values for each model. A decisive first local maximum indicates strong evidence for a model (parametrization + number of clusters).
Paper:

Parallel algorithms for hierarchical clustering

Clark F. Olson Computer Science Department, Cornell University

Journal: Parallel Computing, Volume 21, Issue 8, Pages 1313-1325, ISSN 0167-8191,
https://doi.org/10.1016/0167-8191(95)00017-I

date of the publication: 1995
Parallel algorithms for hierarchical clustering

Clustering of multidimensional data is required in many fields. One popular method of performing such clustering is hierarchical clustering. This method starts with a set of distinct points, each of which is considered a separate cluster. The two clusters that are closest according to some metric are agglomerated. This is repeated until all of the points belong to one hierarchically constructed cluster. The final hierarchical cluster structure is called a dendrogram. Which is simply a tree that shows which clusters were agglomerated at each step.

A dendrogram shows how the clusters are merged hierarchically. [1]
some metric to determine the distance between pairs of clusters:

1. Graph metrics: Consider a completely connected graph where the vertices are the points we wish to cluster and the edges have a cost function that is the Euclidean distance between the points:
   - Single link
   - Average link
   - Complete link
2. Geometric metrics: These metrics define a cluster center for each cluster and use these cluster centers to determine the distances between clusters:

- Centroid
- Median
- Minimum variance