Cluster analysis

Author: Otakar Trunda

Cluster analysis (CA) is one of the disciplines studied in the field of data analysis and data mining in general. The goal of cluster analysis is to discover similarities in a given dataset and to find groups of similar records in the data.

Cluster analysis can be used to find trends and patterns and to identify groups of "typical" records called clusters.

Results of clustering can then be used for example to understand behavior of customers, anticipate their needs and offer appropriate services.

Example: A supermarket keeps record of its customers by means of loyalty cards. Every purchase is stored in the database and identified with specific customer. How can the company benefit from cluster analysis?

Solution: There are many ways how to use CA in this scenario. Company may identify groups of records (sets of goods purchased together) based on the type of customer – e.g. a cluster of "vegetarian purchases", cluster of "family purchases", "buying goods for children", cluster of "people with celiac disease" and so on. Another type of grouping can be done using the relative price of goods. E.g. cluster of people that often purchase expensive goods versus cluster of customers that usually go for the cheapest alternative. Yet another clustering may involve time when the purchase took place which can discover group of customers that prefer shopping in the morning, other group of those who go shopping in the evening, in the night, at weekends and so on.

This information can be used to better address the advertisement, optimize the layout of types of goods in the shop, anticipate demand, optimize the reserve management and so on.

Formal description

There are many variants of the problem; we will only describe it in the most basic form:

Input:

- A set of m **features** $f_1,\ldots, f_m$
    - Features can be discrete or continues
    - Discrete features can be either numerical or categorical. Numerical features are ordered, categorical aren't.
    - For each discrete feature f_i we are given a set of its possible values V_i
    - For each continues feature f_i we are given a lower bound l_i and an upper bound u_i. Set of possible values V_i is an interval $V_i=<l_i, u_i>$
    - Set of features constitutes a **feature space F** as a Cartesian product of their possible values $F = \prod_{i=1}^{m} V_i$
- A set of n points in the feature space $a_1, \ldots, a_n \in F$
    - Points $a_1, \ldots, a_n$ are called data or **records**.
    - The set of all records is called **dataset**, denoted by $R = \{a_1, \ldots, a_n\}$
- A metric *d* on the feature space
    - $d: F \times F \to \mathbb{R}^+$
    - Metric is a map taking two points in the feature space and returning their distance (non-negative real number)
    - It can be scaled such that its maximal value is 1 by this transform
    - $d': F \times F \to < 0,1 >,\ d'(x,y) = {d(x,y)} \Big/ {\max_{k,l \in F} d(k,l)}$
    - Then a similarity measure can be introduced as follows
    - $s: F \times F \to < 0,1 >, s(x,y) = 1 - d(x,y)$
    - s measures the similarity of points. Higher values of $s(x,y)$ indicates higher similarity between x and y
- Number of clusters to be found denoted by k

Output:

- A map $c: R \to \{1, \ldots, k\}$ denoting a membership in clusters such that records in the same cluster are similar to each other while those in different clusters are not similar.
    - $c(a_i) = j$ means that a_i belongs to the j-th cluster

Example:

Suppose we study results of school tests. For each student we are given his/her

- Height (100 – 200cm)
- Color of eyes (green, blue, brown)
- Grades in math and literature (A,B,C,D,E)

Height is a continues feature, with the lowest value 100cm and the highest value 200cm. Color of eyes is a discrete categorical feature (there is no ordering on the colors). Grades are two discrete numerical features (there is an ordering on the values).

The feature space in this case looks like this:
$F =< 100{,}200 > \times \{green, blue, brown\} \times \{\mathrm{A, B, C, D, E}\} \times \{\mathrm{A, B, C, D, E}\}$

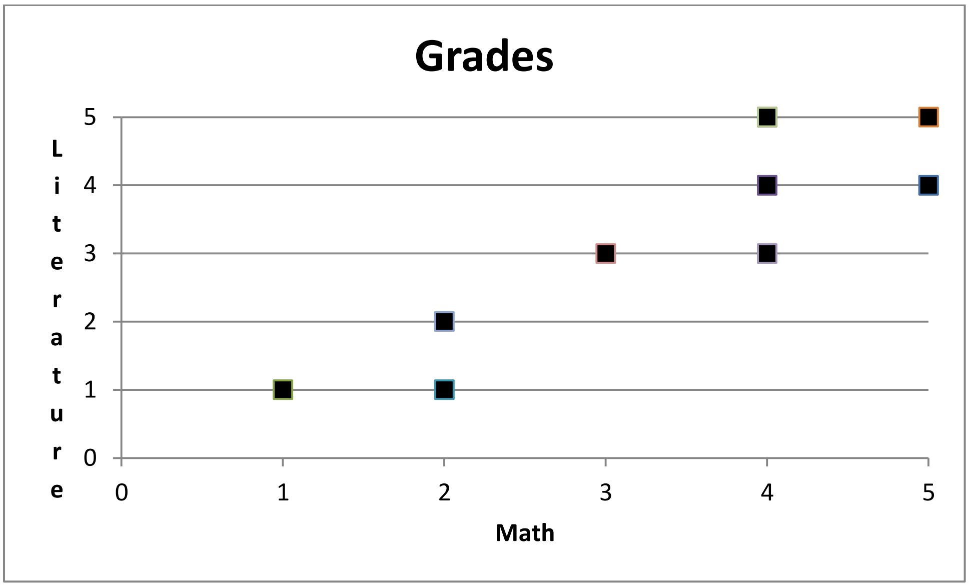
We are given the following data (grades are converted to numbers)

height	eyes	Math	Literature
178	green	5	4
178	brown	4	5
167	brown	1	1
154	brown	4	4
151	green	2	1
179	green	5	5
157	brown	2	2
160	blue	3	3
160	brown	4	5
151	blue	4	3

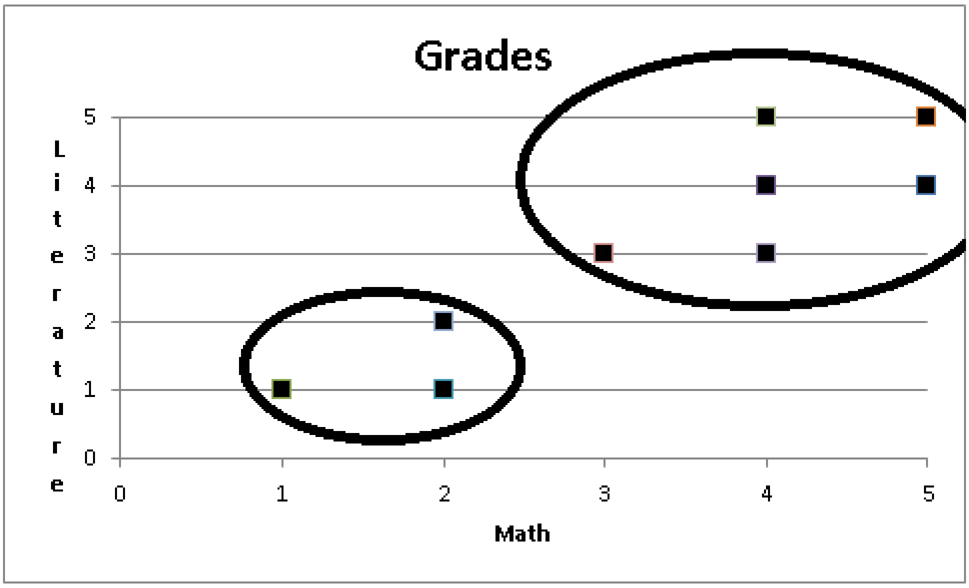
We have 10 data points from the feature space.

As a metric, we will use $d(x,y) = \sqrt{(x.Math - y.Math)^2 + (x.Lit - y.Lit)^2}$ this metric doesn't take into account information about height and eyes color so we will only cluster the data according to grades. Let's say we want to find 2 clusters, so we set the parameter *k* to 2.

The following graph shows the position of records in the feature space (showing only relevant features).



Possible result of the clustering: records are grouped into 2 clusters. The first cluster contains 3 points and represents students with good grades while the second cluster represents "bad grades students" and contains 7 points.



Note that the condition on "good" clustering is somewhat vaguely defined. Specific variants of CA use different measures to estimate the quality of clustering but often the only real measure is the opinion of an expert. This means that there is no "correct" clustering algorithm – there are several different algorithms, not all of them are suitable for all datasets. CA is not a fully automated procedure. The users often use the trial-error method to find proper algorithm, metric, parameters and data preprocessing before they are happy with the results.

Methods of cluster analysis

There has been published over 100 CA algorithms and there is still extensive research in the field. We will discuss here only a few basic CA algorithms.

K-means

K-means is one of the most popular CA algorithms. It tries to place k *centroids* into the feature space in order to minimize the distance between points and their nearest centroids. For each record, the nearest centroid is found and their distance is calculated. These distances are squared and summed over all dataset to get the penalty of clustering. Centroids are placed such that this penalty is minimized. Formally

$$c_1^*, \dots c_k^* = argmin_{c_1,\dots c_k} \sum_{i=1}^{n} \left[\min_{j=1\dots k} \left(d\left(a_i, c_j\right) \right)^2 \right]$$

Where $c_1^*, \dots c_k^*$ are the centroids.

Variants:

- k-medoids – only points from the dataset can be used as centroids.
- Unknown number of clusters – algorithm determines the proper number of clusters on its own

Advantages:

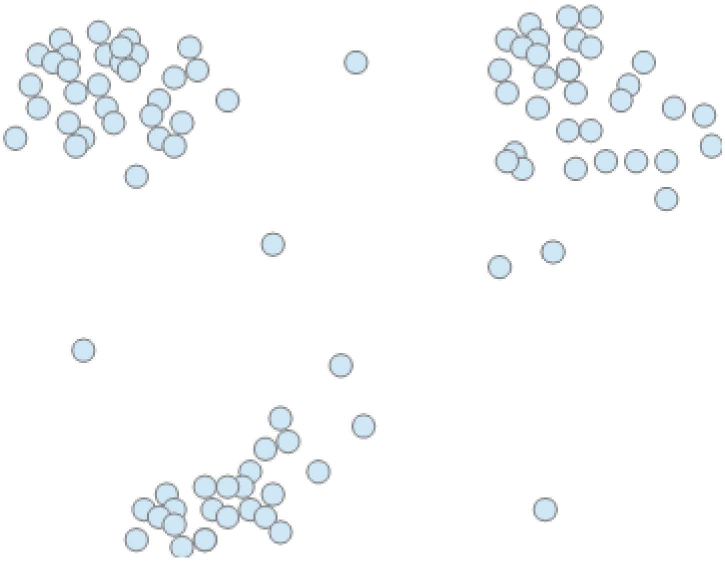
- Easy definition and interpretation of the clusters
- Works well when the clusters have similar size and width

Disadvantages:

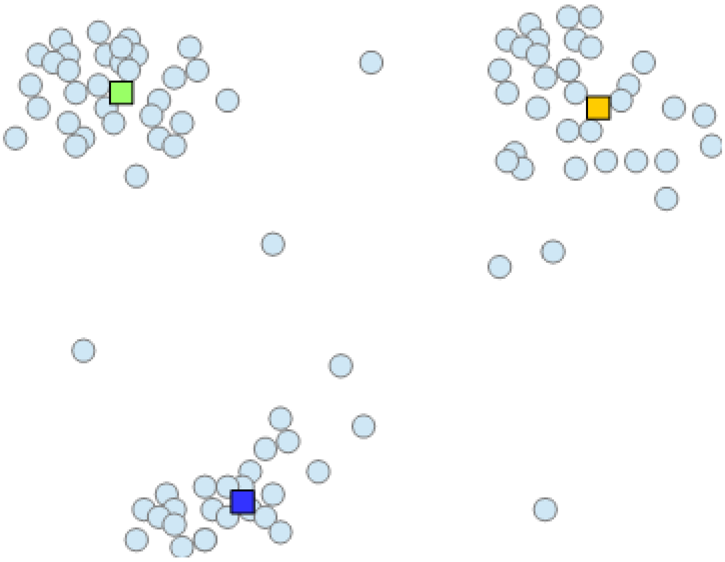
- Can’t handle clusters of different sizes and density-based clusters
- Works only with convex clusters
- Minimizing the penalty criterion is too complex to guarantee global optimality, suboptimal strategies are used to find some local minimum

Examples

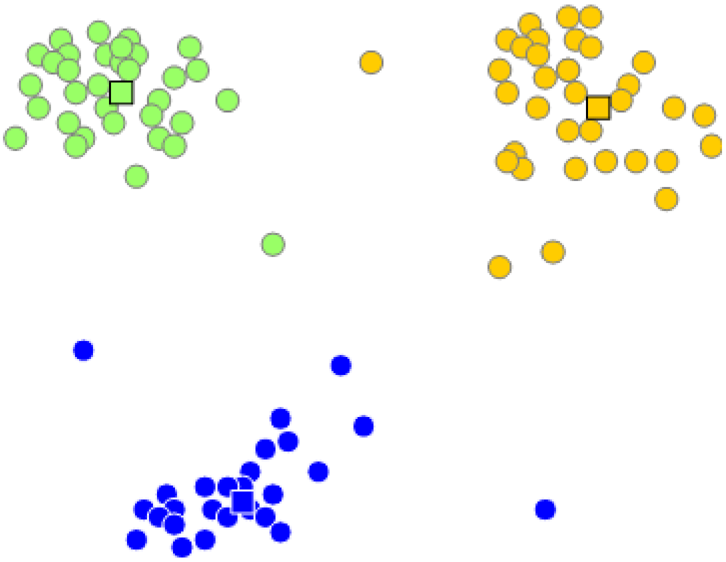
An input dataset in a 2D feature space can look like this:



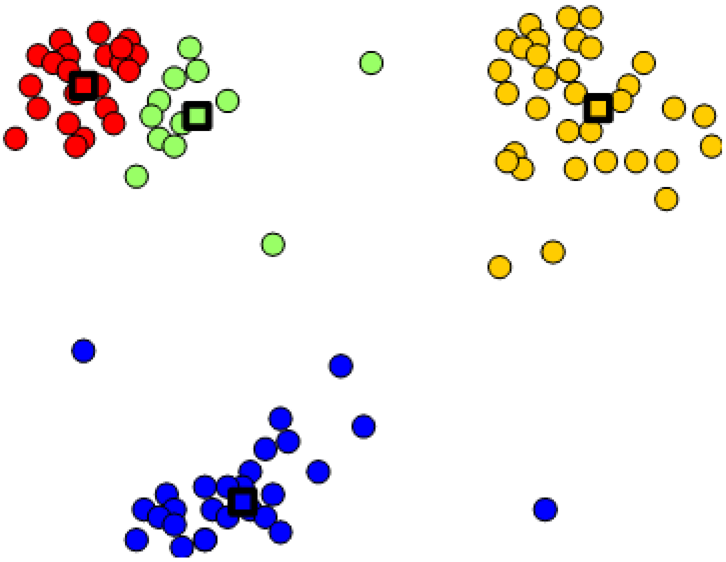
The k-means algorithm can then find optimal positions of 3 means like this:



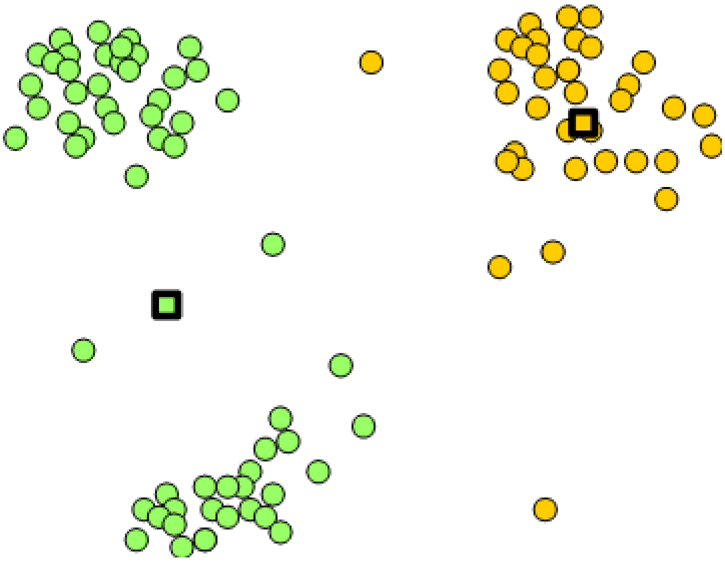
The division into clusters is then done by finding the nearest mean for each record. The result of the clustering looks like this:



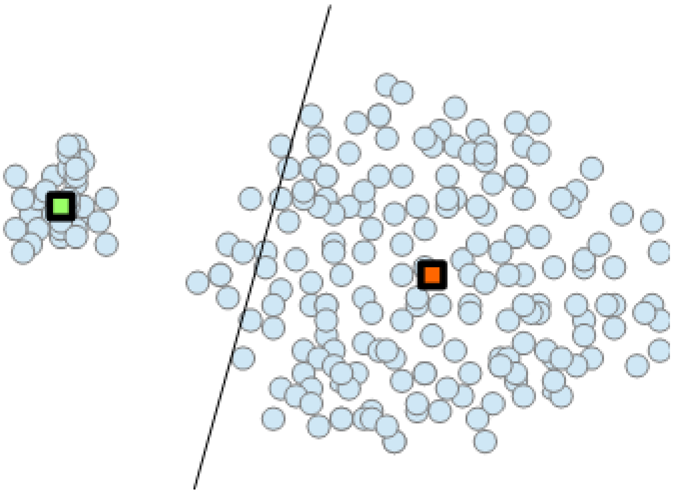
In the previous scenario we assumed that the correct number of cluster is given. If it is not given, the algorithm has to find out the correct number on its own. If it doesn't find the correct number, we might get odd results. If for example the algorithm tried to find 4 clusters instead of 3 the results might look like this.



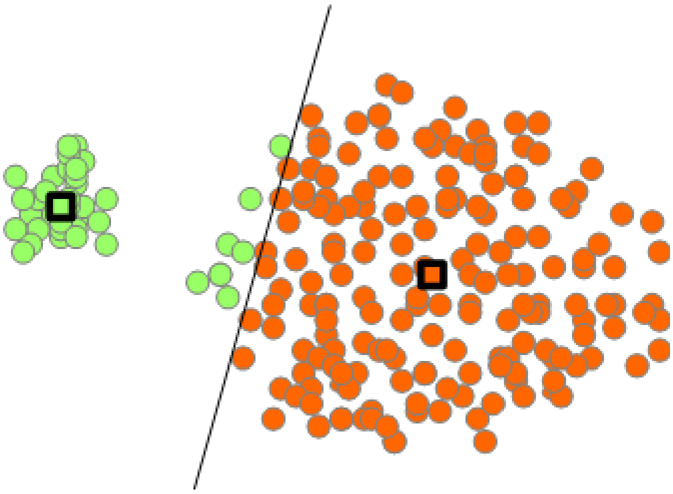
On the other hand, if the number of clusters were too small we might get these results:



That is one of the disadvantages of the algorithm. Another problem might be caused by different size of the clusters. Consider the following example:



The membership of records into clusters is determined by the nearest mean, which then gives the following result. This problem can be solved by Probability-distribution clustering.



Probability-distribution clustering

Clusters are defined as probability distributions (e.g. Gaussians with specific means and variances) and the data is considered a random sample. The parameters of the distributions determine the location and shape of the clusters.

For given data point and given probability distribution (including its parameters), we can calculate the probability of the point being generated by the distribution.

Let's denote ϑ_i the parameters of the i-th distribution. For a data point, we calculate the likelihood of the point being generated by every cluster and select the cluster with the highest probability. This is done for every point and the probabilities are summed. The result measures the quality of the clustering – the highest value, the better. Parameters of the distributions are set in order to maximize this criterion.

Formally:

$$\vartheta_1^*, \dots, \vartheta_k^* = argmax_{\vartheta_1,\dots\vartheta_k}\left[\sum_{i=1}^{n} \max_{j=1,\dots,k}(P[a_i|\vartheta_j])\right]$$

Values of the parameters can be found by the EM-algorithm.

Advantages:

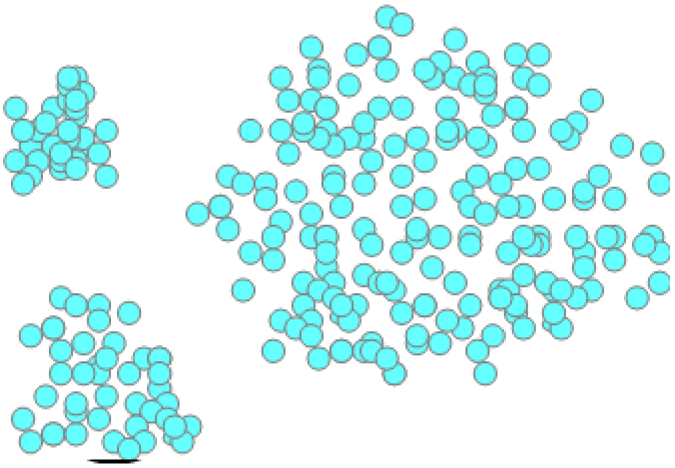
- Can correctly handle clusters with different size and width
- Straightforward generalization of k-means

Disadvantages

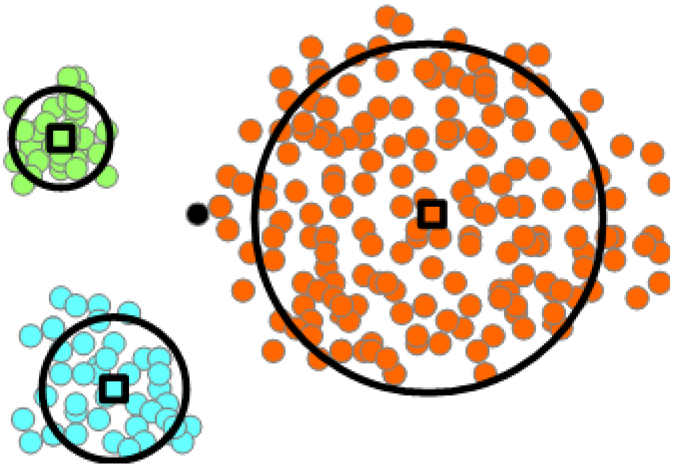
- Works only with convex clusters
- Can't handle density-based clusters
- Almost impossible to find optimal values for parameters, approximations are used

Examples

Consider the following example of the input dataset.

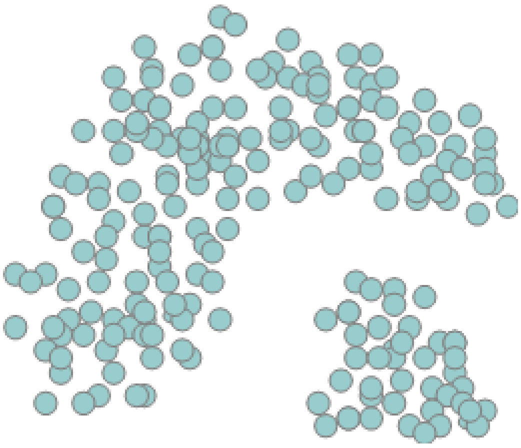


If we use distributions with different variances, we can correctly assign records to cluster. The result is given in the following picture. The circles represent variances while squares represent means of the distributions.

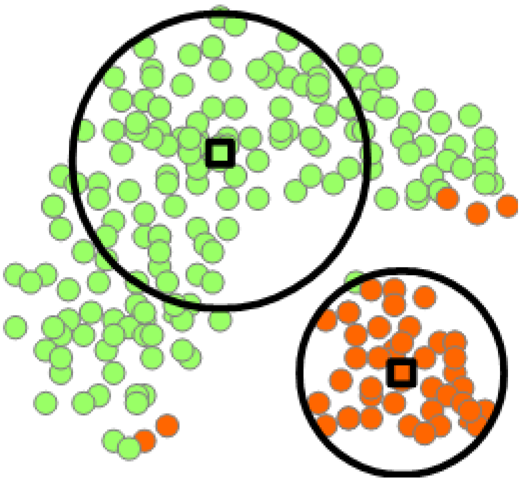


The black marked record might be closer to the green and blue mean, but it will still be assigned to the orange cluster since the orange distribution with its high variance has much greater probability to generate such point that the other two distributions.

The distribution clustering however still only works with convex clusters. Consider the following example:



We see that there should be two clusters, but even the distribution clustering cannot find then correctly. It gives the following results:



Hierarchical clustering

Once the "regular" clustering is done, we can go further and perform clustering inside each cluster to distinguish the records even more precisely. This procedure can be repeated until each mini-cluster contains only one record. The method is similar to biological taxonomy where organisms are divided into categories by their species then each species is again divided into subspecies and so on.

The resulting diagram of hierarchical categories is called dendrogram.

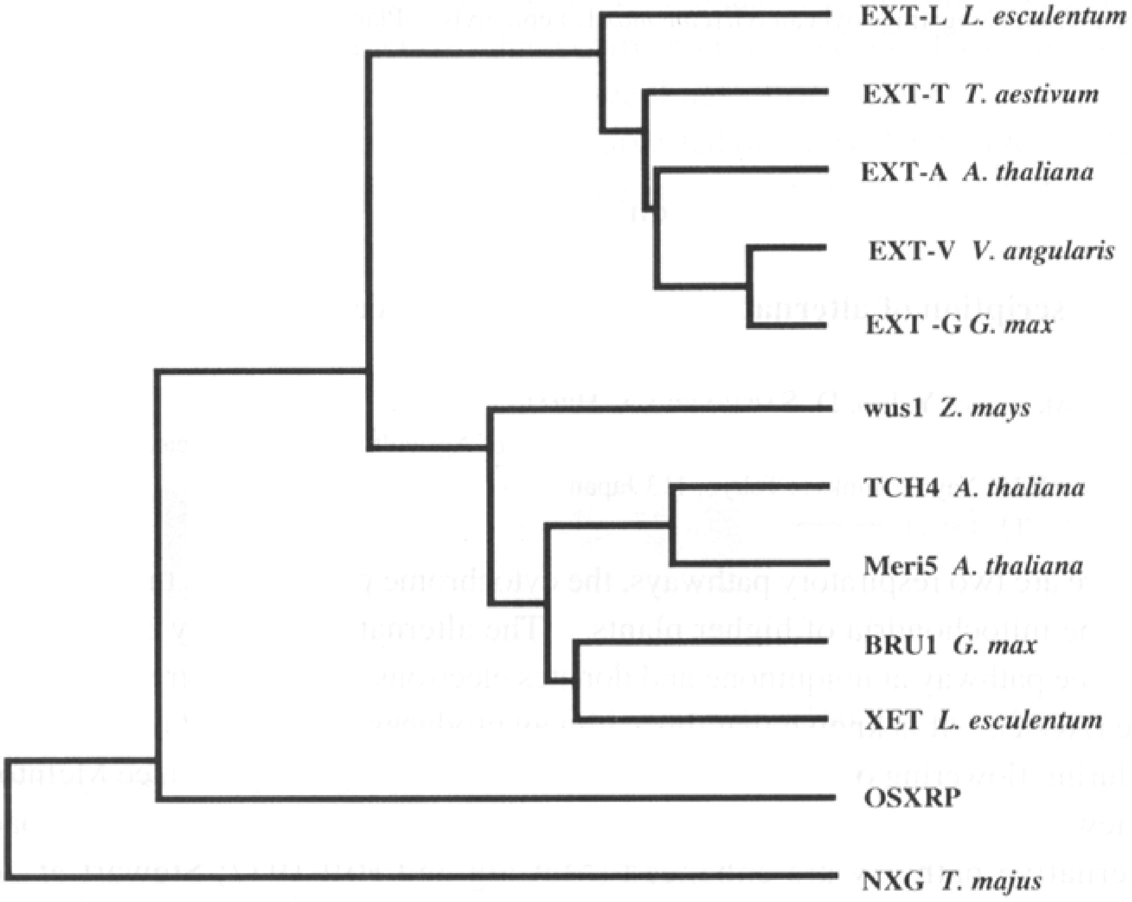
Method:

There are two basic approaches to Hierarchical clustering

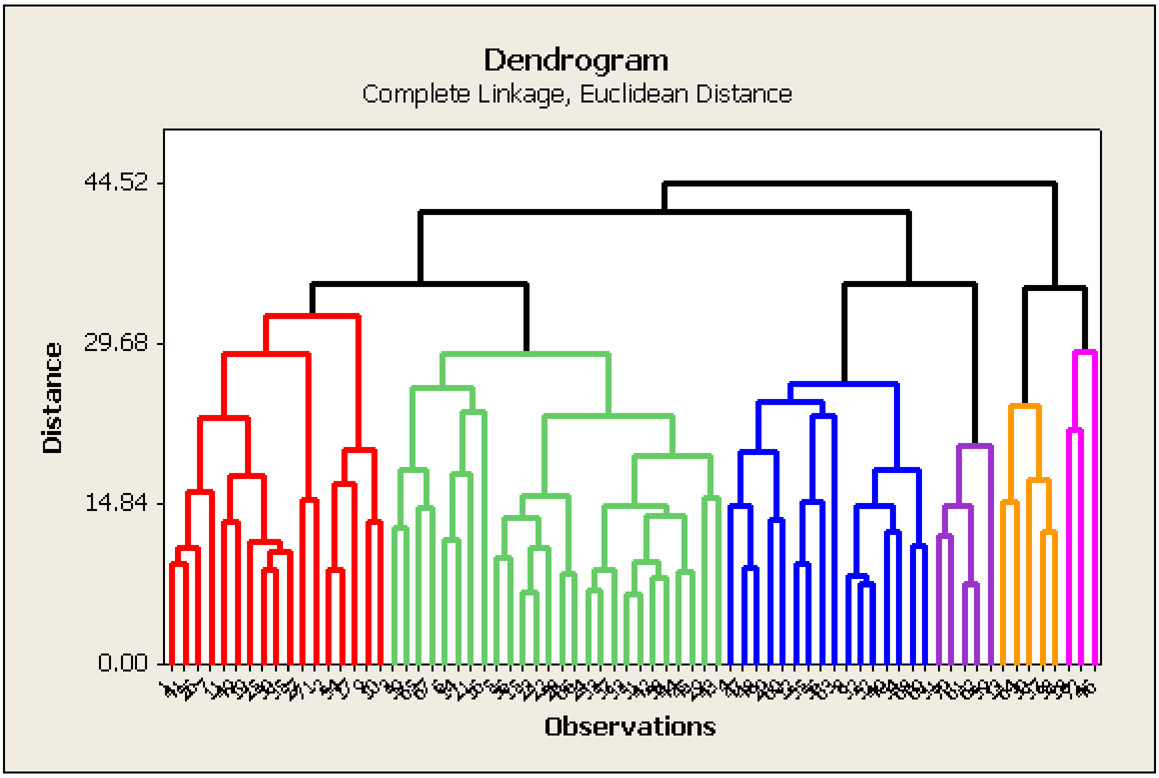
- Top-down approach: starts with one large cluster and successively divides it into smaller ones by separating the most heterogeneous group into two groups The process can go on until there is only one record in each group or it can be stopped when the granularity is sufficient
- Bottom-up approach starts with individual records and successively unifies groups that are most similar.

Example:

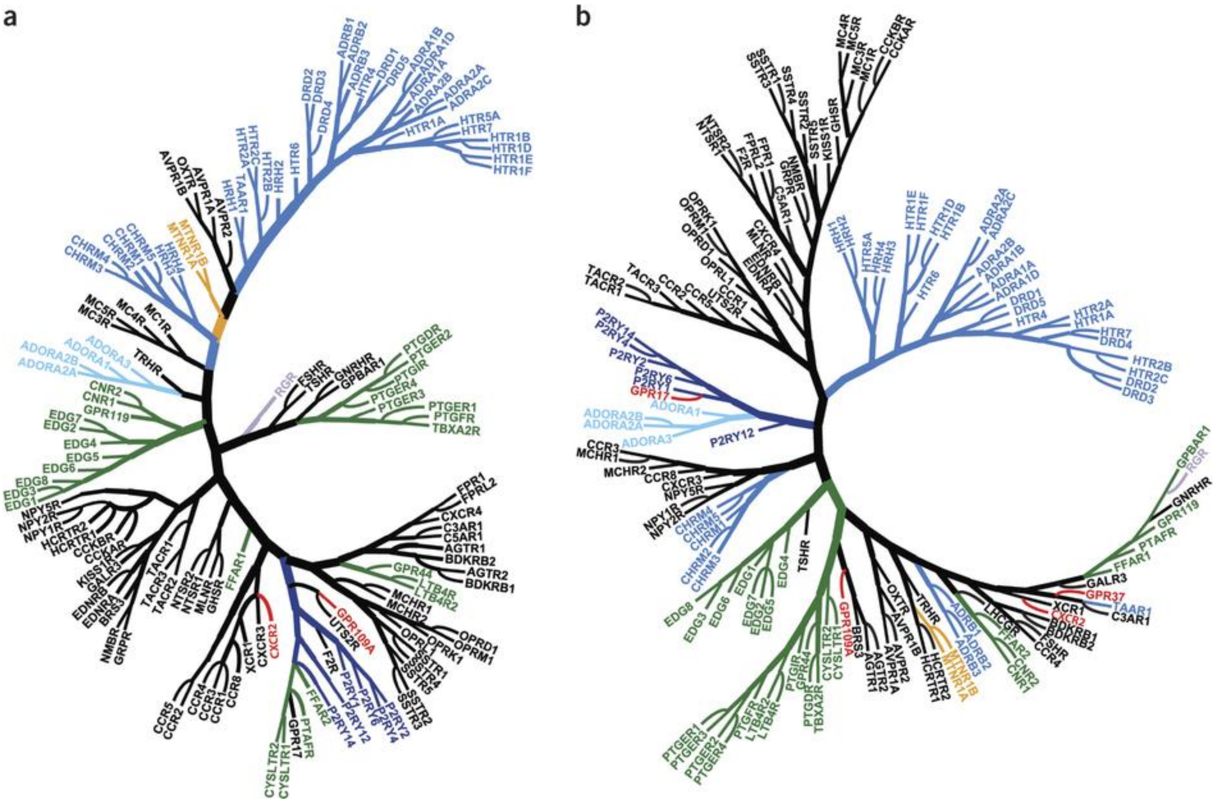
A dendrogram might look like this



The resulting clusters can be visualized for example like this:



There are several different ways to visualize a dendrogram. One of the more exotic ways is depicted in the following picture.



Advantages:

- Doesn’t need the parameter k
- Can handle clusters of different sizes
- Provides more information than regular clustering (relations between clusters)

Disadvantages:

- Has problems with outliers
- Is not always appropriate

Quality measures of clustering

In order to find a good clustering we have to be able to measure its quality. There two basic ways to measure the quality of the clustering.

External measures use the opinion of an expert to evaluate the clustering. This is often the case with benchmark problems to compare the quality of algorithms. In a typical scenario, we are given the data points together with the correct division into clusters. The quality of a clustering is then evaluated based on its similarity with this given results.

Internal measures don’t rely on an expert to determine the quality of clustering. An index is computed based on similarity of records within the clusters and distance of those across clusters. The clustering is considered good if the distances within clusters are low and distances across clusters are high.

There are several formulae for the specific computation of the quality index. We will not present them here, but an interested reader can find them in the bibliography.

Conclusions

We have introduced the area of cluster analysis, presented the basic methods used to solve clustering problems and pointed out their advantages and drawbacks. We also mentioned several means to measure the quality of the clustering.

There is still an extensive research in the area and we have only presented the basics here. Interested reader can find more information in the mentioned bibliography.

Bibliography

Brian S. Everitt, S. L. (2011). *Cluster Analysis.* John Wiley & Sons.

Mirkin, B. (1996). *Mathematical Classification and Clustering.* Springer Science & Business Media.

Romesburg, H. C. (2004). *Cluster Analysis for Researchers.* Lulu.com.