MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 10: Clustering

Pauliina Ilmonen

Contents

Clustering

Agglomerative Hierarchical Algorithms

Moving Centers Method

Words of Warning

References

Clustering

Clustering

Let $x_1, x_2, ..., x_n$ be measurements of p variables on n objects that are believed to be heterogeneous. The aim in cluster analysis is to group these objects into k homogeneous classes. The number of classes, k, is also often unknown (but usually assumed to be a lot smaller than n).

In multisample problem one has m samples and the aim is to group the m samples into k homogeneous classes.

Clustering

Clustering methods rely on two (separate) issues:

- The choice of a distance or dissimilarity measure between objects.
- The choice of a group building algorithm.

Clustering

Cluster analysis is a difficult problem in a general framework.

An intuitively appealing approach:

1. Define all the possible partitions of the n $p-$variate data points into k classes, $k = 1, 2, ..., n$.
2. For each obtained partition, compute the value of a chosen criterion.
3. Select the partition that optimizes the criterion.

Problem: The number of combinations that have to be computed (even when n is small) is huge! For example, for $n = 12$, the number of possible partitions is over 4 millions.

Agglomerative Hierarchical Algorithms

Agglomerative Hierarchical Algorithms

Agglomerative hierarchical algorithms are methods that start from n classes and go step by step to $n-1, n-2, \ldots$ nested classes.

Agglomerative Hierarchical Algorithms

1. Start from the finest partition: n clusters, each containing one data point x_i
2. Calculate distances $d_{ij} = d(x_i, x_j)$, where d is an appropriate distance between individuals.
3. Find the minimal distance and group together the corresponding individuals.
4. Compute distances between obtained groups using an appropriate linkage function.
5. Find the minimal distance and group together the corresponding closest groups.
6. Repeat steps 4 and 5 until you have one single group.

How Many Clusters to Choose?

In agglomerative hierarchical algorithms, the minimal value in step (3 and) 5 provides the so called aggregation level for each step. (In other words: the aggregation level is the distance between the clusters that were grouped.) The number of clusters is chosen based on the aggregation level. High level indicates grouping of heterogeneous clusters. Thus one can decide to "cut" at a desired level.

Which Linkage Functions to Use?

There are several ways to measure the distance between groups:

- The minimum linkage:

$$d(A, B) = \min_{x_i \in A, x_j \in B} d(x_i, x_j).$$

- The maximum linkage:

$$d(A, B) = \max_{x_i \in A, x_j \in B} d(x_i, x_j).$$

- The average linkage:

$$d(A, B) = \frac{1}{n_A n_B} \sum_{x_i \in A} \sum_{x_j \in B} d(x_i, x_j).$$

- The Ward linkage:

$$d(A, B) = \frac{n_A n_B}{n_A + n_B} d(c_A, c_B),$$

  where c_A and c_B are the centers of the clusters A and B, respectively.

Linkage functions

Minimum linkage is simple, but the problem is that quite different groups could be clustered together just for having two close elements (chaining). This approach is still very often used in practice. Also in maximum linkage, the problem is that the decision is based on single points. If one wishes to avoid these problems, the average linkage provides a safer choice.

How to Choose the Distance?

With quantitative data, the euclidian distance

$$d^2(x_i, x_j) = (x_i - x_j)^T (x_i - x_j)$$

is a classical choice.

How to Choose the Distance?

Also principal component metric is quite popular choice. Then

$$d^2(x_i, x_j) = (x_i - x_j)^T D^{-1}(x_i - x_j),$$

where $D = diag(s_1^2, ..., s_p^2)$ and s_t^2 is the variance of the tth component of x.

How to Choose the Distance?

There are plenty of other choices: Manhattan distance, Maximum distance, ...

How to Choose the Distance?

With qualitative data, if one wishes to perform a cluster analysis of the row profiles of a contingency table, one could use the chi-square distances between the row profiles (as in MCA).

How to Choose the Distance?

Also context related distances can be used. For example, if one of the variables is considered being more important than the other variables, then one can put more weight on that.

Hierarchical Clustering Methods

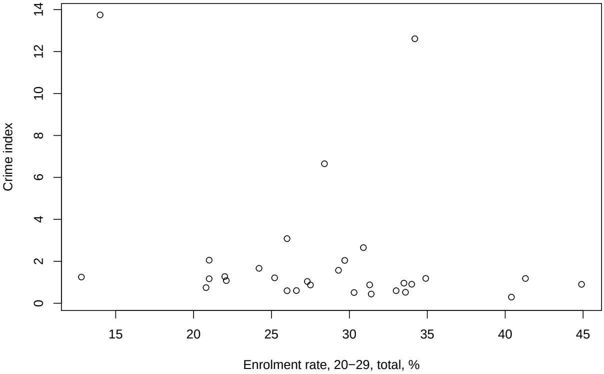
Agglomerative hierarchical clustering algorithms are "bottom up" methods. One may also start from one cluster and split step by step. These "top down" methods are called divisive hierarchical clustering algorithms.

Agglomerative Hierarchical Clustering, Example

The data consists of the education enrolment rate of 20–29-year-olds in the OECD countries (where data was available) and a crime index based on the UN homicide and robbery rates. Agglomerative hierarchical clustering was applied to the data set. The metric used was Euclidean distance and linkage criteria was the average linkage.

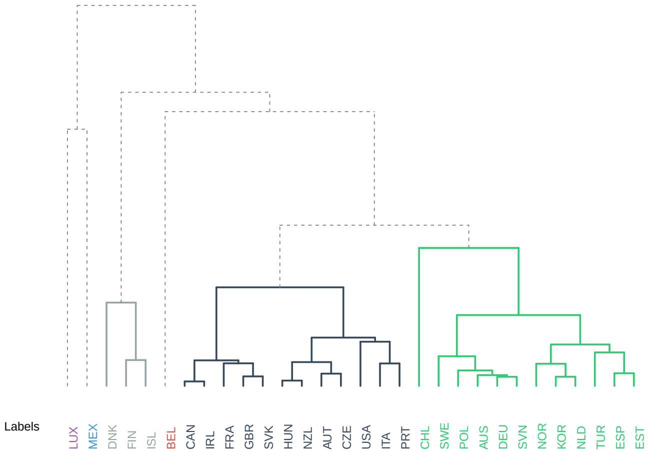
Agglomerative Hierarchical Clustering, Example

Scatter plot of the data:



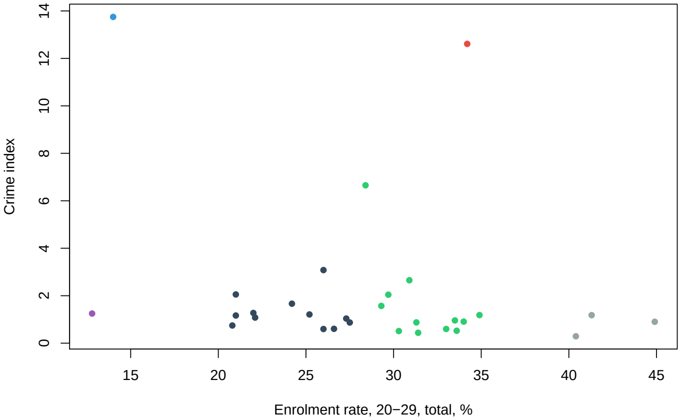
Agglomerative Hierarchical Clustering, Example Continues

The results of the cluster analysis presented as a classification tree:



Agglomerative Hierarchical Clustering, Example Continues

Scatter plot using the cluster colours from the previous slide.



Nonhierarchical Clustering Methods

Agglomerative hierarchical clustering algorithms and divisive hierarchical clustering algorithms are popular, but there exists several other clustering methods. The use of nonhierarchical clustering methods usually requires knowledge of the number of clusters.

Moving Centers Method

Moving Centers Method (*k*-means clustering)

Moving centers clustering method is based on calculating distances from "centers". The method requires knowing the number of clusters k.

Moving Centers Method (k-means clustering)

1. Choose randomly k data points $c_1, ..., c_k$ out of $x_1, ..., x_n$.
2. Define k sets $A_1, ..., A_k$ such that $A_t = \{x_i \mid d(x_i, c_t) \leq d(x_i, c_j), \text{ for } j \neq t\}$.
3. Calculate new centers $c_1, ..., c_k$ (for example sample means) of the sets $A_1, ..., A_k$.
4. Repeat steps 2 and 3 until convergence.

Distances

As when applying hierarchical clustering, also when applying moving centers clustering, one has to consider the context and decide what distance is (the most) appropriate. One also has to decide how to define the "center." Usually sample mean is used, but other locations can be used as well.

Moving Centers Method, Problems and Solutions

- Problem: Possible empty clusters in step 2. Solution: Choose one new center randomly.
- Problem: The algorithm always converges, but sometimes to a local optimum and sometimes very slowly. Solution: Choose the initial points wisely.

Moving Centers Method, Choosing the Initial Points Wisely

k-means++ initialization

Words of Warning

Some Words of Warning

- Optimal clustering methods are computationally very very heavy — in general not doable using standard computers and software.
- Different clustering methods can produce different solutions.
- The chosen distances, and methods to calculate distances between sets, may have an effect on the outcome.

Next Week

Next week we will talk about some very recently developed multivariate methods.

References

References I

📕 K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

L. Simar, An Introduction to Multivariate Data Analysis, Université Catholique de Louvain Press, 2008.