K-means and Hierarchical clustering

Juan Camilo Rivera. [1] Hugo Andres Dorado. [1]

[1] Big data and site-specific agriculture
Decision and Policy Analysis
Centro Internacional de Agricultura Tropical

Data analysis course, 2018

Outline

1. K-means
    - Overview of K-means
    - K-means algorithm
    - Required R packages and functions
    - Case of study
    - Pros vs Cons
2. Hierarchical
    - Overview of Hierarchical
    - Algorithm
    - Case of study
    - Pros vs Cons

Overview of K-means

- MacQueen 1967
- Unsupervised machine learning for partitioning
- Different and similar cluster

Definition

An object is considered to be in a particular cluster if it is closer that cluster's centroid than any centroid.

K-means algorithm

Step by step

1. Specific the number cluster k.
2. Select randomly k objects from the data set as initial cluster
3. Assigns each observation to their closest centroid, using Euclidean distance.
4. Update the centroid by calculating the mean values
5. Iteratively minimize the total with sum of square

Required R packages and functions

libraries

- stats
- factorextra

functions

`kmeans`(x, centers, iter,max=10, nstart =1)

- x = numeric matrix or data frame.
- centers = possible values are the number of cluster
- iter.max = the maximum number of iterations allowed
- nstart = the number of random starting partitions when centers is a number

Case of study

Violent Crime Rates by US State

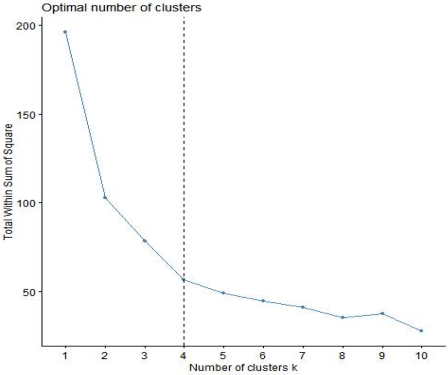
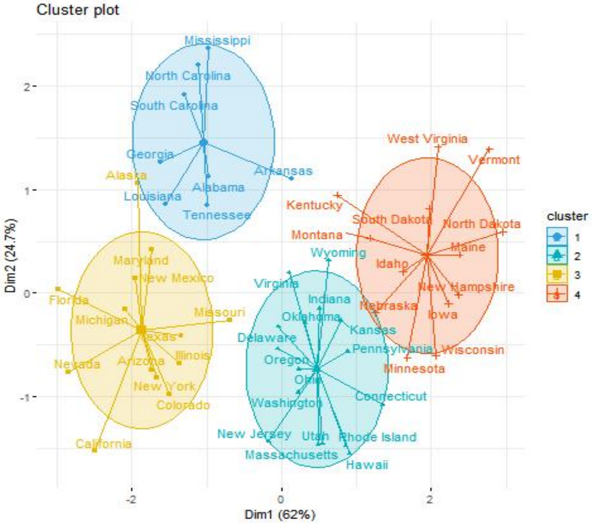


Figure: number of cluster

Case of study

Violent Crime Rates by US State



Pros vs Cons

Pros

- Very simple and fast algorithm
- Easy to interpret the clustering results
- Large amount data

Cons

- It assumes prior knowledge of data and analyst requires to choose appropriate number cluster.
- It's sensitive to outliers.
- The final result obtained is sensitive to initial random selection.

Overview of Hierarchical

Hierarchical clustering can be subdivided into two groups:

Agglomerative

Each observation is initially considered as a cluster or its own (leaf). Then, the most similar cluster are successively merged until there is just one single big cluster (root).

Divise

An inverse of agglomerative clustering, begin with the root, in which all objects are included in one cluster. Then the most heterogeneous clusters are successively divided until all observation are in their own cluster.

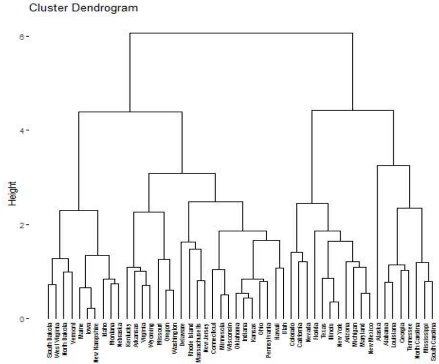
Algorithm

1. Preparing data
2. Computing (dis) similarity information between every pair of objects in the data set
3. Using linkage function to group objects into hierarchical cluster tree, based on the distance information at step 1.
4. Determining where to cut the hierarchical tree into clusters. This creates a partition of the data.

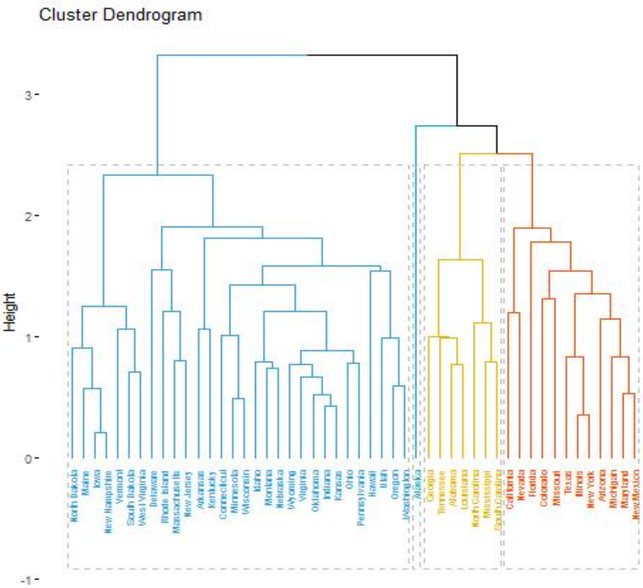
Case of study

USArrests
Linkage

1. **Complete**, Maximum
2. **Single**, Minimum
3. **Average**
4. **Centroid**



Case of study



Pros vs Cons

Pros

- Non number of cluster to be specified
- Easy to implement

Cons

- Slow
- Sometimes it is difficult to identify the correct number of cluster by the dendogram .
- No uncertainty about the tree structure
- Algorithm can never undo what was done previously

For Further Reading I

- Pierre Lafaye de Micheaux.
  The R Software,
  Fundamentals of programming and statistical analysis
  Springer, 2013.

- William Sullivan
  Machine Learning for Beginners Guide Algorithms

- Giuseppe Ciaburro
  Balaji Venkateswaran
  Neural Networks with R