Hierarchical Clustering Analysis

Belias Michael

2 April 2016

Contents

1 Introduction

The Hierarchical algorithm is separated into 2 methods, the agglomerative and the divisive. In agglomerative algorithm begins with all observations as individuals and connects the 2 being most similar (least disimilar) and forms a first cluster then repeats the process till all are connected into 1 cluster with similarity 0.

Agglomerative algorithms :

- single — nearest neighbor
- complete — furthest neighbor or compact
- ward.D2 — Ward's minimum variance method ()
- mcquitty — McQuitty's method (WPGMA : Weighted Pair Group Method with Arithmetic Mean)
- average — average similarity (UPGMA : Unweighted Pair Group Method with Arithmetic Mean)
- median — median (as opposed to average) similarity
- centroid — geometric centroid
- flexible — flexible Beta

While the most popular distance metrics are :

- manhattan (Absolute distance between the two vectors : $\sum |x_i - y_i|$)
- euclidean (The Usual distance between the two vectors : $\sqrt{\sum(x_i - y_i)^2}$)
- minkowski (The minkowski is the generalization of the 2 above $(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$, as we can see for p = 1 we get Manhattan and for p=2 Euclidean)
- maximum (for $p \to \infty$ we get the Chebyshev or Maximum distance distance)
- canberra $(\sum(|x_i - y_i|/|x_i + y_i|))$
- binary

2 Hierarchical Cluster Analysis

In R, we typically use the *hclust()* function to perform hierarchical cluster analysis.

With *hclust()* we will calculate a cluster analysis from either a similarity or dissimilarity matrix, but plots better when working from a dissimilarity matrix. We have to provide a *dist()* object

```
df<- read.table("C:/Users/Mike/Desktop/My Complete Book In R/K-means/k-means.txt")
```

```
set.seed(6970222)
```

```
hc= hclust(dist(df[,-8]), method = "average")
# clustering plot will not be over crowded

hc$labels = df$species[as.numeric(hc$labels)]



plot(hc,hang = -1, cex = 0.5, xlab = "UPGMA method", main = " Hierarchical UPGMA dendrog

rect.hclust(hc, k=3, border="red")
```

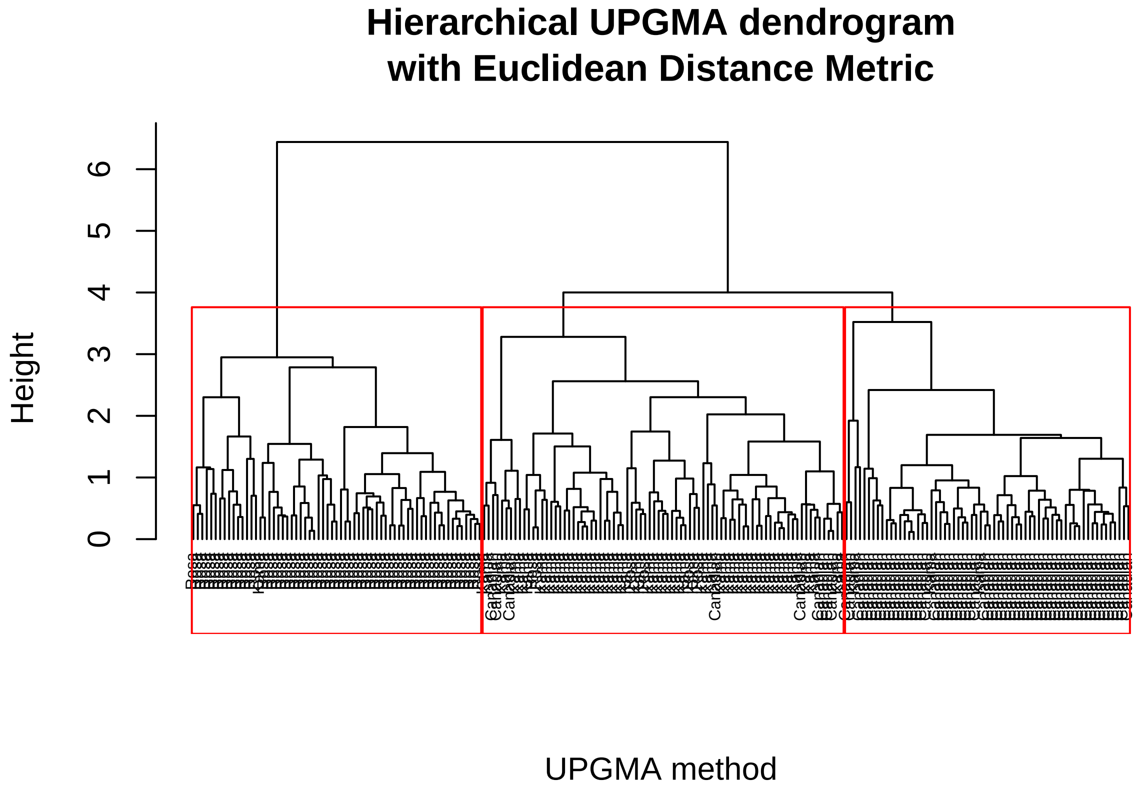

Figure.1

```
groups<- cutree(hc, 3)
table(groups, df$species)
```

```
##
## groups Canadian Kama Rosa
##      1        9   66    6
##      2       61    3    0
##      3        0    1   64
```