hw06

April 16, 2020

#

Heirarchial vs K-Means/K-Modes Clustering

0.0.1 Author & Notes

Skylar Trendley

Data Mining

Homework 6

0.0.2 Concept Description:

Using the silhouette method, an optimal number of clusters will be determined for the particular dataset. Then, using k-Means/Modes as well as heirarchial clustering, a system will be constructed to cluster unknown animals into categories based on their attributes. The system will take numeric values and group them based on their distance from a center point.

0.0.3 Data Collection:

The data has been provided by Perry B. Koob. It consists of 18 attributes: a descriptor attribute (name), 15 boolean attributes, and two numeric attributes.

0.0.4 Example Description:

Attribute Information: 1. animal name: Nominal attribute that indicates the animal. Unique for each instance.
2. hair: Nominal boolean attribute that describes if the animal has hair. 3. feathers: Nominal boolean attribute that describes if the animal has feathers. 4. eggs: Nominal boolean attribute that describes if the animal lays eggs.
5. milk: Nominal boolean attribute that describes if the animal produces milk.
6. airborne: Nominal boolean attribute that describes if the animal has the capability of flight.
7. aquatic: Nominal boolean attribute that describes if the animal has the capability to breathe under water.
8. predator: Nominal boolean attribute that describes if the animal is a predator.
9. toothed: Nominal boolean attribute that describes if the animal has teeth.
10. backbone: Nominal boolean attribute that describes if the animal has a backbone.
11. breathes: Nominal boolean attribute that describes if the animal breathes air.
12. venomous: Nominal boolean attribute that describes if the animal produces venom.
13. fins: Nominal boolean attribute that describes if the animal has fins.

14. legs: Numeric classification that describes the number of legs an animal has (set of values: {0,2,4,5,6,8})
15. tail: Nominal boolean attribute that describes if the animal has a tail.
16. domestic: Nominal boolean attribute that describes if the animal has been domesticated.
17. catsize: Nominal boolean attribute that describes if the animal is catsized.
18. gestation: Numeric classification that describes if the animal has live birth (in days)

All attributes that are currently listed as nominal will need to be converted into numeric values. This means replacing the true/false boolean attributes with 0 for false, and 1 for true.

0.0.5 Data Import and Wrangling:

[23]:
```
#import libraries
library(klaR) #kmodes

#import the dataset
data <-read.csv(file ='../src-data/animal-taxonomy.csv',stringsAsFactors=TRUE)

#set gestation and type to null as they are not clusterable attributes
data$gestation <- NULL
data$type <- NULL

#display the dataset
head(data,5)
```

A data.frame: 5 × 17

	animal.name <fct>	hair <lgl>	feathers <lgl>	eggs <lgl>	milk <lgl>	airborne <lgl>	aquatic <lgl>	predator <lgl>	t
1	aardvark	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	T
2	antelope	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	T
3	bass	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	T
4	bear	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	T
5	boar	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	T

[24]:
```
#convert the nominal data to numeric
cols <- sapply(data, is.logical)
data[,cols] <- lapply(data[,cols], as.numeric)
head(data,5)
```

A data.frame: 5 × 17

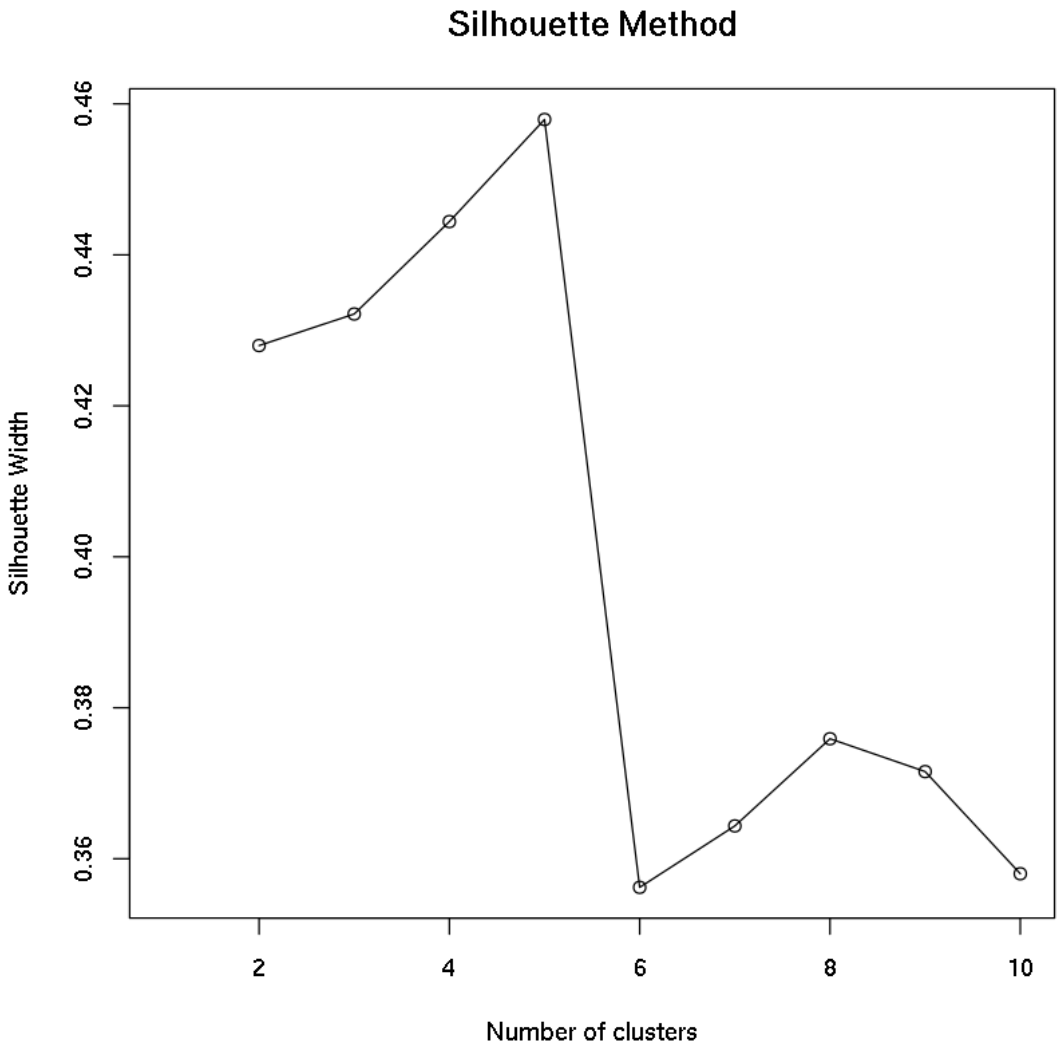
	animal.name <fct>	hair <dbl>	feathers <dbl>	eggs <dbl>	milk <dbl>	airborne <dbl>	aquatic <dbl>	predator <dbl>	to
1	aardvark	1	0	0	1	0	0	1	1
2	antelope	1	0	0	1	0	0	0	1
3	bass	0	0	1	0	0	1	1	1
4	bear	1	0	0	1	0	0	1	1
5	boar	1	0	0	1	0	0	1	1

0.0.6 Mining or Analytics:

The first step we will take is to determine the optimal amount of clusters. Using the silhouette method, a validation metric is utiliized which allows us to find an aggregated measure of the similiarities between neighboring clusters. This metric ranges -1 to 1, where higher values are better.

[86]:
```
#PAM
gower_dist <- daisy(data, metric = "gower",type = list(logratio = 3))

sil_width <- c(NA)

for(i in 2:10){

  pam_fit <- pam(gower_dist,
                 diss = TRUE,
                 k = i)

  sil_width[i] <- pam_fit$silinfo$avg.width

}

# Plot sihouette width (higher is better)

plot(1:10, sil_width,
     xlab = "Number of clusters",
     ylab = "Silhouette Width",
     main = "Silhouette Method")
lines(1:10, sil_width)
```

```
Warning message in daisy(data, metric = "gower", type = list(logratio = 3)):
"binary variable(s) 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17 treated
as interval scaled"
```



After calculating silhouette width for the clusters, we see that 5 clusters yields the highest value. First, we will use the k-Modes method. k-Modes functions similarly to k-Means, but it is optimized to better deal with catagorical data. Then, we will compare it to heirarchial clustering to see which model best suits the dataset.

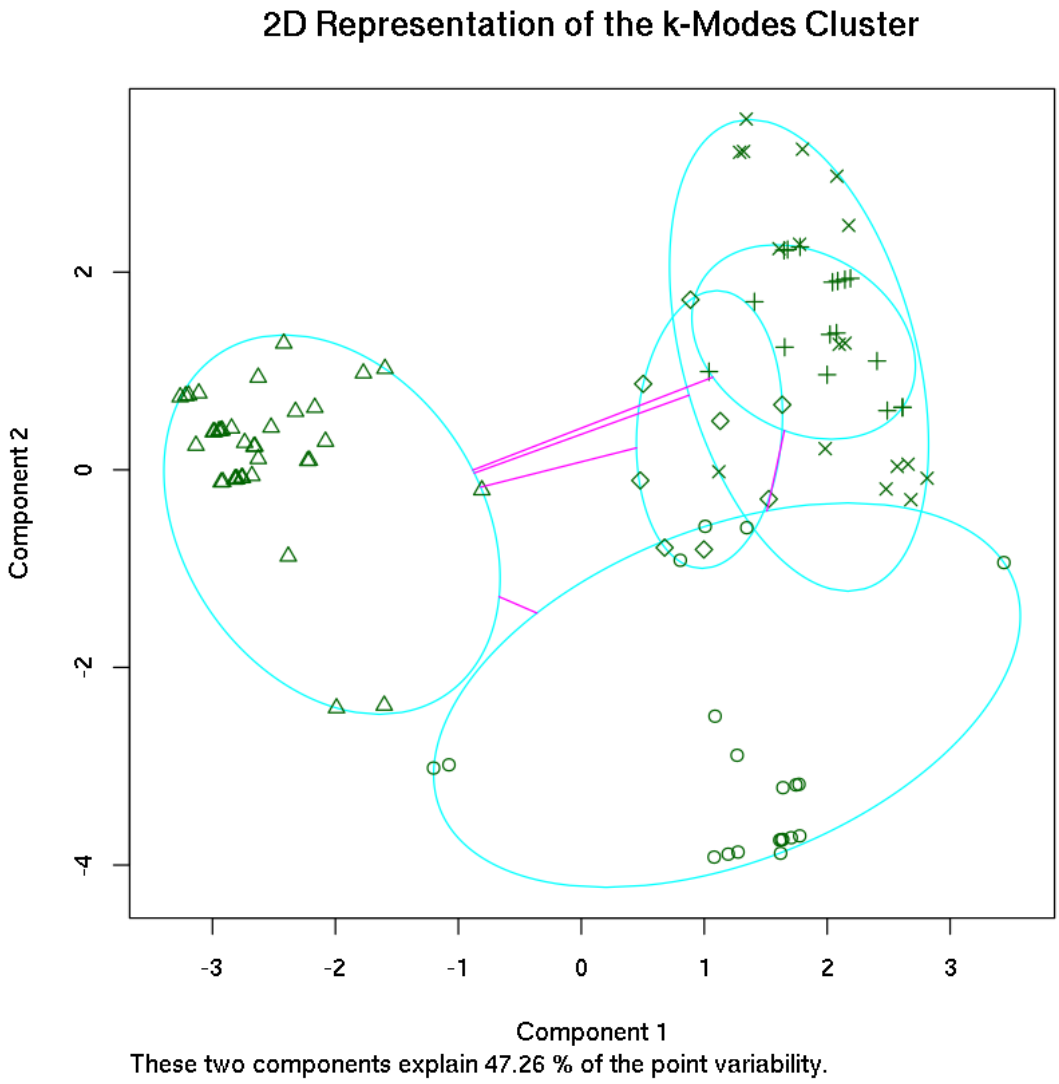
0.0.7 Evaluation: k-Modes (Part I)

[85]:
```
#set a seed to make it reproducable
set.seed(1)

#use kmods with a cluster size of 4
cluster.results <-kmodes(data[,2:17], 5, iter.max = 10, weighted = FALSE )
```

```
#show the results of the kmodes cluster
cluster.results
```

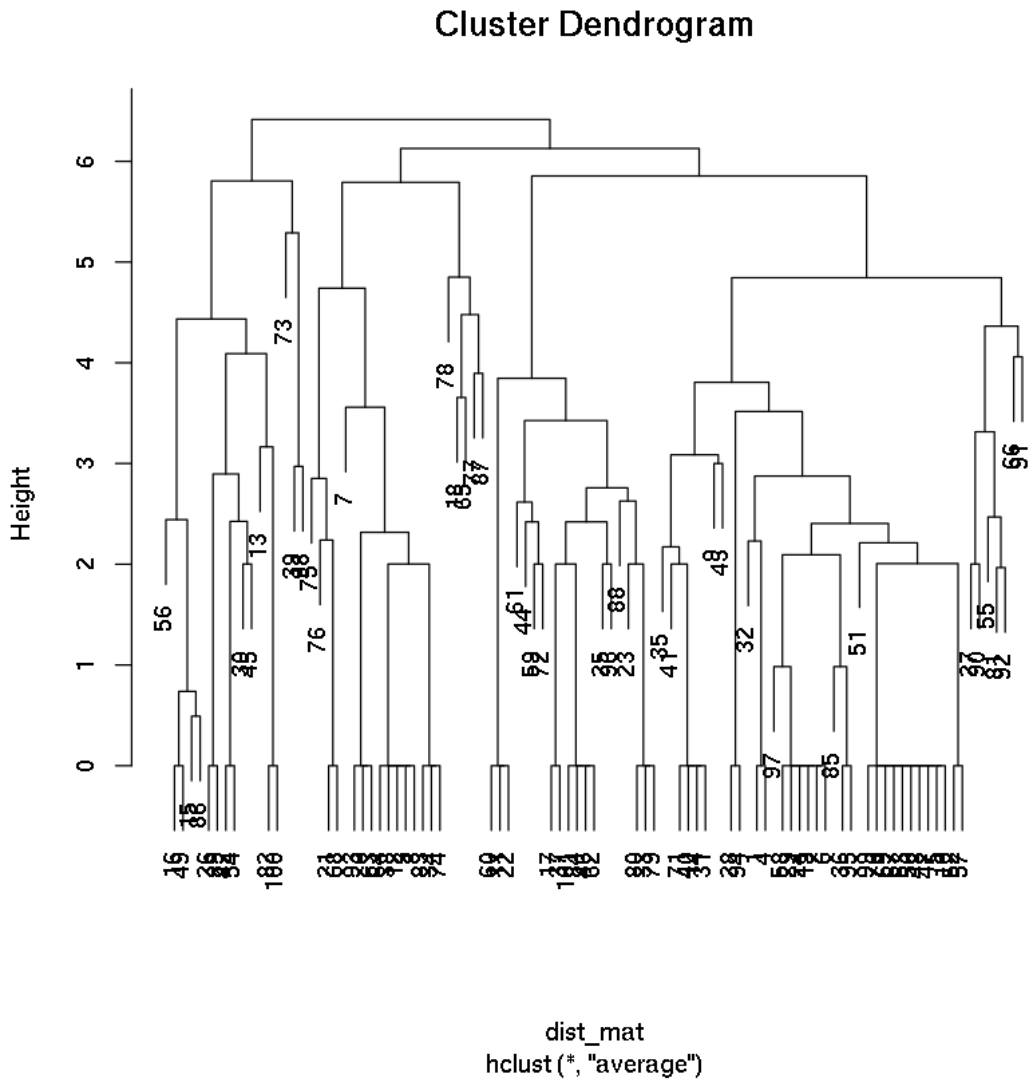
```
K-modes clustering with 5 clusters of sizes 20, 39, 17, 17, 8

Cluster modes:
  hair feathers eggs milk airborne aquatic predator toothed backbone breathes
1    0        0    1    0        0       1        1       1        1        0
2    1        0    0    1        0       0        1       1        1        1
3    0        1    1    0        1       0        0       0        1        1
4    0        0    1    0        0       0        0       0        0        1
5    0        0    1    0        0       0        1       0        1        1
  venomous fins legs tail domestic catsize
1        0    1    0    1        0       0
2        0    0    4    1        0       1
3        0    0    2    1        0       0
4        0    0    6    0        0       0
5        0    0    2    1        0       0

Clustering vector:
  [1] 2 2 1 2 2 2 1 1 2 2 3 1 4 2 4 4 3 1 2 1 1 3 3 2 3 4 1 2 2 4 2 2 3 1 2 2 3
 [38] 1 4 2 2 4 2 5 4 3 2 2 4 2 2 2 2 4 1 4 2 2 3 3 5 3 1 1 5 2 2 1 2 2 2 5 5 1
 [75] 2 2 1 1 3 3 5 4 1 3 2 4 1 3 4 4 5 5 1 2 2 3 2 4 2 4 3

Within cluster simple-matching distance by cluster:
[1] 38 60 19 42 20

Available components:
[1] "cluster"    "size"       "modes"      "withindiff" "iterations"
[6] "weighted"
```

[92]:
```
# plot the result
library(cluster)
clusplot(data,cluster.results$cluster, main='2D Representation of the k-Modes␣
 ↪Cluster')
```



0.0.8 Evaluation: Heirarchial Clustering (Part II)

[87]:
```
#preprocess the data
seeds_df_sc <- as.data.frame(scale(data[,2:17]))

#create a distance matrix via euclidean
dist_mat <- dist(seeds_df_sc, method = 'euclidean')

#build the dendogram
hclust_avg <- hclust(dist_mat, method = 'average')
plot(hclust_avg)
```

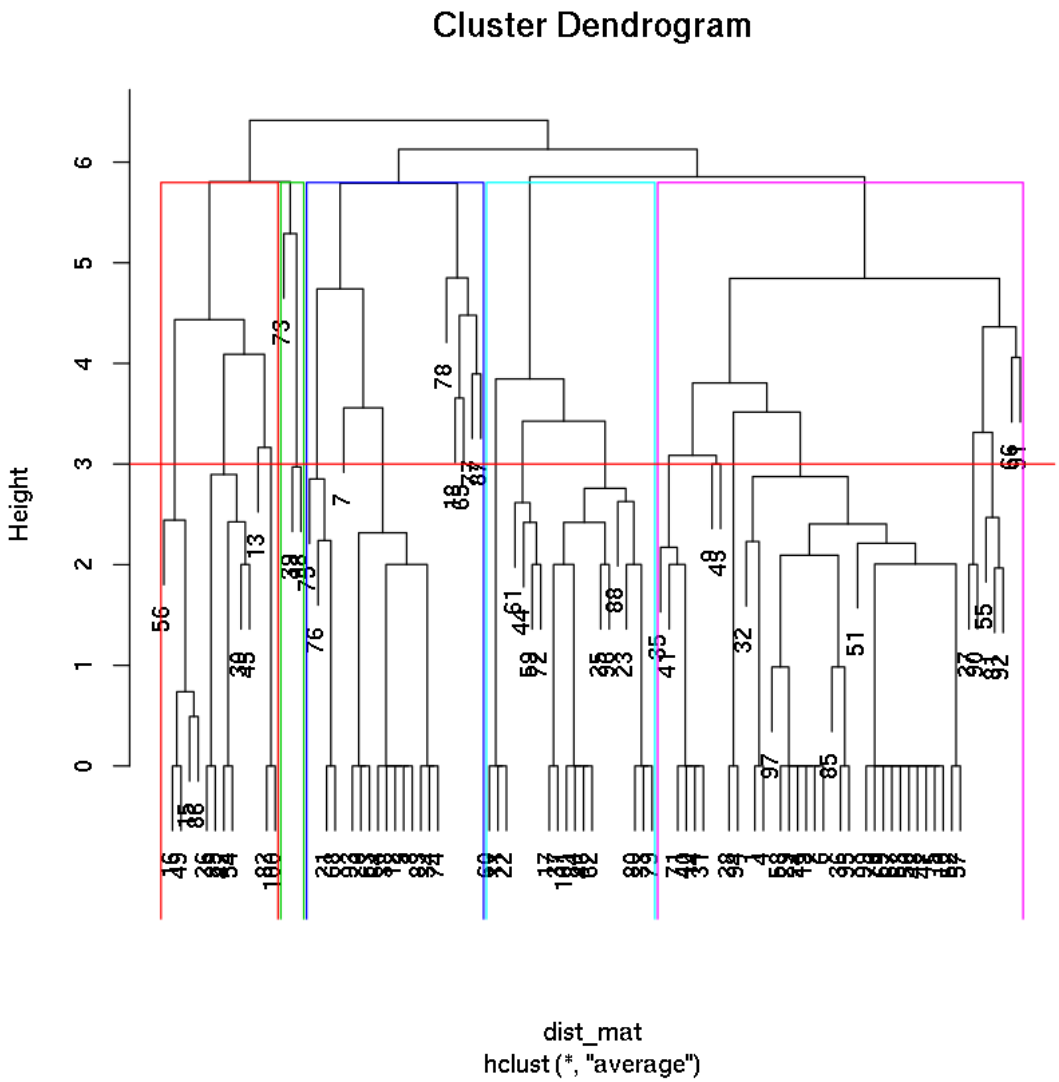


[90]:
```
#cut the dendogram into 5 clusters
cut_avg <- cutree(hclust_avg, k = 5)

#mutate the data to append the cut with specific cluster amount
suppressPackageStartupMessages(library(dplyr))
seeds_df_cl <- mutate(data[,2:17], cluster = cut_avg)
count(seeds_df_cl,cluster)
```

A tibble: 5 × 2

cluster <int>	n <int>
1	43
2	21
3	20
4	14
5	3

[91]:
```
#show the results with the desired cluster cut
plot(hclust_avg)
rect.hclust(hclust_avg , k = 5, border = 2:6)
abline(h = 3, col = 'red')
```



0.0.9 Results:

When comparing the two clustering algorithms, both have their benefits and flaws. k-Means/Modes is typically reliant on euclidean distances while heirarchial can handle virtually any distance, k-Means/Modes requires a seed and can therefore produce differing results if not reproducable while heirarchial is consistent, and k-Means/Modes is less computationally taxing on larger datasets than its heirarchial counterpart. In my findings, I believe that k-Means/Modes performed better for this particular dataset as the results were easier to interpret and it performed better computationally.

0.0.10 References

https://www.r-bloggers.com/clustering-mixed-data-types-in-r https://dabblingwithdata.wordpress.com/2016/10/categorical-data-with-r/ https://www.datacamp.com/community/tutorials/hierarchical-clustering-R https://medium.com/@davidmasse8/unsupervised-learning-for-categorical-data-dd7e497033ae

[ ]: