Clustering

Vladimir Sokolov

10/11/2022

This notebook explores smoke detector data from Kaggle.

Load the smoke_detection_iot.csv. Simplify variables

```
df <- read.csv("smoke_detection_iot.csv")
df <- df[,c(3:14,16)]
df<-df[-c(8,9)]
str(df)
```

```
## 'data.frame':    62630 obs. of  11 variables:
##  $ Temperature.C.: num  20 20 20 20 20.1 ...
##  $ Humidity...   : num  57.4 56.7 56 55.3 54.7 ...
##  $ TVOC.ppb.     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ eCO2.ppm.     : int  400 400 400 400 400 400 400 400 400 400 ...
##  $ Raw.H2        : int  12306 12345 12374 12390 12403 12419 12432 12439 12448 12453 ...
##  $ Raw.Ethanol   : int  18520 18651 18764 18849 18921 18998 19058 19114 19155 19195 ...
##  $ Pressure.hPa. : num  940 940 940 940 940 ...
##  $ NC0.5         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NC1.0         : num  0 0 0 0 0 ...
##  $ NC2.5         : num  0 0 0 0 0 0 0 0 0 2.78 ...
##  $ Fire.Alarm    : int  0 0 0 0 0 0 0 0 0 0 ...
```

Check for null values.

```
sapply(df, function(x) sum(is.na(x)))
```

```
## Temperature.C.    Humidity...      TVOC.ppb.      eCO2.ppm.         Raw.H2
##              0              0              0              0              0
##    Raw.Ethanol  Pressure.hPa.          NC0.5          NC1.0          NC2.5
##              0              0              0              0              0
##     Fire.Alarm
##              0
```

#Kmeans Clustering #{r} `library(NbClust) set.seed(12345) i <- sample(1:nrow(df), nrow(df)*.95, replace=FALSE) train <- df[i,] test <- df[-i,] nc <- NbClust(test, min.nc=2, max.nc=6, method="kmeans") #`

```
library(flexclust)
```

```
## Loading required package: grid
## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4
```

```
set.seed(12345)
fit1 <- kmeans(df[,1:10], 3, nstart=25)
fit1$size
```

```
## [1]   966 61079   585
```

```
fit1$centers
```

```
##   Temperature.C. Humidity...   TVOC.ppb.   eCO2.ppm.   Raw.H2 Raw.Ethanol
## 1       20.12066    20.61642 59805.6729 10452.3996 11517.50    16761.76
## 2       15.76063    49.18435   807.0404   509.6108 12967.31    19809.75
## 3       31.02127    27.32017 24898.3932  1264.7573 12700.53    18901.50
##   Pressure.hPa.       NC0.5       NC1.0       NC2.5
## 1      936.8059 24914.24143 9108.930282 3310.48498
## 2      938.6847    33.53831    8.517547    2.14639
## 3      935.6751  7973.88434 5865.216810 2879.39092
```

```
ct<- table(df$Fire.Alarm, fit1$cluster)
ct
```

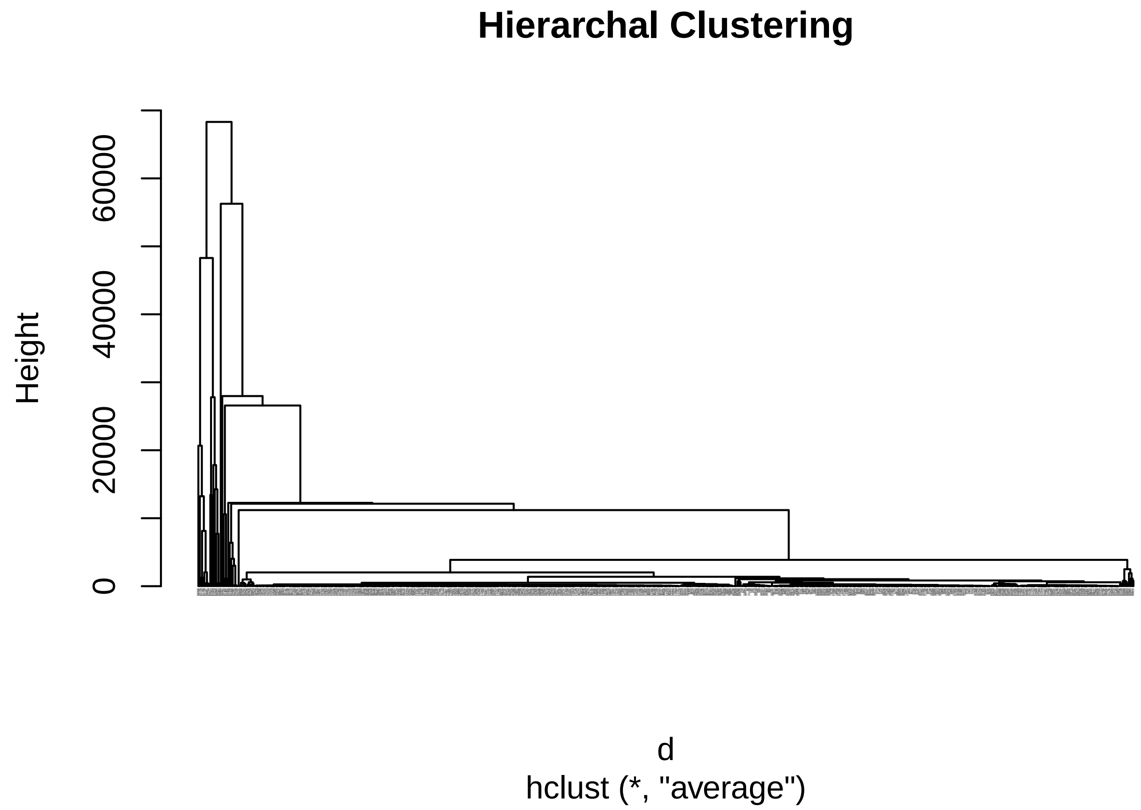
```
##
##         1     2     3
##   0   966 16445   462
##   1     0 44634   123
```

```
randIndex(ct)
```

```
##        ARI
## 0.06607624
```

#Hierarchal Clustering

```
set.seed(12345)
i <- sample(1:nrow(df), nrow(df)*.99, replace=FALSE)
tst <- df[-i,]
d <- dist(tst)
fit.average <-hclust(d, method="average")
plot(fit.average, hang=-1, cex=.1, main="Hierarchal Clustering")
```



#Model-Based Clustering

#`{r} library(mclust) fit3 <-Mclust(tst) plot(fit3) summary(fit3) #`

#Results

knn identified 3 clusters which was interesting. Hierarchical clustering wasn't very insightful based on this data set and was difficult to see anything useful from it. Model based did not like to finish loading on my computer even when given a small amount of data so nothing was learned from it. The one time it worked it produced a set of graphs showing relationships between variables which could have been useful.