K-Means and K-Medoids Clustering on Simulated Data

Halil Bisgin

Optional: Can you try to determine the shape of a data point based on its real class label and color based on its cluster membership? I thought I could kind of determine the shape of a data point based on its real class label and its color based on its cluster membership, but I am uncertain based on the the two plots fro k-means and k-medoids where the smaller clusters and colors don't quite seem to follow the same pattern. I thin I'll have to observe more data.

Introduction

In this analysis, we simulate data to apply both **K-Means** and **K-Medoids** clustering algorithms, compare their results, and visualize the clustering outcomes.

We also measure the time taken for each clustering method.

Load Required Libraries

```
# Install libraries if not already installed
# install.packages("cluster")
# install.packages("factoextra")
# install.packages("ggplot2")
# install.packages("tictoc")

# Load necessary libraries
library(cluster)      # For K-Medoids (PAM)
library(factoextra)   # For visualization
library(ggplot2)      # For plotting
library(tictoc)       # For measuring execution time
```

```
# Set random seed for reproducibility
set.seed(42)

# Create 3 clusters with normal distribution
# Change n from 300 to 1500 in increments of 200
n <- 1500  # Total number of data points
cluster1 <- data.frame(x = rnorm(n/3, mean = 2, sd = 0.5), y = rnorm(n/3, mean = 2, sd = 0.5))
cluster2 <- data.frame(x = rnorm(n/3, mean = 6, sd = 0.5), y = rnorm(n/3, mean = 6, sd = 0.5))
cluster3 <- data.frame(x = rnorm(n/3, mean = 10, sd = 0.5), y = rnorm(n/3, mean = 2, sd = 0.5))

# Combine clusters into one dataset
data <- rbind(cluster1, cluster2, cluster3)
colnames(data) <- c("feature1", "feature2")
```

```r
# Normalize the data
scaled_data <- scale(data)
```

Apply K-Means Clustering

```r
tic("K-Means")  # Start timer
kmeans_result <- kmeans(scaled_data, centers = 6, nstart = 10)
toc()  # Stop timer
```

```
## K-Means: 0 sec elapsed
```

Apply K-Medoids Clustering

```r
tic("K-Medoids")  # Start timer
kmedoids_result <- pam(scaled_data, 5)
toc()  # Stop timer
```

```
## K-Medoids: 0.73 sec elapsed
```

Prep data for visualization

```r
data$KMeans_Cluster <- as.factor(kmeans_result$cluster)
data$KMedoids_Cluster <- as.factor(kmedoids_result$clustering)
```

Plotting k-means clusters

```r
p1 <- ggplot(data, aes(x = feature1, y = feature2, color = KMeans_Cluster)) +
  geom_point(size = 3) +
  ggtitle("K-Means Clustering") +
  theme_minimal()

# Print the K-Means plot
print(p1)
```

Plotting k-medoids clusters

```
p2 <- ggplot(data, aes(x = feature1, y = feature2, color = KMedoids_Cluster)) +
  geom_point(size = 3) +
  ggtitle("K-Medoids Clustering") +   theme_minimal()

# Print the K-Means plot
print(p2)
```


```
# Load the Iris dataset
library(datasets)
data(iris)
summary(iris)
```

```
##    Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```
df <- iris
scaled_data <-df

# I'm afraid from this point on I did something wrong and didn't know how to incorporate the Iris data
```

Apply K-Means Clustering for Iris

```
# Apply k-means clustering with k = 3 (for the three species of Iris)
set.seed(20) # for reproducibility
iris_cluster <- kmeans(iris[, 1:4], centers = 3)

# Add the cluster assignments to the iris dataset
iris$Cluster <- as.factor(iris_cluster$cluster)
```

Apply K-Medoids Clustering for Iris

```
tic("K-Medoids")  # Start timer
kmedoids_result <- pam(scaled_data, 3)
toc()  # Stop timer
```

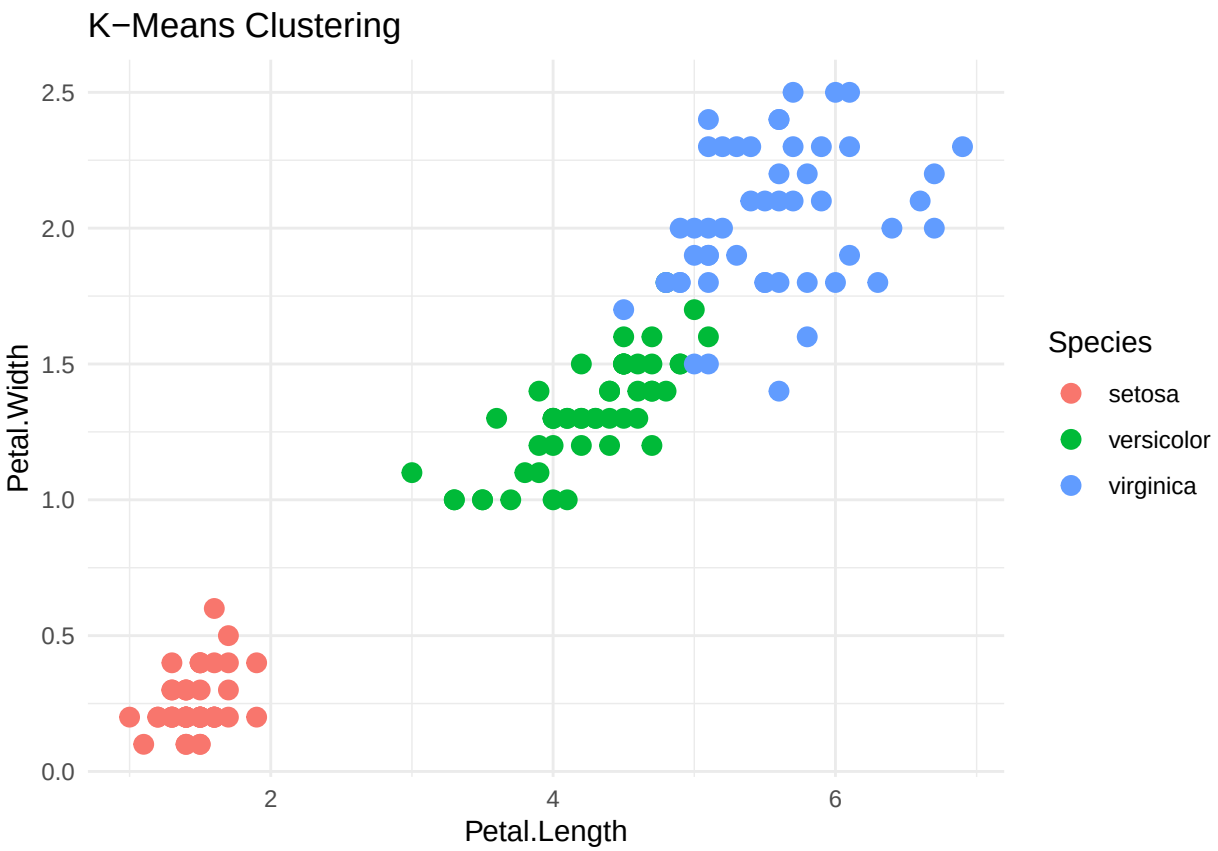
```
## K-Medoids: 0 sec elapsed
```

Prep data for visualization for Iris

```
data$KMeans_Cluster <- as.factor(kmeans_result$cluster)
data$KMedoids_Cluster <- as.factor(kmedoids_result$clustering)
```

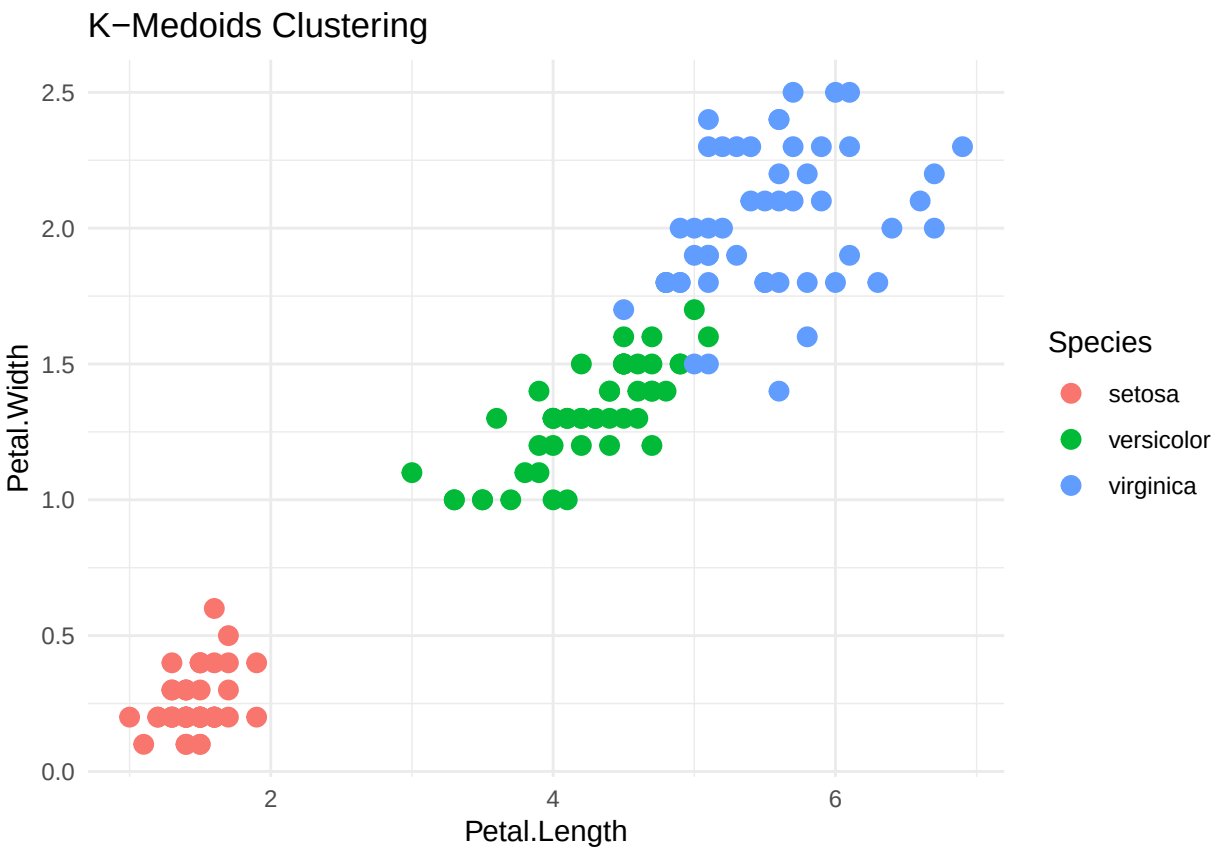
Plotting k-means clusters fro Iris

```
p1 <- ggplot(scaled_data, aes(x = Petal.Length, y = Petal.Width, color = Species)) +
  geom_point(size = 3) +
  ggtitle("K-Means Clustering") +
  theme_minimal()

# Print the K-Means plot
print(p1)
```



Plotting k-medoids clusters for Iris

```
p2 <- ggplot(scaled_data, aes(x = Petal.Length, y = Petal.Width, color = Species)) +
  geom_point(size = 3) +
  ggtitle("K-Medoids Clustering") +
  theme_minimal()

# Print the K-Medoids plot
print(p2)
```


K−Medoids Clustering
Petal.Width
Petal.Length
2.5
2.0
1.5
1.0
0.5
0.0
2
4
6
Species
setosa
versicolor
virginica