САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ им. ПЕТРА ВЕЛИКОГО

Институт прикладной математики и механики

Высшая школа прикладной математики и вычислительной физики

Отчёт

по лабораторной работе №4 «Кластеризация» по дисциплине «Системы искусственного интеллекта»

Студентка гр. 3630201/70101 _____ О. В. Саксина

Преподаватель _____ Л. В. Уткин

Санкт-Петербург,
2021г.

Содержание

1 Задание 1

1.1 Постановка задачи

Разбейте множество объектов из набора данных pluton в пакете «cluster» на 3 кластера методом центров тяжести (kmeans). Сравните качество разбиения в зависимости от максимального числа итераций алгоритма.

1.2 Реализация

Была обучена модель методов центров тяжести при максимальном числе итераций равном 3, 5, 10, 20. В результате каждая модель давала разбиение с отношением межкластерной дисперсии к внутрикластерной равной 92,1 %, не зависящим от данного параметра, т.к. алгоритм разбиения останавливался после второй итерации.

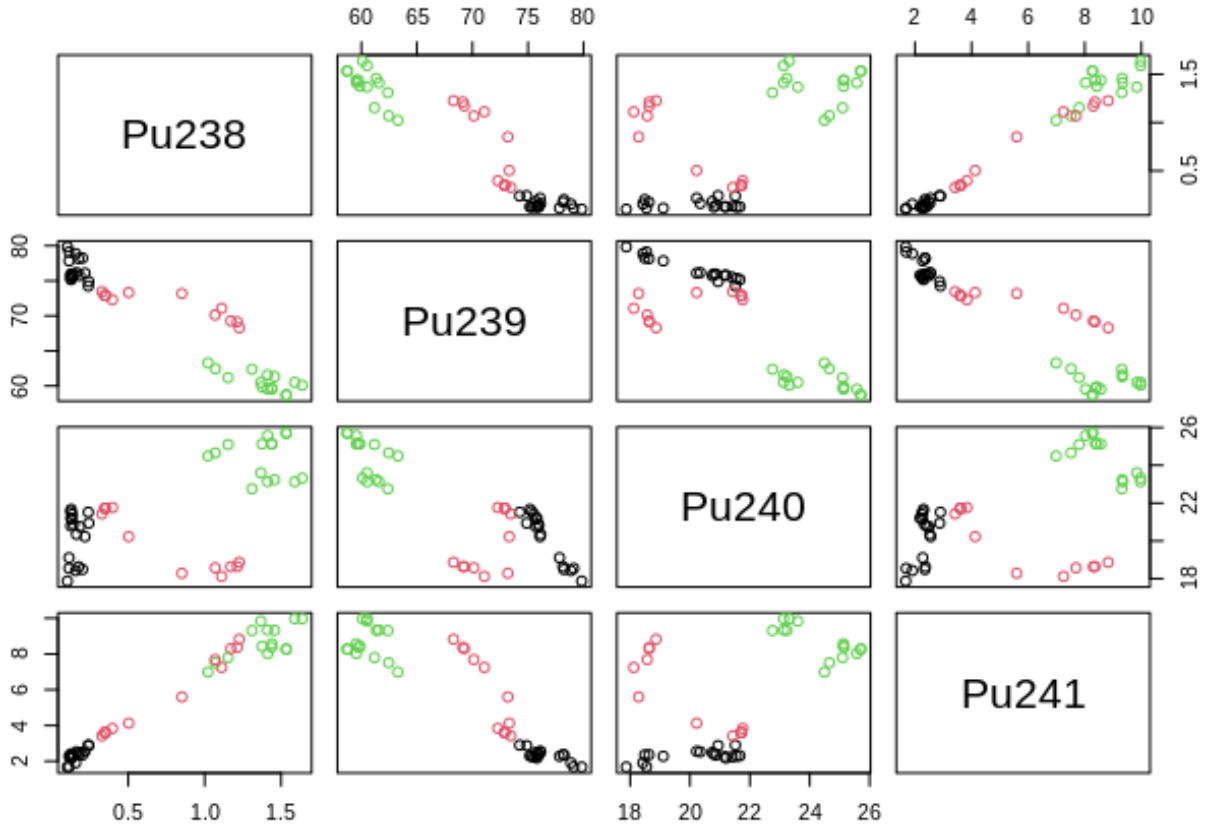

Рис. 1: Разбиение множества объектов на кластеры

2 Задание 2

2.1 Постановка задачи

Сгенерируйте набор данных в двумерном пространстве, состоящий из 3 кластеров, каждый из которых сильно “вытянут” вдоль одной из осей. Исследуйте качество кластеризации методом clara в зависимости от 1) использования стандартизации; 2) типа метрики. Объясните полученные результаты.

2.2 Реализация

Был сгенерирован набор данных, представленный на рис. 2. Качество кластеризации при разных параметрах модели, определяемое как ширина силуэта, показано в таблице 1. Значение силуэта показывает, насколько объект похож на свой кластер по сравнению с другими кластерами. Для всей кластерной структуры чем ближе данная оценка к 1, тем лучше. Видно, что евклидово и манхэттонское расстояния по качеству примерно одинаковы и оценки не зависят от использования стандартизации (последнее можно объяснить характером

распределения: положение точек задавалось по нормальному распределению с одинаковыми параметрами для всех трёх кластеров), жаккардово расстояние показало худшие результаты.

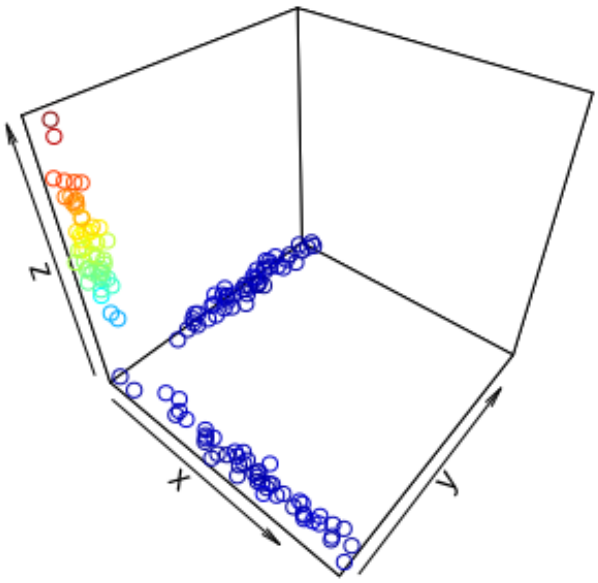

Рис. 2: Распределение данных

	Тип метрики		
Использование стандартизации	euclidean	manhattan	jaccard
TRUE	0.73	0.77	0.52
FALSE	0.73	0.77	0.01

Таблица 1: Зависимость ширины силуэта от типа метрики и использования стандартизации

3 Задание 3

3.1 Постановка задачи

Постройте дендрограмму для набора данных votes.repub в пакете «cluster» (число голосов, поданных за республиканцев на выборах с 1856 по 1976 год). Строки представляют 50 штатов, а столбцы - годы выборов (31). Проинтерпретируйте полученный результат.

3.2 Реализация

Построенная дендограмма представлена на рис. 3. Дендрограмма показывает степень близости отдельных объектов и кластеров, а также наглядно демонстрирует в графическом виде последовательность их объединения или разделения. В данном случае первое разделение всего множества объектов выделяет в левое поддерево группу штатов, где было отдано меньше всего голосов (в штате Миссисипи в среднем за все годы было отдано 22.33407 голосов, что является наименьшим значением), а в правое – штаты, где больше голосов (в штате Вермонт в среднем за все годы было отдано 66.31677 голосов, что является наибольшим значением).

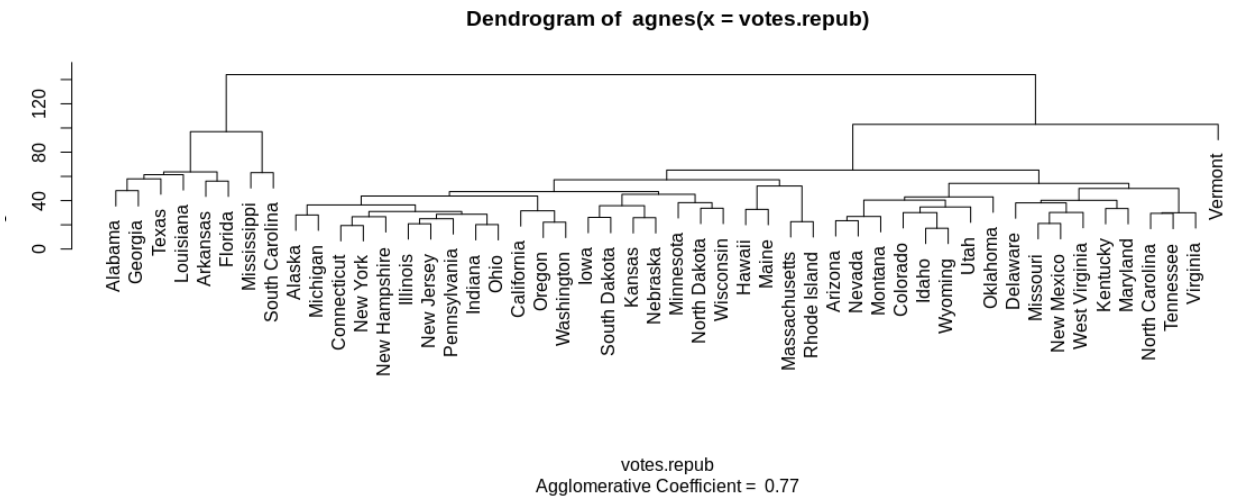

Рис. 3: Дендрограмма для набора данных votes.repub

4 Задание 4

4.1 Постановка задачи

Постройте дендрограмму для набора данных animals в пакете «cluster». Данные содержат 6 двоичных признаков для 20 животных. Переменные
[, 1] war теплокровные;
[, 2] fly летающие;
[, 3] ver позвоночные;
[, 4] end вымирающие;
[, 5] gro живущие в группе;
[, 6] hai имеющие волосяной покров.
Проинтерпретируйте полученный результат.

4.2 Реализация

Построенная дендограмма представлена на рис. 4. Каждое поддерево объединяет максимально похожих друг на друга животных. Например на втором разбиении в один кластер попали bee (пчела) и fly (муха), а в другой – cpl (гусеница) и spi (паук). Можно рассматривать эту дендограмму, как подобие филогенетического дерева.

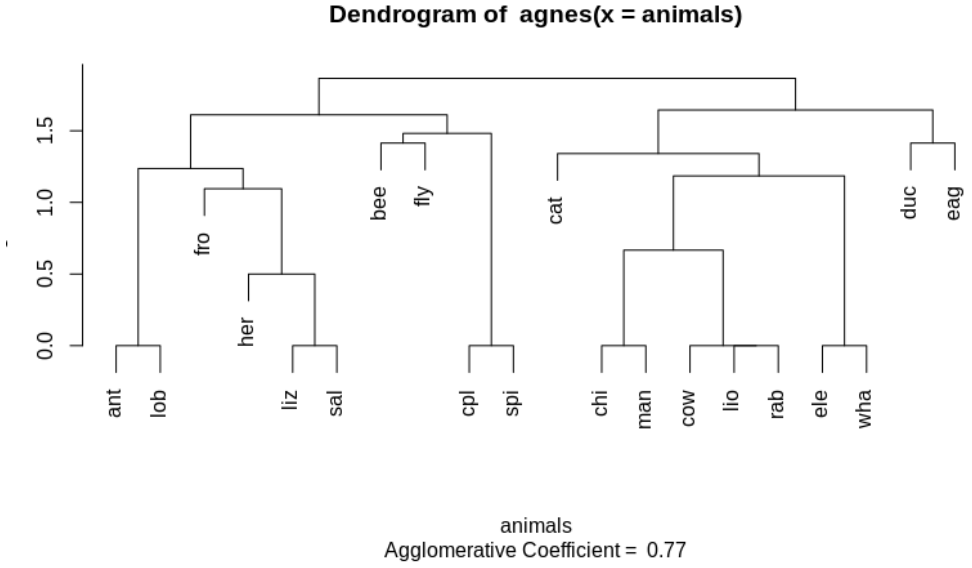

Рис. 4: Дендрограмма для набора данных animals

5 Задание 5

5.1 Постановка задачи

Рассмотрите данные из файла seeds_dataset.txt, который содержит описание зерен трех сортов пшеницы: Kama, Rosa and Canadian. Признаки:
1. область A,
2. периметр P,
3. компактность C $= 4 * \pi * A/P^2$,
4. длина зерна,
5. ширина зерна,
6. коэффициент ассиметрии,
7. длина колоска.

5.2 Реализация

Посмотрев на матрицу рассеяния для данных, представленную на рис. 5, можно увидеть, что есть параметры, которые лучше других объясняют разделение зёрен на сорта. Применив метод главных компонент, получим Компоненту 1, которая объясняет 71% дисперсии исходных данных, и Компоненту 2, объясняющую 17% дисперсии исходных данных. Иллюстрация этого приведена на рис. 6.

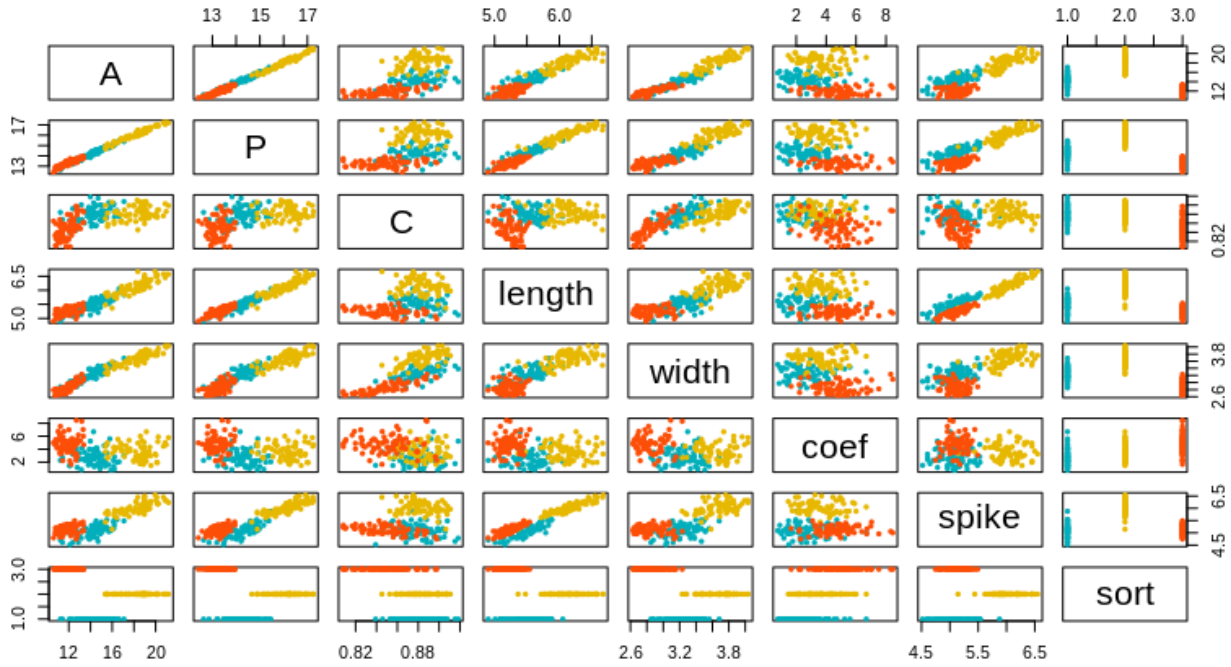

Рис. 5: Матрица рассеяния для данных seeds_dataset

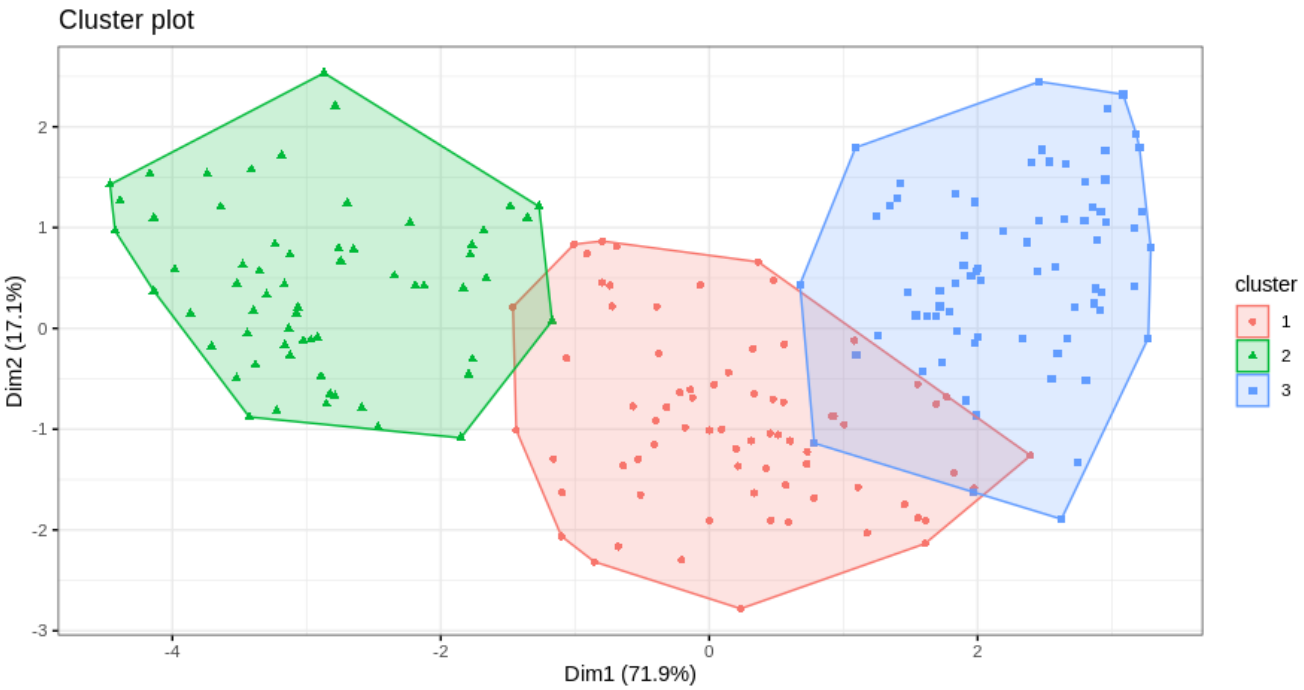

Рис. 6: Кластеризация по главным компонентам

Приложение

Задание 1

```
library(cluster)
library(factoextra)

c1 = kmeans(pluton, centers = 3, iter.max = 1)
c2 = kmeans(pluton, centers = 3, iter.max = 2)
c3 = kmeans(pluton, centers = 3, iter.max = 3,
            algorithm = "Hartigan-Wong",trace=TRUE )
summary(c3)
fviz_cluster(c2, data = pluton,
             geom = "point",
             ggtheme = theme_bw()
)
plot(pluton, col = c2$cluster)
```

Задание 2

```
library(cluster)
library(scatterplot3d)
library("plot3D")
library(factoextra)

n = 50
mmax = 150
mmin = 0
sdmax = 50
sdmin = 5

data = data.frame(x = c(rnorm(n, mmax, sdmax), rnorm(n, mmin, sdmin), rnorm(n, mmin, sdmin)),
                  y = c(rnorm(n, mmin, sdmin), rnorm(n, mmax, sdmax), rnorm(n, mmin, sdmin)),
                  z = c(rnorm(n, mmin, sdmin), rnorm(n, mmin, sdmin), rnorm(n, mmax, sdmax)))
scatterplot3d(x = data$x, y = data$y, z = data$z)
scatter3D(x = data$x, y = data$y, z = data$z)

cl_ef = clara(data, k = 3, metric = "euclidean", stand = FALSE)
cl_mf = clara(data, k = 3, metric = "manhattan", stand = FALSE)
cl_et = clara(data, k = 3, metric = "euclidean", stand = TRUE)
cl_mt = clara(data, k = 3, metric = "manhattan", stand = TRUE)
cl_jf = clara(data, k = 3, metric = "jaccard", stand = FALSE)
cl_jt = clara(data, k = 3, metric = "jaccard", stand = TRUE)
fviz_cluster(cl_ef, data = data,
             geom = "point",
             ggtheme = theme_bw()
)
plot(cl_ef)
plot(cl_mf)
plot(cl_et)
plot(cl_mt)
plot(cl_jf)
plot(cl_jt)
plot(data, col=cl_ef$clustering)
```

Задание 3

```
library(cluster)

data = votes.repub
plot(agnes(votes.repub))
states = row.names(data)
means = c()
```

```
for (i in 1:length(states)){
  means[i] = mean(unlist(data[states[i],]), na.rm=TRUE)
}
rank = mapply(c, states, means , SIMPLIFY = FALSE )
df = data.frame(states=states, means=means)
df = df[order(df$means),]
```

Задание 4

```
library(cluster)

plot(agnes(animals))
mona(animals)
```

Задание 5

```
library(cluster)
library(factoextra)

setwd("/home/olga/MyProjects/Polikek/ML/Cluster/datasets")

data = read.delim("seeds_dataset.txt")
my_cols <- c("#00AFBB", "#E7B800", "#FC4E07")
pairs(data[,1:8], pch = 19,  cex = 0.5,
      col = my_cols[data$sort])

sorts = data[,ncol(data)]
data = data[,-ncol(data)]

cl = clara(data, k = 3)
fviz_cluster(cl, data = data,
             geom = "point",
             ggtheme = theme_bw()
)
plot(cl)

set.seed(12345)
data2 = subset(data, select = -c(A,P))
k2 = kmeans(data2, centers = 3)
k2$betweenss/k2$totss * 100
tab = table(k2$cluster, sorts)
tab
cl = clara(data, k = 3)
fviz_cluster(cl, data = data,
             geom = "point",
             ggtheme = theme_bw()
)
pca <- prcomp(data, center = TRUE, scale = TRUE)
summary(pca)
pca$rotation
biplot(pca)
```