**Лабораторная работа №4**

**“Кластеризация”**

по дисциплине “Машинное обучение”

|  |  |
| --- | --- |
| Выполнил  студент гр. 33504/2 | Лелюхин Д.О. |
| Руководитель | Селин И.А. |

**Оглавление**

[**Первое задание** 3](#_Toc513726654)

[**Код программы** 3](#_Toc513726655)

[**Результаты** 3](#_Toc513726656)

[**Второе задание** 7](#_Toc513726657)

[**Код программы** 7](#_Toc513726658)

[**Результаты** 8](#_Toc513726659)

[**Третье задание** 10](#_Toc513726660)

[**Код программы** 10](#_Toc513726661)

[**Результаты** 11](#_Toc513726662)

[**Четвертое задание** 11](#_Toc513726663)

[**Код программы** 11](#_Toc513726664)

[**Результаты** 12](#_Toc513726665)

[**Пятое задание** 12](#_Toc513726666)

[**Код программы** 12](#_Toc513726667)

[**Результаты** 13](#_Toc513726668)

# **Первое задание**

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.

## **Код программы**

library(cluster)

data("pluton")

for(i in seq(2, 16, by = 2))

{

cl <- kmeans(pluton, 3, iter.max = i)

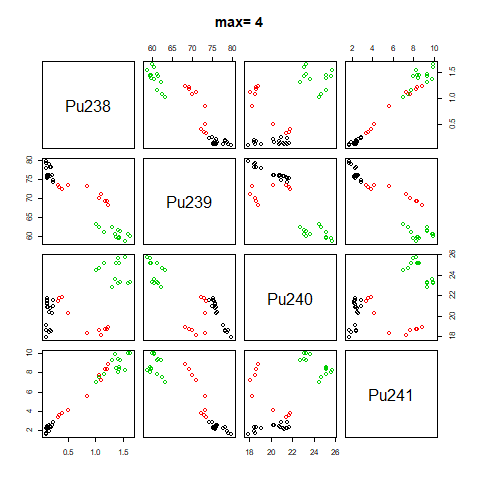
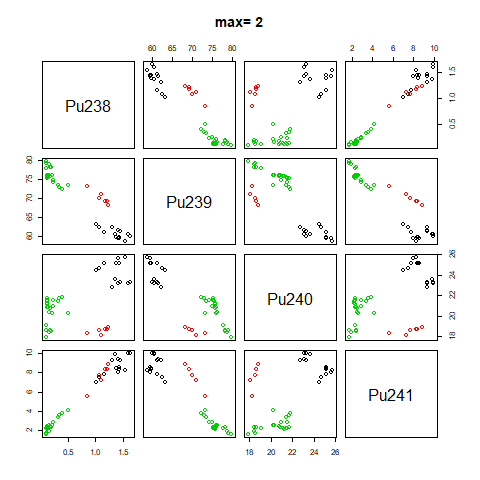
png(filename = paste(toString(i),'claster.png'))

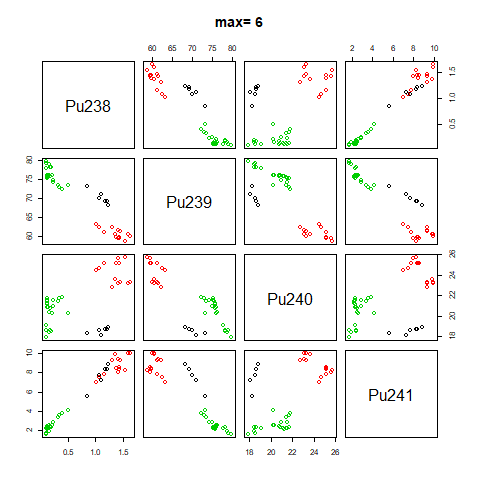
plot(pluton, col = cl$cluster, main=paste('max=',toString(i)))

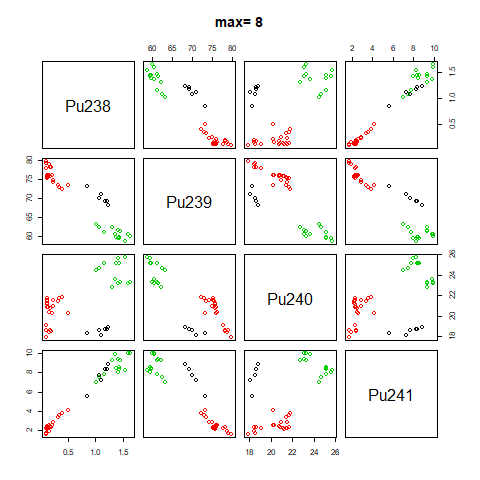
points(cl$centers, col = 1:3, pch = 8, cex=2)

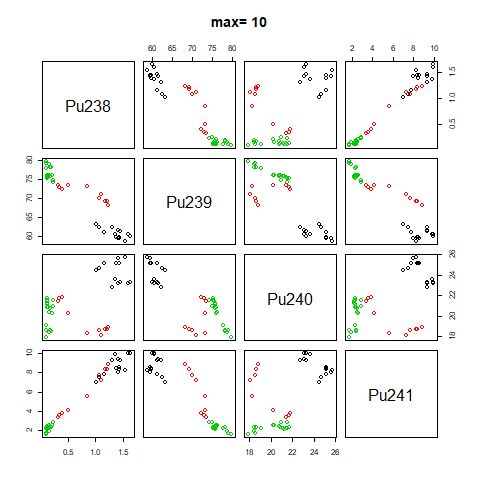
}

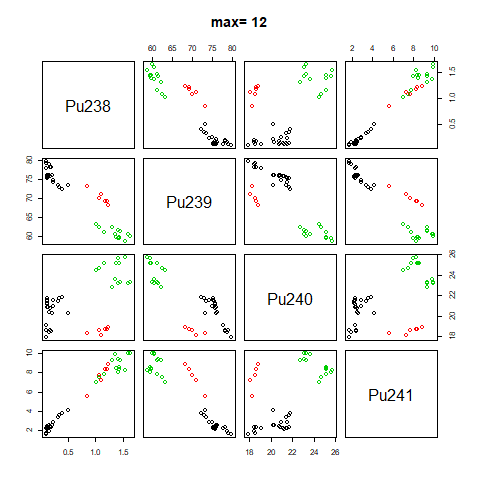
## **Результаты**

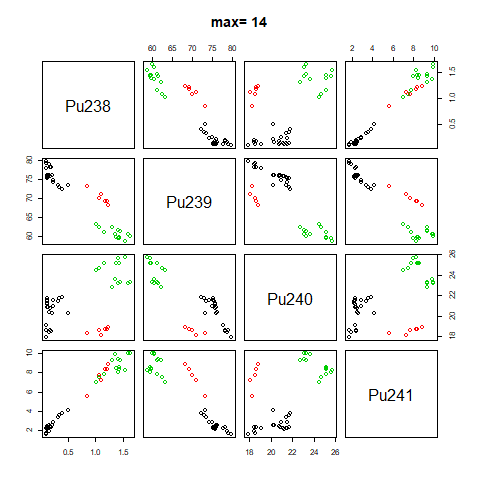


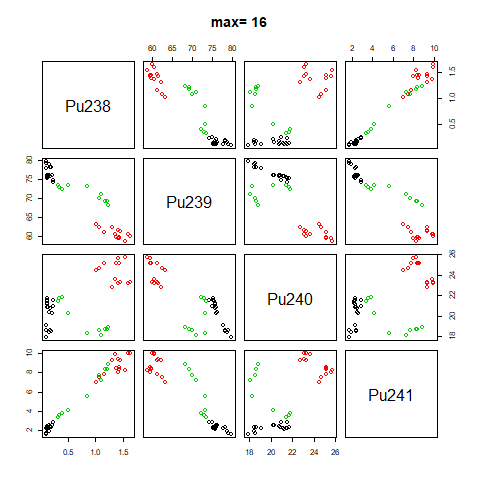












Вывод: Качество разбиения при увеличении max числа итераций увеличивается.

# **Второе задание**

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.

## **Код программы**

library(cluster)

cl1 <- c()

cl2 <- c()

for(i in seq(1,50, by=2))

{

cl1 <- c(cl1,runif(1, min=10, max=11.5))

cl1 <- c(cl1,runif(1, min=18, max=18.5))

cl1 <- c(cl1,runif(1, min=25, max=26))

cl2 <- c(cl2, i)

}

fr<-data.frame(cl2,cl1)

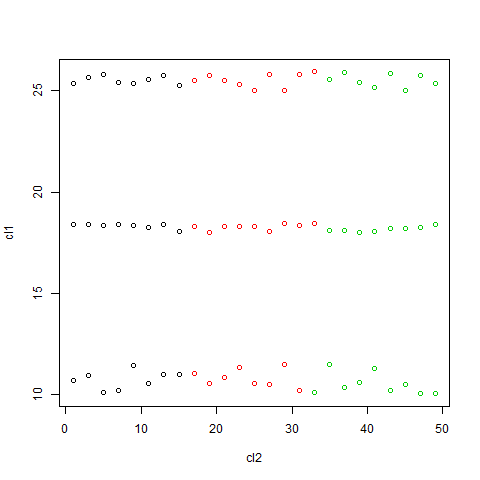
res<-clara(fr, 3,metric = "manhattan", stand = FALSE)

png(file = 'cl.jpg')

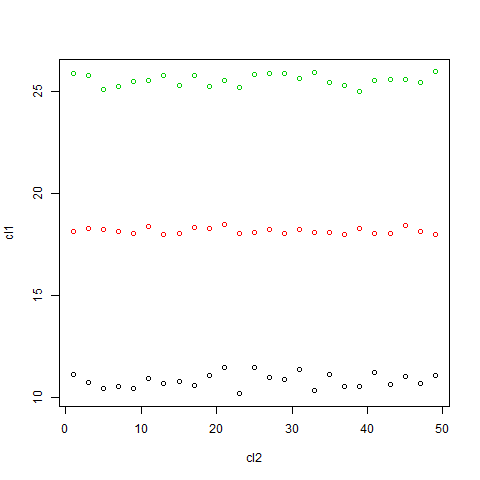
plot(fr, col=res$clustering)

## **Результаты**

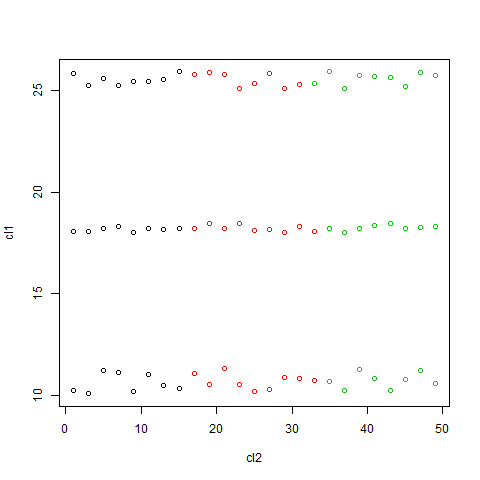
metric = "manhattan", stand = FALSE



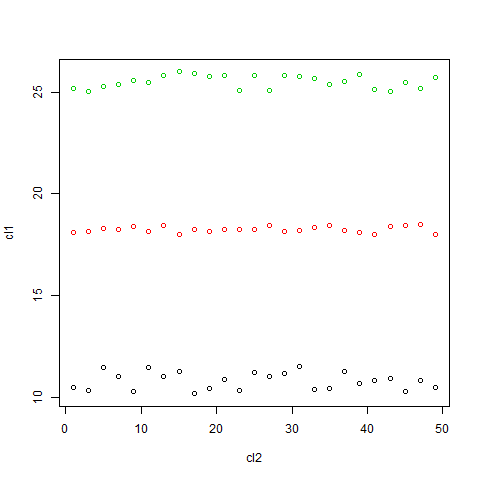
metric = "manhattan", stand = TRUE



metric = "euclidean", stand = FALSE



metric = "euclidean", stand = TRUE



Вывод: На сгенерированной выборке метод clara работает лучше всего при метрике "manhattan" и

"euclidean" и стандартизации TRUE.

# **Третье задание**

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.

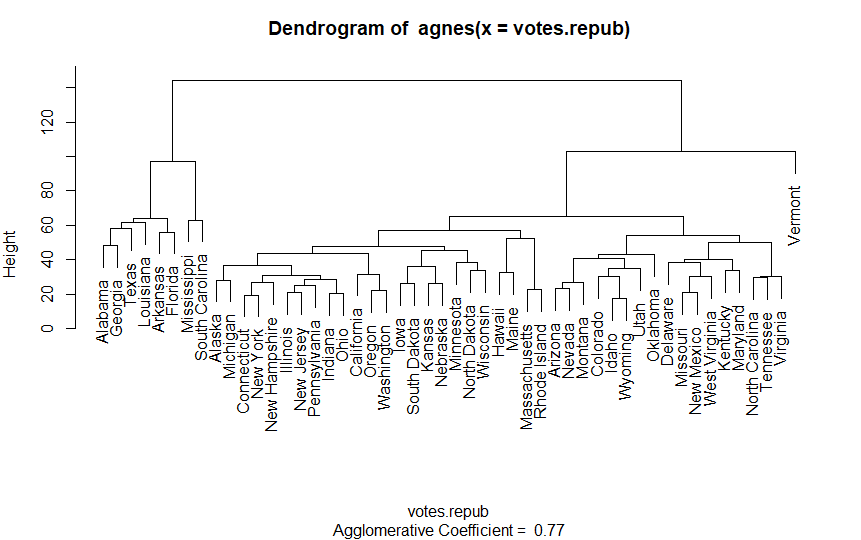
## **Код программы**

library(cluster)

data(votes.repub)

plot(agnes(votes.repub))

## **Результаты**



# **Четвертое задание**

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.

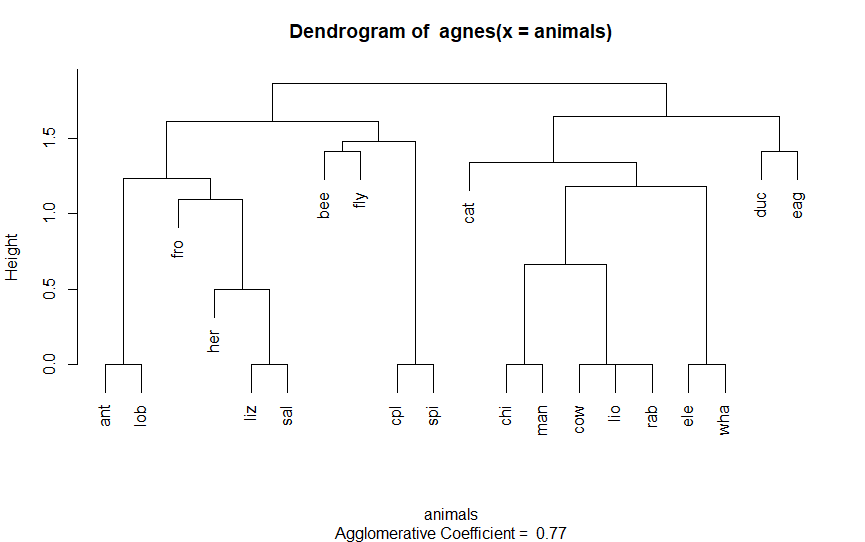
## **Код программы**

library(cluster)

data(animals)

plot(agnes(animals))

## **Результаты**



# **Пятое задание**

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.

## **Код программы**

library(cluster)

A\_raw <- read.table("seeds\_dataset.txt", stringsAsFactors = TRUE)

cl <- kmeans(A\_raw[,-8], 3, iter.max = 20)

plot(A\_raw, col = cl$cluster)

## **Результаты**

