Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский Политехнический Университет Петра Великого
—
Институт компьютерных наук и технологий
Высшая школа искусственного интеллекта

ЛАБОРАТОРНАЯ РАБОТА №5

«Кластеризация»

по дисциплине «Машинное обучение, часть1»

Выполнил: студент группы
3540201/20302 С.А. Ляхова

<подпись>

Проверил: Л.В. Уткин
д.т.н., профессор

<подпись>

Санкт-Петербург
2022

Содержание

1. Цель работы

Исследовать методы kmeans, clara, agnes пакета cluster языка R, реализующие методы k-средних, k-медоидов и агломеративной иерархической кластеризации соответственно, выполнив поставленные задачи и проанализировав результаты.

2. Формулировка задания

1) 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.

2) 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.

3) 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.

4) 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.

5) 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.

3. Ход работы

Задание №1

Данные pluton представлены 45 объектами, каждый из которых имеет четыре вещественных признака: Pu238, Pu239, Pu240 и Pu241. На рисунках 1-3 представлены кластеры с различной длительностью работы алгоритма.

Данные разбиты на 3 кластера методом центров тяжести (kmeans). Каждый квадрат на графике - результат кластеризации по двум различным признакам. С увеличением числа итераций работы алгоритма увеличивается точность построения кластеров, однако в некоторый момент улучшение кластеризации с увеличением числа итераций не наблюдается.

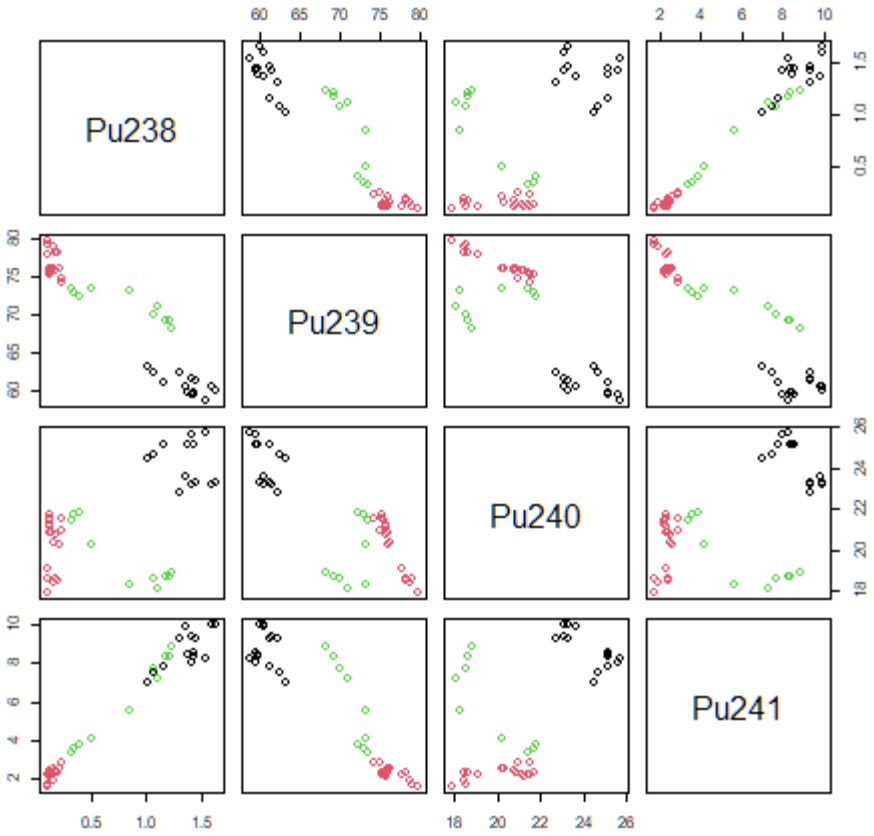

Рисунок 1. Результат кластеризации датасета pluton с 3 итерациями

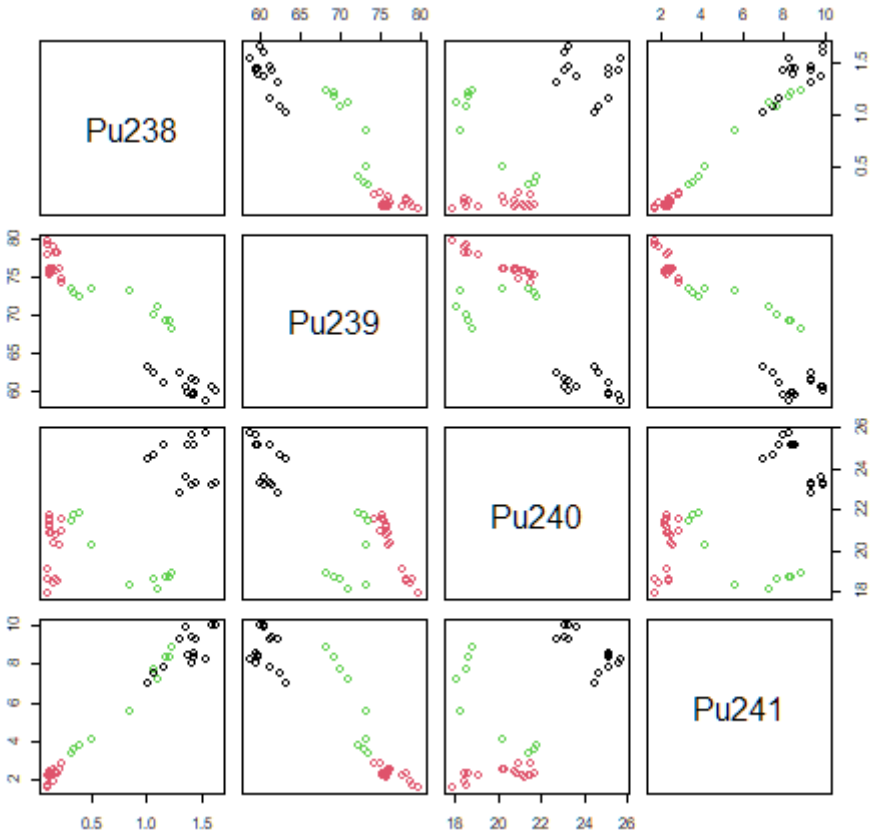

Рисунок 2. Результат кластеризации датасета pluton с 8 итерациями

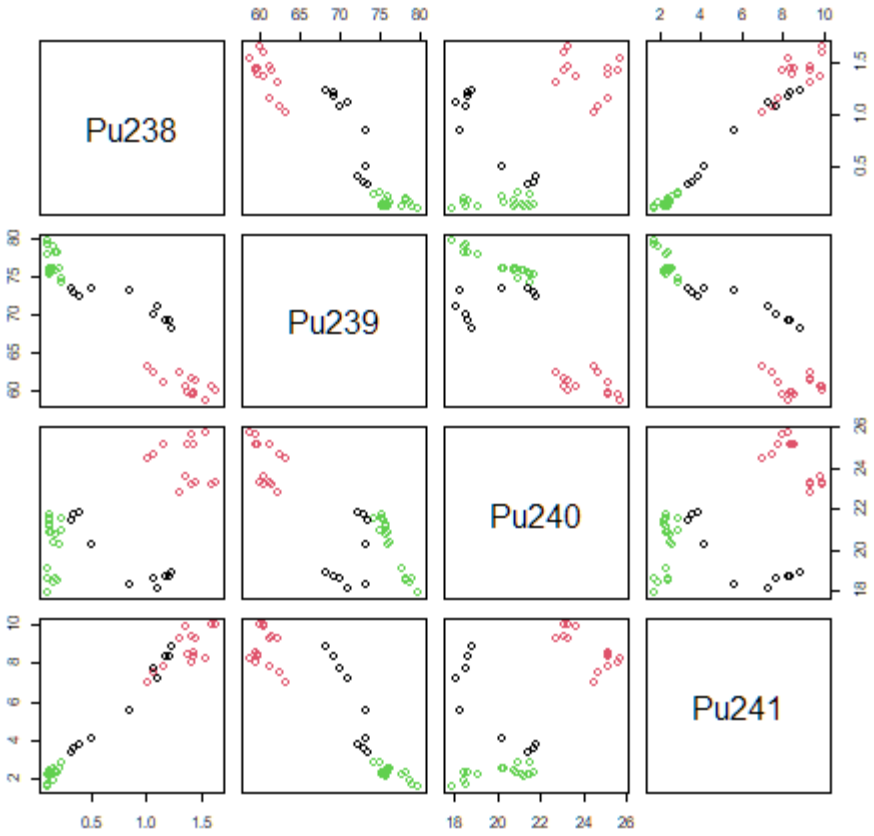

Рисунок 3. Результат кластеризации датасета pluton с 13 итерациями

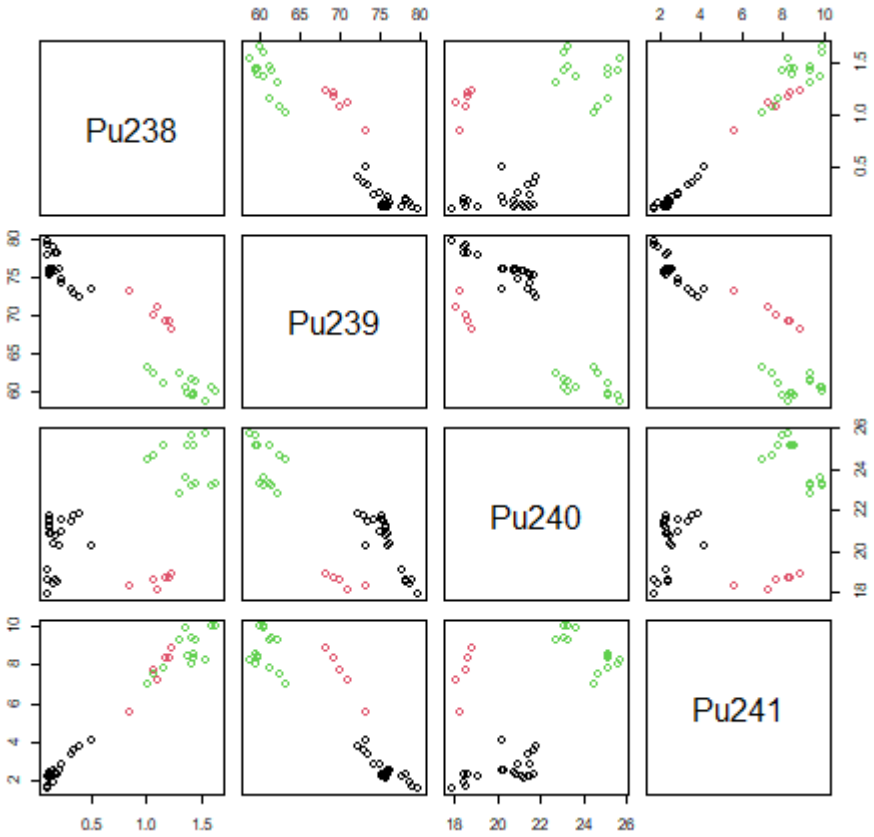

Рисунок 4. Результат кластеризации датасета pluton с 18 итерациями

С увеличением числа итераций работы алгоритма увеличивается точность построения кластеров, однако в некоторый момент улучшение кластеризации с увеличением числа итераций не наблюдается. При числе итераций =18 алгоритм показывает другие классы, на мой взгляд, более верные и логичные. Качество кластеризации при числе итераций =18 составило 91.3%

Задание №2

Генерация данных:

```
size = 100
x1 <- matrix(c(rnorm(size, mean = -40, sd = 10), rnorm(size, mean = 0, sd = 2)), ncol = 2)
x2 <- matrix(c(rnorm(size, mean = -30, sd = 5), rnorm(size, mean = 0, sd = 5)), ncol = 2)
x3 <- matrix(c(rnorm(size, mean = -20, sd = 2), rnorm(size, mean = 0, sd = 10)), ncol = 2)
```

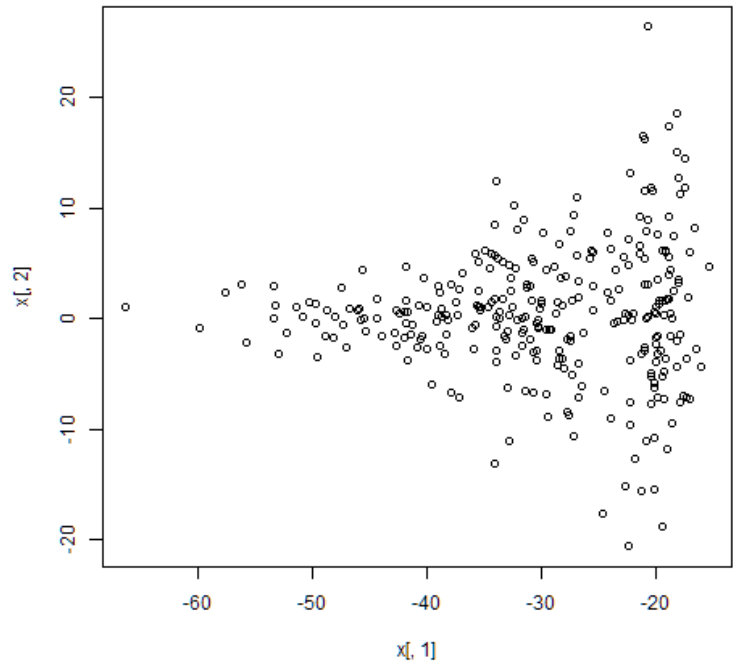

Рисунок 5. Визуализация сгенерированных данных в двумерном пространстве

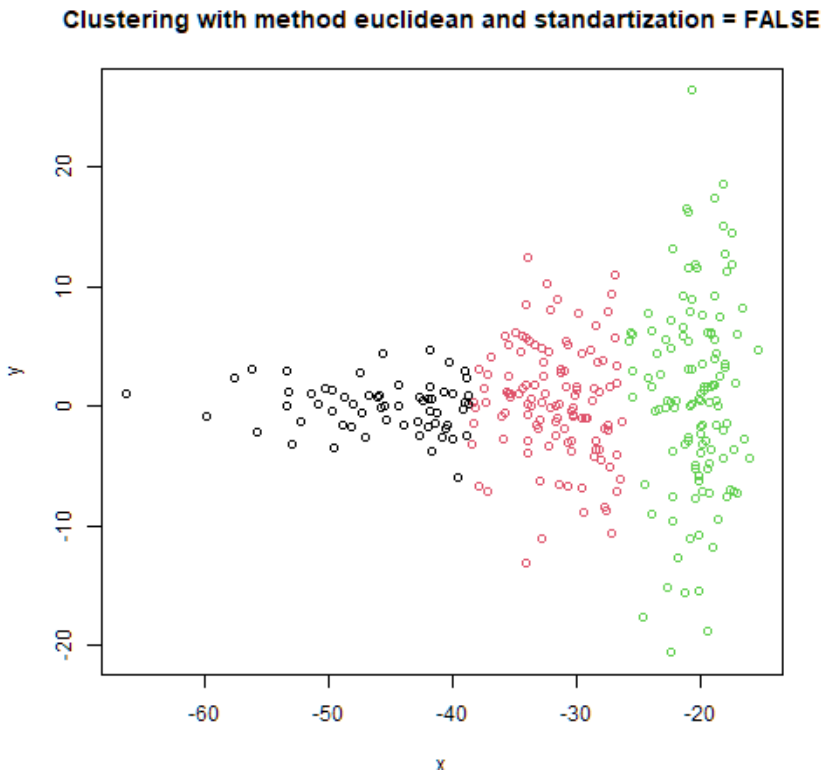

Рисунок 6. Результаты кластеризации без использования стандартизации и евклидовой метрики

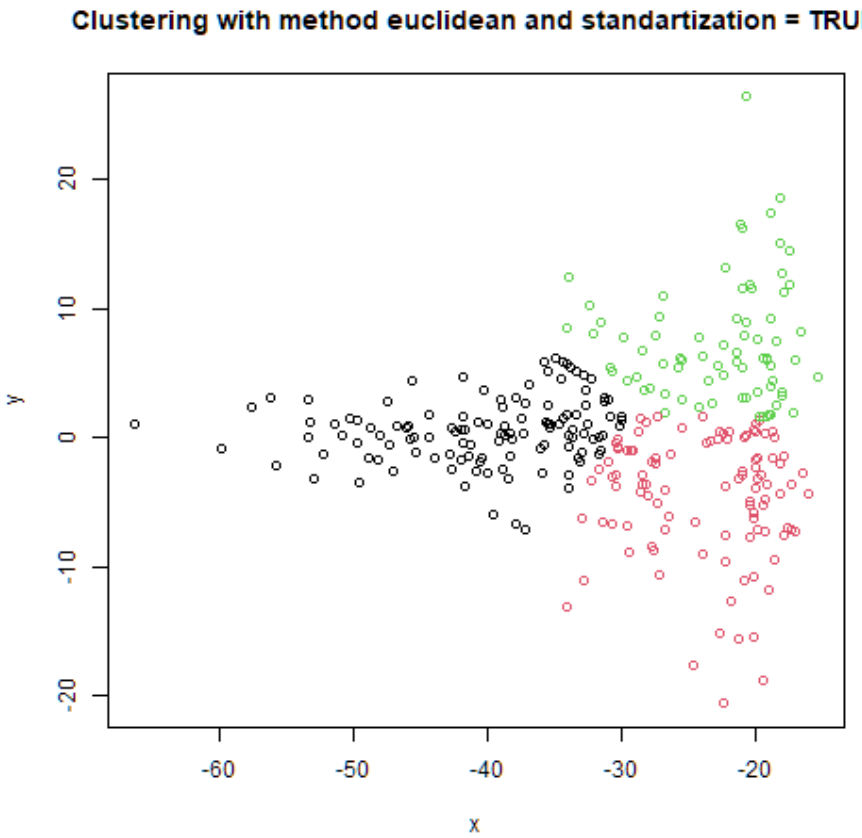

Рисунок 7. Результаты кластеризации с использованием стандартизации и евклидовой метрики

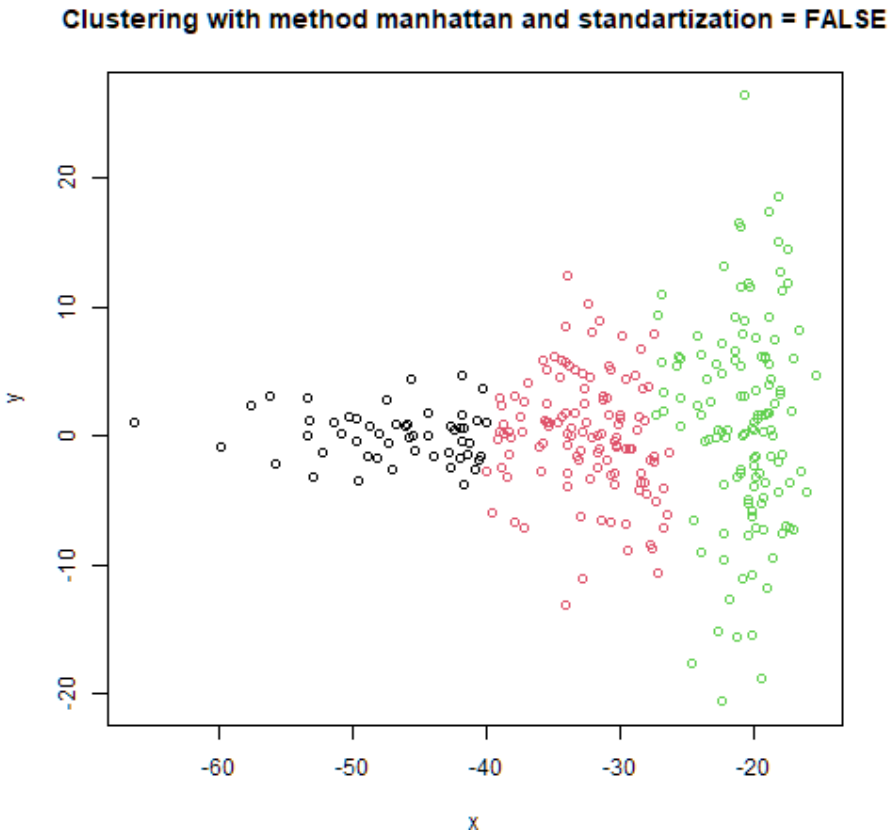

Рисунок 8. Результаты кластеризации без использования стандартизации и манхэттенской метрики

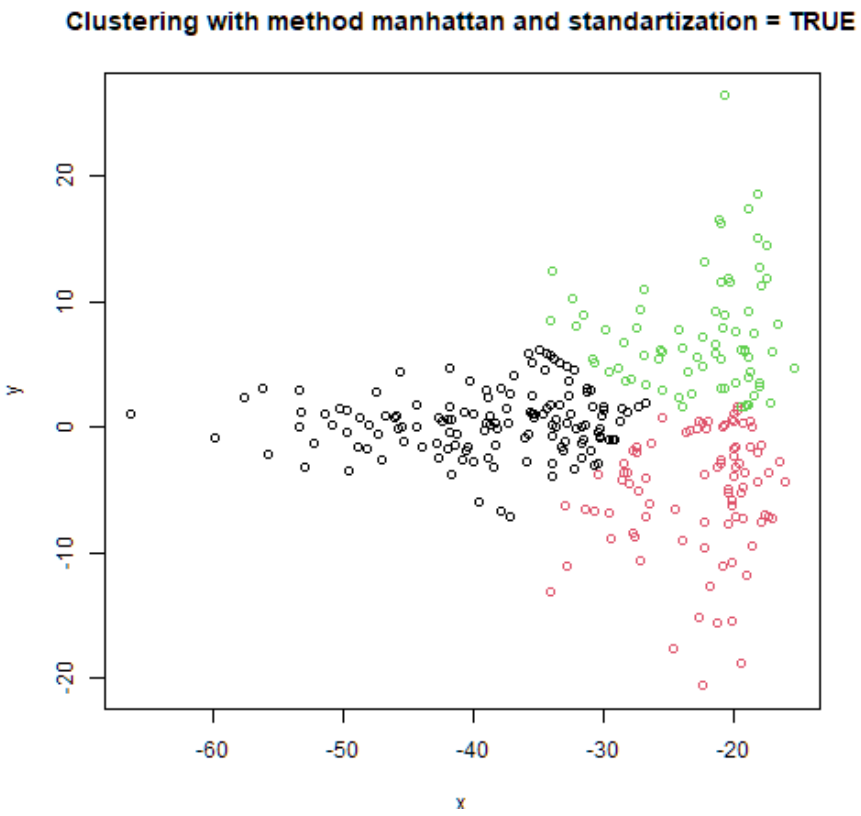

Рисунок 9. Результаты кластеризации с использованием стандартизации и манхэттенской метрики

Судя по результатам использование параметра стандартизации заметно снижает качество кластеризации. Из рисунков 6 и 8 можно сделать вывод, что кластеризация с использованием евклидовой метрики выходит немного точнее, чем с использованием манхэттенской.

Задание №3

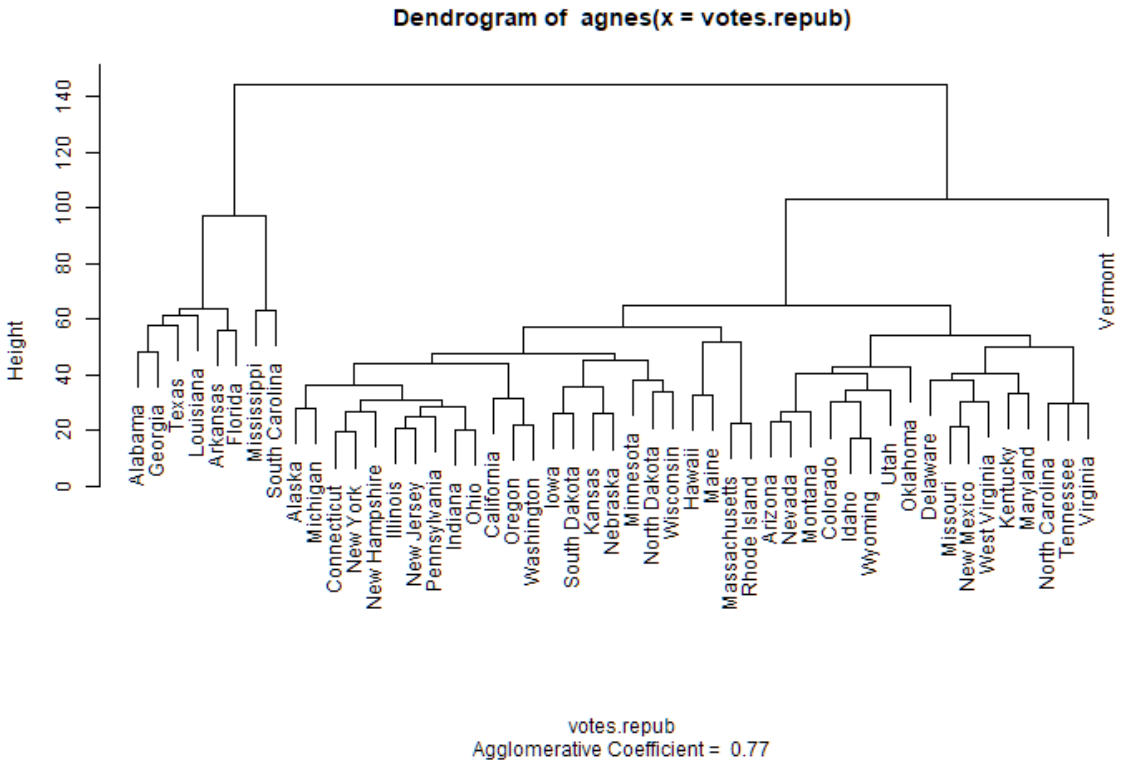

Рисунок 10. Дендрограмма датасета votes

На дендрограмме на рисунке 10 результаты кластеризации по увеличению числа голосов (по штатам), которые были отданы за республиканскую партию.

Из дендограммы видно, что наибольшую поддержку республиканцев с 1856 по 1997 года оказывал Вермонт, он выделен в отдельную ветвь. Штаты Алабама, Джорджия Техас и Луизиана были очень похожи (по поставленному выше критерию), однако Луизиана поддерживала республиканцев больше, чем Техас, а Техас поддерживал республиканцев больше, чем Алабама и Джорджия. И так далее.

Задание №4

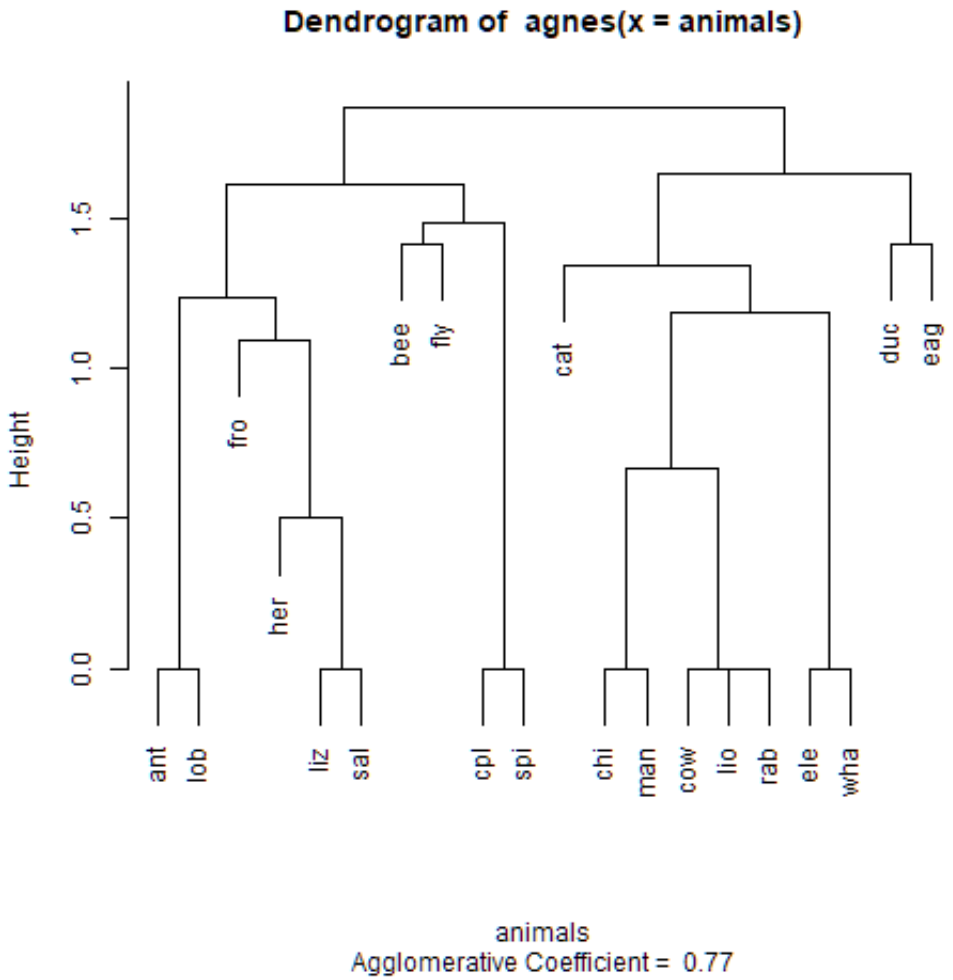

Рисунок 11. Дендрограмма, построенная на основании датасета animals

Так как данные представляют собой описание видов животных, то данная дендограмма показывает видовую близость тех или иных животных (исходя из заданных признаков).

Задание №5

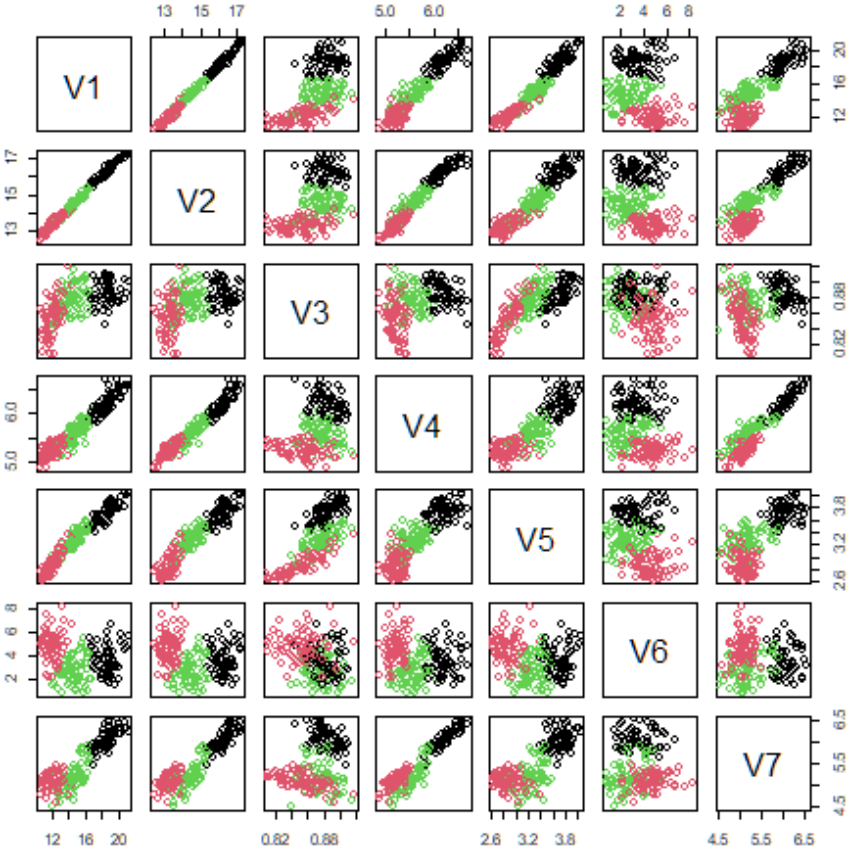

Рисунок 12. Результат кластеризации методом kmeans

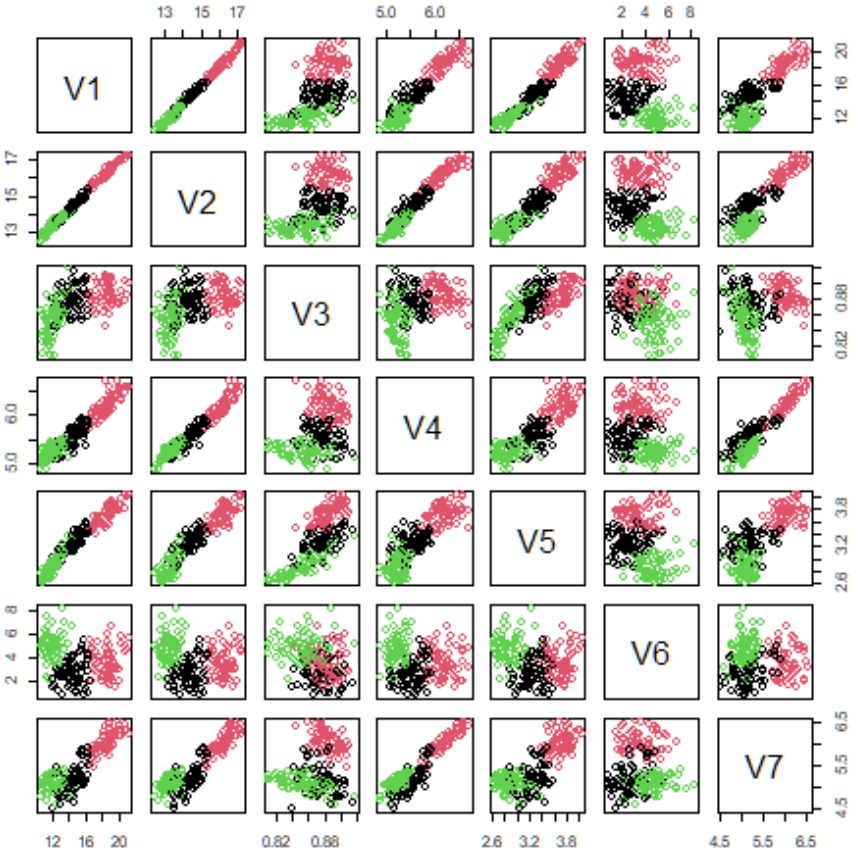

Рисунок 13. Результат кластеризации методом clara

Из рисунков видно, что визуализация данных может быть очень различна в зависимости от исследуемых признаков.

Результаты кластеризации двумя методами сильно похожи, однако имеют незначительные различия. Точность обоих методов составила 89%.

4. Вывод

Были исследованы методы кластеризации.

Метод k-средних наиболее прост в использовании, однако нуждается в ручном устранении своих недостатков, а именно – неизвестности оптимального выбора исходных центров кластеров, отсутствие гарантии достижения глобального минимума суммарного квадратичного отклонения и необходимость в заведомом известном числе кластеров.

Метод k-медоидов несколько сложнее в настройке, но лишён недостатков метода k-средних.

Иерархическая кластеризация предоставляет возможность ручного выделения кластеров на нужном уровне дендрограммы, а также не нуждается в числе кластеров в качестве параметра.

Приложение 1

```
# Задание 1 ##########################################
path <- "C:\\Users\\Софья\\Desktop\\1_курс_1_семестр\\Машинное
обучение\\Лабораторная 5. Cluster\\"

library(cluster)
data(pluton)

iterations <- c(3, 8, 13, 18)

for (i in iterations) {
  cl <- kmeans(pluton, iter.max = i, centers = 3)
  png(paste(path, "Kmeans function with", i, "iterations.png"))
  plot(pluton,
       col = cl$cluster,
       main = paste("Kmeans function with", i, "iterations"))
  dev.off()
}
```

Приложение 2

```
# Задание 2 ##########################################
#-------------------------------------------------------
library(cluster)
size = 100
x1 <- matrix(c(rnorm(size, mean = -40, sd = 10), rnorm(size, mean = 0, sd = 2)), ncol = 2)
x2 <- matrix(c(rnorm(size, mean = -30, sd = 5), rnorm(size, mean = 0, sd = 5)), ncol = 2)
x3 <- matrix(c(rnorm(size, mean = -20, sd = 2), rnorm(size, mean = 0, sd = 10)), ncol = 2)

colnames(x) <- c("x", "y")
x <- rbind(x1, x2, x3)

png(paste(path, "Data.png"))
plot(x[, 1], x[, 2])
dev.off()

for (i in c(TRUE, FALSE)) {
  metrics <-  c("euclidean", "manhattan")

  for (m in metrics) {
    cl <- clara(x, 3, stand = i, metric = m)
    png(paste(path, "Clustering with method", m, "and", i, ".png"))
    plot(x, col = cl$cluster, main = paste("Clustering with method", m, "and standartization =", i))
    dev.off()
  }
}
```

Приложение 3

```
# Задание 3 ##########################################
#-------------------------------------------------------------
library(cluster)
data(votes.repub)
png(paste(path, "Votes.png"), width = 720)
plot(agnes(votes.repub))
dev.off()
```

Приложение 4

```
# Задание 4 ##########################################
#-------------------------------------------------------
library(cluster)
data(animals)
png(paste(path, "Animals.png"))
plot(agnes(animals))
dev.off()
```

Приложение 5

```
# Задание 5 ##########################################
#------------------------------------------------------
library(cluster)
seed <- read.csv(paste(path, "seeds_dataset.txt", sep = ""), header = FALSE, sep = "\t")
seed <- na.omit(seed)
result <- seed[, 8]
seed <- seed[,-8]

km <- kmeans(seed, centers = 3)
png(paste(path, "Cluster.png"))
plot(seed, col = km$cluster)
dev.off()

tbl1 <- table(km$cluster, result)

print(sum(diag(tbl1)) / sum(tbl1))

cl <- clara(seed, 3)
png(paste(path, "Clara.png"))
plot(seed, col = cl$cluster)
dev.off()
tbl <- table(cl$cluster, result)

print(sum(diag(tbl)) / sum(tbl))
```