Министерство образования и науки РФ
Санкт-Петербургский Политехнический университет Петра Великого
Институт компьютерных наук и технологий
Высшая школа искусственного интеллекта

ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ № 6

«Кластеризация»
по дисциплине «Машинное обучение»

Выполнила:
студентка гр. 3540201/20301 _____ Климова О. А.
подпись, дата

Проверил:
д.т.н., проф. _____ Уткин Л. В.
подпись, дата

Санкт-Петербург
2022

Содержание

Постановка задачи

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.

1 Задание 1

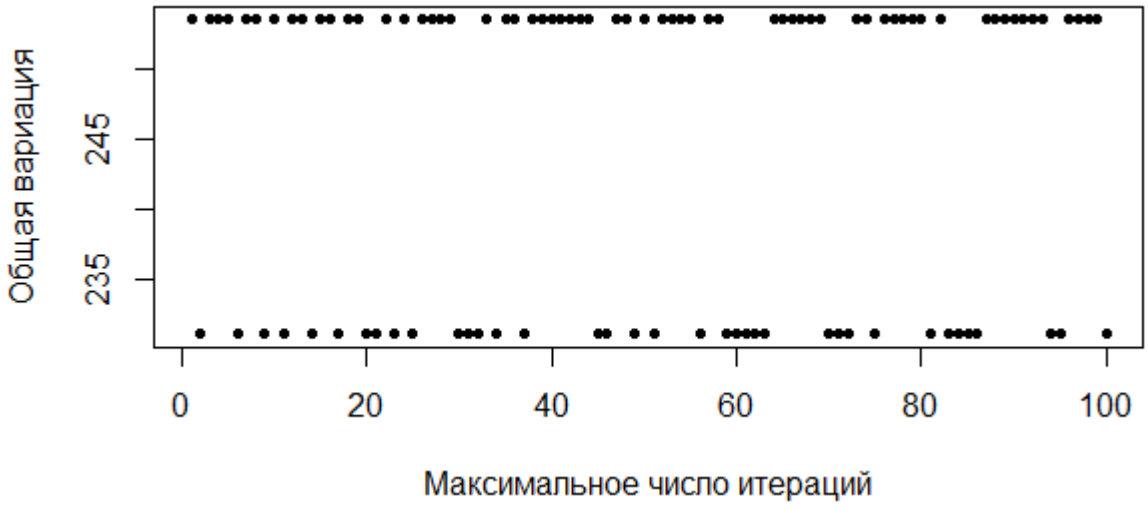
В данном задании было разбито множество объектов из набора данных pluton в пакете «cluster» на 3 кластера методом центров тяжести (kmeans).

Для оценки используем Локтевой метод, то есть будем оценивать общую вариацию внутри кластера (общую сумму квадратов внутри кластера) и говорить, что наименьшему значению данного параметра соответствует наилучшая кластеризация:

$$minimize\left(\sum_{k=1}^{k} W(C_k)\right)$$

Общую сумму квадратов внутри кластера можно получить с помощью параметра tot.withinss среди выходных данных функции kmeans.

При проведении эксперимента при изменении наибольшего числа итераций от 1 до 100, то получим:

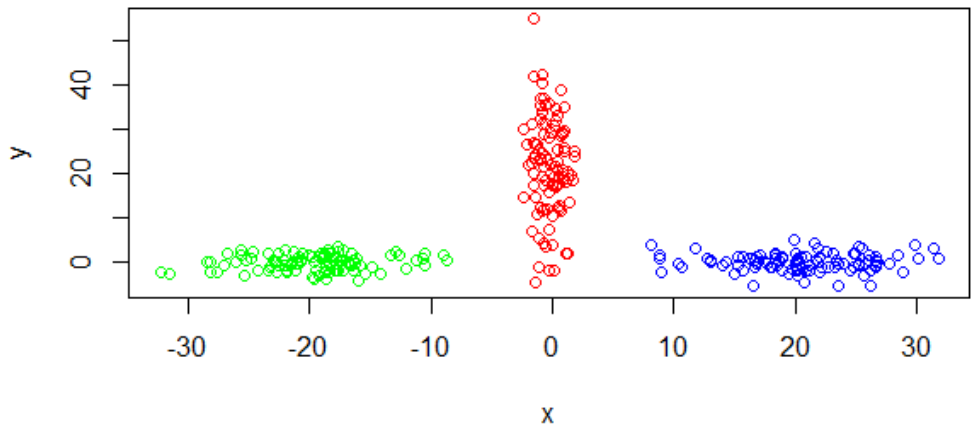


Можно видеть, что значение общей вариации принимает два значения: 253.4651 или 231.1747. Так как min(253.4651 ; 231.1747) = 231.1747, то при различном максимальном числе итераций, соответствующему данному значению, качество кластеризации самое высокое.

Код представлен в Приложении 1.

2 Задание 2

В данном задании был сгенерирован набор данных в двумерном пространстве, состоящий из 3 кластеров, каждый из которых сильно “вытянут” вдоль одной из осей:

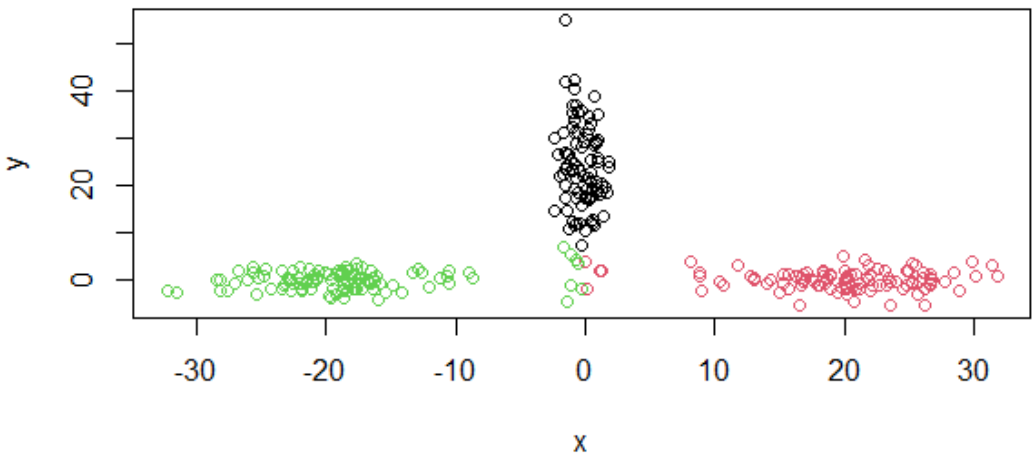


Для генерации этих данных было использовано нормальное распределение с различными параметрами матожиданий и среднеквадротичных отклонений.

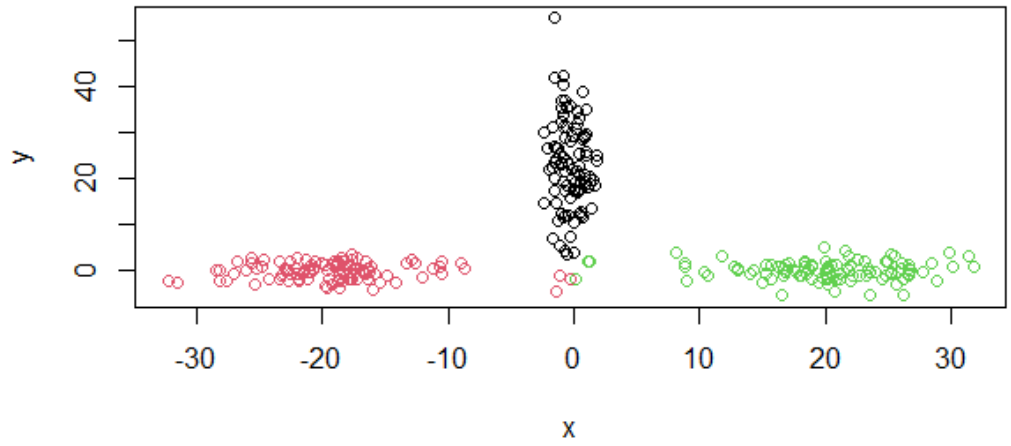
Качество кластеризации было оценено методом clara с использованием параметра avg.width - средней ширины силуэта, который определяет насколько хорошо каждый объект лежит в своем кластере (высокому значению средней ширины силуэта соответствует высокое качество кластеризации) :

1) использование стандартизации (метрика «euclidean»):

используется (avg.width = 0.665**):**



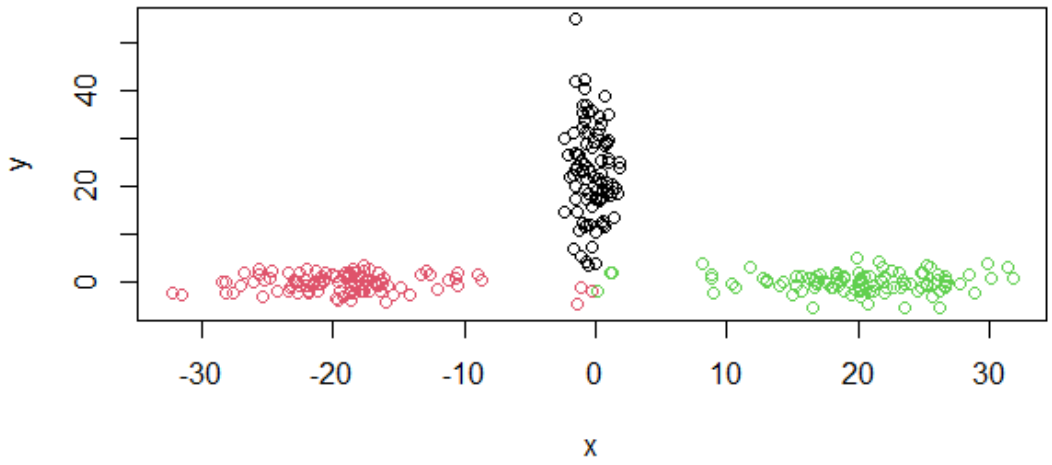
не используется (avg.width = 0.683**):**



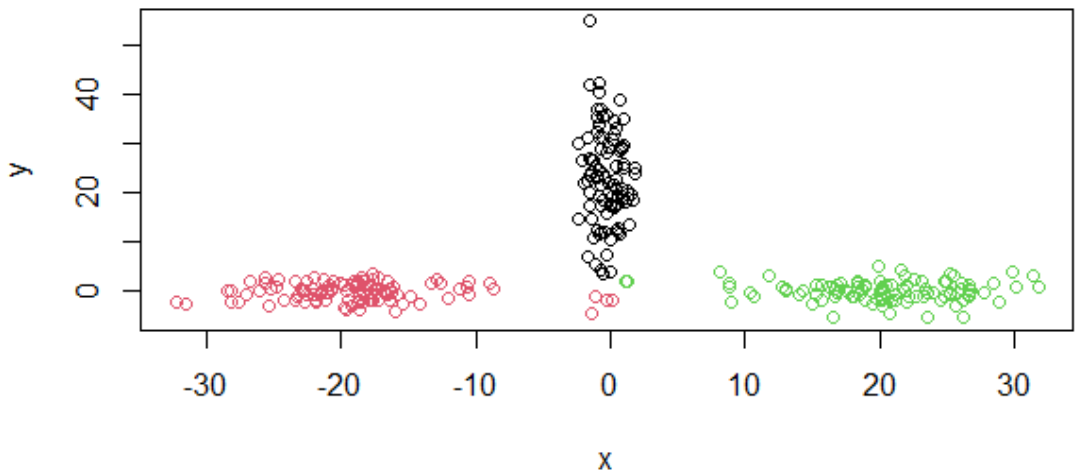
Можно видеть, что большее качество кластеризации получаем при отсутствии использования стандартизации.

2) тип метрики (не используем стандартизацию):

euclidean (avg.width = 0.683**):**

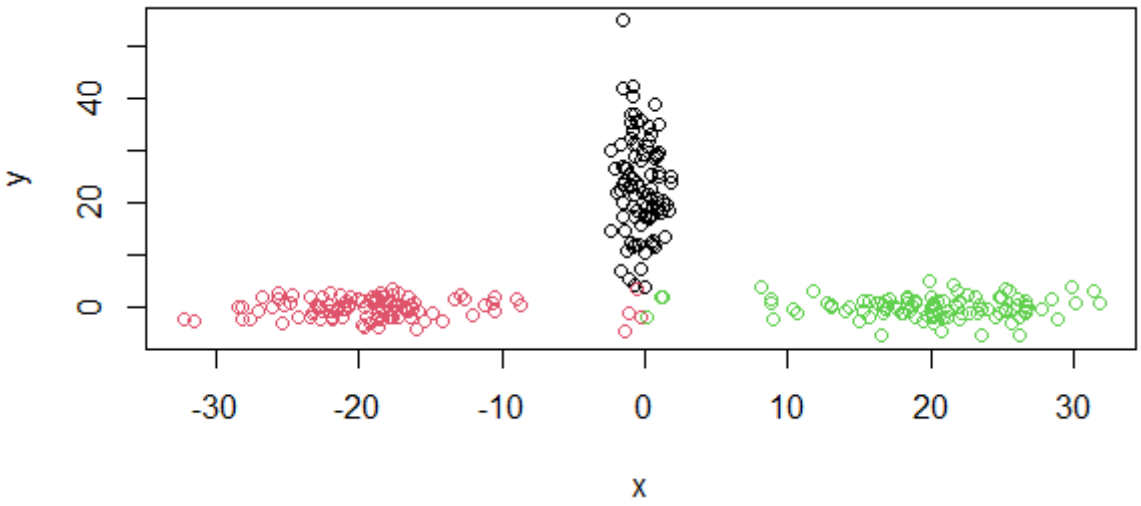


manhattan (avg.width = 0.72**):**



Можно видеть, что большее качество кластеризации получаем при использования метрики manhattan.

Посмотрим на результат использования метрики manhattan и стандартизации:



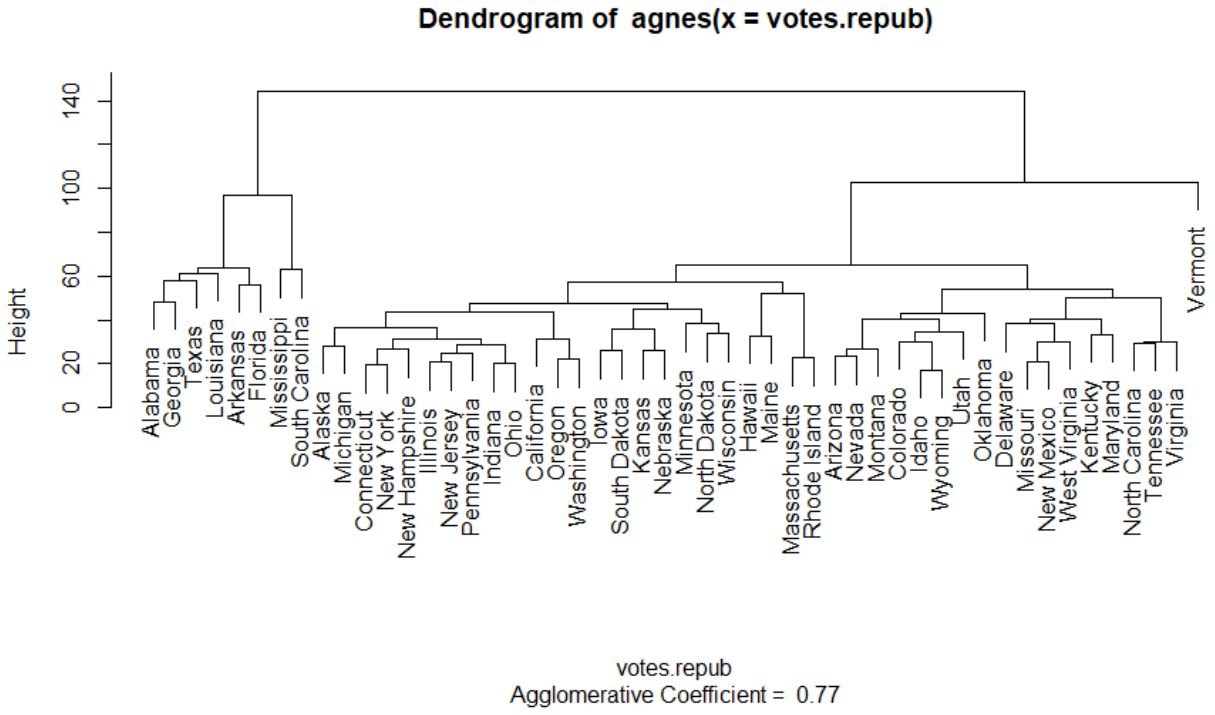
Как и можно было предположить, при таком выборе параметров точность классификации падает, так как avg.width становится равной 0.6938.

Таким образом, можно сделать вывод, что для кластеризации созданного набора данных методом clara оптимальнее всего использовать метрику manhattan и не использовать стандартизацию.

Код представлен в Приложении 2.

3 Задание 3

В данном задании была построена дендрограмма для набора данных votes.repub в пакете «cluster» (число голосов, поданных за республиканцев на выборах с 1856 по 1976 год), где строки представляют 50 штатов, а столбцы - годы выборов (31):

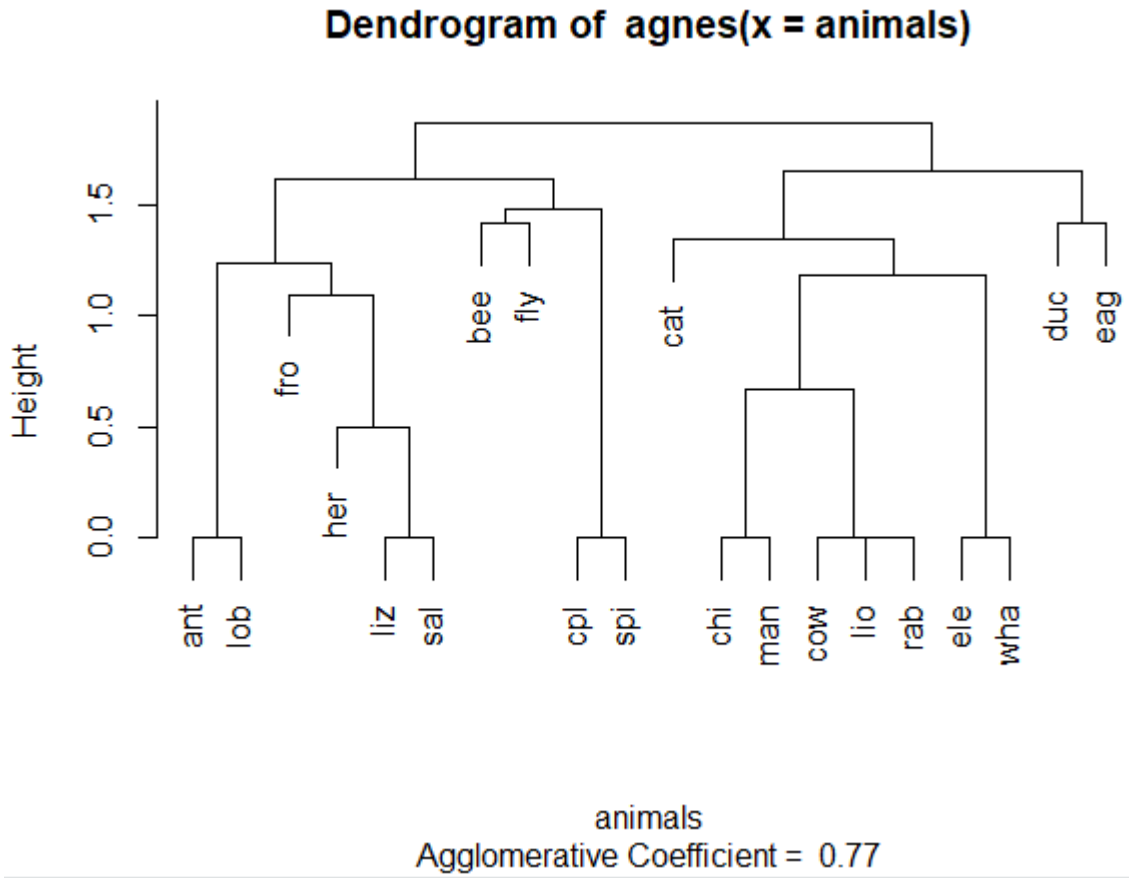


Дендрограмма представляет собой дерево, содержащее уровни, каждый из которых соответствует одному из шагов последовательного укрупнения кластеров.

По полученной дендограмме можно видеть, что при разделении на 2 кластера, в первый войдут Alabama, Georgia, Texas, Louirsiana, Arkansas, Florida, Mississippi, South Carolina, а во второй все остальные. Данное разделение обусловлено количеством голосов в каждом штате. В левый кластер вошли те штаты, среднее количество голосов в которых за все годы не очень велико.

4 Задание 4

В данном задании была построена дендрограмма 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:



По полученной дендрограмме можно видеть, что при разбиении данных на два класса в один войдут: ant, lob, fro, her, liz, sal, bee, fly, cpi, spi, а в другой все остальные. Данное разделение обусловлено схожестью парметров, все животные выделенные в первый кластер не теплокровные (war = 1), а во второй – все теплокровные (war = 2). И далее можно видеть, что чем больше схожих параметров у объектов, тем в более узкие кластеры они объединяются. Например, на нижнем уровне ant и lob в одном кластере, потому что у них расходится из 6-ти параметров только один – gro.

Код представлен в Приложении 4.

5 Задание 5

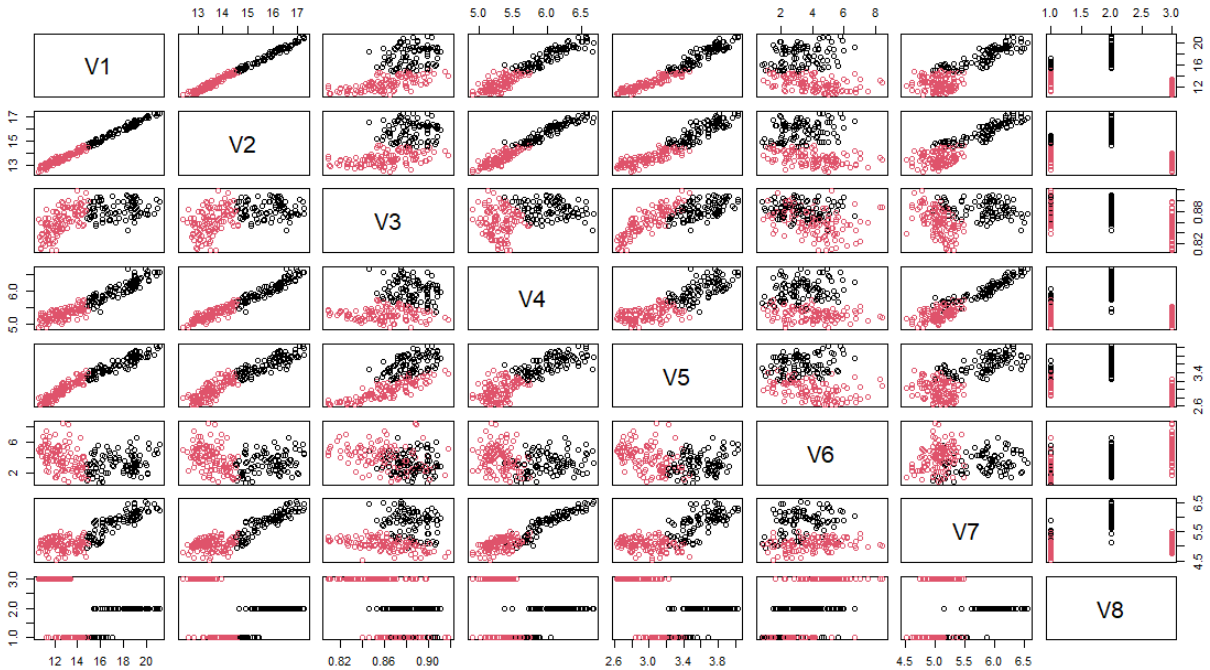
В данном задании были рассмотрены данные из файла seeds_dataset.txt, который содержит описание зерен трех сортов пшеницы: Kama, Rosa and Canadian с признаками: 1. область A, 2. периметр P, 3. компактность C = 4*pi*A/P^2, 4. длина зерна, 5. ширина зерна, 6. коэффициент ассиметрии, 7. длина колоска.

Данные имеют вид:

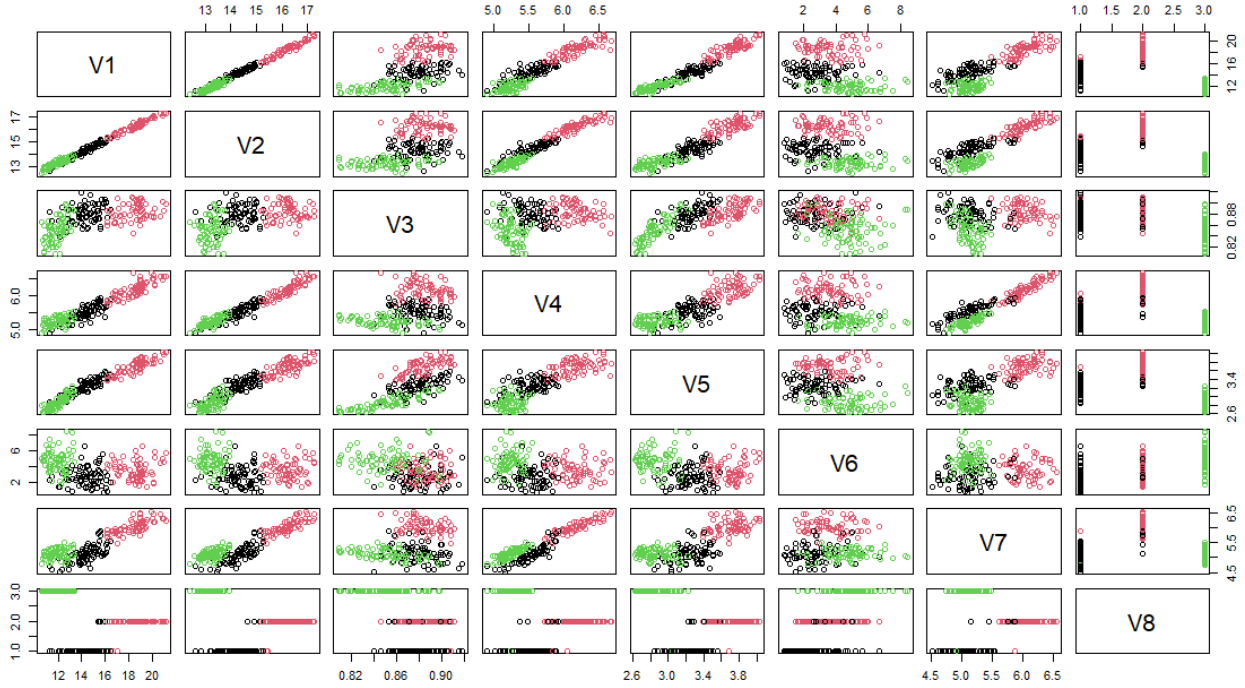
```
> seeds
      V1    V2     V3    V4    V5     V6    V7 V8
1  15.26 14.84 0.8710 5.763 3.312 2.2210 5.220  1
2  14.88 14.57 0.8811 5.554 3.333 1.0180 4.956  1
3  14.29 14.09 0.9050 5.291 3.337 2.6990 4.825  1
4  13.84 13.94 0.8955 5.324 3.379 2.2590 4.805  1
5  16.14 14.99 0.9034 5.658 3.562 1.3550 5.175  1
6  14.38 14.21 0.8951 5.386 3.312 2.4620 4.956  1
7  14.69 14.49 0.8799 5.563 3.259 3.5860 5.219  1
8  14.11 14.10 0.8911 5.420 3.302 2.7000 5.000  1
9  16.63 15.46 0.8747 6.053 3.465 2.0400 5.877  1
10 16.44 15.25 0.8880 5.884 3.505 1.9690 5.533  1
11 15.26 14.85 0.8696 5.714 3.242 4.5430 5.314  1
12 14.03 14.16 0.8796 5.438 3.201 1.7170 5.001  1
13 13.89 14.02 0.8880 5.439 3.199 3.9860 4.738  1
14 13.78 14.06 0.8759 5.479 3.156 3.1360 4.872  1
15 13.74 14.05 0.8744 5.482 3.114 2.9320 4.825  1
```

Сначала данные были разбиты на кластеры с помощью K-медоидов:

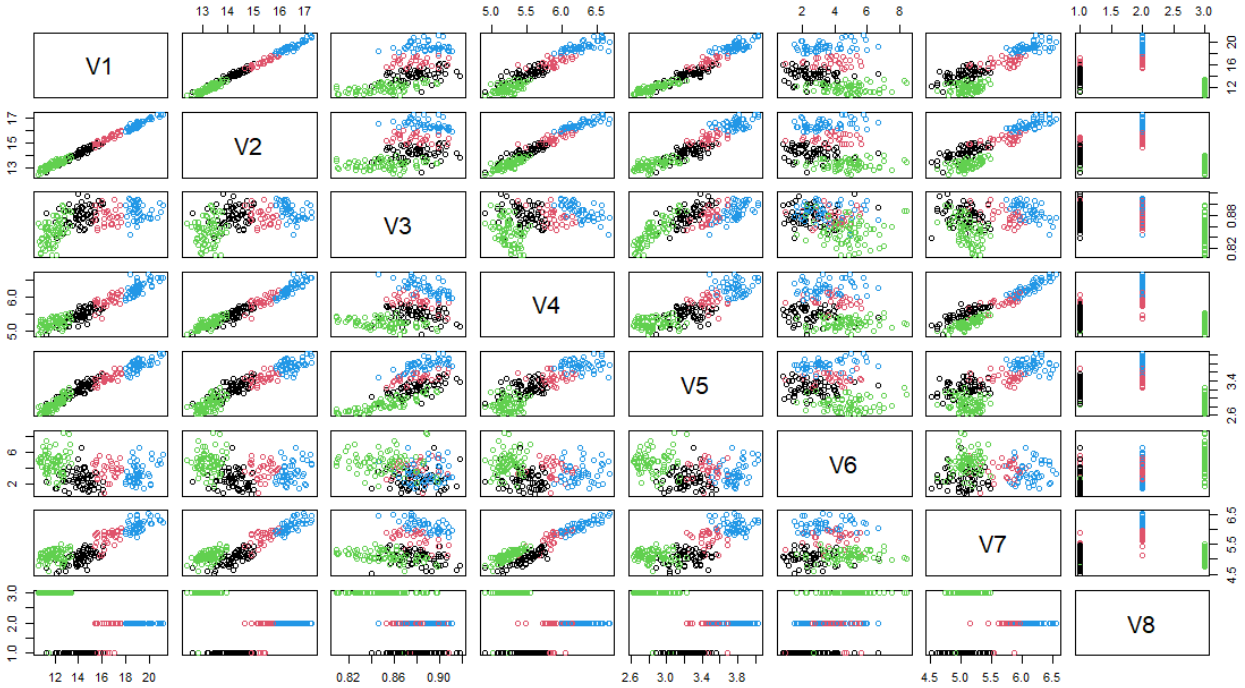
2 кластера (avg.width = 0.5145091):



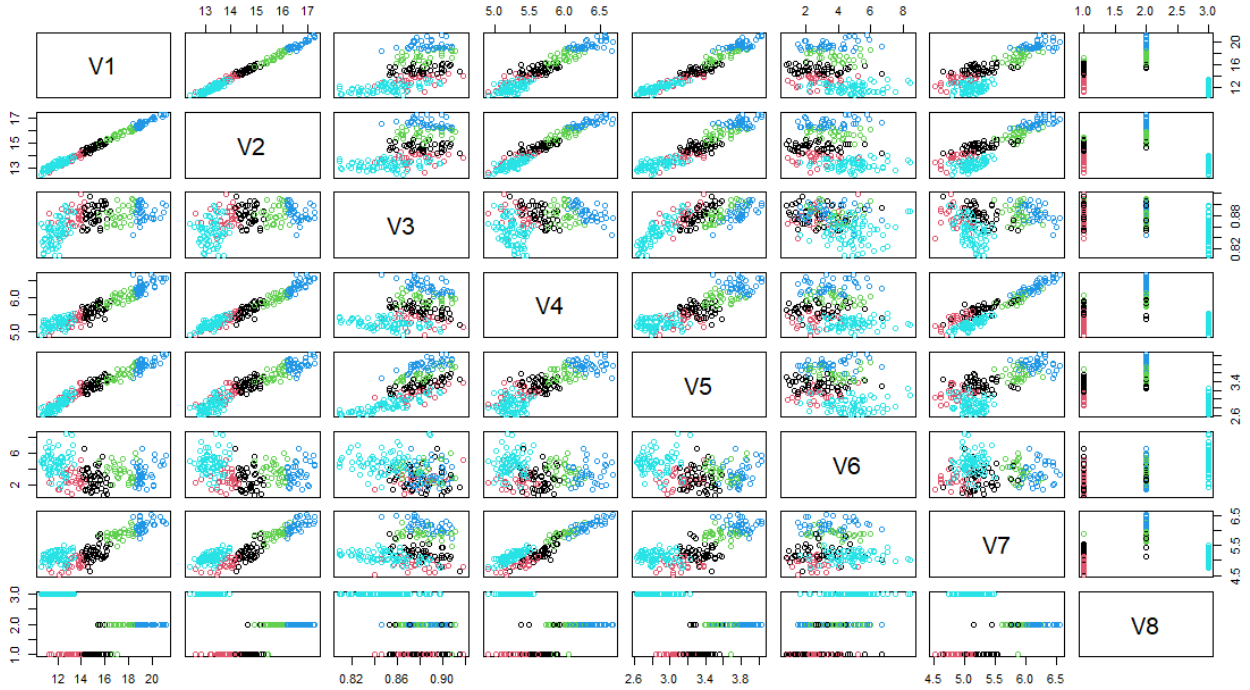
3 кластера (avg.width = 0.5612057):



4 кластера (avg.width = 0.5047038):

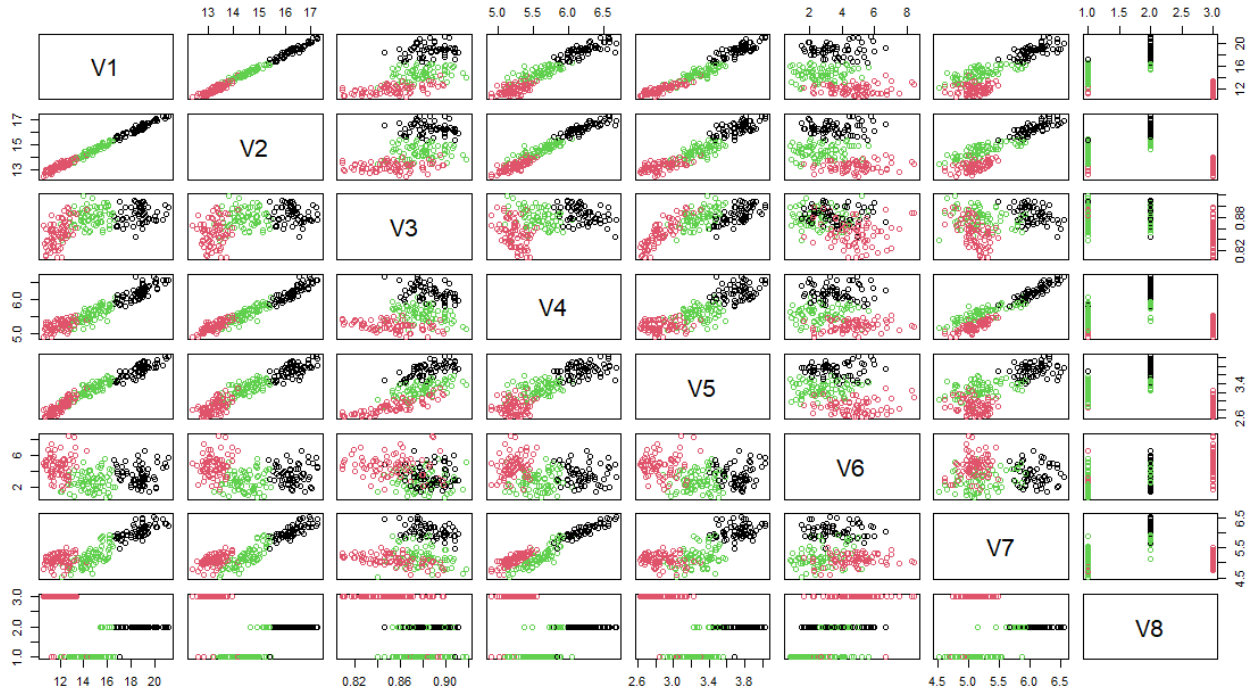


5 кластеров (avg.width = 0.4616301):



Так как при k = 3 наибольшая средняя ширина силуэта, значит точнее всего разбивать множество на три кластера.

Далее данные были разбиты на три кластера с помощью K-средних:



Значение общей вариации для трех классов tot.withinss = 625.72.

Код представлен в Приложении 5.

Приложение 1. Код для задания 1

```
#ЗАДАНИЕ 1
library(cluster)
data(pluton)
pluton
#кластеризация на три класса методом kmeans
cl_pluton <- kmeans(pluton, 3)
plot(pluton, col = cl_pluton$cluster)
#создаем вектор максимального количества итераций (от 1 до 100)
iter_max = c()
for(i in 1:100){
  iter_max <- append(iter_max, i)
}
cl_pluton
#общая вариация внутри кластера (сумма cl_pluton$withinss)
cl_pluton$tot.withinss

#вектор общих вариаций внутри кластера при различных iter.max
totwithinss = c()
for(i in iter_max){
  cl_pluton <- kmeans(pluton, 3, iter.max = iter_max[i])
  totwithinss <- append(totwithinss, cl_pluton$tot.withinss)
}
#построение графика
plot(iter_max, totwithinss, pch = 20, xlab="Максимальное число итераций",
     ylab="Общая вариация",
     main = "")
totwithinss
```

Приложение 2. Код для задания 2

```
#ЗАДАНИЕ 2
#генерация трех кластеров, сильно вытянутых вдоль оси (с помощью норм. распр-я)
data <- data.frame(x = c(rnorm(100, 0, 1), rnorm(100, 20, 5), rnorm(100, -20, 5)),
                   y = c(rnorm(100, 20, 10), rnorm(100, 0, 2), rnorm(100, 0, 2)))

col = c(rep("red", 100), rep("blue", 100), rep("green", 100))

#объединение вектора чисел и вектора светов
data_col <- cbind(data, col)
plot(data_col$x, data_col$y, xlab="x", ylab="y", col = data_col$col)

#использование стандартизации
#используется
cl_eucl1 = clara(data, k = 3, metric = "euclidean", stand = TRUE)
plot(data, col = cl_eucl1$clustering, xlab = "x", ylab = "y")
#не используется
cl_eucl2 = clara(data, k = 3, metric = "euclidean", stand = FALSE)
plot(data, col = cl_eucl2$clustering, xlab = "x", ylab = "y")

#использование метрик
#euclidean
cl_eucl2 = clara(data, k = 3, metric = "euclidean", stand = FALSE)
plot(data, col = cl_eucl2$clustering, xlab = "x", ylab = "y")

#manhattan
cl_manh1 = clara(data, k = 3, metric = "manhattan", stand = FALSE)
plot(data, col = cl_manh1$clustering, xlab = "x", ylab = "y")

#manhattan
cl_manh2 = clara(data, k = 3, metric = "manhattan", stand = TRUE)
plot(data, col = cl_manh2$clustering, xlab = "x", ylab = "y")

cl_eucl1$silinfo$avg.width
cl_eucl2$silinfo$avg.width
cl_manh1$silinfo$avg.width
cl_manh2$silinfo$avg.width
```

Приложение 3. Код для задания 3

```
#ЗАДАНИЕ 3
library(cluster)
data("votes.repub")
votes.repub
#построение дэндрограммы
plot(agnes(votes.repub))
```

Приложение 4. Код для задания 4

```
#ЗАДАНИЕ 4
library(cluster)
data("animals")
animals
plot(agnes(animals))
```

Приложение 5. Код для задания 5

```
#ЗАДАНИЕ 5
library(cluster)
seeds <- read.table("C:/Users/Unicorn/Desktop/Машинное Обучение/Лабы/seeds_dataset.txt",
sep = "", stringsAsFactors = TRUE)
seeds
#слишком большой получается
plot(agnes(seeds))
#кластеризация (на 3 класстера)
cl_clara1 = clara(seeds, k = 3, metric = "manhattan", stand = FALSE)
cl_clara1
plot(seeds, col = cl_clara1$clustering)
cl_clara1$silinfo$avg.width

#кластеризация на три класса методом kmeans
cl_kmeans1 <- kmeans(seeds, 3)
plot(seeds, col = cl_kmeans1$cluster)
cl_kmeans1$tot.withinss
```