Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Пермский государственный аграрно-технологический университет

имени академика Д.Н. Прянишникова»

Кафедра информационных технологий и программной инженерии

**ЛАБОРАТОРНАЯ РАБОТА № 3**

Кластеризация.

|  |  |
| --- | --- |
| Выполнил: | группа ПИНб-4  П.С. Плотников |
| Проверил: | доцент каф. ИТиПИ,  Т.А. Казаченко |

Пермь-2024 г

# Ответы на вопросы

Таблица 1

|  |  |  |
| --- | --- | --- |
| **№** | **Вопрос** | **Ответ** |
|  | Как называются методы, изучаемые в лабораторной 3? | Метод k-средних, иерархическая кластеризация, метод главных компонент |
|  | К какой категории методов ML они относятся? | Обучение без учителя |
|  | Что может прогнозировать алгоритм, обученный данным методом? | Алгоритм прогнозирует группы похожих друг на друга объектов. |
|  | В чем принципиальное отличие задачи классификации от задачи кластеризации? | Для задачи классификации необходимы заранее известные ответы (обучение с учителем). При кластеризации ответы не известны (обучение без учителя) |
|  | Что значит «метод сошелся» для метода K-means? | Это значит, что центроиды не меняют своего положения при последующих итерациях, т.е. объекты разделились по группам. |
|  | Чем могут отличаться задачи кластеризации? | Кластеризация может быть частью более сложного решения, а не основной задачей. Может быть «мягкой» и «жесткой». Количество центроидов может быть выбрано заранее или определяться алгоритмом. Могут отличаться размеры кластеров, также кластеры могут быть вложенными. |
|  | Приведите примеры «жесткой» и «мягкой» кластеризаций (не из указанных в лекции). | Мягкая: Гауссовые смешанные модели (GMM). Мягкая кластеризация допускает, что объект может быть частью нескольких кластеров. Жесткая: Иерархическая кластеризация. В жесткой кластеризации объект может принадлежать только одному кластеру. |

**Формулировка задачи**

Решите задачу кластеризации на данных pluton с использованием стандартных пакетов в среде R.

Разбейте множество объектов из набора данных pluton на 3 кластера методом центров тяжести (К-means). Сравните качество разбиения в зависимости от макисмального числа итераций алгоритма.

**Описание признаков используемого датасета**

Исходный датасет pluton содержит в себе четыре количественных признака, каждый из которых означает процент содержания конкретного изотопа плутония в партии замеров. Признаки:

1. Pu238: всегда меньше 2%;
2. Pu239: находится в промежутке 60-80%;
3. Pu240: находится в промежутке 18-26%;
4. Pu241: всегда меньше 10%.

Всего в файле 45 объектов, то есть 45 партий замеров изотопов.

**Программный код**

# Подключение библиотеки визуализации

library(factoextra)

# Чтение данных

pluton = read.table("./pluton.txt", header = TRUE)

# Кол-во кластеров

k = 3

# Кол-во итераций

# При iters < 6 – нечеткое разделение

# При iters >= 6 – ровное разделение на 3 кластера

iters = 6

# Кластеризация методом k-средних

clusters = kmeans(pluton, k, nstart = iters)

# Отображение результата кластеризации на исходном датасете

fviz\_cluster(clusters, pluton, main = "Кластеризация плутона",

xlab = FALSE, ylab = FALSE)

**Результат вычислений**

На рисунке 1 показан результат кластеризации при трех итерациях, на рисунке 2 – при шести.

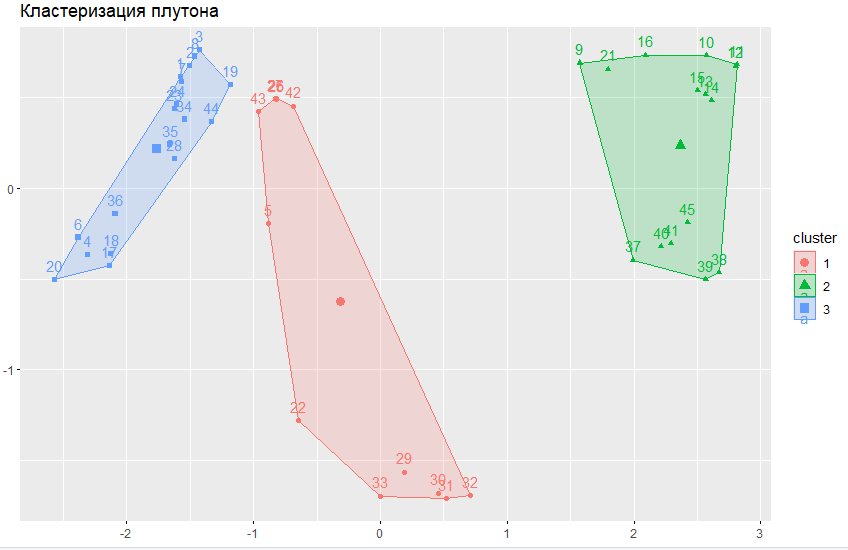


Рисунок 1 – Результат кластеризации при трех итерациях

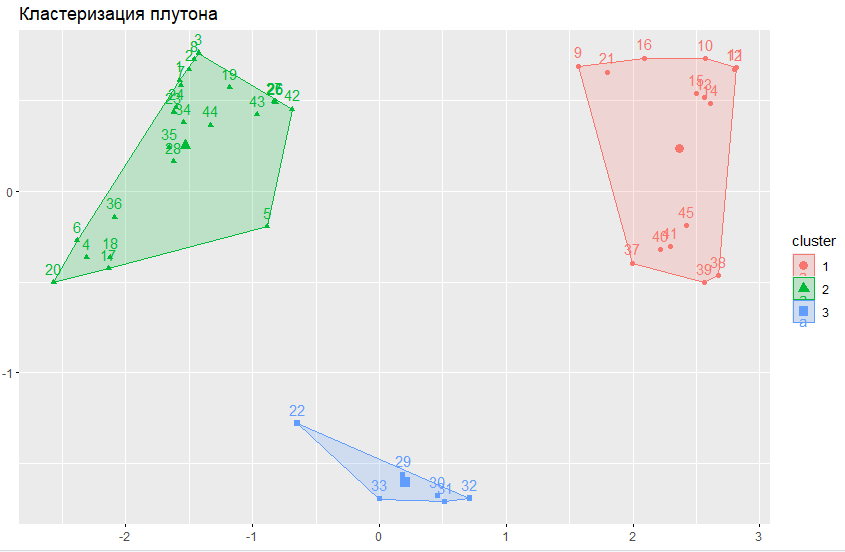


Рисунок 2 – Результат кластеризации при шести итерациях

**Пояснение результатов**

Как можно заметить на рисунке 1, некоторые объекты очевидно находятся не в своем кластере, а именно объекты под номерами: 5, 26, 27, 42 и 43. Связано это с тем, что центроиды кластеров выбираются случайным образом. Более того, количества итераций недостаточно, чтобы четко разбить облака данных на кластеры.

На рисунке 2, при шести итерациях, объекты четко разделяются на три кластера.

В кластер 1 попали изотопы, у которых количество Pu239 минимально. В верхней части кластера находятся объекты, у которых, в добавок, максимально Pu240. В нижней – максимально Pu241.

В кластер 2 попали объекты, у которых максимальны и Pu239, и Pu240.

В кластере 3 значение Pu240 минимально, остальные изотопы имеют приблизительно средние значений от своих возможных.

**Формулировка задачи**

Решите задачу кластеризации на данных votes.repub с использованием стандартных пакетов в среде R.

Постройте дендрограмму для набора данных votes.repub (число голосов, поданных за республиканцев на выборах с 1856 по 1976 год). Проинтерпретируйте полученный результат.

**Описание признаков используемого датасета**

В используемом датасете votes.repub имеется два вида признаков: номинальный (штат) и количественные (процент проголосовавших). Для построения дендрограммы требуются только количественные признаки.

Количественные признаки делятся на года, в каждом годе имеется процент проголовавших за республиканцев.

**Прграммный код**

# Чтение данных

votes = read.table("./votes.repub.txt", header = TRUE)

# Избавляемся от NA

votes[is.na(votes)] = 0

# Заменим номера строк названиями шататов

rownames(votes) = votes$States

# Делаем срез, т.к. для дендрограммы важны голоса, а не штаты

votes = votes[,2:32]

# Получаем расстояния между точками

votes\_d = dist(votes)^2

# Получаем кластеризацию

votes\_h = hclust(votes\_d, method = "ward.D")

plot(votes\_h, main = "Дендрограма votes.repub")

# Выделим на графике 3 кластера

rect.hclust(votes\_h, k = 3, border="red")

**Результаты вычислений**

На рисунке 3 представлена дендрограмма иерархической кластеризации.

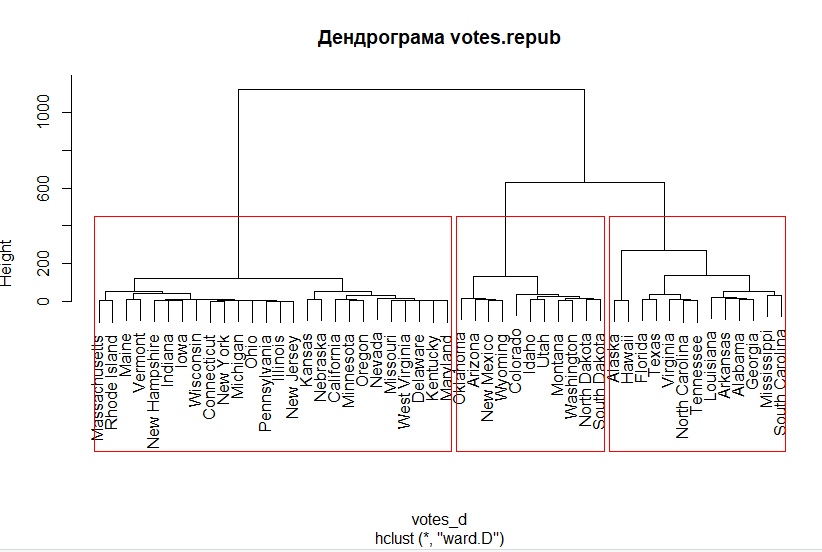


Рисунок 3 – Дендрограмма решения

**Пояснение результатов**

Дендрограмма показывает, как штаты группируются на основе похожести голосов за республиканцев.

Длинные ветви указывают на кластеры с низкой степенью сходства между штатами.

**Формулировка задачи**

Решите задачу кластеризации на данных и swiss с использованием стандартных пакетов в среде R.

Рассчитайте, какое минимальное число главных компонент объясняет больше 90% изменчивости в исходных данных swiss и добавлять значения этих компонент в исходный файл данных в виде новых переменных.

**Программный код**

# Загрузка набора данных swiss

data(swiss)

# Стандартизация данных

swiss\_scaled <- scale(swiss)

# Вычисление главных компонент

pca\_model <- prcomp(swiss\_scaled, scale. = TRUE)

# Процент объяснённой дисперсии

explained\_variance <- cumsum(pca\_model$sdev^2) / sum(pca\_model$sdev^2)

# Минимальное число компонент, объясняющих >90% изменчивости

min\_components <- which(explained\_variance > 0.9)[1]

cat("Минимальное число компонент:", min\_components, "\n")

# Построение графика объяснённой дисперсии

fviz\_eig(pca\_model, addlabels = TRUE, xlab = "Признаки",

main = "Диаграма зависимости компонент",

ylab = "Процент объяснения изменчивости")

# Добавление значений компонент в исходный набор данных

principal\_components <- as.data.frame(pca\_model$x[, 1:min\_components])

colnames(principal\_components) <- paste0("PC", 1:min\_components)

swiss\_with\_pcs <- cbind(swiss, principal\_components)

**Результаты вычислений**

На рисунке 4 показана диаграмма, на которой отображено процент объяснения изменчивости каждого признака. На рисунке 5 показан вывод в консоли, в котором посчитано, сколько нужно признаков. На рисунке 6 показан новый датасет, состоящий из исходных данных и выявленных признаков.

Изображение выглядит как текст, диаграмма, снимок экрана, График

Контент, сгенерированный ИИ, может содержать ошибки.

Рисунок 4 – Диаграмма зависимости компонент

Изображение выглядит как текст, Шрифт, снимок экрана, Цвет электрик

Контент, сгенерированный ИИ, может содержать ошибки.

Рисунок 5 – Вывод в консоли

Изображение выглядит как текст, число, снимок экрана, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Рисунок 6 – Новый датасет

**Пояснение результатов**

Модель показала, что для объяснения больше 90% изменчивости, необходимо минимум 4 компоненты. Сложим проценты изменчивости первых четырех компонент диаграммы.

|  |  |
| --- | --- |
|  | (1) |

Сложив первые четыре признака с диаграммы, можно понять, что действительно четырех компонент достаточно четырех компонент.