Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования «Кубанский государственный университет»

Кафедра вычислительных технологий

**ОТЧЕТ**

о выполнении лабораторной работы № 5\_1

по дисциплине «Обработка больших данных»

Тема:задачи классификации и кластеризации

.

Выполнил: ст. гр. 36/2

Агаджанян А. С.

Проверил: преп. кафедры ВТ

Яхонтов А. А.

Краснодар

2025

Набор данных при изначальном прочтении выглядел как показано на рисунке 1.

Изображение выглядит как текст, снимок экрана, меню, число

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.1 – изначальный набор данных

Далее была проведена нормализация (масштабирование) и набор данных стал выглядеть так, как показано на рисунке 2.

Изображение выглядит как текст, снимок экрана, число, меню

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.2 – измененный набор данных

Был проведен дескриптивный анализ датасета, результаты которого видны на рисунке 3 и 4.

Изображение выглядит как текст, Шрифт, число, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.3 Дескриптиынй анализ (1 часть)

Изображение выглядит как текст, снимок экрана, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.4 Дескриптивный анализ (2 часть)

Для выявления оптимального количества кластеров, на которые надо разбить датасет, были применены следующие методы: “метод локтя” (рисунок 5), ”метод силуэта” (рисунок 6), ”метод статистики разрыва” (рисунок 7), ”алгоритм на основе консенсуса” (рисунок 8).

Изображение выглядит как текст, линия, График, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.5 – метод локтя

Метод локтя показывает, что оптимальное число кластеров – 4, потому что дальше график перестает существенно уменьшаться.

Изображение выглядит как текст, линия, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 6 – метод силуэта

Метод силуэта показывает, что оптимальное число кластеров – 2.

Изображение выглядит как текст, диаграмма, линия, График

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.7 – метод статистики разрыва

Метод статистики разрыва показывает, что оптимальное число кластеров – 5.

Изображение выглядит как текст, снимок экрана, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.8 – алгоритм на основе консенсуса

Алгоритм на основе консенсуса показывает, что оптимальное число кластеров – 5.

Разобьем набор данных на 5 кластеров и нарисуем дендрограмму (рис. 9)

Изображение выглядит как текст, диаграмма, Технический чертеж, План

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 9 – дендрограмма

Изображение выглядит как текст, снимок экрана, диаграмма, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 10 – разделили выборку на 5 класстеров

Выполним класстеризацию по k-means:

Изображение выглядит как текст, снимок экрана, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 11 – 5 групп класстеров с легендой

Изображение выглядит как текст, диаграмма, График, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 12 – изменена палитра легенды и эллипсоидное очертание

Скаттерплоты, отражающие характеристики классов ирисов:

Изображение выглядит как текст, диаграмма, чек

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, снимок экрана, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Построение трехмерной кластеризации по scatterplot3d (были выбраны 3 характеристики из всей выборки – это “education”, “tradeshare”, “rgdp60”):

Изображение выглядит как диаграмма, линия, График, Параллельный

Контент, сгенерированный ИИ, может содержать ошибки.

Вывод: были закреплены знания об алгоритмах классификации и кластеризации данных, ознакомлен с некоторыми функциями языка R, осуществляющими этот вид анализа. Научился визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты

Листинг:

library(parameters)

library (factoextra)

library (cluster)

library(scatterplot3d)

library(lattice)

file\_path <- 'D:/Универ/3 курс/6 семестр/Бигдата/DataSets\_Lab\_5/09\_темпы роста\_стран/GrowthSW.csv'

# Читаем файл как текст

file\_lines <- readLines(file\_path, encoding = "UTF-8")

# Разделяем строки на столбцы

data\_list <- lapply(file\_lines, function(line) {

strsplit(line, ",")[[1]] # Разделяем строку по запятой

})

# Получаем имена столбцов из первой строки

column\_names <- data\_list[[1]]

# Извлекаем данные, начиная со второй строки

data\_rows <- data\_list[-1]

# Создаем матрицу данных

data\_matrix <- do.call(rbind, data\_rows)

dataset <- as.data.frame(data\_matrix, stringsAsFactors = FALSE)

# Назначаем имена столбцов

colnames(dataset) <- column\_names

#Преобразуем необходимые столбцы в числовой формат (кроме первого со странами)

for (col in 2:ncol(dataset)){

dataset[[col]] <- as.numeric(dataset[[col]])

}

dataset = dataset[2:7]

# 1. Выполнить дескриптивный анализ данных

mean(as.matrix(dataset), na.rm = TRUE)

median(as.matrix(dataset), na.rm = TRUE)

moda <- function(x) {

tab <- tabulate(x)

is\_max <- tab == max(tab)

mean(which(is\_max == TRUE))

}

moda(as.matrix(dataset))

results <- data.frame(

Минимум = apply(dataset, 2, min, na.rm = TRUE),

Q1 = apply(dataset, 2, quantile, probs = 0.25, na.rm = TRUE),

Медиана = apply(dataset, 2, median, na.rm = TRUE),

Q3 = apply(dataset, 2, quantile, probs = 0.75, na.rm = TRUE),

Максимум = apply(dataset, 2, max, na.rm = TRUE),

IQR = apply(dataset, 2, function(x) IQR(x, na.rm = TRUE)),

Дисперсия=apply(dataset, 2, var, na.rm = TRUE),

Среднекв.отклонение=apply(dataset, 2, sd, na.rm = TRUE)

)

for (col in 1:ncol(dataset)) {

if(any(is.na(dataset[,col]))){ # Проверяем, есть ли NA в столбце

median\_val <- median(dataset[, col], na.rm = TRUE)

dataset[is.na(dataset[, col]), col] <- median\_val

}

}

# 2. Оценить оптимальное число кластеров, для этого построить диаграмму

# "Метод силуэта", “Метод локтя”, "Статистику разрыва" и Алгоритм консенсуса

dataset\_scaled <- data.frame(scale(dataset))

dist.dataset\_scaled <- dist(dataset\_scaled) # матрица попарных расстояний

clust.dataset\_scaled <- hclust(dist.dataset\_scaled, "ward.D")

# “Метод локтя”

fviz\_nbclust(dataset\_scaled, kmeans, method = "wss")

# "Метод силуэта"

fviz\_nbclust(dataset\_scaled, kmeans, method = "silhouette") +

labs(subtitle = "Silhouette method")

# "Статистика разрыва"

gap\_stat <- clusGap(dataset\_scaled, FUN = kmeans, nstart = 5,K.max =5, B = 5)

fviz\_gap\_stat(gap\_stat)

# Алгоритм консенсуса

n\_clust <- n\_clusters(dataset\_scaled, package = c("easystats", "NbClust", "mclust"), standardize = FALSE)

n\_clust

plot(n\_clust)

# 3. Выполнить иерархическую кластеризацию вашего набора данных, построив дендрограмму

clusters = 5

plot(clust.dataset\_scaled, cex=0.5)

rect.hclust(clust.dataset\_scaled, k=clusters, border="red")

# 4. Построить диаграмму со столбчатыми диаграммами и боксплотами групп

groups <- cutree(clust.dataset\_scaled, k=5)

g1<-colMeans(dataset\_scaled[groups==1,])

g2<-colMeans(dataset\_scaled[groups==2,])

g3<-colMeans(dataset\_scaled[groups==3,])

g4<-colMeans(dataset\_scaled[groups==4,])

g5<-colMeans(dataset\_scaled[groups==5,])

df<-data.frame(g1,g2,g3,g4,g5)

df1<-t(df)

df<-t(df1)

barplot(df, ylim=c(0,12), axes = FALSE,

col=c("red","green","blue","yellow","pink"), beside=TRUE)

axis(2, at = 0:5, labels = 0:5)

legend("top", legend = rownames(df), col=c("red","green","blue","yellow","pink"), lwd=10, bty = "n")

# 5. Выполнить кластеризацию своего датасета по k-means.

km.res <- kmeans(as.matrix(dataset\_scaled), 5, nstart = 10)

fviz\_cluster(km.res, dataset\_scaled[, -5], ellipse.type = "norm")

fviz\_cluster(km.res, dataset\_scaled[, -5],

palette = "Set2", ggtheme = theme\_minimal())

# 6. Выполнить построение scatterplot с помощью функций plot или pairs

pairs(dataset\_scaled)

colors <- c("red","green","blue","yellow","purple")

pairs(dataset\_scaled,pch = 19, cex = 0.8,

col = colors[groups],

lower.panel=NULL)

# 7. Построить трехмерную кластеризацию по scatterplot3d

groups <- cutree(clust.dataset\_scaled, k=3)

colors <- c("yellow", "green", "orange")

colors\_groups <- colors[groups]

s3d <- scatterplot3d(dataset\_scaled[2:4], pch=16, color=colors\_groups)