Министерство науки и высшего образования Российской Федерации

Новосибирский государственный технический университет

Кафедра Вычислительной техники



Лабораторная работа №2

По дисциплине «Методы анализа данных»

По теме

«Решение задач интеллектуального анализа данных (ИАД): кластеризация объектов средствами интегрированной системы Statistica, языка R».

Факультет: АВТ

Группа: АВТ-819

Студент: Ванин К.Е.

Преподаватель: Пешков А.В.

Новосибирск

2022

# Цель лабораторной работы

* изучить алгоритмы и методы кластерного анализа данных на примере решения конкретной задачи ИАД;

- исследовать эффективность использования различных алгоритмов и методов кластерного анализа данных для решения прикладной задачи;

- ознакомиться и получить практические навыки работы с модулями интегрированной статистической системы *Statistica*, языка *R,*реализующими решение задачи кластеризации объектов.

# Задание к лабораторной работе

1. Подготовьте файл данных в системе *Statistica* для проведения интеллектуального анализа в соответствии с вариантом (данные по вариантам в конце файла).
2. Проведите кластеризацию объектов, используя алгоритмы и методы кластерного анализа. Перед кластеризацией выполните нормировку количественных признаков.

Выполните кластеризацию объектов с помощью иерархических агломеративных методов. В качестве исходных данных для решения задачи возьмите количественные показатели. В табл. 2. для каждого варианта приведены названия иерархических методов и меры расстояний, которые необходимо использовать. Для каждого метода и меры необходимо выполнить следующее (всего 2 комбинации метода и меры):

- построить дендрограмму иерархического метода;

* + проанализировать матрицу расстояний и схему объединения, построить график зависимости числа кластеров от коэффициента слияния;
  + на основе визуального анализа построенного графика выдвинуть гипотезу о числе кластеров в данных;
  + внести полученные результаты в табл. 3.

1. Выполните кластеризацию объектов на основе использования качественных признаков. Воспользуйтесь одним из предложенных в варианте иерархическим агломеративным методом и процентом несогласия в качестве меры расстояния между объектами. Выполните шаги п. 2. Внесите полученные результаты в табл. 3.
2. Проведите кластеризацию объектов, используя метод k-средних и количественные показатели в качестве исходных данных. Выполнить следующие шаги:
   * задать число кластеров, определенное в п. 2;
   * рассчитать евклидово расстояние между кластерами;
   * рассчитать значимость влияния каждого из признаков на выделение кластеров на основе *F*-статистики;
   * построить график средних значений признаков по каждому из кластеров;
   * определить объекты, относящиеся к каждому кластеру;
   * построить диаграмму рассеяния в пространстве трех признаков, которые вносят наибольший вклад в выделение кластеров, на диаграмме выделить полученные кластеры;
   * внести полученные результаты в табл. 3.
   * по результатам анализа заполнить табл. 3.
3. Выполните кластеризацию объектов средствами R (используйте один из иерархических методов с одной из мер расстояний и метод k-средних)

- постройте дендрограмму, дендрограмму+тепловую карту, рассчитайте кофенетическую корреляцию по результатам работы иерархического метода. Сопоставьте с результатами, полученными в п. 2.;

- сопоставьте результаты кластеризации данных с помощью метода k-средних с результатами п. 4.;

- постройте графики зависимости суммы квадратов отклонений внутри кластера от количества кластеров и суммы квадратов отклонений между кластерами от количества кластеров для метода k-средних. Сделайте выводы.

1. Опишите полученные кластеры в терминах предметной области, дайте каждому кластеру условное название.
2. На основе проведенного статистического анализа ответьте на следующие вопросы:
   * С помощью какого иерархического метода было наиболее точно определено количество кластеров? Какая при этом использовалась мера расстояния между объектами? Какой метод оказался наименее точным? Объясните полученный результат.
   * Какой из иерархических методов дал наиболее точное разбиение объектов на кластеры? Какая при этом использовалась мера расстояния между объектами? Какой метод оказался наименее точным? Объясните полученный результат.
   * На основе использования, какого типа исходных данных (количественные, качественные) получено более точное разбиение объектов на классы и почему?
   * Какие признаки вносят наибольший вклад в разделение объектов на кластеры и почему? Проранжируйте признаки в порядке значимости.
   * Какие признаки можно не рассматривать при кластеризации объектов и почему?
   * Выделите два кластера объектов наиболее и наименее схожих по совокупности рассматриваемых признаков. Обоснуйте свой выбор.

# **Постановка задачи ИАД. Варианты 2.**

**Задача 2. (вариант 2, 5, 8, 11, 14)** Исследуются промышленныепредприятия г. Новосибирска. Необходимо разделить предприятия на группы (классы) в соответствии с рядом устойчивых показателей эффективности работы предприятия. Разделение предприятий на группы проводится с целью установления нормативных показателей для каждой выделенной группы. Также требуется построить правило отнесения предприятия к одной из выделенных групп (классов).

Исходные данные для проведения статистического анализа представлены в табл. 1. В статистическое исследование включены данные по 40 промышленным предприятиям г. Новосибирска. Зарегистрированы следующие показатели эффективности работы предприятия за 2018 г.

* объем выработки, в млн. руб.;
* фонд заработной платы, в млн. руб.;
* интегральный показатель использования рабочего времени;
* прибыль, в млн. руб.;
* уровень автоматизации производства (1 – ниже среднего; 2 – средний; 3- выше среднего);
* текучесть кадров (1 – ниже среднего уровня; 2 – средний уровень; 3 – выше среднего).

**вариант 8**

*Таблица 1.*

# Исходные данные

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **номер п/п** | **объем производства** | **прибыль** | **фонд зар. платы** | **интегр. показ.  исп. раб. времени** | **уровень  автом.** | **текуч. кадров** |
| 1 | 36,12 | 5,00 | 3,00 | 0,30 | 2 | 2 |
| 2 | 38,08 | 15,00 | 4,37 | 0,64 | 2 | 1 |
| 3 | 32,19 | 3,76 | 1,50 | 0,35 | 1 | 3 |
| 4 | 43,27 | 14,00 | 4,00 | 0,80 | 3 | 1 |
| 5 | 34,08 | 13,70 | 4,14 | 0,65 | 2 | 1 |
| 6 | 41,98 | 12,56 | 4,72 | 0,66 | 2 | 1 |
| 7 | 40,09 | 18,04 | 5,38 | 0,92 | 3 | 1 |
| 8 | 33,03 | 6,67 | 2,22 | 0,45 | 2 | 2 |
| 9 | 37,65 | 8,44 | 2,56 | 0,44 | 2 | 2 |
| 10 | 33,20 | 8,00 | 0,75 | 0,50 | 1 | 3 |
| 11 | 32,03 | 4,38 | 0,81 | 0,60 | 1 | 3 |
| 12 | 40,60 | 17,15 | 5,61 | 0,87 | 3 | 1 |
| 13 | 42,56 | 12,26 | 4,62 | 0,65 | 2 | 1 |
| 14 | 36,99 | 4,06 | 0,67 | 0,34 | 1 | 3 |
| 15 | 26,93 | 3,61 | 0,65 | 0,32 | 1 | 3 |
| 16 | 34,19 | 10,00 | 0,73 | 0,60 | 1 | 3 |
| 17 | 30,81 | 7,84 | 2,53 | 0,45 | 2 | 2 |
| 18 | 42,36 | 16,18 | 5,67 | 0,87 | 3 | 1 |
| 19 | 39,87 | 19,76 | 5,44 | 0,89 | 3 | 1 |
| 20 | 24,90 | 3,63 | 0,78 | 0,33 | 1 | 3 |
| 21 | 38,73 | 11,86 | 4,39 | 0,65 | 2 | 1 |
| 22 | 37,10 | 12,68 | 4,41 | 0,65 | 2 | 1 |
| 23 | 35,20 | 12,97 | 4,20 | 0,65 | 2 | 2 |
| 24 | 34,76 | 3,77 | 0,78 | 0,31 | 1 | 3 |
| 25 | 35,59 | 12,61 | 4,52 | 0,61 | 2 | 1 |
| 26 | 34,93 | 17,40 | 5,47 | 0,89 | 3 | 1 |
| 27 | 39,49 | 17,54 | 5,53 | 0,90 | 3 | 1 |
| 28 | 37,24 | 4,54 | 0,66 | 0,34 | 1 | 3 |
| 29 | 30,46 | 6,98 | 2,44 | 0,44 | 2 | 1 |
| 30 | 41,50 | 19,87 | 5,50 | 0,94 | 3 | 1 |
| 31 | 40,92 | 12,00 | 4,00 | 0,70 | 3 | 1 |
| 32 | 38,36 | 7,89 | 2,58 | 0,44 | 2 | 1 |
| 33 | 33,99 | 6,77 | 2,69 | 0,44 | 2 | 2 |
| 34 | 36,03 | 8,00 | 4,48 | 0,40 | 2 | 1 |
| 35 | 34,56 | 7,07 | 2,45 | 0,45 | 2 | 2 |
| 36 | 40,60 | 17,72 | 5,51 | 0,88 | 3 | 1 |
| 37 | 35,61 | 14,28 | 4,53 | 0,64 | 2 | 2 |
| 38 | 34,77 | 8,09 | 2,42 | 0,47 | 2 | 2 |
| 39 | 38,58 | 7,86 | 2,68 | 0,46 | 2 | 2 |
| 40 | 33,31 | 4,28 | 0,72 | 0,36 | 1 | 3 |
| 41 | 39,15 | 15,81 | 5,48 | 0,91 | 3 | 1 |
| 42 | 42,75 | 18,70 | 2,00 | 0,87 | 3 | 1 |
| 43 | 43,27 | 16,73 | 5,69 | 0,93 | 3 | 1 |
| 44 | 39,13 | 14,76 | 4,34 | 0,40 | 2 | 2 |
| 45 | 40,04 | 13,80 | 4,30 | 0,64 | 2 | 1 |
| 46 | 39,30 | 12,55 | 4,67 | 0,65 | 2 | 2 |
| 47 | 32,75 | 6,57 | 2,65 | 0,46 | 2 | 1 |
| 48 | 32,24 | 10,00 | 0,90 | 0,46 | 2 | 2 |
| 49 | 34,70 | 4,23 | 0,77 | 0,38 | 1 | 3 |
| 50 | 29,42 | 4,57 | 0,85 | 0,35 | 1 | 3 |

Таблица 2.

|  |  |  |
| --- | --- | --- |
| **Номер варианта** | **Иерархический метод** | **Мера расстояния** |
| 8. | *1. Single Linkage – одиночная связь*  *2. Ward's method*  *- метод Варда* | *1. Euclidean distances – евклидово расстояние*  *2. 1-Pearson r – коэффициент корреляции Пирсона* |

# Ход работы

1. Подготовили файл данных в системе *Statistica* для проведения интеллектуального анализа в соответствии с вариантом.

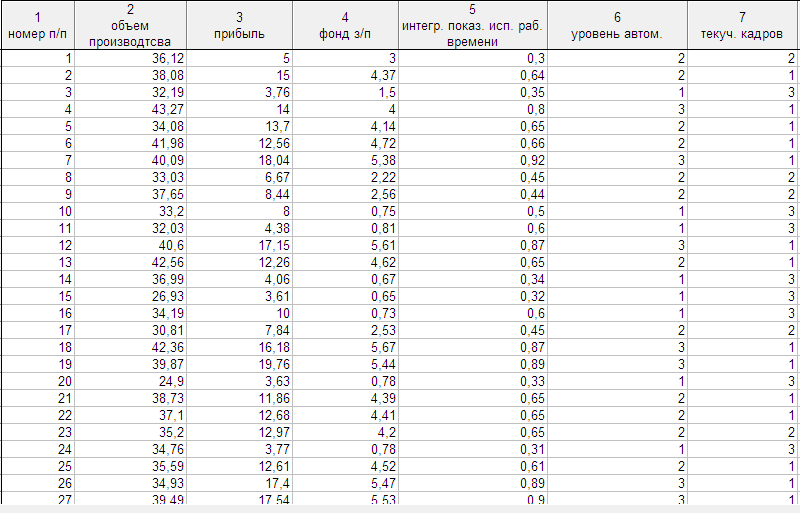


Рисунок 1 – Подготовленные исходные данные

1. Выполнили нормировку количественных признаков.

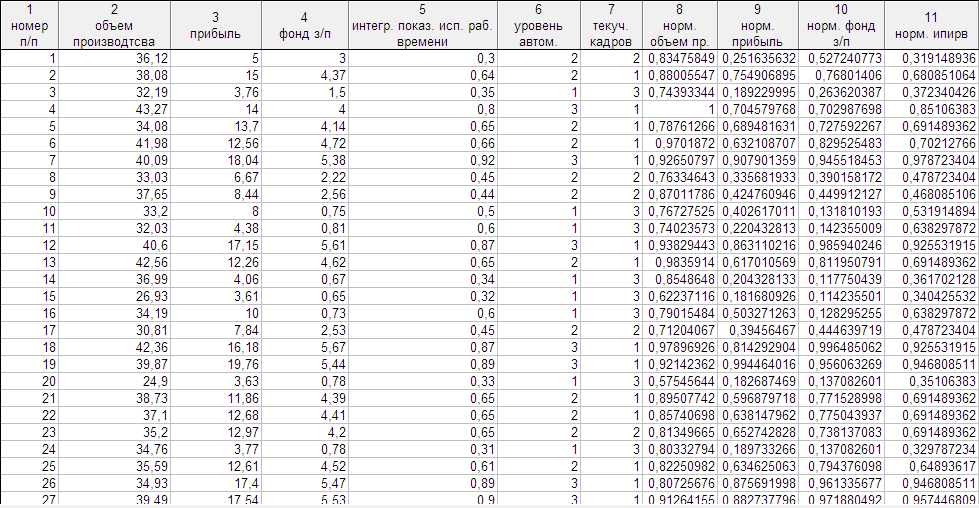


Рисунок 2 – Нормированные количественные признаки

Была выполнена кластеризация объектов с помощью иерархических агломеративных методов.

Иерархический метод: *Single Linkage – одиночная связь*

Мера расстояний: *Euclidean distances – евклидово расстояние*

Данные на рис. 3-7 демонстрируют использование полной связи с мерой расстояния евклидова расстояния для количественных параметров.

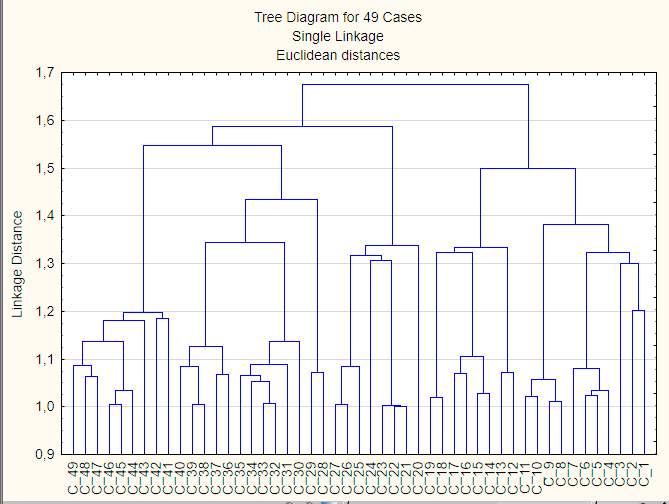


Рисунок 3 – Вертикальная дендрограмма иерархического метода

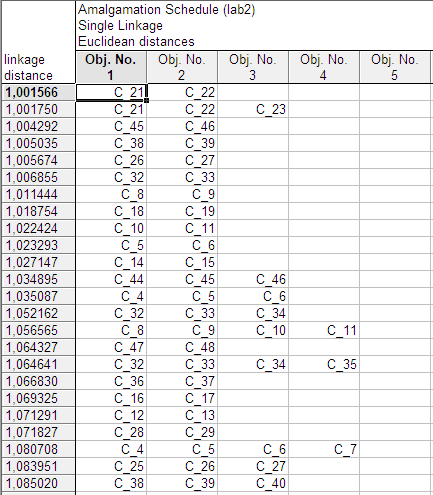


Рисунок 4 - Схема объединения

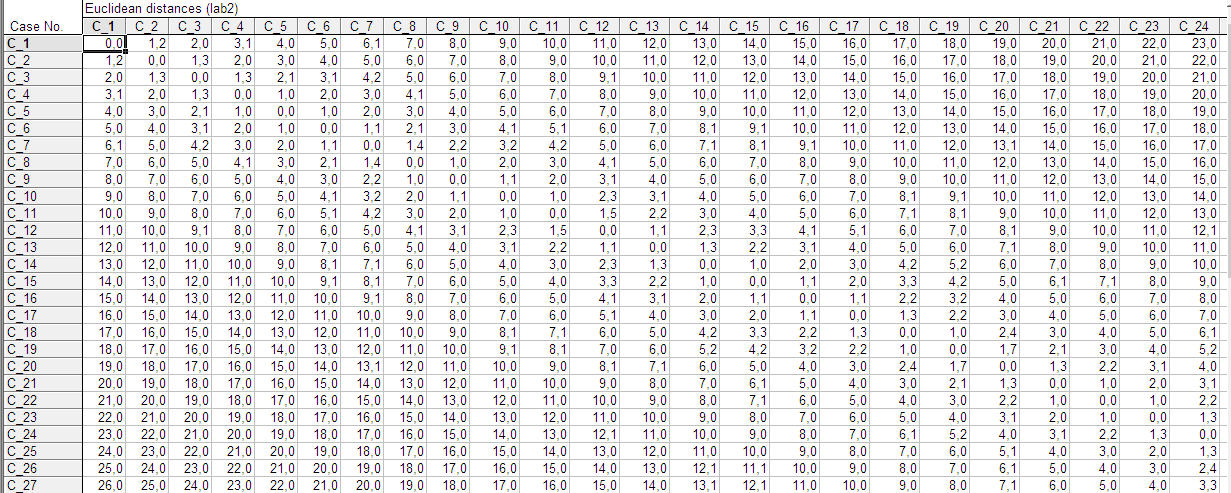


Рисунок 5 - Матрица расстояний

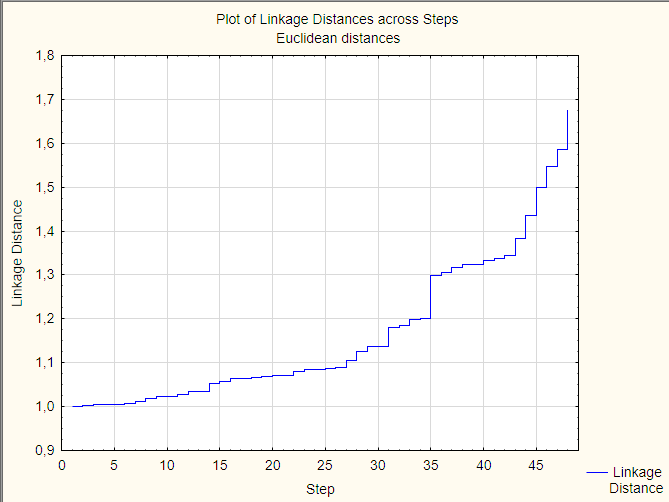


Рисунок 6 - График зависимости шагов от коэффициента слияния

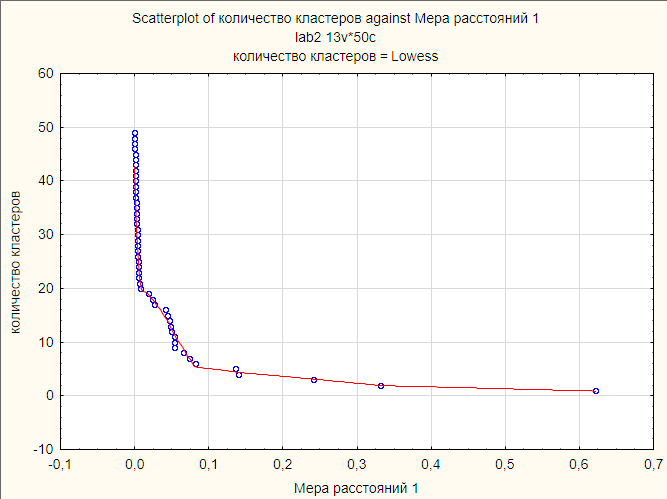
**

Рисунок 7 – График зависимости числа кластеров от коэффициента слияния

Предположительное число кластеров 4.

Иерархический метод: *Ward's method – метод Варда*

Мера расстояний: *1-Pearson r – коэффициент корреляции Пирсона*

Данные на рис. 8-12 демонстрируют использование не взвешенного попарного среднего с мерой расстояния Чебышева для количественных параметров.

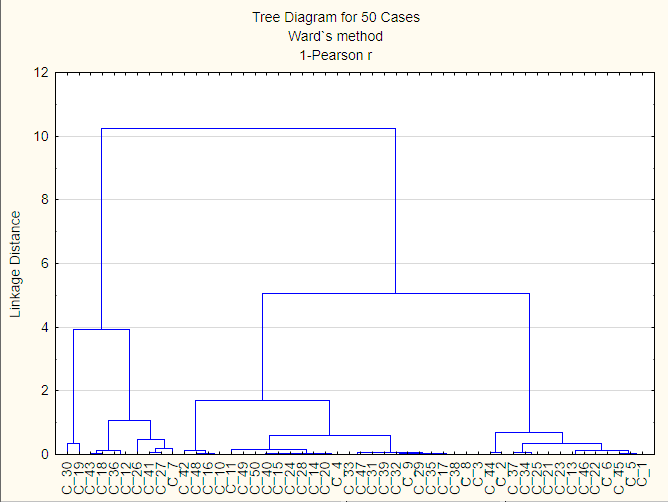


Рисунок 8 – Вертикальная дендрограмма иерархического метода

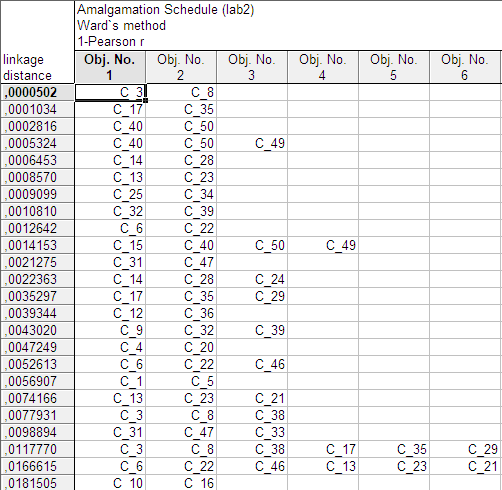


Рисунок 9 - Схема объединения

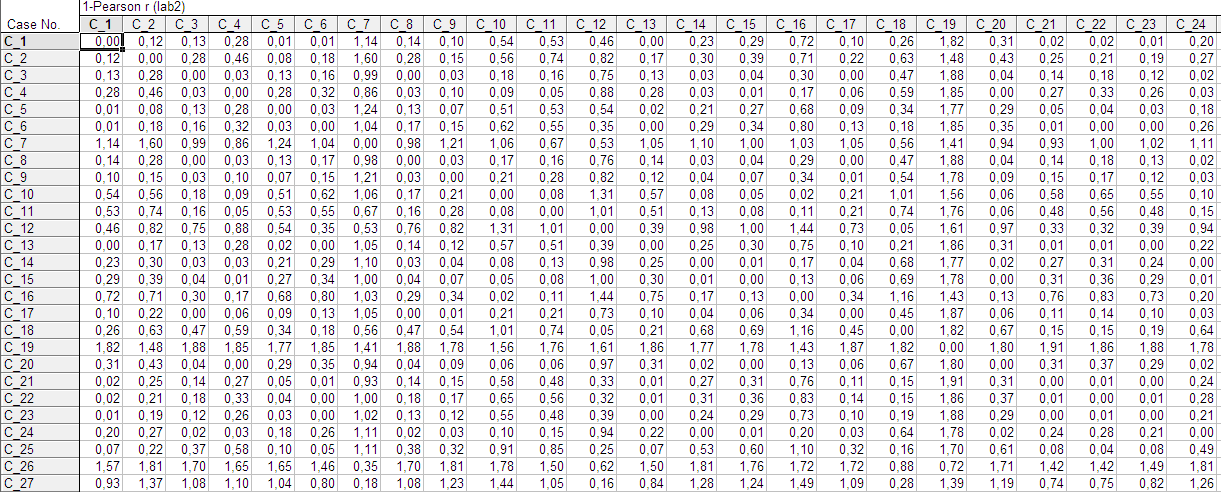
**

Рисунок 10 - Матрица расстояний

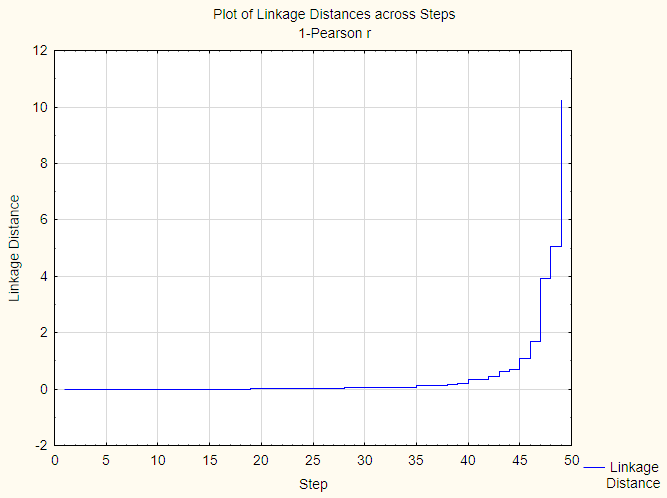


Рисунок 11 - График зависимости шагов от коэффициента слияния

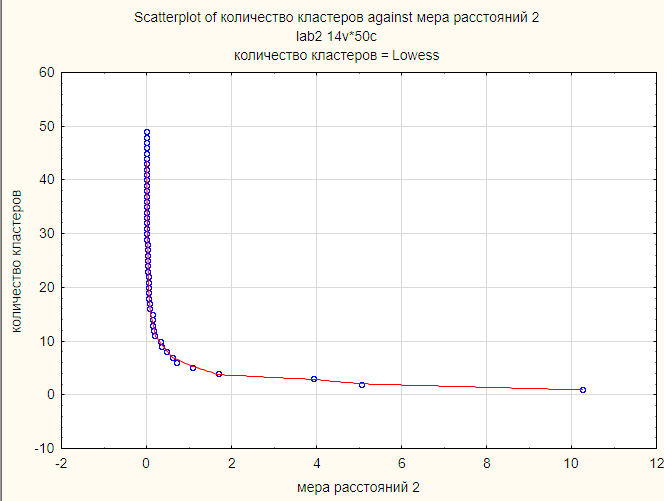
**

Рисунок 12 – График зависимости числа кластеров от коэффициента слияния

Предположительное число кластеров 4.

1. Нами была выполнена кластеризация объектов на основе использования качественных признаков.

Был использован один из предложенных в варианте методов:

Иерархический метод: *Ward's method – метод Варда*

Мера расстояний: *Percent disagreement – процент несогласия*

Данные на рис. 13-16 демонстрируют использование не взвешенного попарного среднего с мерой расстояния процента несогласия для качественных параметров.

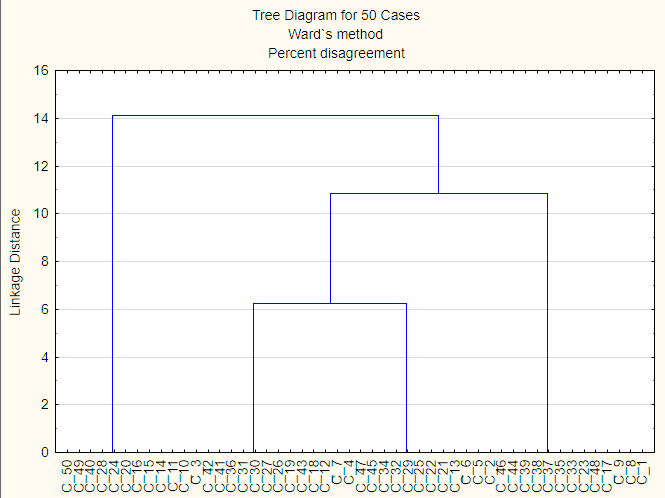


Рисунок 13 – Вертикальная дендрограмма иерархического метода

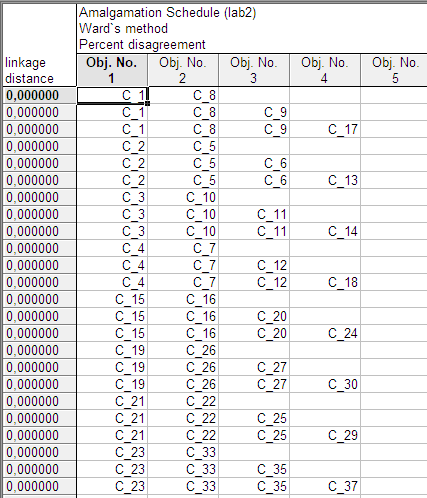


Рисунок 14 - Схема объединения

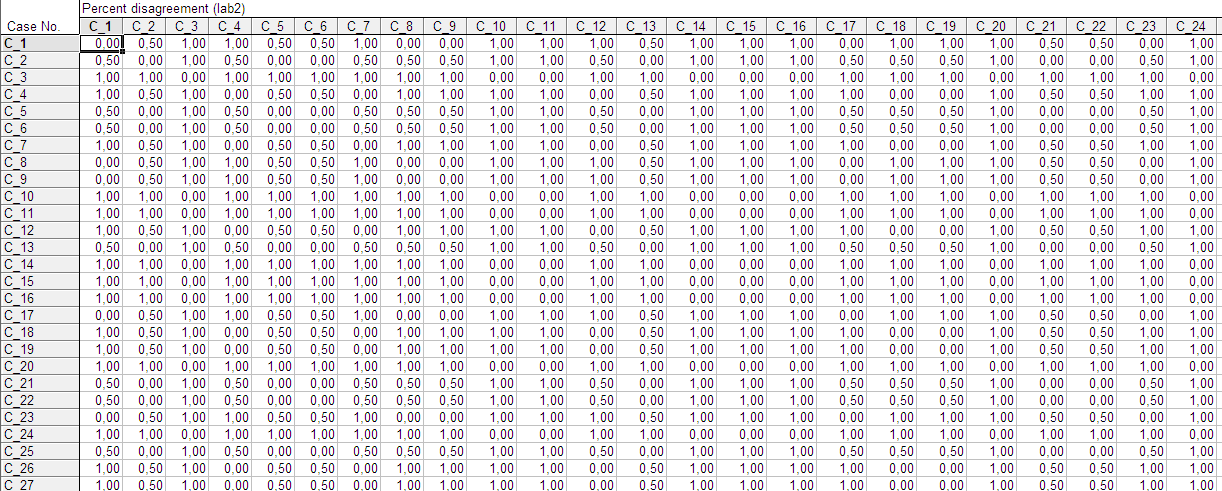
**

Рисунок 15 - Матрица расстояний

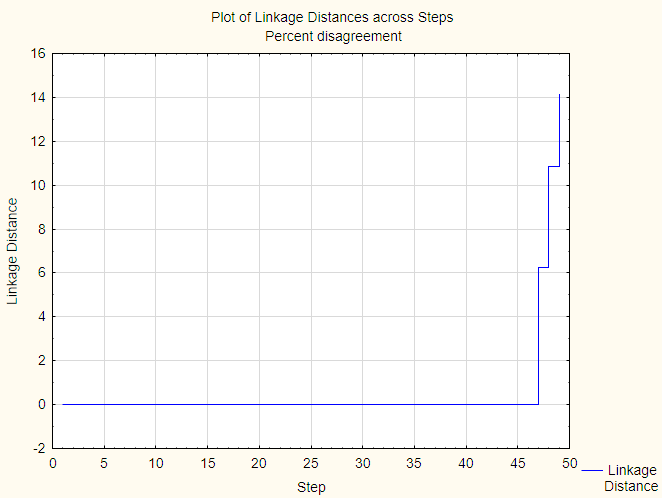


Рисунок 16 - График зависимости шагов от коэффициента слияния

Предположительное число кластеров 2.

1. Провели кластеризацию объектов, используя метод k-средних и количественные показатели в качестве исходных данных. Выполнили следующие шаги:

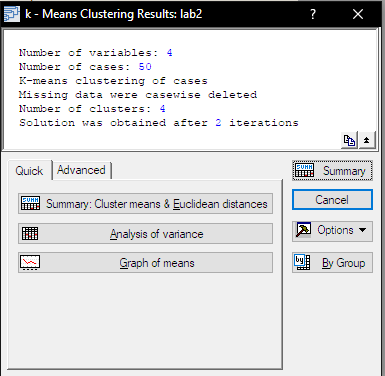


Рисунок 17 - задать число кластеров, определенное в п. 2

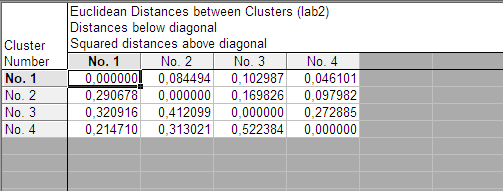


Рисунок 18 - Евклидово расстояние между кластерами

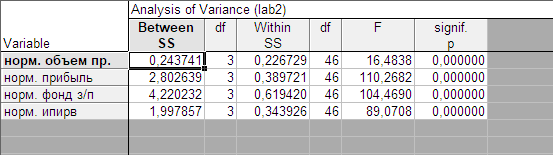


Рисунок 19 - Значимость влияния признаков на выделение кластеров на основе F-статистики

Из таблицы, представленной на рисунке 19, можно сказать, что все показатели являются статистически значимыми, так как p-значение незначительно.

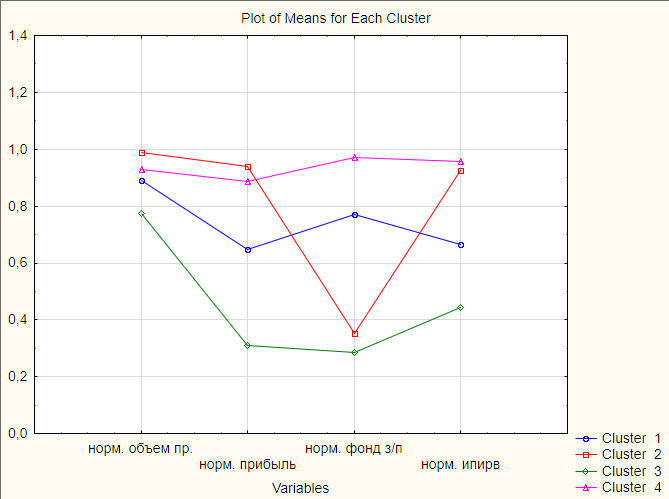


Рисунок 20 - График средних значений

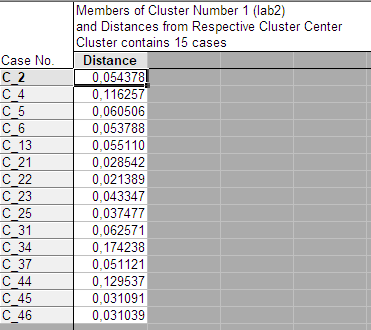
**

Рисунок 21- Объекты, входящие в кластер 1 и расстояние от ожидаемого кластерного центра

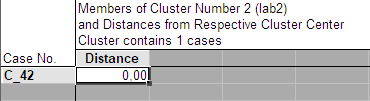
**

Рисунок 22- Объекты, входящие в кластер 2 и расстояние от ожидаемого кластерного центра

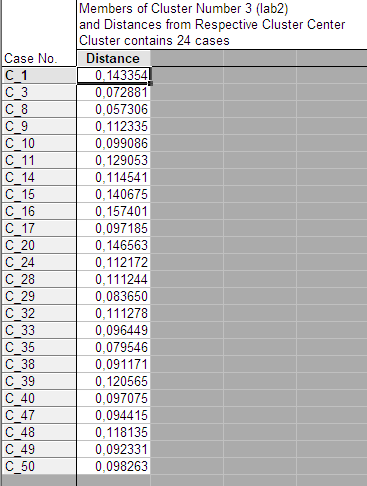
**

Рисунок 23- Объекты, входящие в кластер 3 и расстояние от ожидаемого кластерного центра

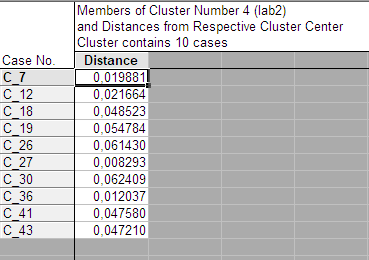
**

Рисунок 24- Объекты, входящие в кластер 4 и расстояние от ожидаемого кластерного центра

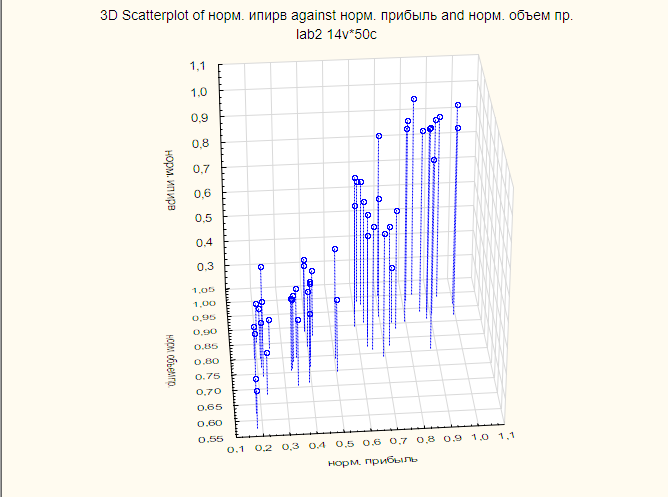
**

Рисунок 25 - Диаграмма рассеяния в трехмерном пространстве

*Таблица 3*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **метод** | **мера расстояния** | **количество кластеров** | **Количество объектов в кластере** | | | |
| **кластер 1** | **кластер 2** | **кластер 3** | **кластер 4** |
| *Single Linkage – одиночная связь* | *Euclidean distances – евклидово расстояние* | 4 | 22 | 9 | 8 | 11 |
| *Ward's method – метод Варда* | *1-Pearson r – коэффициент корреляции Пирсона* | 4 | 2 | 10 | 24 | 14 |
| *Ward's method – метод Варда* | *Percent disagreement – процент несогласия* | 2 | 13 | 13 | - | - |
| *k-средних* | *-* | 4 | 15 | 1 | 24 | 10 |

*.*

1. Столбцы с нормированными данными были сохранены в .csv-файл, запятые в десятичных числах заменены на точки. Программный код на языке R приведен в приложении.

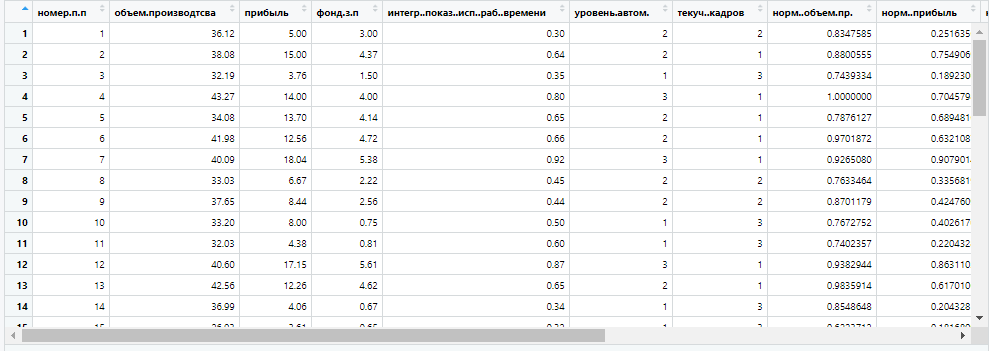


Рисунок 26 – Загруженные данные

Выполнили кластеризацию объектов средствами R, используя метод:

Иерархический метод: *Single Linkage – одиночная связь*

Мера расстояний: *Euclidean distances – евклидово расстояние*

Создали матрицу попарных расстояний (матрица расстояний Евклида). Провели кластерный анализ метод полной связи (*complete*). Результаты записали в список clust.prog.

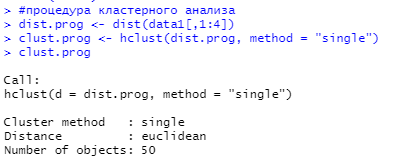
**

Рисунок 27 – Информация о входных данных для кластерного анализа

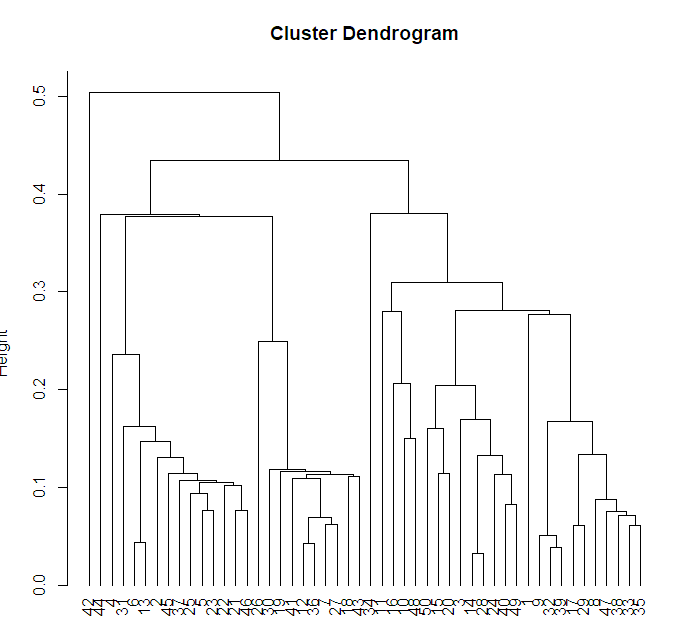
**

Рисунок 28 – Вертикальная дендрограмма

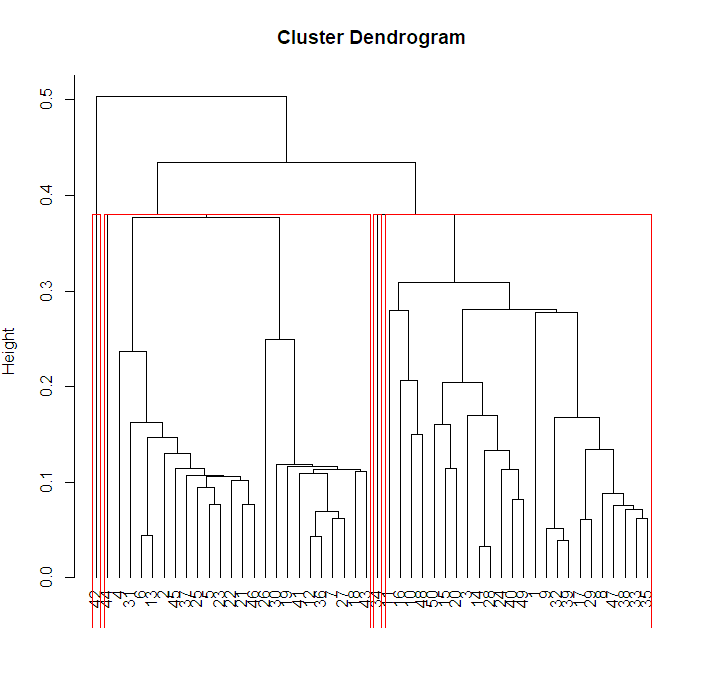
**

Рисунок 29 – Вертикальная дендрограмма с разделенными кластерами

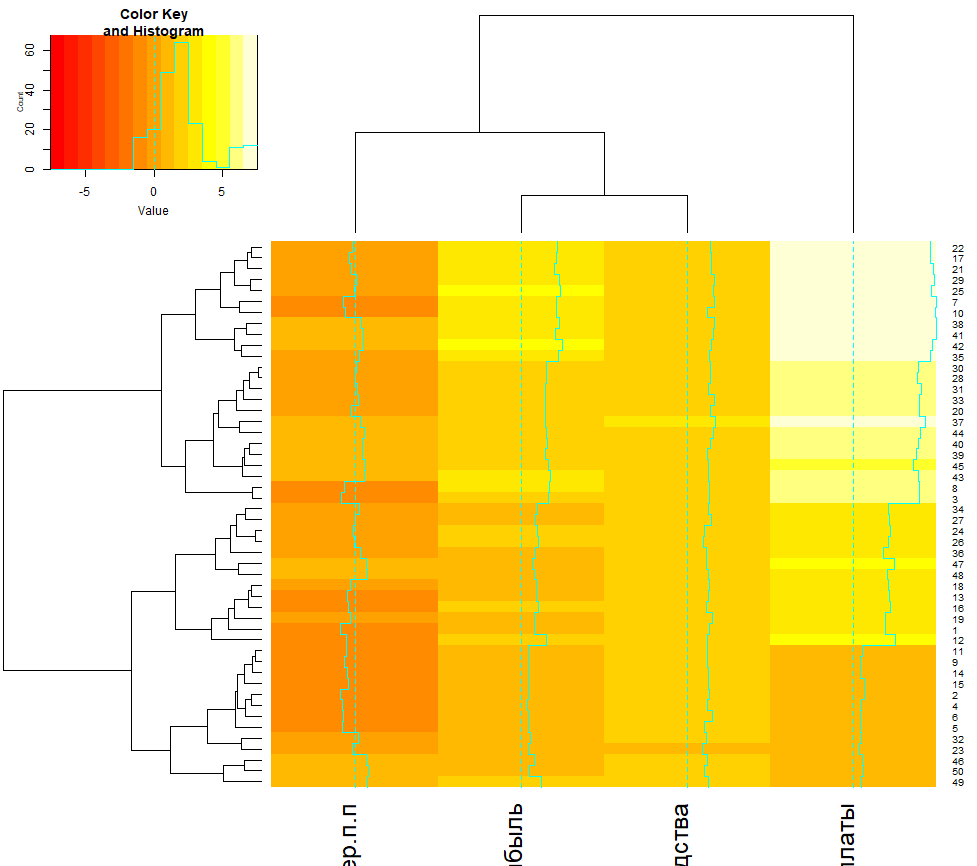


Рисунок 30- Дендрограмма + тепловая карта

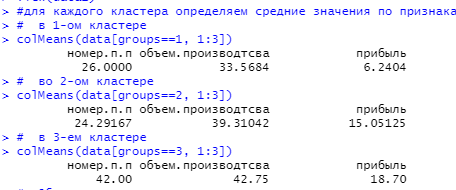
**

Рисунок 31- Средние значения по признакам

Как и в пункте 2, мы получили предполагаемое количество кластеров 3.

Была проведен кластерный анализ с использованием метода k-средних.

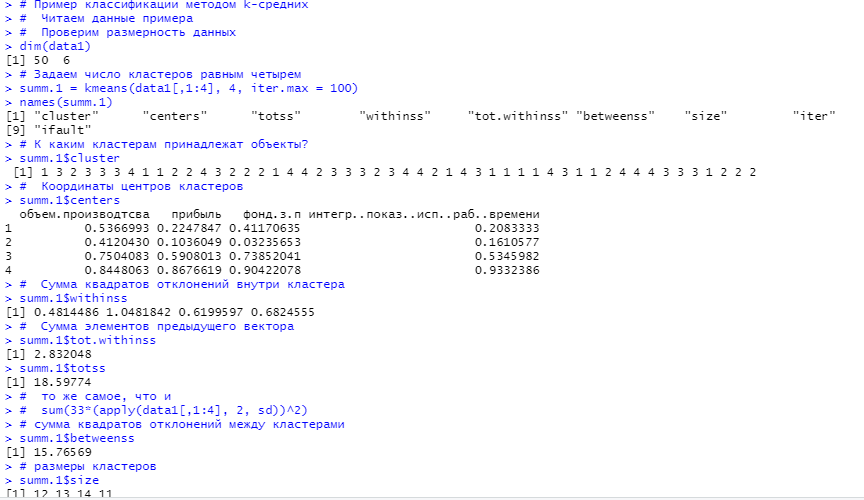


Рисунок 32 – Результаты анализа методом k-средних

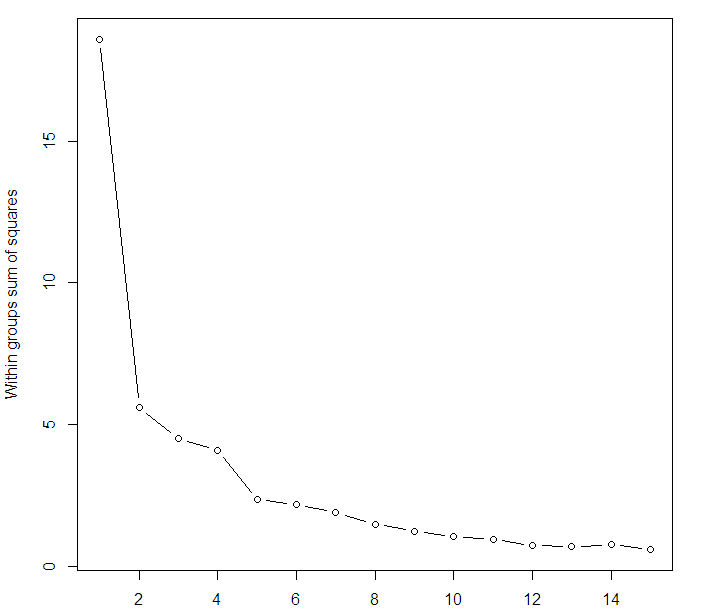


Рисунок 33- График зависимости суммы квадратов отклонений внутри кластера от числа кластеров

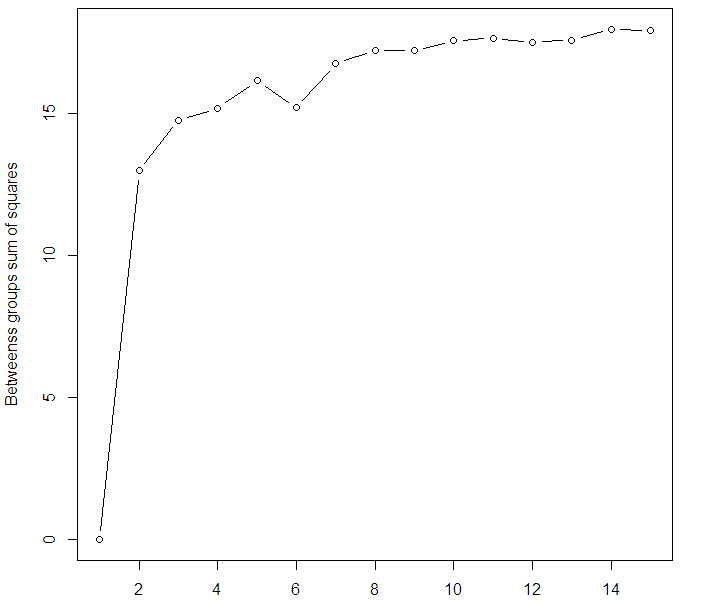
**

Рисунок 35- График зависимости суммы квадратов отклонений между кластерами от числа кластеров

# Вывод

В ходе выполнения лабораторной работы нами были изучены алгоритмы и методы кластерного анализа данных на примере решения конкретной задачи ИАД. Также исследована эффективность использования различных алгоритмов и методов кластерного анализа данных для решения прикладной задачи; Мы ознакомились и получили практические навыки работы с модулями интегрированной статистической системы *Statistica*, языка *R,*реализующими решение задачи кластеризации объектов.

# Приложение

Листинг

#загрузка и просмотр данных варианта

data <- read.table("C:\\Users\\рбт\\Desktop\\учеба\\7 семестр\\мад\\LR2.csv", header=TRUE, sep=",")

View(data)

#удаление пропущенных значений

data1 <- na.omit(data)

#cтандартизация переменных(к минимуму 0 и максимуму 1)

maxs <- apply(data[,2:5], 2, max)

mins <- apply(data[,2:5], 2, min)

data1 <- scale(data[,2:5], center = mins, scale = maxs - mins)

View(data1)

#процедура кластерного анализа

dist.prog <- dist(data1[,1:4])

clust.prog <- hclust(dist.prog, method = "single")

clust.prog

#построение дендрограммы, оценка качества кластеризации

plot(clust.prog, hang = -1)

# на дендрограмме красными прямоугольниками выделим 4 кластера

rect.hclust(clust.prog, k=4, border="red")

#вычисления кофенетической корреляции между исходной матрицей расстояний и повторной (после кластеризации)

d2<-cophenetic(clust.prog)

d2

cor(dist.prog,d2)

#определение числа кластеров

#на 3 кластера

#Вектор groups содержит номер кластера, в который попал классифицируемый объект

groups <- cutree(clust.prog, k=3)

groups

#добавить в таблицу с исходными данными

data1 <- cbind(data1, groups)

View(data1)

#для каждого кластера определяем средние значения по признакам

# в 1-ом кластере

colMeans(data[groups==1, 1:3])

# во 2-ом кластере

colMeans(data[groups==2, 1:3])

# в 3-ем кластере

colMeans(data[groups==3, 1:3])

# Обзор результатов процедуры кластерного анализа

#===================================================

# какие результаты хранятся в списке clust.prog?

names(clust.prog)

# история объединения кластеров

clust.prog$merge

# менее удобный способ получения того же результата

clust.prog[1]

# коэффициенты слияния на каждом шаге

clust.prog$height

# менее удобный способ получения того же результата

clust.prog[2]

# порядок следования объектов на дендрограмме

# позволяет определить объекты, входящие в каждый кластер.

clust.prog$order

# менее удобный способ получения того же результата

clust.prog[3]

# иерархический метод объед инения

clust.prog$method

# менее удобный способ получения того же результата

clust.prog[5]

# текст выполняемой команды

clust.prog$call

# менее удобный способ получения того же результата

clust.prog[6]

# мера расстояния

clust.prog$dist.method

# менее удобный способ получения того же результата

clust.prog[7]

#===================================================

#совмещенные тепловые карты и дендрограммы.

install.packages("gplots")

library(gplots)

dist.prog1 <- function(x) dist(x, method="euclidean")

clust.prog1 <- function(x) hclust(x, "complete")

hv3 <- heatmap.2(data3, distfun=dist.prog1, hclustfun=clust.prog1)

# Пример классификации методом k-средних

# Читаем данные примера

# Проверим размерность данных

dim(data1)

# Задаем число кластеров равным четырем

summ.1 = kmeans(data1[,1:4], 4, iter.max = 100)

names(summ.1)

# К каким кластерам принадлежат объекты?

summ.1$cluster

# Координаты центров кластеров

summ.1$centers

# Сумма квадратов отклонений внутри кластера

summ.1$withinss

# Сумма элементов предыдущего вектора

summ.1$tot.withinss

summ.1$totss

# то же самое, что и

# sum(33\*(apply(data1[,1:4], 2, sd))^2)

# сумма квадратов отклонений между кластерами

summ.1$betweenss

# размеры кластеров

summ.1$size

# Построение графика зависимости суммы квадратов отклонений внутри кластера от числа кластеров

wss <- (nrow(data1[,1:4])-1)\*sum(apply(data1[,1:4],2,var))

for (i in 2:15) wss[i] <- kmeans(data1[,1:4],

centers=i)$tot.withinss

plot(1:15, wss, type="b", xlab="Number of Clusters",

ylab="Within groups sum of squares")

# Построение графика зависимости суммы квадратов отклонений между кластерами от числа кластеров

wss <- 0

for (i in 2:15) wss[i] <- kmeans(data1[,1:4],

centers=i)$betweenss

plot(1:15, wss, type="b", xlab="Number of Clusters",

ylab="Betweenss groups sum of squares")