Министерство образования и науки Российской Федерации

Севастопольский государственный университет

Кафедра ИС

Отчет

По дисциплине: “Интеллектуальный анализ данных”

Лабораторная работа №4

“Кластерный анализ. Основные этапы и задачи

кластерного анализа данных”

Выполнил:

ст.гр. ИС/б-32

Долженко И.А.

Проверила:

Сырых О.А.

Севастополь

2020

1 ЦЕЛЬ РАБОТЫ

Закрепить теоретические знания и приобрести практические навыки в проведении кластерного анализа по экспериментальным данным. Исследовать возможности языка R для проведения кластерного анализа.

2 ХОД РАБОТЫ

**1.** **Проведем кластерный анализ методом k-средних в R.** Для этого разобьем данные на 2 кластера.

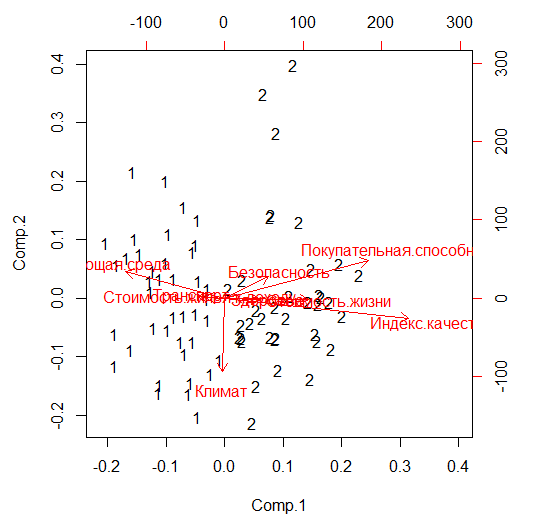


Рисунок 1 – Разбиение данных на 2 кластера

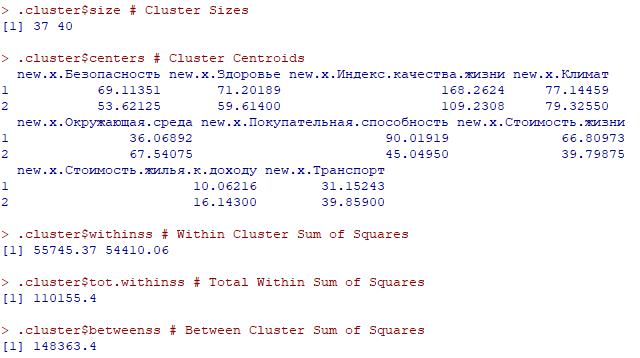


Рисунок 2 – Результат выполненного анализа

Для выбора лучшей группировки в смысле критерия минимума отношений средних внутри кластерных и меж кластерных расстояний было проведено деление на 2 – 10 кластеров и заполнена таблица.

Таблица 1 – Расчет численного показателя меры качества классификации

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Кластеры** | **Dii** | **m** | **Dij** | **J1** | **J2** | **J** |
| **2** | 110155,4 | 2 | 148363,4 | 110155,4 | 74181,7 | 1,48494 |
| **3** | 87885,72 | 3 | 170633,1 | 29295,24 | 56877,7 | 0,515057 |
| **4** | 75016,2 | 4 | 183502,7 | 12502,7 | 45875,68 | 0,272534 |
| **5** | 65351,49 | 5 | 193167,4 | 6535,149 | 38633,48 | 0,169158 |
| **6** | 58619,82 | 6 | 199899 | 3907,988 | 33316,5 | 0,117299 |
| **7** | 54212,6 | 7 | 204306,3 | 2581,552 | 29186,61 | 0,08845 |
| **8** | 48644,94 | 8 | 209873,9 | 1737,319 | 26234,24 | 0,066223 |
| **9** | 44085,24 | 9 | 214433,6 | 1224,59 | 23825,96 | 0,051397 |
| **10** | 41639,31 | 10 | 216879,5 | 925,318 | 21687,95 | 0,042665 |

Графически значения данного показателя качества классификации представлено на рис 3. Для этого построена диаграмма, на которой по оси Х – количество кластеров, по оси Y – значения показателя J.

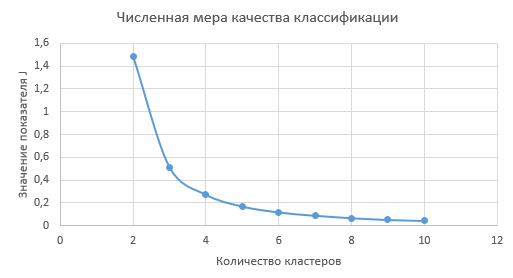


Рисунок 3 – Диаграмма численной меры качества классификации

В соответствии с этим критерием оптимальным разбиением экспериментальных данных является разбиение на 3 кластера.

**2.** **Проведем иерархический кластерный анализ в среде R.**

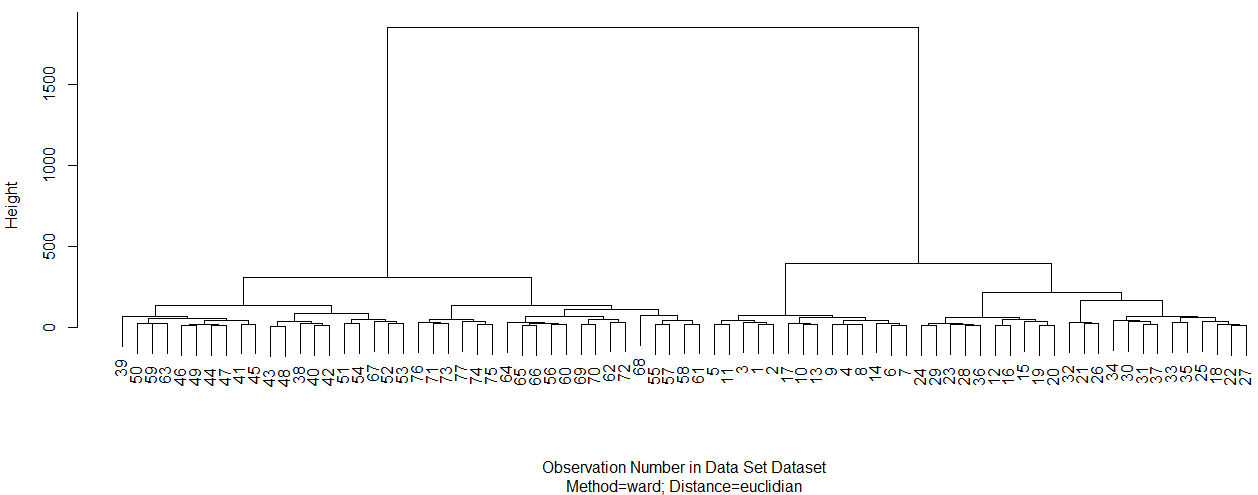


Рисунок 4 – Метод Уорда

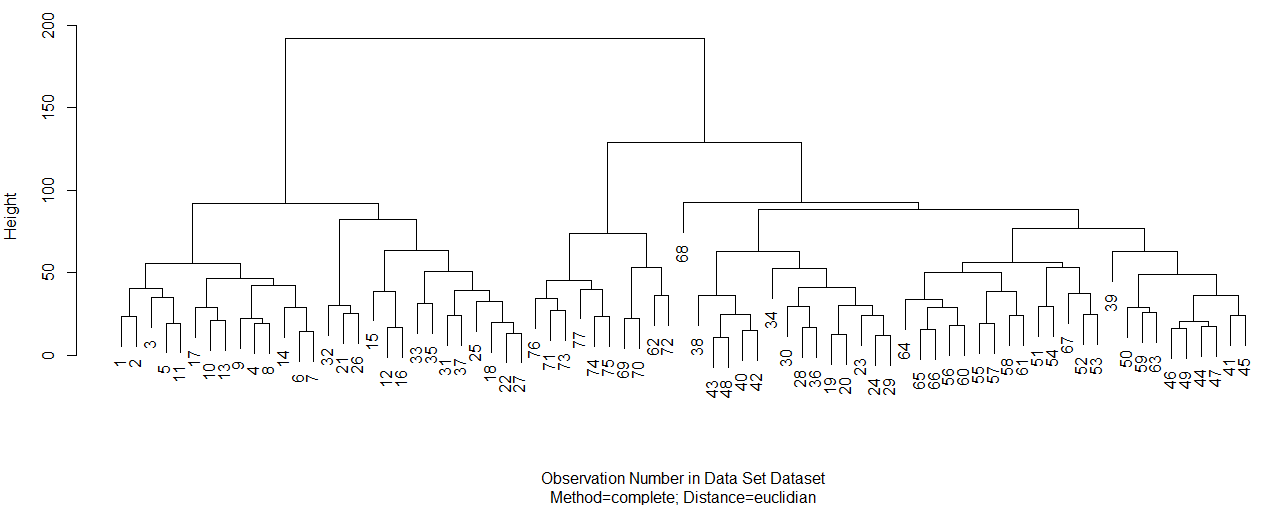


Рисунок 5 – Метод полной связи

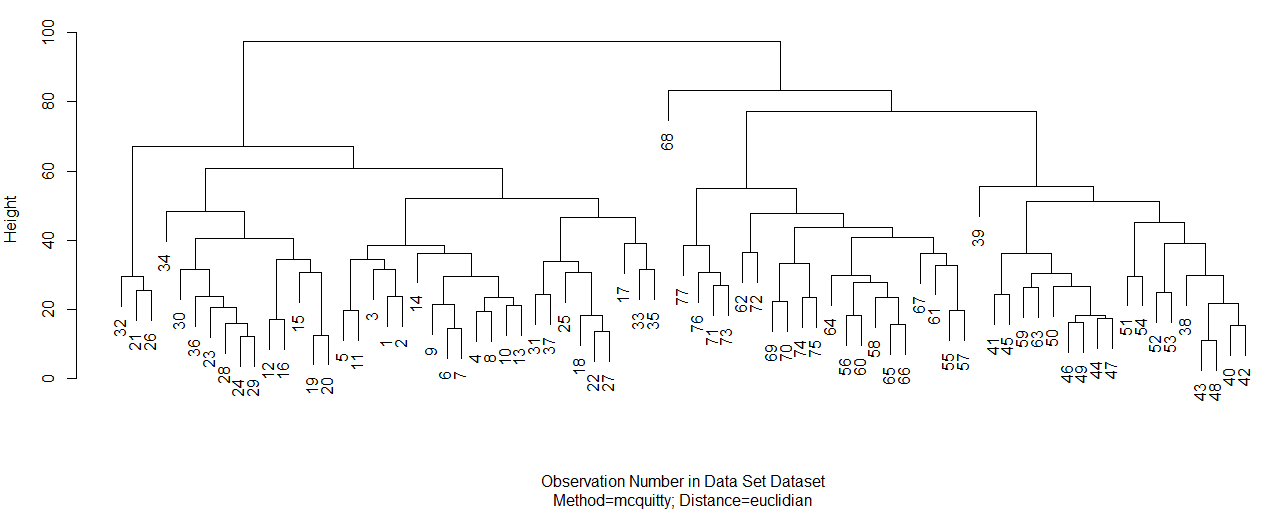


Рисунок 6 – Метод Мак-Квитти

ВЫВОД

В ходе выполнения лабораторной работы были закреплены теоретические знания и приобретены практические навыки в проведении кластерного анализа по экспериментальным данным. Исследованы возможности языка R для проведения кластерного анализа.

Для выбора лучшей группировки в смысле критерия минимума отношений средних внутри кластерных и меж кластерных расстояний было проведено деление на 2 – 10 кластеров, проведена таблица и графически было выявлено, что оптимальным разбиением экспериментальных данных является разбиение на 3 кластера.