

Министерство науки и высшего образования Российской Федерации  
Севастопольский государственный университет  
Кафедра ИС

Отчет  
по лабораторной работе №4  
«Кластерный анализ. Основные этапы и задачи кластерного анализа данных»  
по дисциплине  
«ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»

Выполнил студент группы ИС/б-17-2-о

Горбенко К. Н.

Проверил

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.

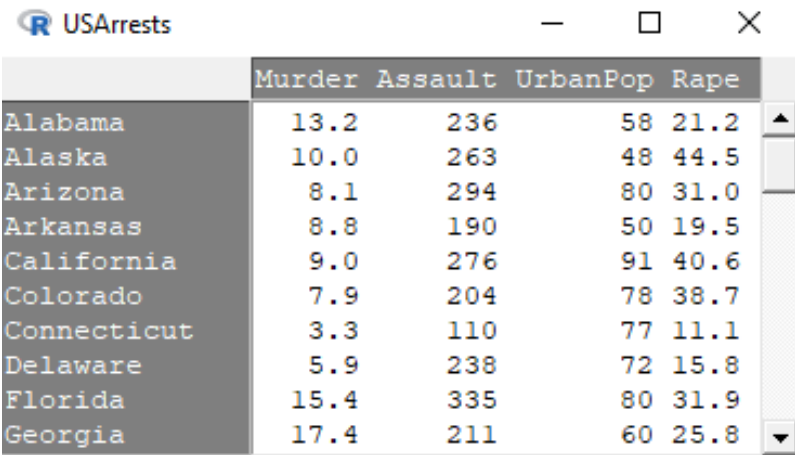
## 2 ЗАДАНИЕ НА РАБОТУ

1. Создать файл с исходными данными.
2. Провести кластерный анализ экспериментальных данных.
3. Проведя процедуру кластеризации несколько раз при различных значениях числа кластеров (от 2-х до 10 кластеров), необходимо выбрать лучшую группировку в смысле критерия минимума отношений средних внутри кластерных и меж кластерных расстояний. Полученные результаты оформите в виде таблицы. Изобразить графически значения данного показателя качества классификации. Для этого построить диаграмму, на которой по оси X – количество кластеров, по оси Y – значения показателя J.

## 3 ХОД РАБОТЫ

### 3.1 Кластерный анализ

Выполним иерархический кластерный анализ. Загрузим данные о количестве арестов в США. Для этого воспользуемся известным набором данных из R, название которого «USArrests».



	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8

Рисунок 1 – Загруженные данные

- murder - количество задержаний за убийство;
- assault - количество задержаний за нападение;
- rape - количество задержаний за изнасилование;
- urban pop - процент городского населения.

Функция кластерного анализа в R:

```
kmeans(x, centers, iter.max=10, nstart=1,
algorithm=c("Hartigan-Wong" "Lloyd" "Forgy" "MacQueen"))
```

Выполним разбиение на два кластера по переменным «Количество убийств» и «Количество нападений». Результат разбиения изображен на рисунке 2.

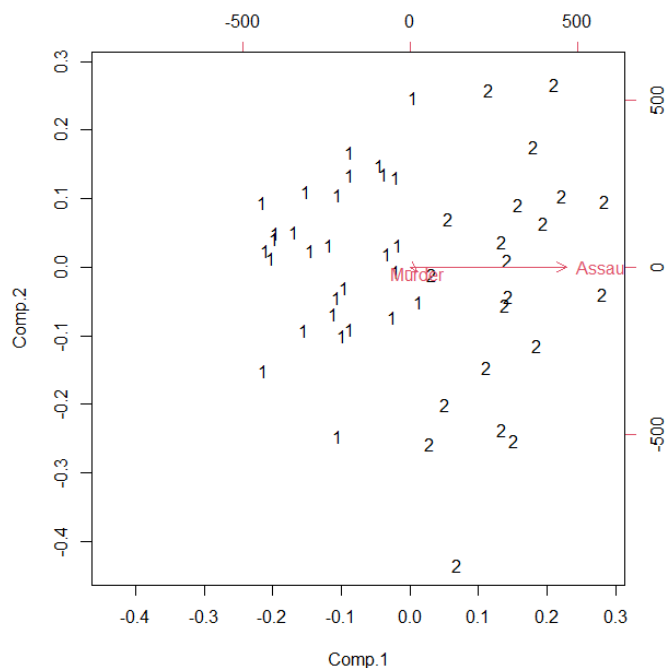


Рисунок 2 – Разбиение данных на 2 кластера

Результаты разбиения представлены на рисунке 3.

```
Rcmdr> .cluster <- kMeans(model.matrix(~1 + Assault + Murder, USArrests), centers = 2, iter.max = 10, num.seeds = 10)
Rcmdr> .cluster$size # cluster Sizes
[1] 29 21
Rcmdr> .cluster$centers # Cluster centroids
      new.x.Assault new.x.Murder
1      109.7586     4.841379
2      255.0000    11.857143
Rcmdr> .cluster$withinss # within Cluster Sum of Squares
[1] 47963.82 35741.53
Rcmdr> .cluster$tot.withinss # Total within Sum of Squares
[1] 83705.35
Rcmdr> .cluster$betweenss # Between Cluster Sum of Squares
[1] 257537.3
Rcmdr> biplot(princomp(model.matrix(~1 + Assault + Murder, USArrests)), xlab = as.character(.cluster$cluster))
Rcmdr> remove(.cluster)
```

Рисунок 3 – Результаты разбиения

- первый кластер содержит 29 элементов, второй – 21;
- сумма квадратов расстояний внутри кластера: 1 – 47963, 2 – 35741;
- общая сумма квадратов расстояний внутри кластеров: 83705;
- сумма квадратов расстояний между кластерами – 257537.

Для выбора лучшей группировки в смысле критерия минимума отношений средних внутри кластерных и меж кластерных расстояний было проведено деление на 3 – 10 кластеров и заполнена таблица в MS Excel.

m	Dii	Dij	J1	J2	J
2	83705	257537	83705	128,769	0.650043
3	36346	304896	12115	101,632	0.119205
4	23960	317281	5990	79,320	0.075517
5	13811	327431	2762	65,486	0.042177
6	8440	332801	1406	55,467	0.025348
7	5306	335936	758	47,991	0.015795
8	4484	336758	560	42,095	0.013303
9	2369	338873	263	37,653	0.006985
10	2271	338970	227	33,897	0.006697

Рисунок 4 – Расчет численного показателя меры качества классификации

Значения данного показателя качества классификации представлено графически на рис 4. Для этого построена диаграмма, на которой по оси X – количество кластеров, по оси Y – значения показателя J.

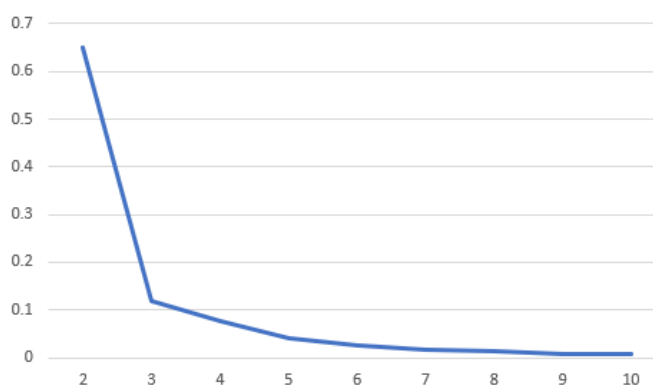


Рисунок 5 – Диаграмма численной меры качества классификации

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

### 3.2 Иерархический анализ

Был проведен иерархический анализ методом Уорда:

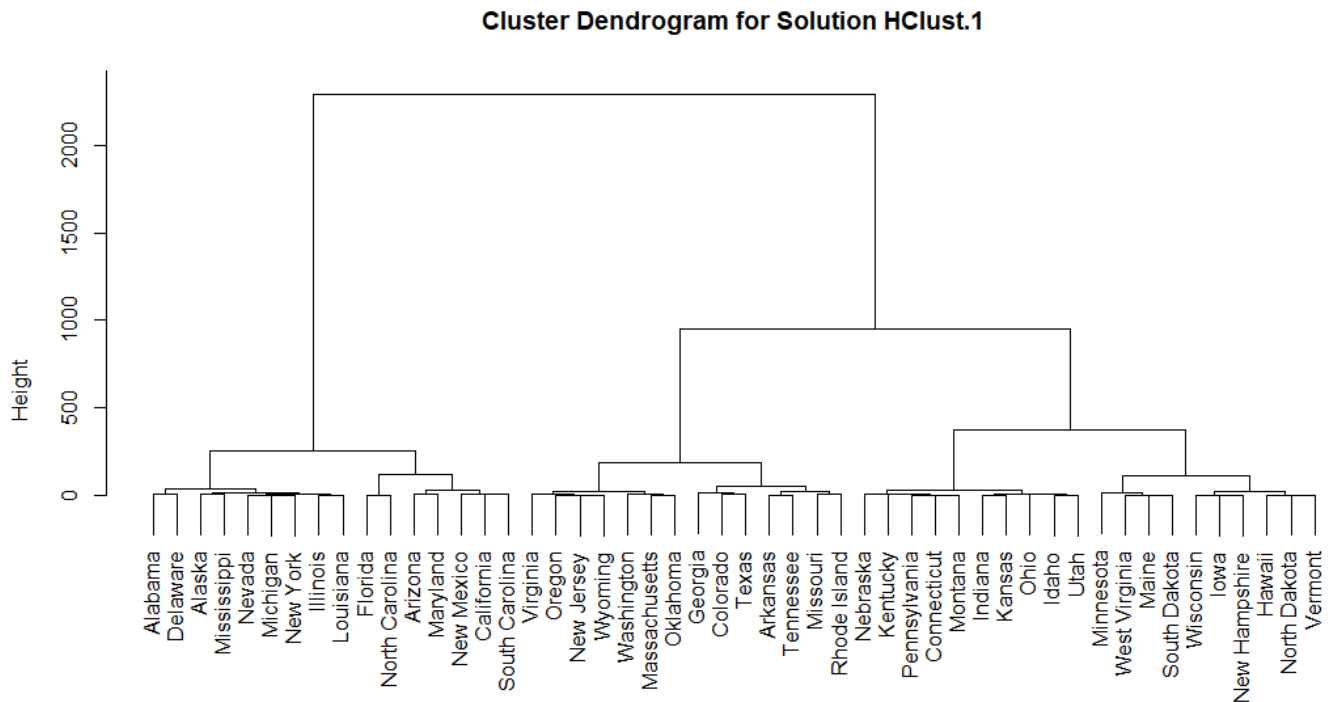


Рисунок 6 – Иерархический анализ методом Уорда

Результатом анализа является количество элементов в каждом из кластеров:

```
1 Rcmdr> summary(as.factor(cutree(HClust.1, k = 3))) # Cluster Sizes
2 1    2    3
3 16  14  20
```

Был проведен иерархический анализ методом простой связи:

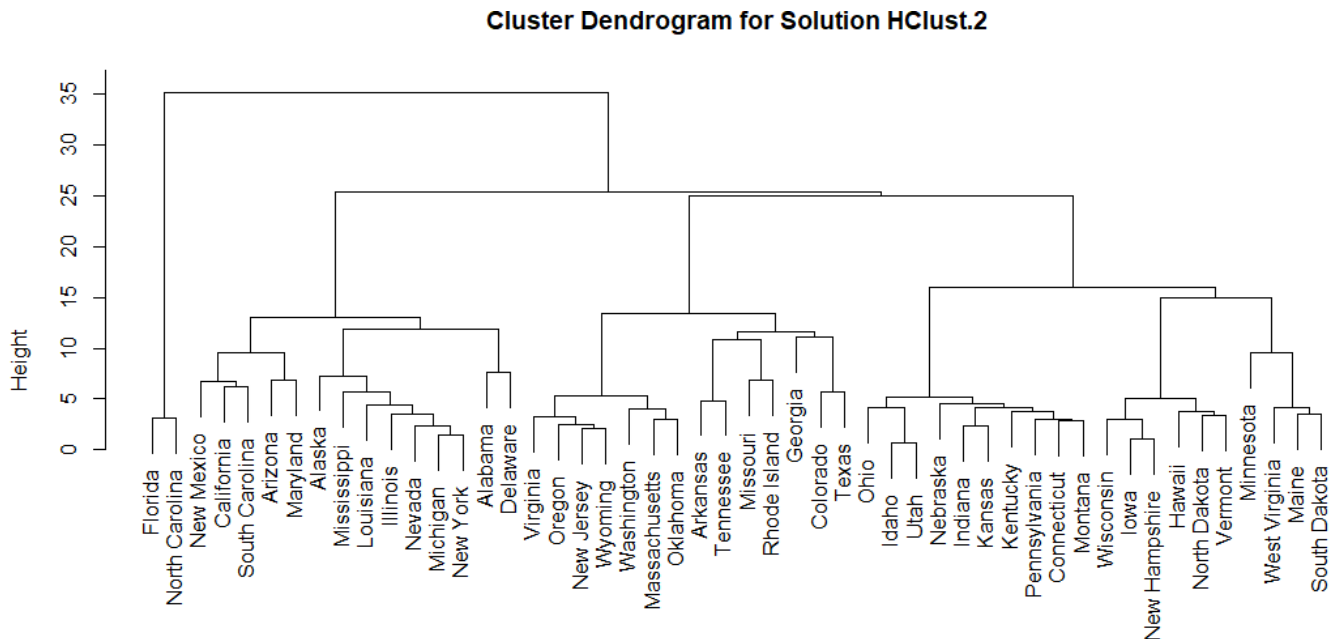


Рисунок 7 – Иерархический анализ методом простой связи

Результатом анализа является количество элементов в каждом из кластеров:

```
1 Rcmdr> summary(as.factor(cutree(HClust.2, k = 3))) # Cluster Sizes
2 1 2 3
3 14 34 2
```

## ВЫВОДЫ

В ходе лабораторной работы провели многомерный анализ данных. Для этого использовали кластеризацию, которая предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек". Целью кластеризации является поиск существующих структур. Кластерный анализ позволяет сокращать размерность данных, делать ее наглядной. Провели иерархический анализ с использованием k-средних. Данный алгоритм является наиболее распространенным.