Многомерный статистический анализ

Кластерный анализ

- **Многомерный статистический анализ** - раздел статистики математической, посвященный математическим методам, направленным на выявление характера и структуры взаимосвязей между компонентами исследуемого многомерного признака и предназначенным для получения научных и практических выводов.

Основные подразделы:

- **Анализ многомерных распределений и их основных характеристик**

- ~~**Анализ характера и структуры взаимосвязей компонент исследуемого многомерного признака:**~~

1. анализ регрессионный,
2. анализ дисперсионный,
3. анализ ковариационнй,
4. анализ факторный,
5. анализ латентно-структурный,
6. анализ логлинейный,
7. поиск взаимодействий

- **Анализ геометрической структуры исследуемой совокупности многомерных наблюдений :**

1. анализ дискриминантный,
2. анализ кластерный,
3. шкалирование многомерное

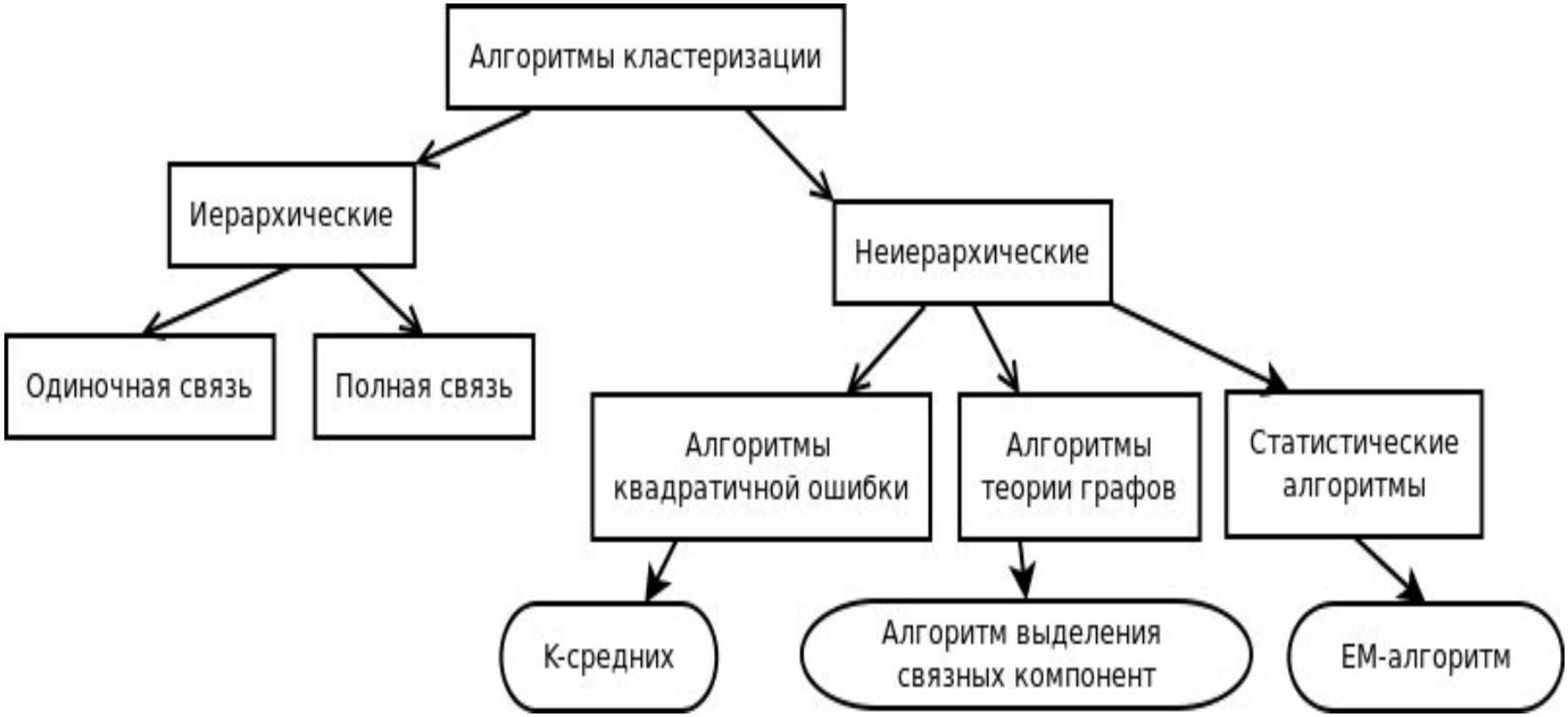
Прикладное значение многомерного статистического анализа:

- - проблемы статистического исследования зависимостей между рассматриваемыми показателями;
- - проблемы классификации элементов (объектов или признаков);
- - проблемы снижения размерности рассматриваемого признакового пространства и отбора наиболее информативных признаков.

Кластерный анализ:

- «Совокупность математических методов, предназначенных для формирования относительно "отдаленных" друг от друга групп "близких" между собой объектов по информации о расстояниях или связях (мерах близости) между ними. По смыслу аналогичен терминам: автоматическая классификация, таксономия, распознавание образов без учителя." ("Статистический словарь»)
- Это обобщенное название достаточно большого набора алгоритмов, используемых при создании классификации. В ряде изданий используются и такие синонимы кластерного анализа, как классификация и разбиение.

Алгоритмы кластеризации

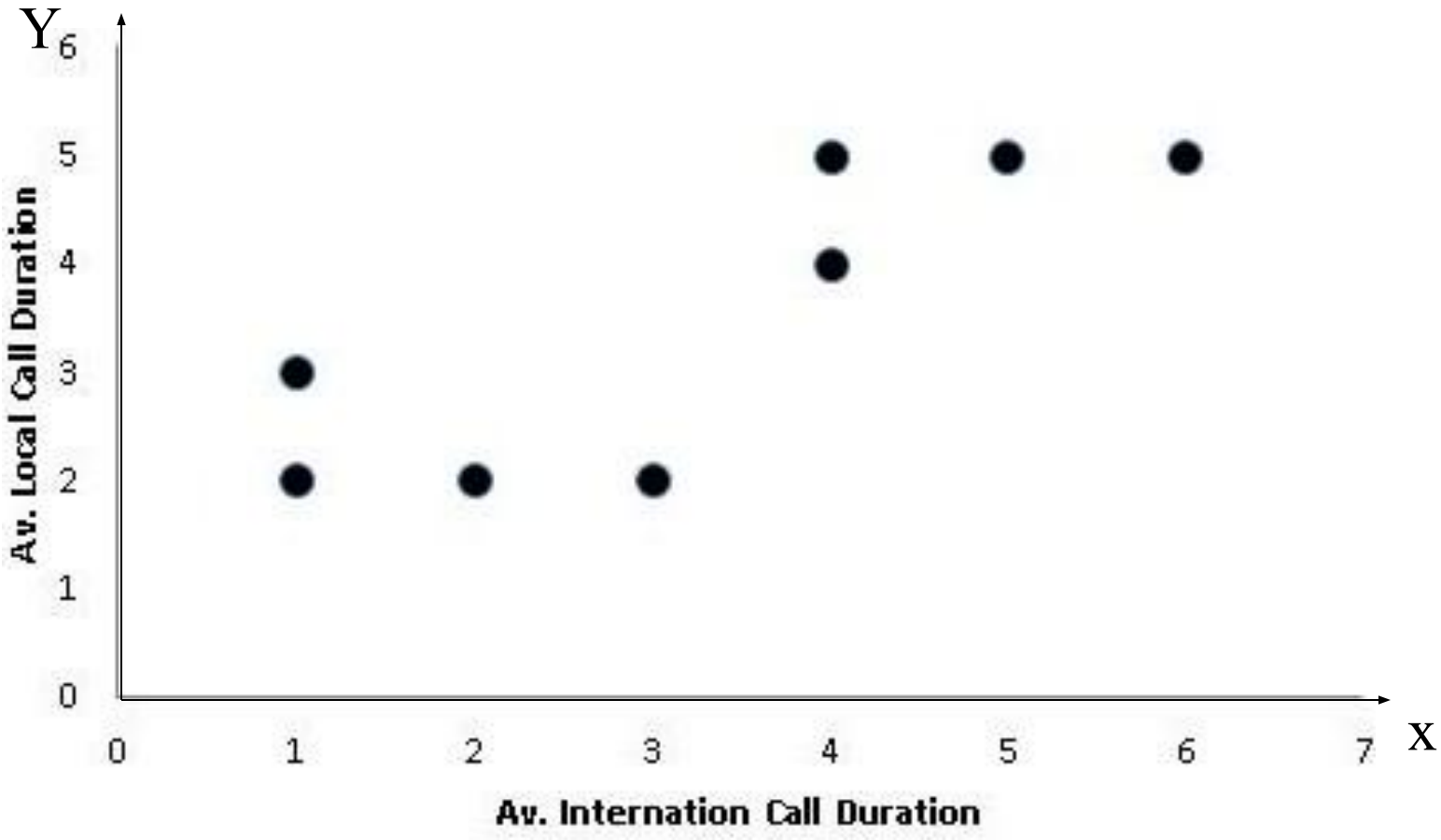


Кластерный анализ (на примере сегментации потребителей)

8 потребителей и средняя продолжительность их разговоров (локальных и международных).

Customer#	Av. Local Call Duration	Av. International Call Duration
1	2	2
2	1	2
3	1	3
4	3	2
5	4	5
6	4	4
7	5	5
8	6	5

x


Y
6
5
4
3
2
1
0
Av. Local Call Duration
0
1
2
3
4
5
6
7
x
Av. Internation Call Duration

Евклидово расстояние для нахождения Центроидов для Кластеров

$$Distance = \sqrt{(X_{centroid\,C_1} - X_i)^2 + (Y_{centroid\,C_1} - Y_i)^2}$$

Расстояние может быть вычислено и по другим формулам:

- **квадрат евклидова расстояния** – для придания веса более отдаленным друг от друга объектам
- **манхэттенское расстояние** – для уменьшения влияния выбросов
- **степенное расстояние** – для увеличения/уменьшения влияния по конкретным координатам
- **процент несогласия** – для категориальных данных

и др.

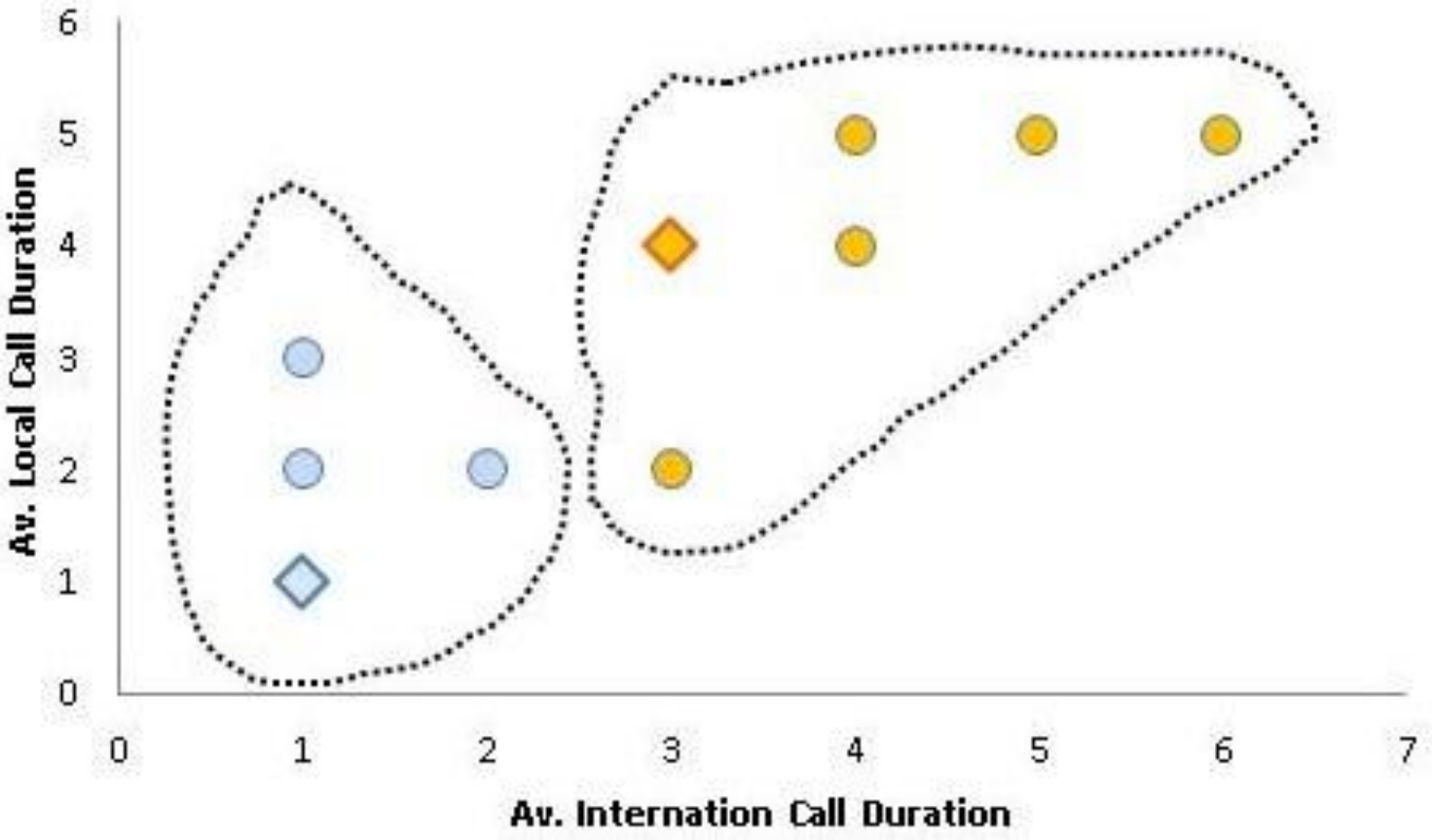
Av. Local Call Duration	Av. International Call Duration	Distance from C_1	Distance from C_2	Cluster Membership
2	2	1.41	2.24	C_1
1	2	1.00	2.83	C_1
1	3	2.00	2.24	C_1
3	2	2.24	2.00	C_2
4	5	5.00	1.41	C_2
4	4	4.24	1.00	C_2
5	5	5.66	2.24	C_2
6	5	6.40	3.16	C_2

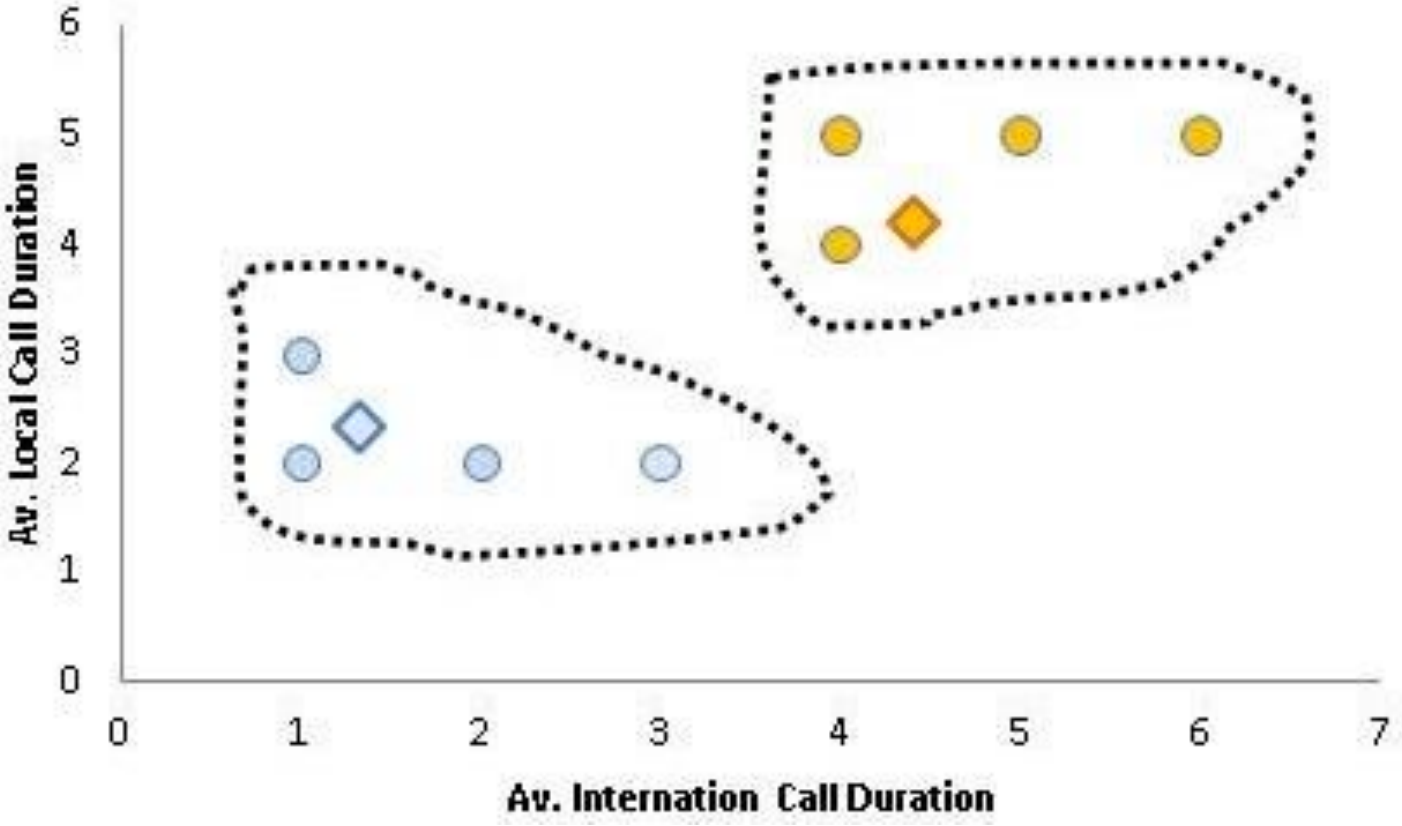
Расстояние до C1 и C2

Пренадлежность к C1 или к C2

Для первого потребителя:

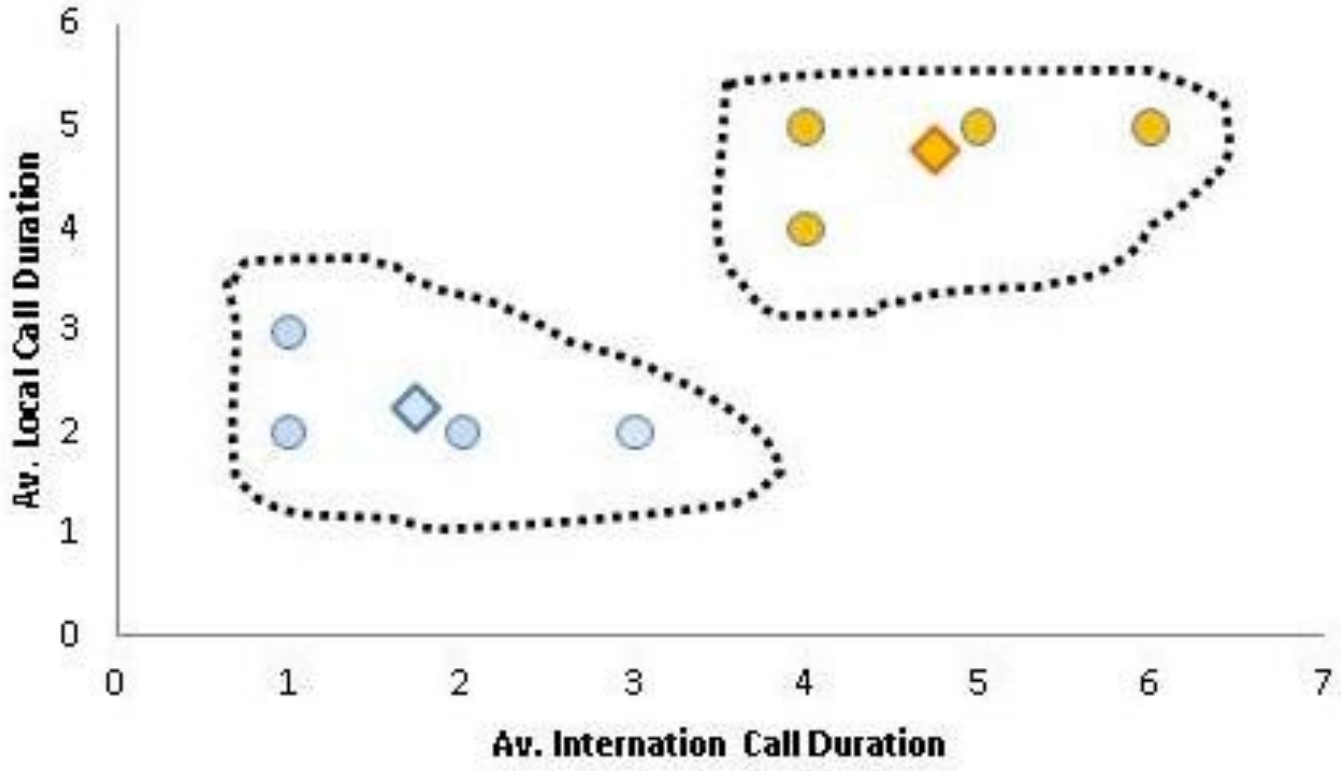
$$Distance\ from\ C_1 = \sqrt{(1-2)^2 + (1-2)^2} = \sqrt{2} = 1.41$$


Av. Local Call Duration
0
1
2
3
4
5
6
0
1
2
3
4
5
6
7
Av. Internation Call Duration



- C1 (1.33, 2.33) и C2 (4.4, 4.2)

C1 (1.75, 2.25) и C2(4.75, 4.75)



C1 (1.75, 2.25) и C2(4.75, 4.75)

Нормализация данных

$$X^* = \frac{X - min(X)}{max(X) - min(X)}$$

$$Y^* = \frac{(90000 - 45000)}{(130000 - 45000)} = 0.529$$