Липецкий государственный технический университет

Кафедра прикладной математики

МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ

Лекция 7

7. Методы кластерного анализа

Составитель - Сысоев А.С., к.т.н., доц.

Липецк – 2021

Outline

7.1. Задачи кластерного анализа

Задача кластеризации (обучения без учителя) заключается в следующем. Имеется обучающая выборка $\mathbf{X^\ell = \{x_1, \ldots, x_\ell\} \subset X}$ и функция расстояния между объектами $\boldsymbol{\rho(x, x')}$. Требуется разбить выборку на непересекающиеся подмножества, называемые *кластерами*, так, чтобы каждый кластер состоял из объектов, близких по метрике $\boldsymbol{\rho}$, а объекты разных кластеров существенно отличались. При этом каждому объекту $\mathbf{x_i \in X^\ell}$ приписывается метка (номер) кластера $\mathbf{y_i}$.

Алгоритм кластеризации — это функция $\mathbf{a: X \to Y}$, которая любому объекту $\mathbf{x \in X}$ ставит в соответствие метку кластера $\mathbf{y \in Y}$. Множество меток $\mathbf{Y}$ в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации.

Решение задачи кластеризации принципиально неоднозначно:

- не существует однозначно наилучшего критерия качества кластеризации,
- число кластеров неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием,
- результат кластеризации существенно зависит от метрики $\boldsymbol{\rho}$.

7.1. Задачи кластерного анализа

Цели кластеризации могут быть различными в зависимости от особенностей конкретной прикладной задачи:

- Понять структуру множества объектов $\mathbf{X}^\ell$, разбив его на группы схожих объектов. Упростить дальнейшую обработку данных и принятия решений, работая с каждым кластером по отдельности (стратегия «разделяй и властвуй»).
- Сократить объём хранимых данных в случае сверхбольшой выборки $\mathbf{X}^\ell$, оставив по одному наиболее типичному представителю от каждого кластера.
- Выделить нетипичные объекты, которые не подходят ни к одному из кластеров. Эту задачу называют одноклассовой классификацией, обнаружением нетипичности или новизны (novelty detection).

7.2. Эвристические графовые алгоритмы

Вершинам графа соответствуют объекты выборки, а рёбрам — попарные расстояния между объектами $\rho_{ij} = \rho(x_i, x_j)$.

Алгоритм выделения связных компонент. Задаётся параметр **R** и в графе удаляются все рёбра **(i, j)**, для которых $\rho_{ij} > R$. Соединёнными остаются только наиболее близкие пары объектов. Идея алгоритма заключается в том, чтобы подобрать такое значение $R \in [\min \rho_{ij}, \max \rho_{ij}]$, при котором граф развалится на несколько связных компонент. Найденные связные компоненты — и есть кластеры.

Связной компонентой графа называется подмножество его вершин, в котором любые две вершины можно соединить путём, целиком лежащим в этом подмножестве. Для поиска связных компонент можно использовать стандартные алгоритмы поиска в ширину (алгоритм Дейкстры) или поиска в глубину.

Недостатки:

- ограниченная применимость,
- плохая управляемость числом кластеров.

7.2. Эвристические графовые алгоритмы

Алгоритм кратчайшего незамкнутого пути строит граф из **$\ell-1$** рёбер так, чтобы они соединяли все **ℓ** точек и обладали минимальной суммарной длиной. Такой граф называется *кратчайшим незамкнутым путём*, минимальным покрывающим деревом или каркасом.

Алгоритм 1.1. Алгоритм кратчайшего незамкнутого пути (КНП)

1: Найти пару точек (i, j) с наименьшим ρ_{ij} и соединить их ребром;
2: **пока** в выборке остаются изолированные точки
3: найти изолированную точку, ближайшую к некоторой неизолированной;
4: соединить эти две точки ребром;
5: удалить $K - 1$ самых длинных рёбер;

Недостатки:

- ограниченная применимость,
- высокая трудоёмкость.

7.2. Эвристические графовые алгоритмы

Алгоритм FOREL (ФОРмальный ЭЛемент)

Алгоритм 1.2. Алгоритм FOREL

1: Инициализировать множество некластеризованных точек:
$U := X^\ell$;
2: **пока** в выборке есть некластеризованные точки, $U \neq \varnothing$:
3: взять произвольную точку $x_0 \in U$ случайным образом;
4: **повторять**
5: образовать кластер — сферу с центром в x_0 и радиусом R:
$K_0 := \{x_i \in U \mid \rho(x_i, x_0) \leqslant R\}$;
6: поместить центр сферы в центр масс кластера:
$x_0 := \frac{1}{|K_0|} \sum_{x_i \in K_0} x_i$;
7: **пока** центр x_0 не стабилизируется;
8: пометить все точки K_0 как кластеризованные:
$U := U \setminus K_0$;
9: применить алгоритм КНП к множеству центров всех найденных кластеров;
10: каждый объект $x_i \in X^\ell$ приписать кластеру с ближайшим центром;

Недостатки:

- чувствителен к выбору начального положения точки $\mathbf{x_0}$ для каждого нового кластера.

7.3. Функционалы качества кластеризации

Задачу кластеризации можно ставить как задачу дискретной оптимизации: необходимо так приписать номера кластеров $\mathbf{y_i}$ объектам $\mathbf{x_i}$, чтобы значение выбранного функционала качества приняло наилучшее значение.

- **Среднее внутрикластерное расстояние должно быть как можно меньше:**

$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to \min .$$

- **Среднее межкластерное расстояние должно быть как можно больше:**

$$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i \neq y_j]} \to \max .$$

- **Сумма средних внутрикластерных расстояний должна быть как можно меньше:**

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:\ y_i = y} \rho^2(x_i, \mu_y) \to \min$$

- **Сумма межкластерных расстояний должна быть как можно больше:**

$$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to \max$$

7.4. Статистические алгоритмы

Статистические алгоритмы основаны на предположении, что кластеры неплохо описываются некоторым семейством вероятностных распределений. Тогда задача кластеризации сводится к разделению смеси распределений по конечной выборке.

EM-алгоритм

Гипотеза о вероятностной природе данных. Объекты выборки $\mathbf{X}^{\ell}$ появляются случайно и независимо согласно вероятностному распределению, представляющему собой смесь распределений

$$p(x) = \sum_{y \in Y} w_y p_y(x), \qquad \sum_{y \in Y} w_y = 1,$$

где $\mathbf{p_y(x)}$ – функция плотности распределения кластера $\mathbf{y}$, $\mathbf{w_y}$ – неизвестная априорная вероятность появления объектов из кластера $\mathbf{y}$.

Гипотеза о форме кластеров. Объекты описываются $\mathbf{n}$ числовыми признаками $\mathbf{f_1(x),...,f_n(x)}$, $\mathbf{X} \in \mathbf{R^n}$. Каждый кластер $\mathbf{y} \in \mathbf{Y}$ описывается $\mathbf{n}$-мерной гауссовской плотностью $\mathbf{p_y(x) = N(x; \mu_y, \Sigma_y)}$ с центром $\mathbf{\mu_y = (\mu_{y1},...,\mu_{yn})}$ и диагональной ковариационной матрицей $\mathbf{\Sigma_y = diag(\sigma^2_{y1},..., \sigma^2_{yn})}$.

На E-шаге по формуле Байеса вычисляются скрытые переменные $\mathbf{g_{iy}}$. Значение $\mathbf{g_{iy}}$ равно вероятности того, что объект $\mathbf{x_i} \in \mathbf{X}^{\ell}$ принадлежит кластеру $\mathbf{y} \in \mathbf{Y}$. На M-шаге уточняются параметры каждого кластера $\mathbf{(\mu y, \Sigma y)}$, при этом существенно используются скрытые переменные $\mathbf{g_{iy}}$.

7.4. Статистические алгоритмы

Алгоритм 1.3. Кластеризация с помощью EM-алгоритма

1: начальное приближение для всех кластеров $y \in Y$:

$w_y := 1/|Y|$;

$\mu_y :=$ случайный объект выборки;

$\sigma^2_{yj} := \frac{1}{\ell|Y|} \sum_{i=1}^{\ell} (f_j(x_i) - \mu_{yj})^2, \;\; j = 1, \ldots, n$;

2: **повторять**

3: E-шаг (expectation):

$$g_{iy} := \frac{w_y p_y(x_i)}{\sum_{z \in Y} w_z p_z(x_i)}, \;\; y \in Y, \;\; i = 1, \ldots, \ell;$$

4: M-шаг (maximization):

$$w_y := \frac{1}{\ell} \sum_{i=1}^{\ell} g_{iy}, \;\; y \in Y;$$

$$\mu_{yj} := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} f_j(x_i), \;\; y \in Y, \;\; j = 1, \ldots, n;$$

$$\sigma^2_{yj} := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} (f_j(x_i) - \mu_{yj})^2, \;\; y \in Y, \;\; j = 1, \ldots, n;$$

5: Отнести объекты к кластерам по байесовскому решающему правилу:

$y_i := \arg\max_{y \in Y} g_{iy}, \;\; i = 1, \ldots, \ell$;

6: **пока** y_i не перестанут изменяться;

7.4. Статистические алгоритмы

Алгоритм 1.4. Кластеризация с помощью алгоритма k-средних

1: сформировать начальное приближение центров всех кластеров $y \in Y$:
μ_y — наиболее удалённые друг от друга объекты выборки;
2: **повторять**
3: отнести каждый объект к ближайшему центру (аналог E-шага):
$y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y), \quad i = 1, \ldots, \ell;$
4: вычислить новое положение центров (аналог M-шага):
$\mu_{yj} := \dfrac{\sum_{i=1}^{\ell} [y_i = y] f_j(x_i)}{\sum_{i=1}^{\ell} [y_i = y]}, \quad y \in Y, \quad j = 1, \ldots, n;$
5: **пока** y_i не перестанут изменяться;

Алгоритм k-means крайне чувствителен к выбору начальных приближений центров. Случайная инициализация центров на шаге 1 может приводить к плохим кластеризациям. Для формирования начального приближения лучше выделить k наиболее удалённых точек выборки: первые две точки выделяются по максимуму всех попарных расстояний; каждая следующая точка выбирается так, чтобы расстояние от неё до ближайшей уже выделенной было максимально.

7.5. Иерархическая кластеризация

Иерархические алгоритмы кластеризации, называемые также алгоритмами *таксономии*, строят не одно разбиение выборки на непересекающиеся классы, а систему вложенных разбиений. Результат таксономии обычно представляется в виде таксономического дерева - *дендрограммы*.

- ***Дивизимные*** или нисходящие алгоритмы разбивают выборку на всё более и более мелкие кластеры.
- ***Агломеративные*** или восходящие алгоритмы, в которых объекты объединяются во всё более и более крупные кластеры.

Сначала каждый объект считается отдельным кластером. Для одноэлементных кластеров естественным образом определяется функция расстояния

$$R(\{x\},\{x'\}) = \rho(x,x').$$

Затем запускается процесс слияний. На каждой итерации вместо пары самых близких кластеров **U** и **V** образуется новый кластер **W = U ∪ V**. Расстояние от нового кластера **W** до любого другого кластера S вычисляется по расстояниям **R(U, V), R(U, S)** и **R(V, S),** которые к этому моменту уже должны быть известны:

$$R(U \cup V, S) = \alpha_U R(U,S) + \alpha_V R(V,S) + \beta R(U,V) + \gamma |R(U,S) - R(V,S)|$$

7.5. Иерархическая кластеризация

Расстояние ближнего соседа:

$$R^{\text{б}}(W,S) = \min_{w\in W, s\in S} \rho(w,s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = -\tfrac{1}{2}.$$

Расстояние дальнего соседа:

$$R^{\text{д}}(W,S) = \max_{w\in W, s\in S} \rho(w,s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = \tfrac{1}{2}.$$

Среднее расстояние:

$$R^{\text{с}}(W,S) = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = \gamma = 0.$$

Расстояние между центрами:

$$R^{\text{ц}}(W,S) = \rho^2\Big(\sum_{w\in W} \frac{w}{|W|}, \sum_{s\in S} \frac{s}{|S|}\Big); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = -\alpha_U\alpha_V,\ \gamma = 0.$$

Расстояние Уорда:

$$R^{\text{у}}(W,S) = \frac{|S||W|}{|S|+|W|}\rho^2\Big(\sum_{w\in W} \frac{w}{|W|}, \sum_{s\in S} \frac{s}{|S|}\Big); \quad \alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \beta = \frac{-|S|}{|S|+|W|},\ \gamma = 0.$$

7.5. Иерархическая кластеризация

Алгоритм 1.5. Агломеративная кластеризация Ланса-Уильямса

1: инициализировать множество кластеров C_1:
$t := 1; \quad C_t = \{\{x_1\}, \ldots, \{x_\ell\}\};$
2: **для всех** $t = 2, \ldots, \ell$ (t — номер итерации):
3: найти в C_{t-1} два ближайших кластера:
$(U, V) := \arg\min_{U \neq V} R(U, V);$
$R_t := R(U, V);$
4: изъять кластеры U и V, добавить слитый кластер $W = U \cup V$:
$C_t := C_{t-1} \cup \{W\} \setminus \{U, V\};$
5: **для всех** $S \in C_t$
6: вычислить расстояние $R(W, S)$ по формуле Ланса-Уильямса;

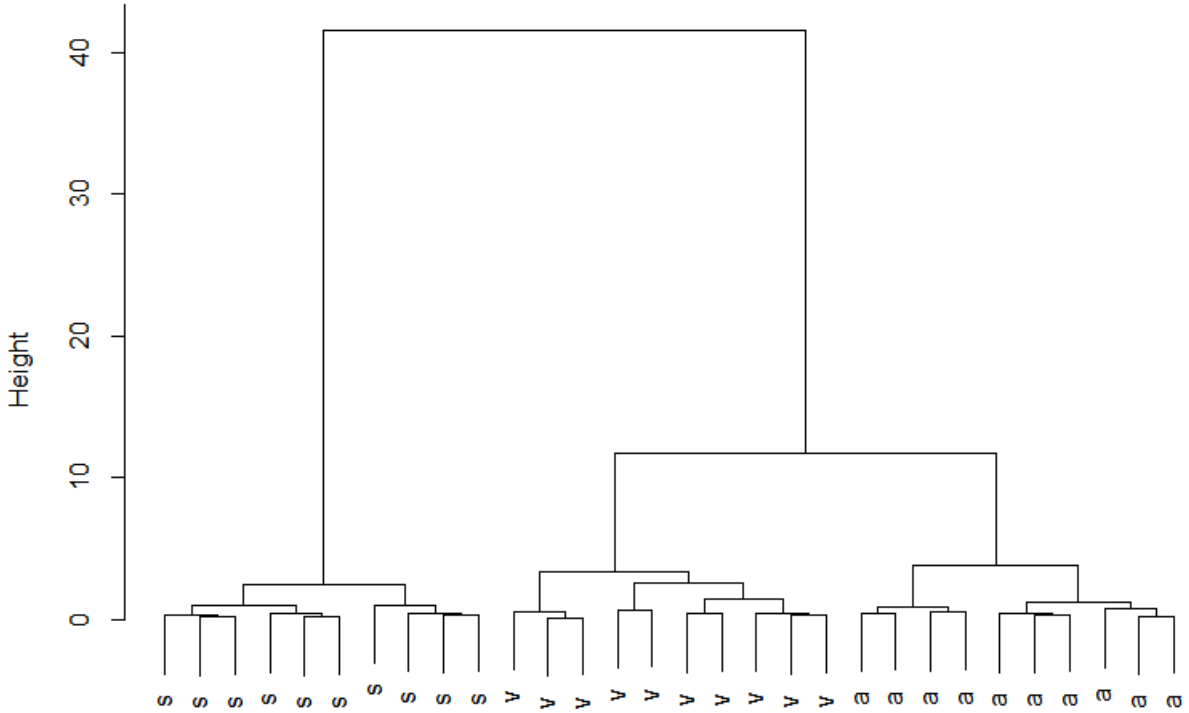
7.6. Определение числа кластеров

На горизонтальной оси находится интервал максимальной длины $|R_{t+1}-R_t|$, и в качестве результирующей кластеризации выдаётся множество кластеров C_t. Число кластеров равно $K = \ell - t + 1$. При необходимости можно задать ограничение на минимальное и максимальное число кластеров $K_0 \le K \le K_1$ и выбирать t, удовлетворяющие ограничениям $\ell - K1 + 1 \le t \le \ell - K0 + 1$.

Во многих прикладных задачах интерес представляет таксономическое дерево целиком, и определять оптимальное число кластеров не имеет особого смысла.

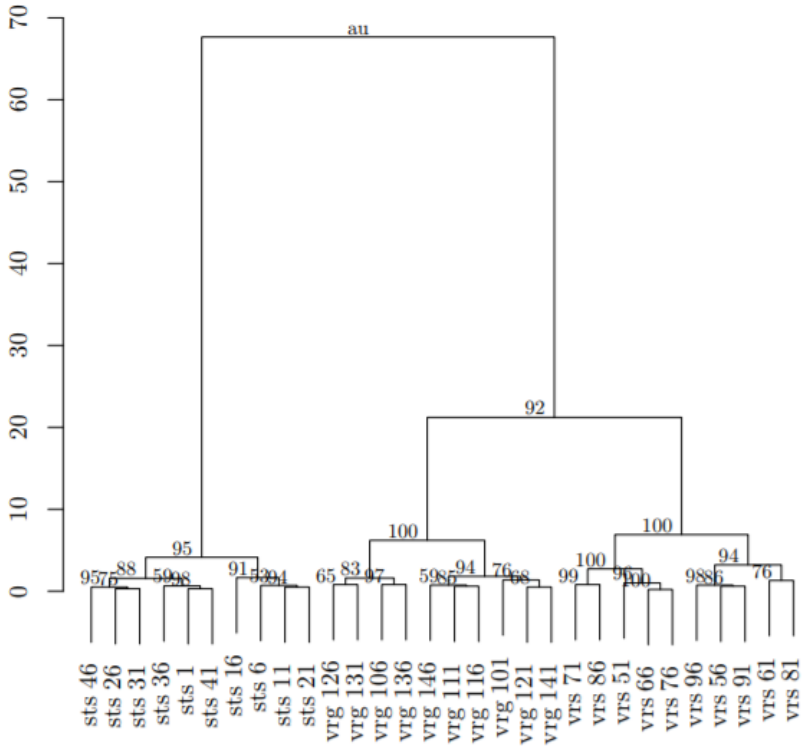
7.7. Кластеризация в R

```
> library(cluster)
> iriss <- iris[seq(1,nrow(iris),5),]
> iriss.dist <- daisy(iriss[,1:4])
> iriss.h <- hclust(iriss.dist, method="ward")
> plot(iriss.h, labels=abbreviate(iriss[,5],1, method="both.sides"), main="")
```



7.7. Кластеризация в R

```
> library(pvclust)
> irisst <- t(iriss[,1:4])
> colnames(irisst) <- paste(abbreviate(iriss[,5], 3),  colnames(irisst))
> iriss.pv<- pvclust(irisst, method.dist="manhattan", method.hclust="ward", nboot=100)
```

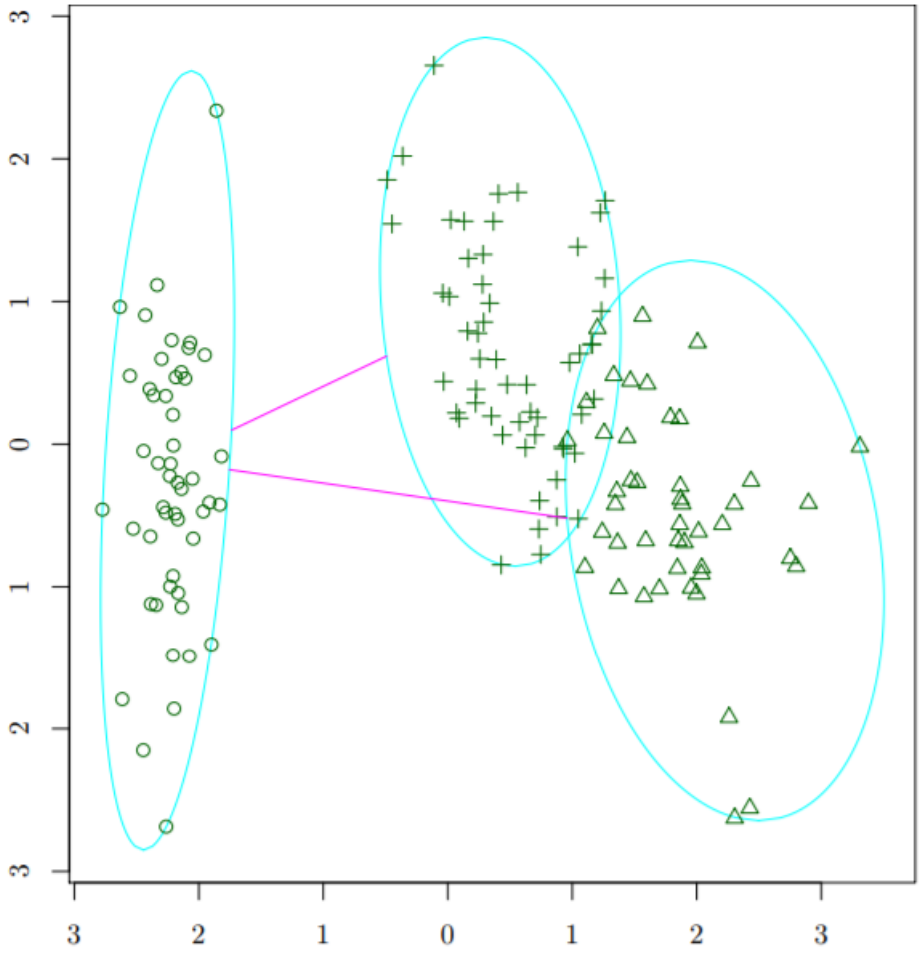


7.7. Кластеризация в R

```
> eq <- read.table("data/eq.txt", h=TRUE)
> eq.k <- kmeans(eq[,-1], 2)
> table(eq.k$cluster, eq$SPECIES)
        arvense         fluviatile
1       37              5
2       1               41

> iris.f <- fanny(iris[,1:4], 3)
> plot(iris.f, which=1, main="")
> head(data.frame(sp=iris[,5], iris.f$membership))
        sp              X1              X2              X3
1       setosa          0.9142273       0.03603116      0.04974153
2       setosa          0.8594576       0.05854637      0.08199602
3       setosa          0.8700857       0.05463714      0.07527719
4       setosa          0.8426296       0.06555926      0.09181118
5       setosa          0.9044503       0.04025288      0.05529687
6       setosa          0.7680227       0.09717445      0.13480286
```

7.7. Кластеризация в R



Литература

Мастицкий С. Э., Шитиков В. К. (2014) Статистический анализ и визуализация данных с помощью R. - Электронная книга, 400 с

Шипунов А. Б., Балдин Е. М., Волкова П. А., Коробейников А. И., Назарова С. А., Петров С. В., Суфиянов В. Г. (2012) Наглядная статистика. Используем R! - М.: ДМК Пресс, 298 с.

Кабаков Р. К. (2014) R в действии. Анализ и визуализация данных на языке R Издательство: ДМК Пресс, 580 с.

Лекции К.В. Воронцова по алгоритмам кластеризации и многомерному шкалированию