Оценка качества кластеризации

- *Внешние критерии*

  Внешние меры основаны на сравнении автоматического разбиения данных с полученным от экспертов «эталонным» разбиением этих же данных
- *Внутренние критерии*

  Оценка компактности и отделимости кластеров без привлечения внешней информации
- *Сравнительные критерии*

  Сопоставление результатов, полученных разными методами кластеризации

Энтропия кластерного решения

- Оценка энтропии кластерного решения (неоднородности кластеров по классам)

$$E(K_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r} * \log\frac{n_r^i}{n_r}$$

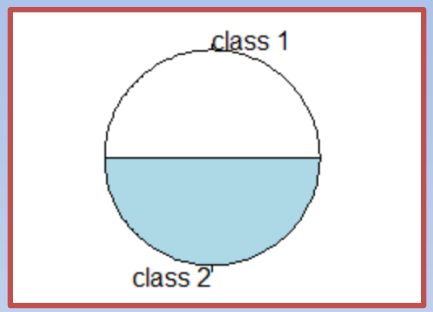
$$Entropy = \sum_{r=1}^{M}\frac{n_r}{n} * E(K_r)$$

n_r – число элементов в *r*-кластере

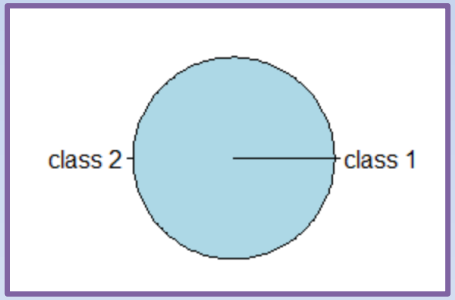
n_r^i - число элементов *i*-того класса внутри кластера *r*

q – общее число классов

M – число кластеров



$$E = -\frac{1}{\log 2}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) = 1$$



$$E = -\frac{1}{\log 2}(0 \cdot \log 0 + 1 \cdot \log 1) = 0$$

Оценка совпадений классов и кластеров объектов

	Same Cluster	Different Cluster
Same Class	f_{11}	f_{10}
Different Class	f_{01}	f_{00}

f_{11} – число пар объектов одного класса, находящихся в одном кластере;
f_{10} – число пар объектов одного класса, находящихся в разных кластерах;
..

$$\text{Rand statistic} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

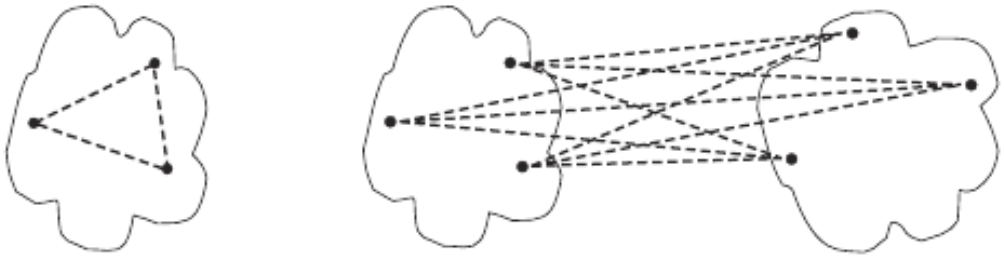
$$\text{Jaccard coefficient} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Внутренние критерии

- Внутренние меры основаны на оценке свойств отделимости (*separation*) и компактности (*cohesion*) полученного разбиения данных

$$overall\ validity = \sum_{i=1}^{K} w_i\ validity(C_i).$$

Отделимость и компактность кластера



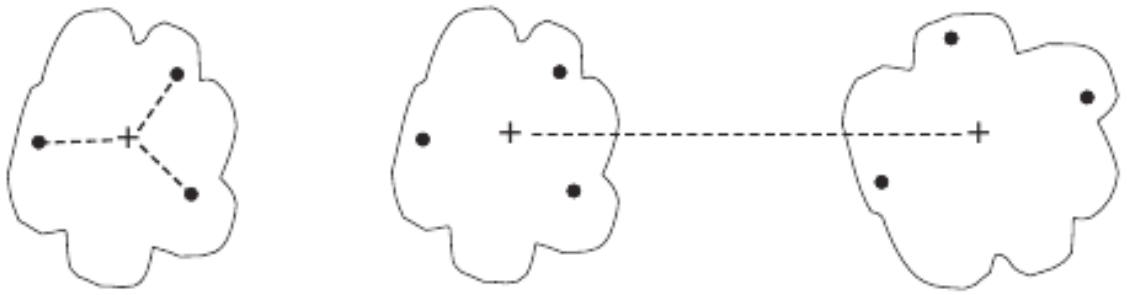
(a) Cohesion.

(b) Separation.

$$cohesion(C_i) = \sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_i}} proximity(\mathbf{x}, \mathbf{y})$$

$$separation(C_i, C_j) = \sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_j}} proximity(\mathbf{x}, \mathbf{y})$$

Отделимость и компактность кластеров (с центрами)

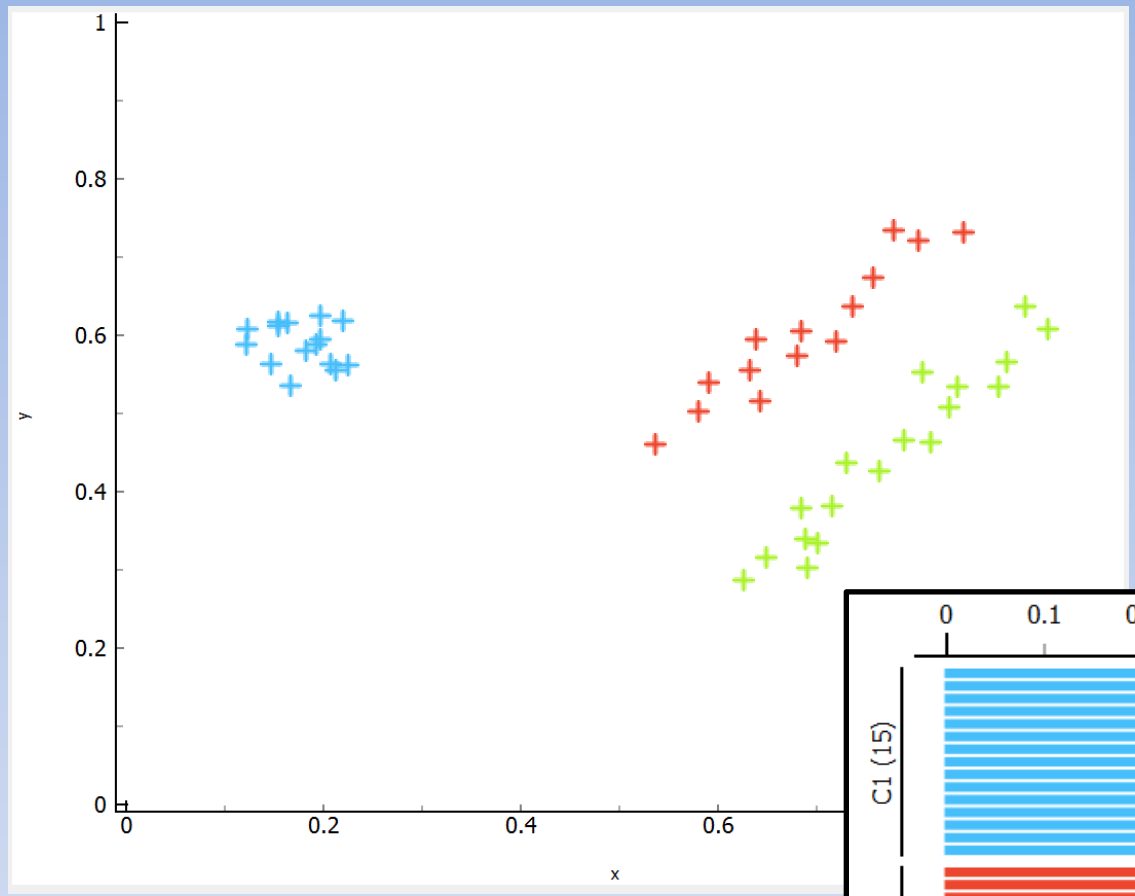


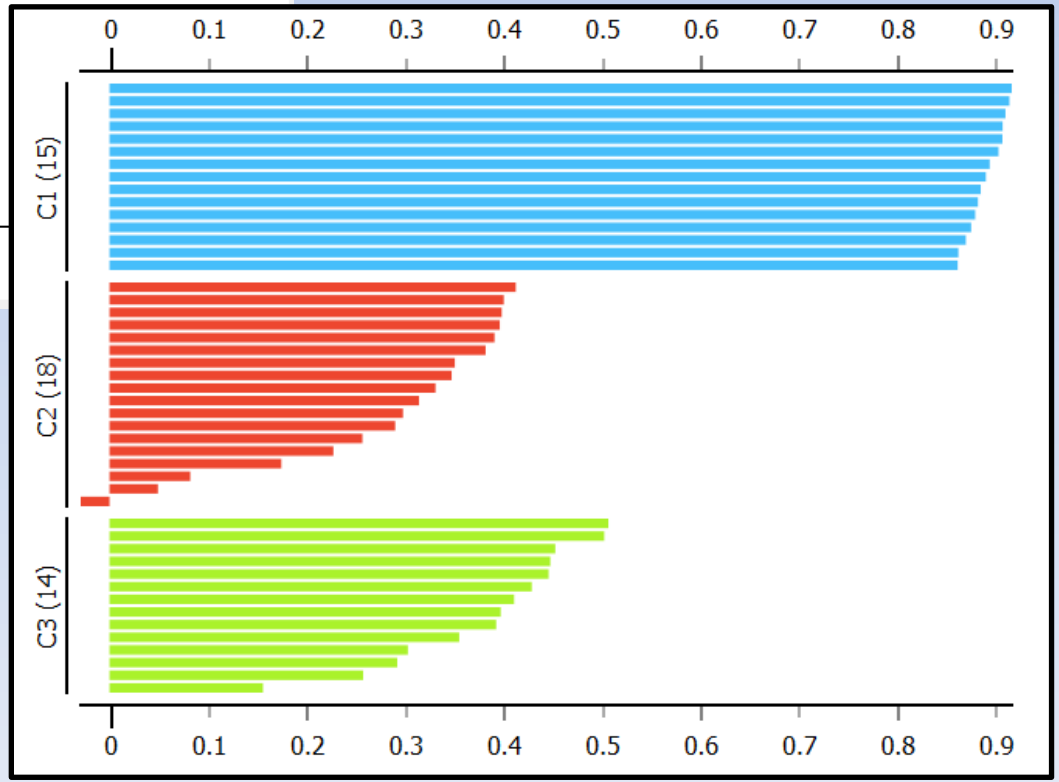
(a) Cohesion. (b) Separation.

$$
\begin{aligned}
cohesion(C_i) &= \sum_{\mathbf{x} \in C_i} proximity(\mathbf{x}, \mathbf{c}_i) \\
separation(C_i, C_j) &= proximity(\mathbf{c}_i, \mathbf{c}_j) \\
separation(C_i) &= proximity(\mathbf{c}_i, \mathbf{c})
\end{aligned}
$$

Коэффициент Silhouette

- Оценивает насколько кластеризация соответствует расстояниям между точками
- Для i-ой точки:
  - a_i – среднее расстояние относительно от всех других точек в своем кластере
  - ***d(i,c)*** – среднее расстояние относительно всех других точек в кластере c
  - b_i = **min** ***d(i,c)***
  - Оценка sw для i-ой точки: $$sw_i = \frac{b_i - a_i}{\max(a_i; b_i)}$$
- Суммарная оценка: $$sw = \frac{1}{n}\sum_{i=1}^{n} sw_i$$


1
0.8
0.6
0.4
0.2
0
y
0
0.2
0.4
0.6
x


0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
C1 (15)
C2 (18)
C3 (14)

Оценка тенденций к кластеризации

- Статистика Хопкинса для набора данных D
  - Случайным образом отбираем p точек из исходного набора (множество *X*)
  - Генерируем *p* – точек равномерно распределенных в пространстве (множество *Y*)
  - Для точек наборов *X* и *Y* вычисляем расстояния до ближайших точек их исходного набора данных D:

    u_i - расстояние от *i*-точки набора *Y* до ближайшей точки *D*;

    w_i – расстояние от *i*-точки набора *X* до ближайшей точки *D*;

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$