Устойчивость кластеризации биологических образцов к малым изменениям данных

19.12.2016
НИС Машинное обучение
Максим Никольский

Задача кластеризации

$\mathbb{X}$ — множество объектов, $\mathbb{Y} = \{1, \ldots, K\}$ — множество меток кластеров, $\rho(x, x')$ — функция расстояния между объектами.

Необходимо по выборке $X \subset \mathbb{X}$ построить алгоритм $a : \mathbb{X} \to \mathbb{Y}$ сопоставляющий каждому объекту метку кластера таким образом, что значение метрики между объектами из одного кластера минимально, а между объектами из разных кластеров — максимально

Задача кластеризации

Внутри каждого кластера объекты наиболее схожи, объекты из разных кластеров наиболее различны

Задача кластеризации

Внутри каждого кластера объекты наиболее схожи, объекты из разных кластеров наиболее различны

Различают плоскую (flat) кластеризацию и иерархическую кластеризацию; жесткую (hard) и мягкую (soft).

Задача кластеризации

Внутри каждого кластера объекты наиболее схожи, объекты из разных кластеров наиболее различны

Различают плоскую (flat) кластеризацию и иерархическую кластеризацию; жесткую (hard) и мягкую (soft).

Плоская кластеризация возвращает “плоское” множество кластеров, не имеющих никакой внутренней структуры, которая соотносила бы кластеры друг другу, иерархическая — иерархию кластеров.

Задача кластеризации

Внутри каждого кластера объекты наиболее схожи, объекты из разных кластеров наиболее различны

Различают плоскую (flat) кластеризацию и иерархическую кластеризацию; жесткую (hard) и мягкую (soft).

Плоская кластеризация возвращает “плоское” множество кластеров, не имеющих никакой внутренней структуры, которая соотносила бы кластеры друг другу, иерархическая — иерархию кластеров.

В жесткой кластеризации каждый объект связан с единственным кластером; в мягкой объект принадлежит каждому кластеру в некоторой степени.

Метрики качества

Достижение высокого значения функции расстояния для объектов из одного класса и низкого для объектов из разных классов — это еще не все.

Метрики качества

Достижение высокого значения функции расстояния для объектов из одного класса и низкого для объектов из разных классов — это еще не все.

Внешний критерий — то, насколько хорошо кластеризация соответствует избранному правилу.

Метрики качества

Достижение высокого значения функции расстояния для объектов из одного класса и низкого для объектов из разных классов — это еще не все.

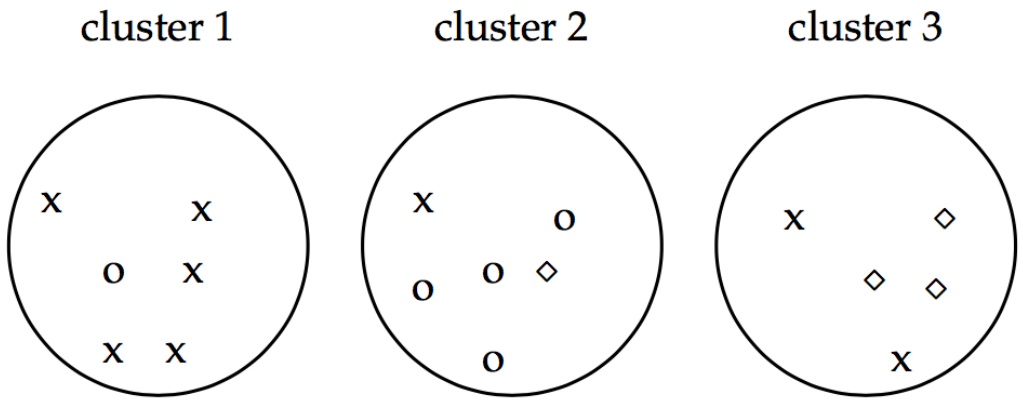
Внешний критерий — то, насколько хорошо кластеризация соответствует избранному правилу.

Допустим, по некоторому избранному правилу каждому объекту присвоен некоторый класс $c \in \mathbb{C}$, где $\mathbb{C} = \{c_1, \ldots, c_m\}$.

Метрики качества

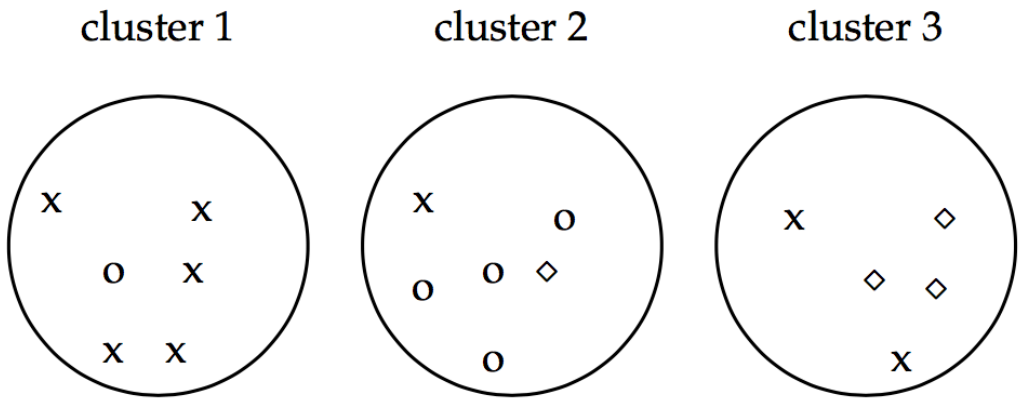
Чистота $purity(\mathbb{Y}, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |y_k \cap c_j|$

Метрики качества



Чистота $purity(\mathbb{Y}, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |y_k \cap c_j|$

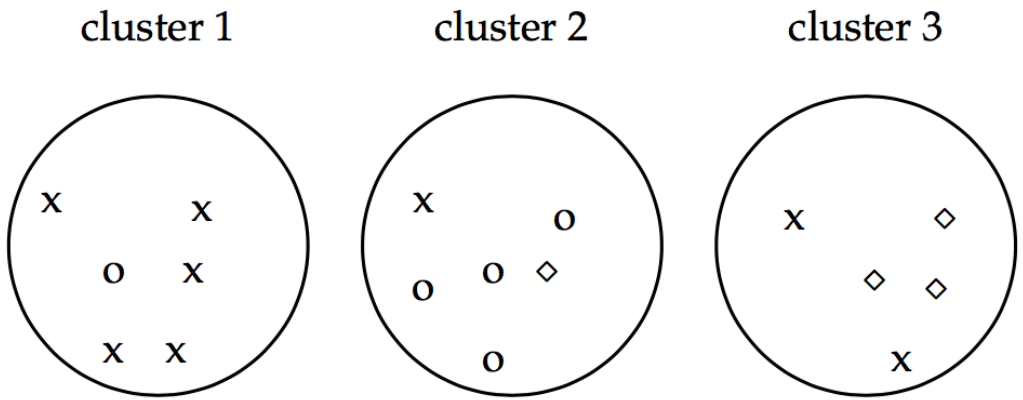
Метрики качества



Чистота $purity(\mathbb{Y}, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |y_k \cap c_j|$

Для данного примера $\frac{1}{17} \cdot (5 + 4 + 3) \approx 0.71.$

Метрики качества

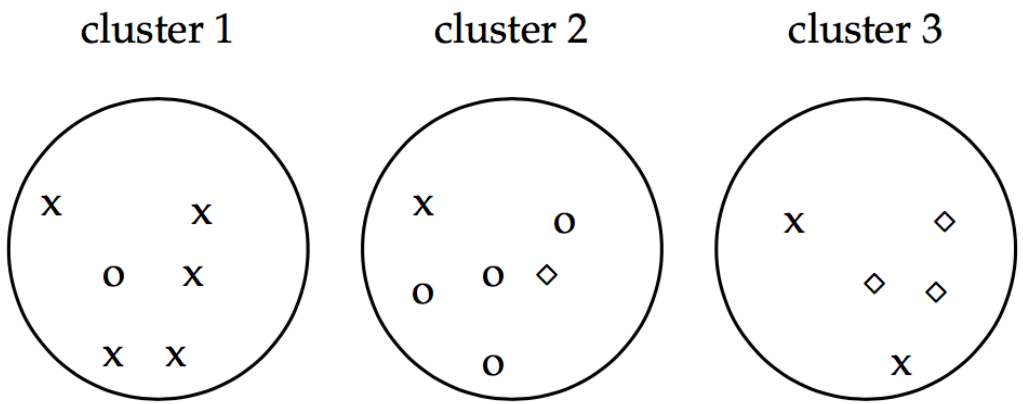


Чистота $purity(\mathbb{Y}, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |y_k \cap c_j|$

Для данного примера $\frac{1}{17} \cdot (5 + 4 + 3) \approx 0.71.$

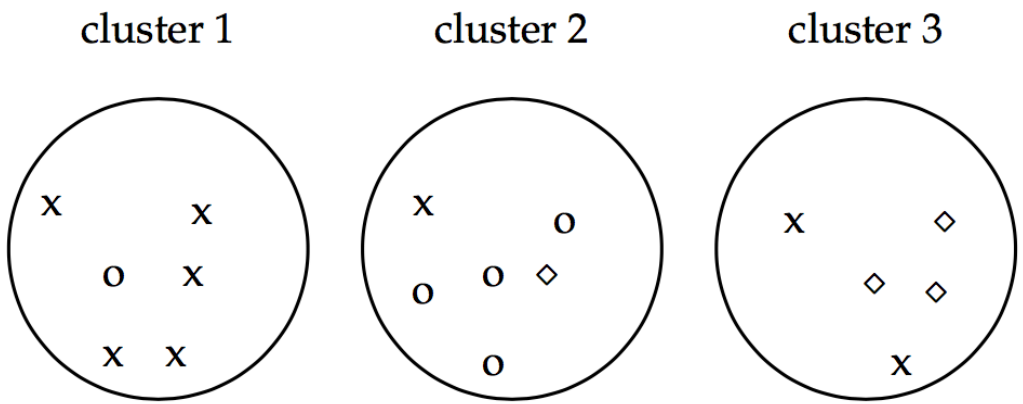
Недостаток — purity нельзя использовать для установления количества кластеров.

Метрики качества



Индекс Рэнда $RI = \frac{TP+TN}{TP+TN+FP+FN}$

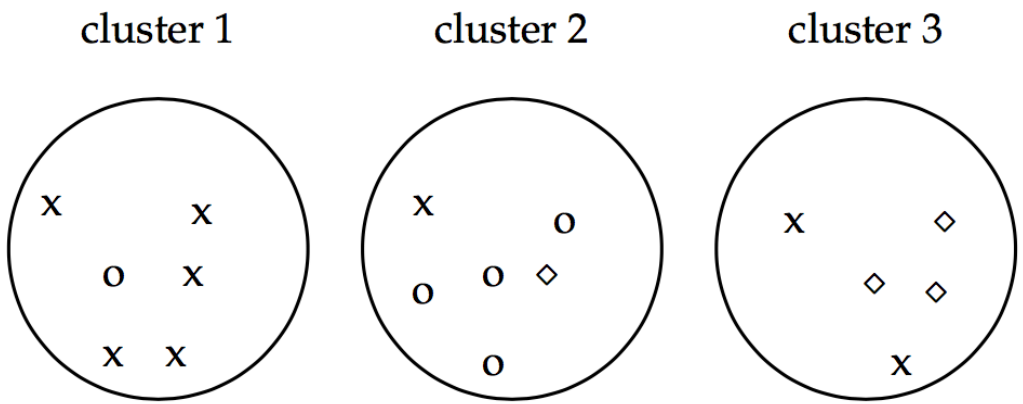
Метрики качества



Индекс Рэнда $RI = \frac{TP+TN}{TP+TN+FP+FN}$

Для данного примера $\frac{20+72}{20+20+24+72} \approx 0.68$

Метрики качества

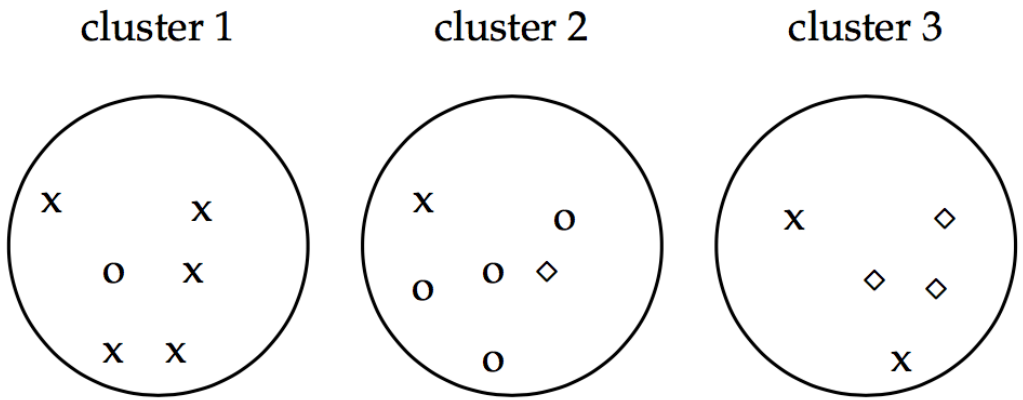


Индекс Рэнда $RI = \frac{TP+TN}{TP+TN+FP+FN}$

Для данного примера $\frac{20+72}{20+20+24+72} \approx 0.68$

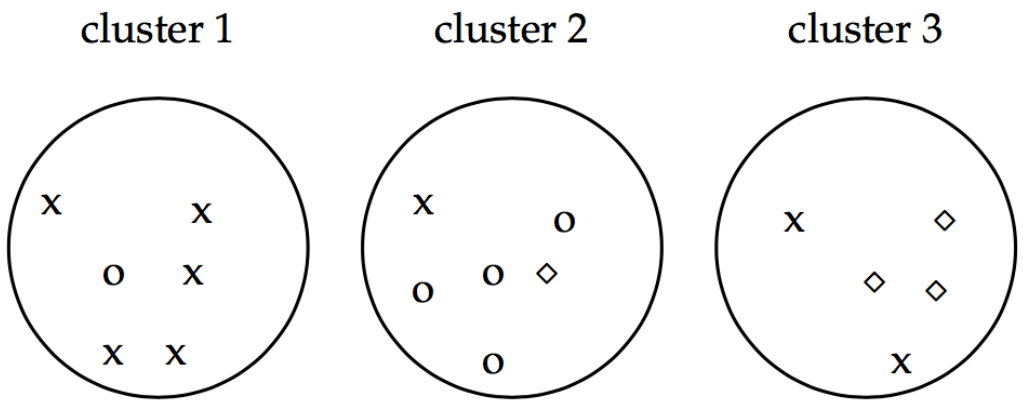
Особенность — у FP и FN одинаковый вес.

Метрики качества



F-мера $F_\beta = \frac{(\beta^2+1)P \cdot R}{\beta^2 P + R}$

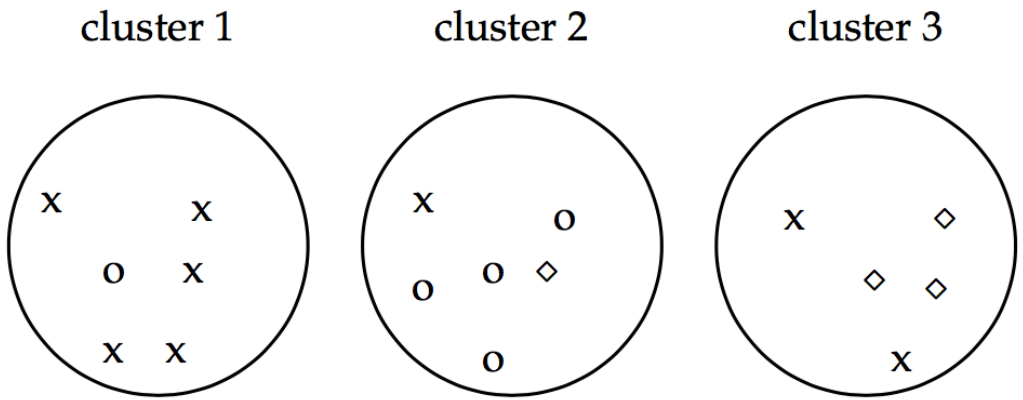
Метрики качества



F-мера $F_\beta = \frac{(\beta^2+1)P \cdot R}{\beta^2 P + R}$

Для данного примера $F_1 \approx 0.48$, $F_5 \approx 0.456$.

Метрики качества



F-мера $F_\beta = \frac{(\beta^2+1)P \cdot R}{\beta^2 P + R}$

Для данного примера $F_1 \approx 0.48$, $F_5 \approx 0.456$.

Особенность — параметр β регулирует вес полноты, т. е. то, насколько сильно алгоритм штрафуется за FN.

Иерархическая кластеризация

Два подхода: восходящая и нисходящая

Восходящая: изначально каждый объект — отдельный кластер

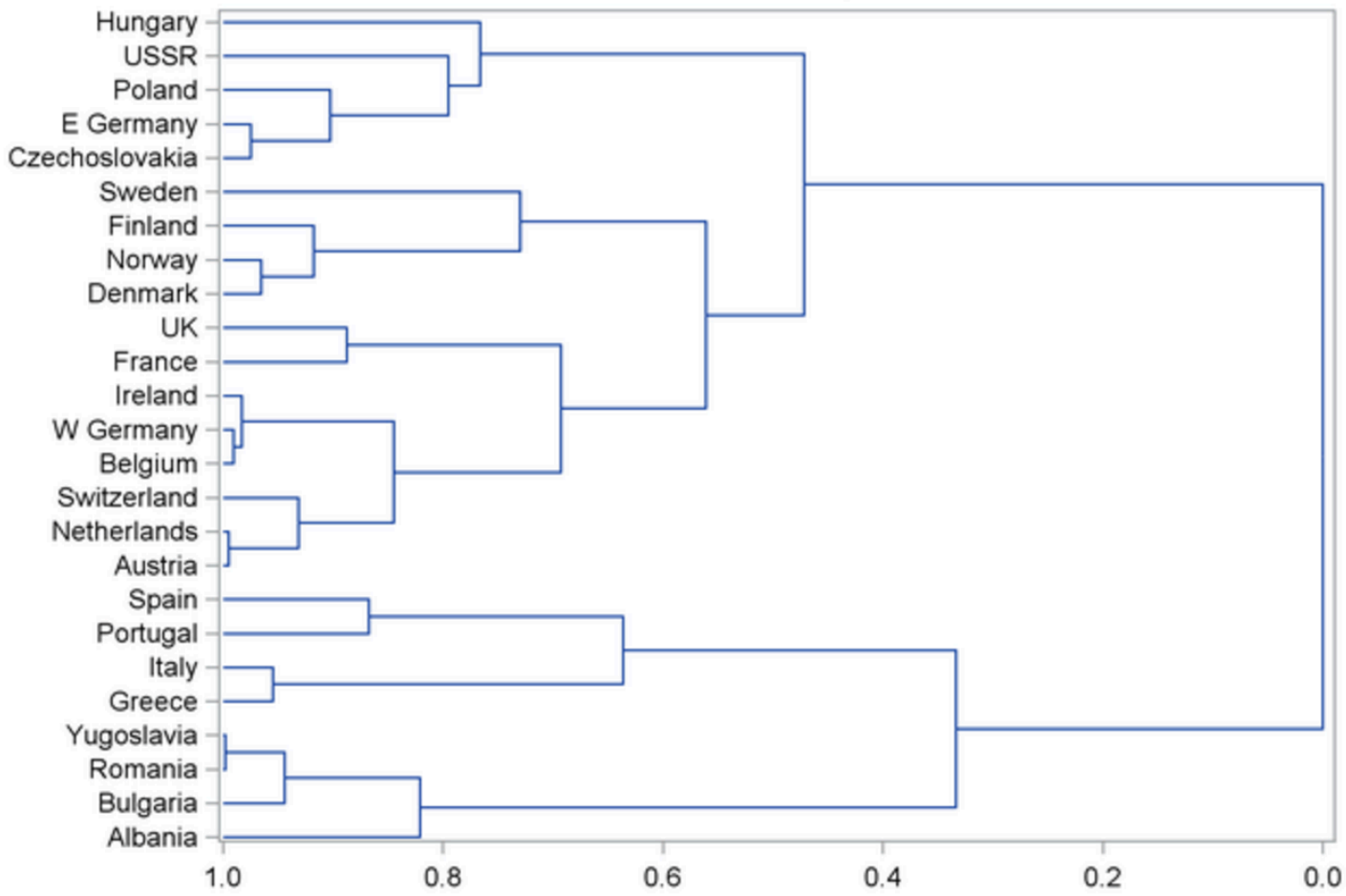
Нисходящая: изначально всего один кластер, содержащий в себе все объекты. На каждом шаге кластер делится некоторым образом на отдельные подкластеры

Иерархическая кластеризация

Удобный инструмент визуализации — дендрограмма

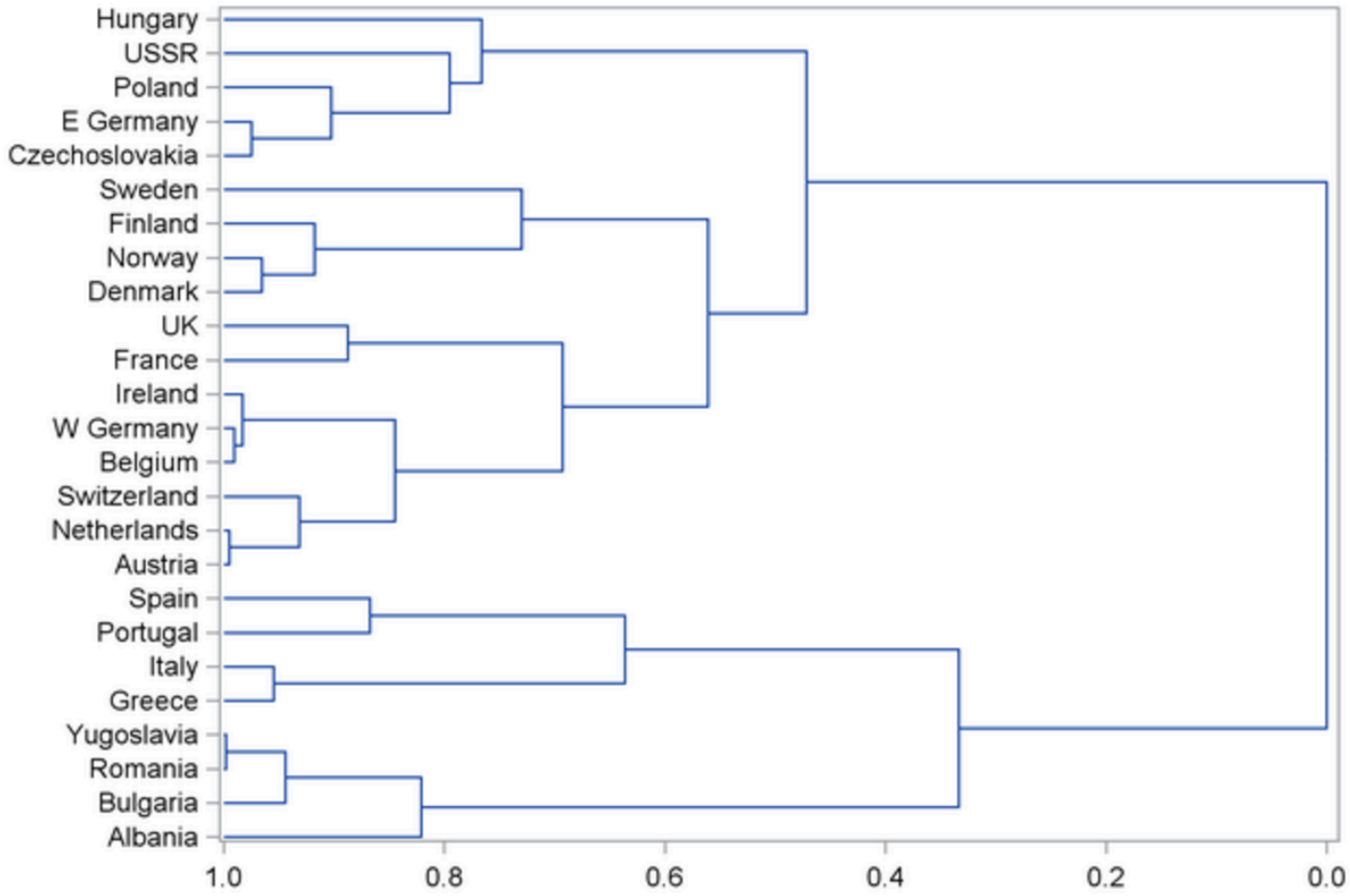
Иерархическая кластеризация

Удобный инструмент визуализации — дендрограмма



Иерархическая кластеризация

Может быть легко превращена в плоскую путем разрезания (cut)



Иерархическая кластеризация

Может быть легко превращена в плоскую путем разрезания (cut)

* Делать разрез на определенном уровне схожести

Иерархическая кластеризация

Может быть легко превращена в плоскую путем разрезания (cut)

* Делать разрез на определенном уровне схожести
* Делать разрез тогда, когда разница между двумя успешными уровнями схожести максимальна

Иерархическая кластеризация

Может быть легко превращена в плоскую путем разрезания (cut)

* Делать разрез на определенном уровне схожести
* Делать разрез тогда, когда разница между двумя успешными уровнями схожести максимальна
* Делать разрез тогда, когда в его результате получится заданное количество кластеров

Сходство кластеров

Сходство кластеров

* **Метод одиночной связи.** Учитываются самые близкие объекты кластеров. Локален (не учитывается общая структура кластеров).

Сходство кластеров

* **Метод одиночной связи.** Учитываются самые близкие объекты кластеров. Локален (не учитывается общая структура кластеров).

* **Метод полной связи.** Учитывается самые дальние объекты кластеров. Нелокален, чувствителен к выбросам, склонен к формированию компактных кластеров.

Сходство кластеров

* **Метод одиночной связи.** Учитываются самые близкие объекты кластеров. Локален (не учитывается общая структура кластеров).

* **Метод полной связи.** Учитывается самые дальние объекты кластеров. Нелокален, чувствителен к выбросам, склонен к формированию компактных кластеров.

* **Метод группового среднего.** Учитывается попарная схожесть всех объектов в обоих кластерах.

Сходство кластеров

* **Метод одиночной связи.** Учитываются самые близкие объекты кластеров. Локален (не учитывается общая структура кластеров).

* **Метод полной связи.** Учитывается самые дальние объекты кластеров. Нелокален, чувствителен к выбросам, склонен к формированию компактных кластеров.

* **Метод группового среднего.** Учитывается попарная схожесть всех объектов в обоих кластерах.

* **Метод центроидов.** Учитывается схожесть центроидов кластеров. Отличается от метода группового среднего тем, что не учитывает схожесть объектов внутри одного кластера.

Нисходящая кластеризация (DIANA)

1) На каждом шаге выбирается кластер с наибольшим диаметром (наименьшая схожесть между парой объектов)
2) В выбранном кластере ищется объект с наименьшей средней схожестью с остальными объектами
3) Наиболее похожие на него объекты формируют новый кластер

Причем тут биология?

Причем тут биология?

Кластеризация — один из наиболее часто используемых методов анализа геномных данных

Причем тут биология?

Кластеризация — один из наиболее часто используемых методов анализа геномных данных

Позволяет биологам установить потенциально значимую взаимосвязь между объектами (ими могут быть гены или эксперименты)

Причем тут биология?

Кластеризация — один из наиболее часто используемых методов анализа геномных данных

Позволяет биологам установить потенциально значимую взаимосвязь между объектами (ими могут быть гены или эксперименты)

Позволяет сгруппировать образцы, имея данные об экспрессивности генов (группировка по фенотипу) и найти подмножество генов, отвечающих за различия

Зачем измерять устойчивость?

Зачем измерять устойчивость?

Не всегда можно собрать данные обо всех потенциально интересных генах

Зачем измерять устойчивость?

Не всегда можно собрать данные обо всех потенциально интересных генах

Собранные данные могут быть грязными

Что за данные?

Данные об экспрессивности генов у раковых больных

Каждый объект — результат эксперимента, каждый признак — экспрессивность конкретного гена

Всего 917 образцов и более 54000 генов

Анализ устойчивости

Анализ устойчивости

Оценка близости результатов различных алгоритмов кластеризации

Анализ устойчивости

Оценка близости результатов различных алгоритмов кластеризации

Анализ устойчивости к отбрасыванию части признаков

Анализ устойчивости

Оценка близости результатов различных алгоритмов кластеризации

Анализ устойчивости к отбрасыванию части признаков

Анализ устойчивости к внесению случайного шума

Анализ устойчивости

Оценка близости результатов различных алгоритмов кластеризации

Анализ устойчивости к отбрасыванию части признаков

Анализ устойчивости к внесению случайного шума

...и все это для разных параметров и метрик

Анализ устойчивости

На первом этапе сравниваются K-Means, несколько видов восходящей иерархической кластеризации (методы полной связи, группового среднего, центроидов) и DIANA

Количество кластеров было фиксировано и равно 4

Для оценки близости результатов различных кластеризаций — индекс Рэнда

Сравнение алгоритмов: размер кластеров

	K-Means	Complete	Average	Centroid	DIANA
1	111	115	2	1	86
2	167	167	69	1	167
3	288	313	167	1	329
4	351	315	679	914	335

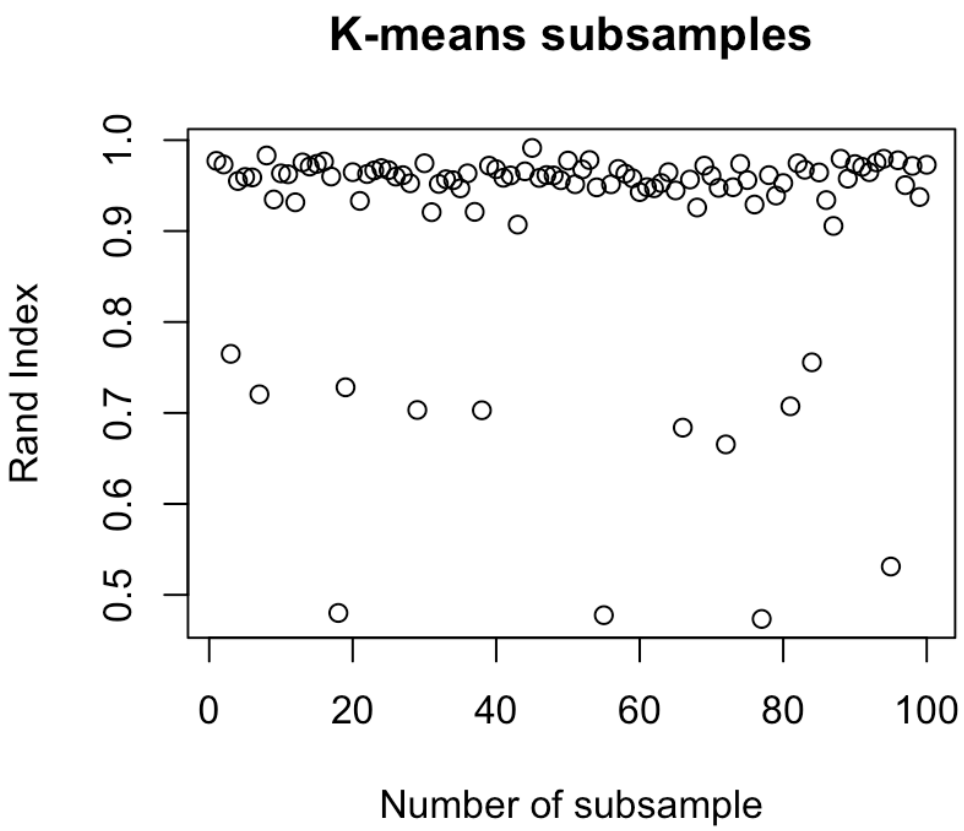
Сравнение алгоритмов: индекс Рэнда

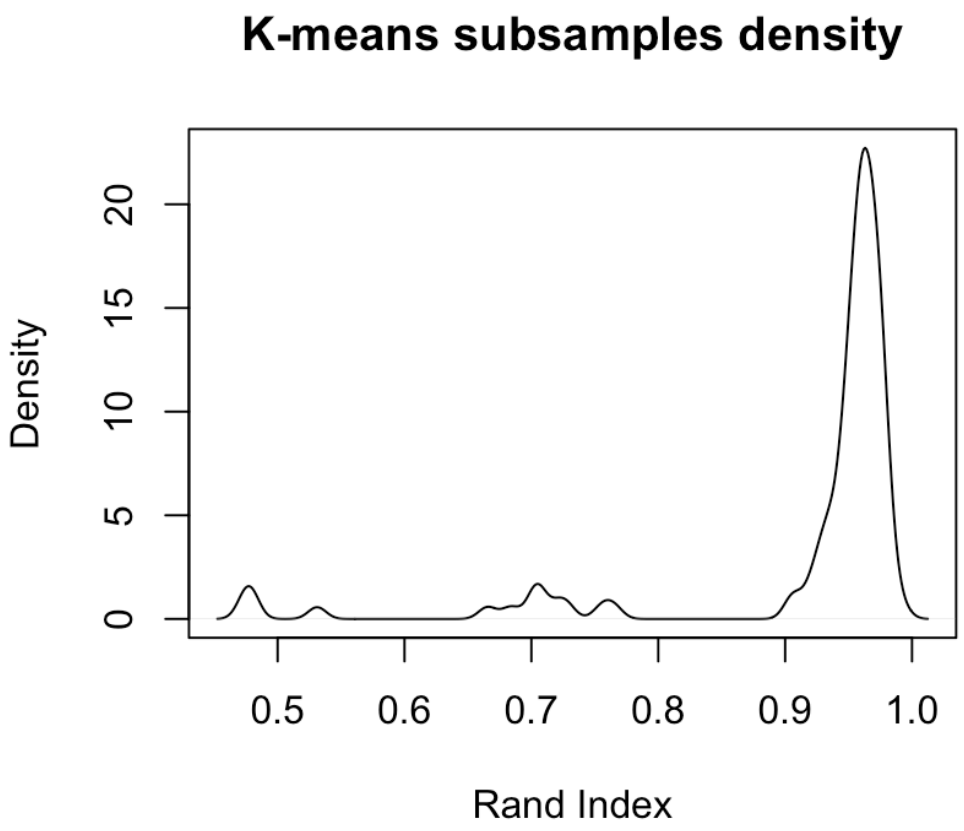
	K-Means	Complete	Average	Centroid	DIANA
K-Means	1	0,732	0,419	0,003	0,865
Complete	0,732	1	0,41	0,003	0,734
Average	0,419	0,41	1	0,015	0,45
Centroid	0,003	0,003	0,015	1	0,003
DIANA	0,865	0,734	0,45	0,003	1

Устойчивость алгоритмов

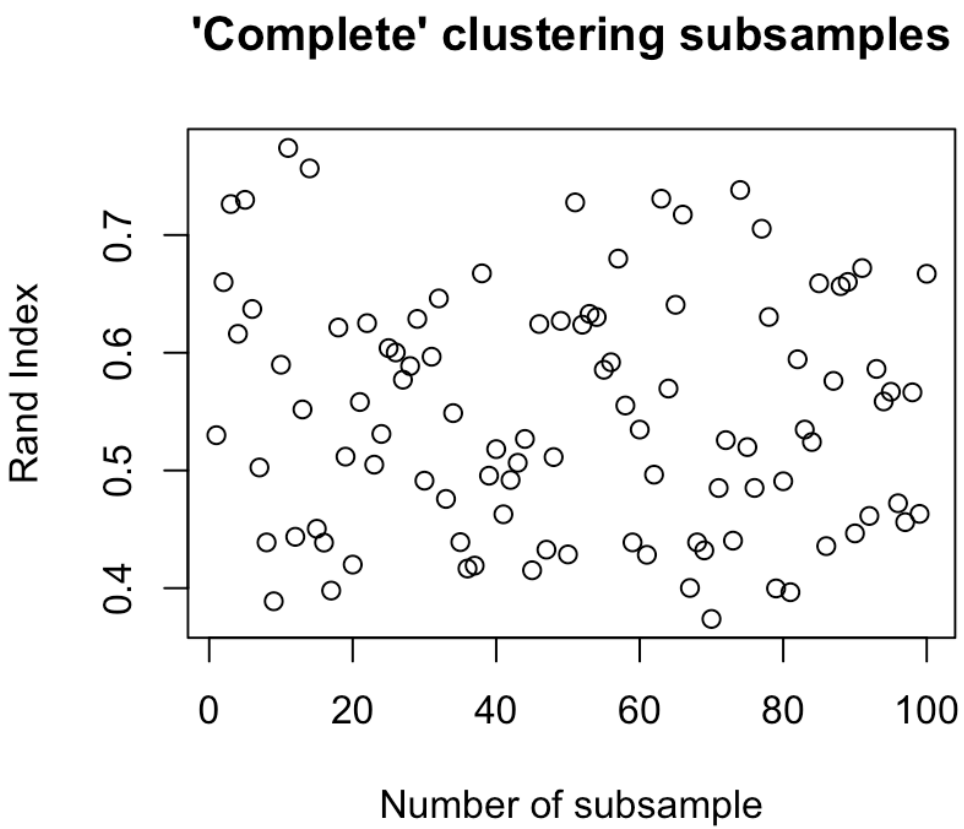
Для каждого алгоритма — 100 повторов кластеризации на данных, из которых было выброшено 90% признаков

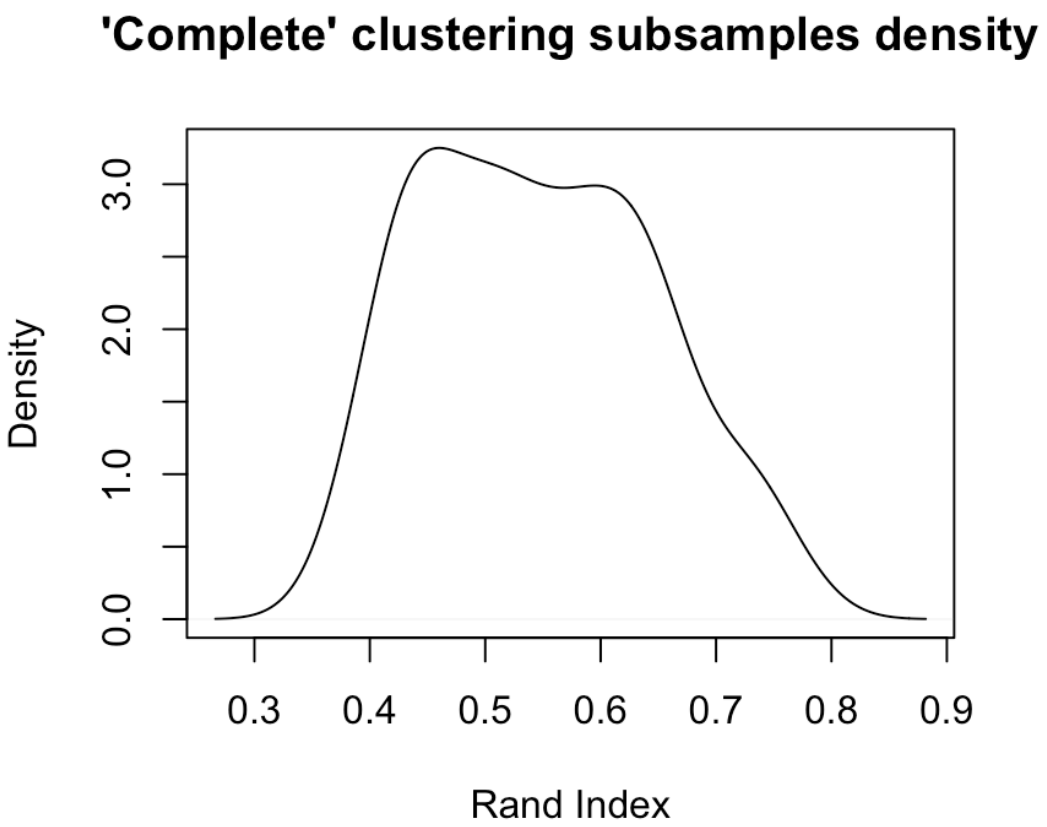
Устойчивость алгоритмов



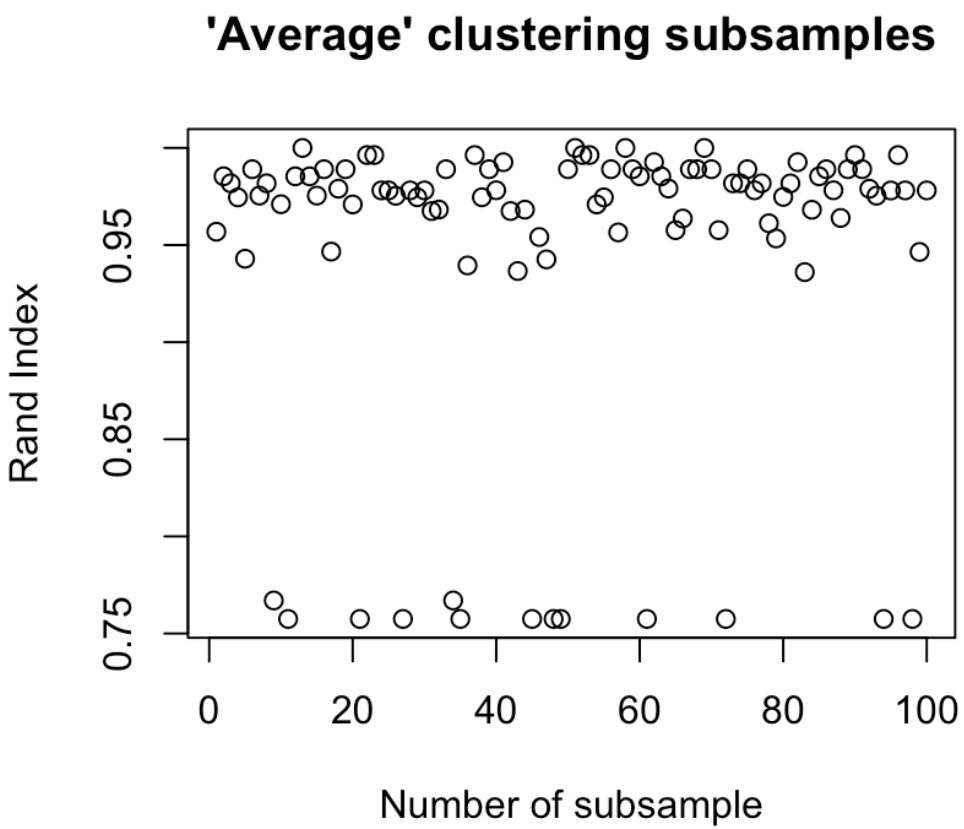


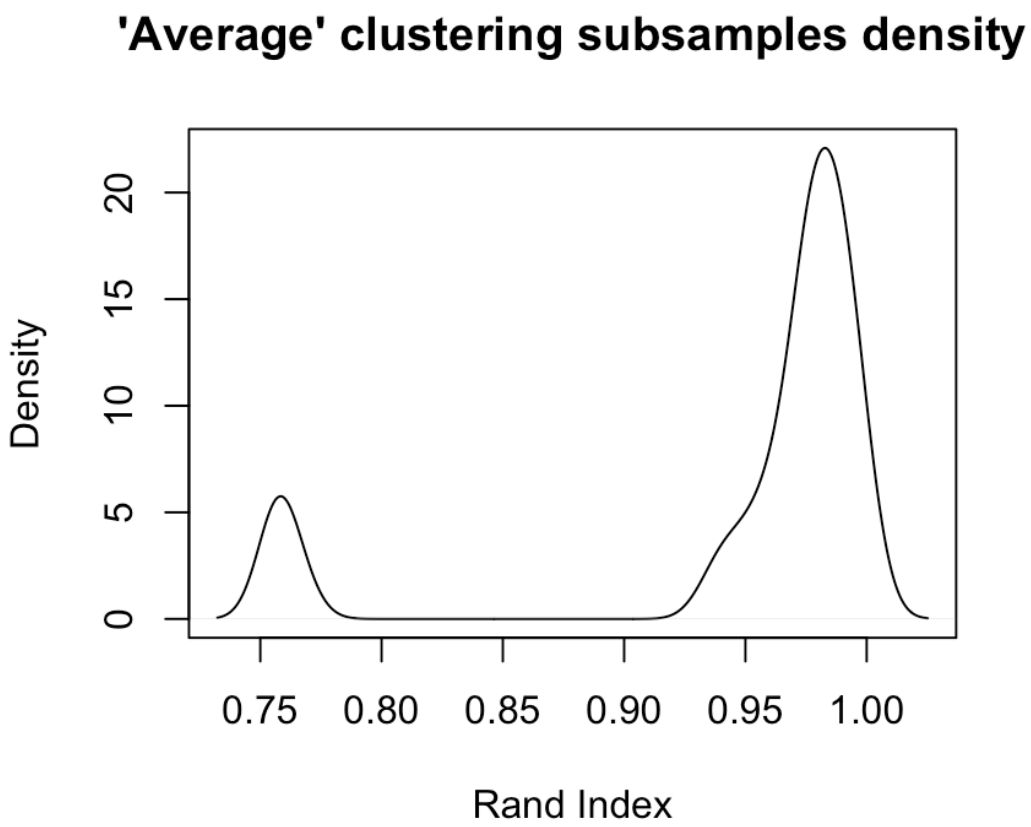
Устойчивость алгоритмов



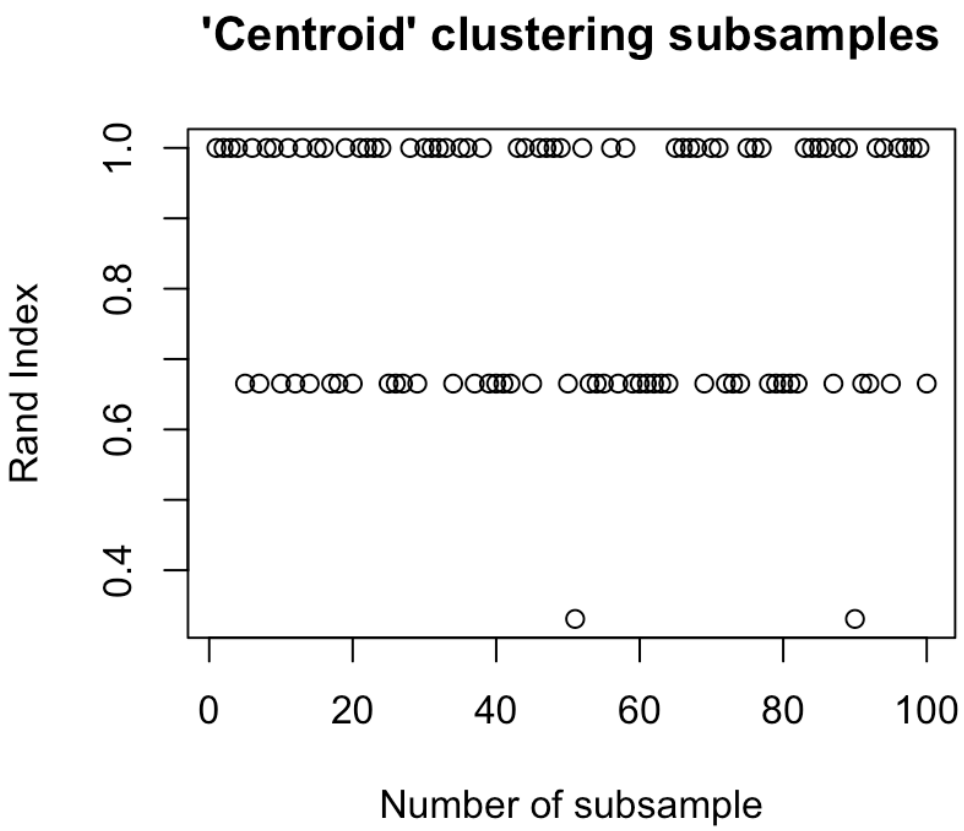


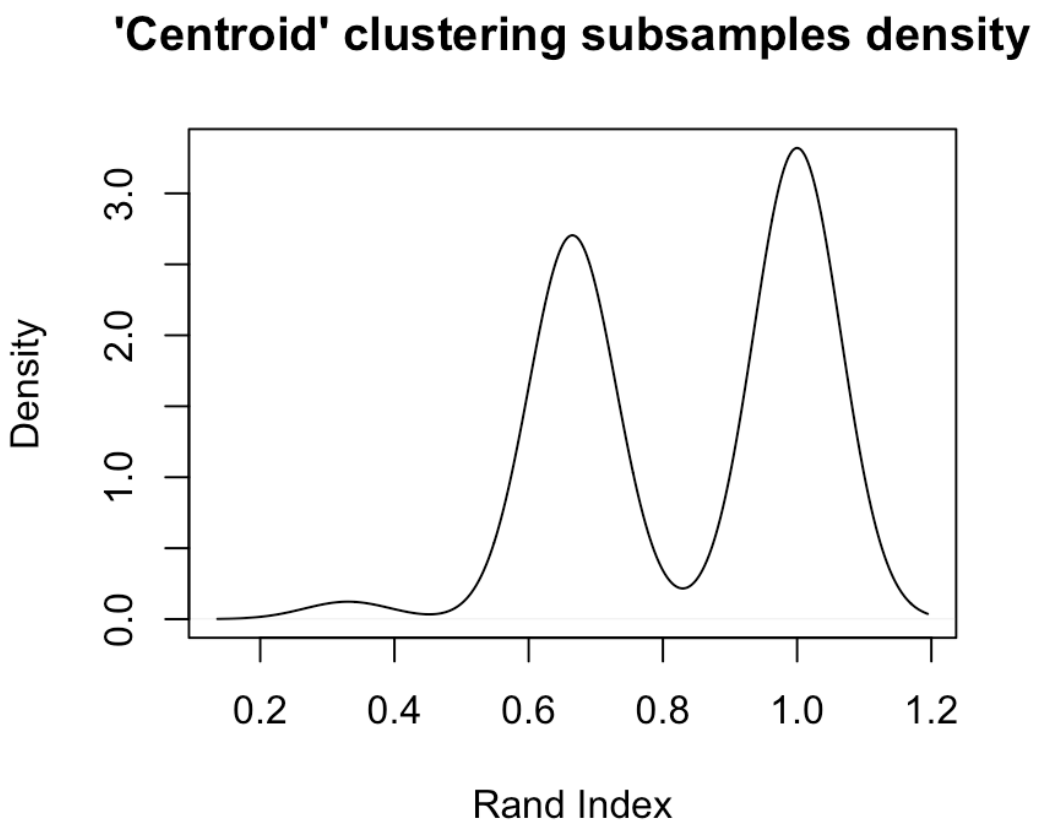
Устойчивость алгоритмов



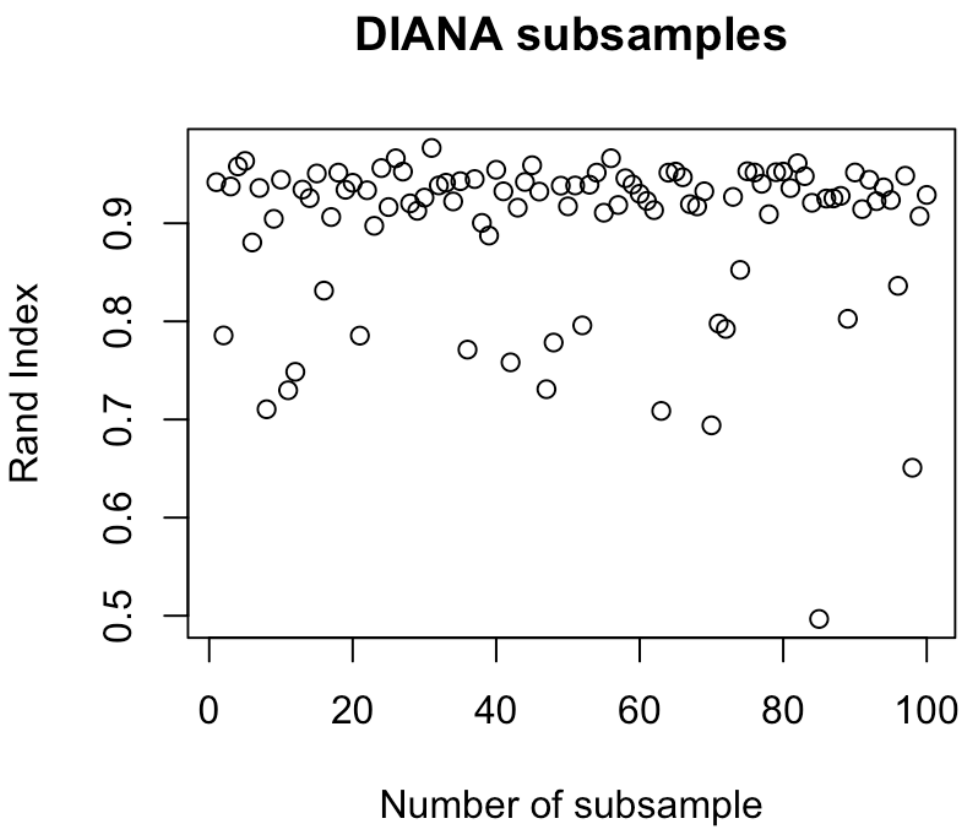


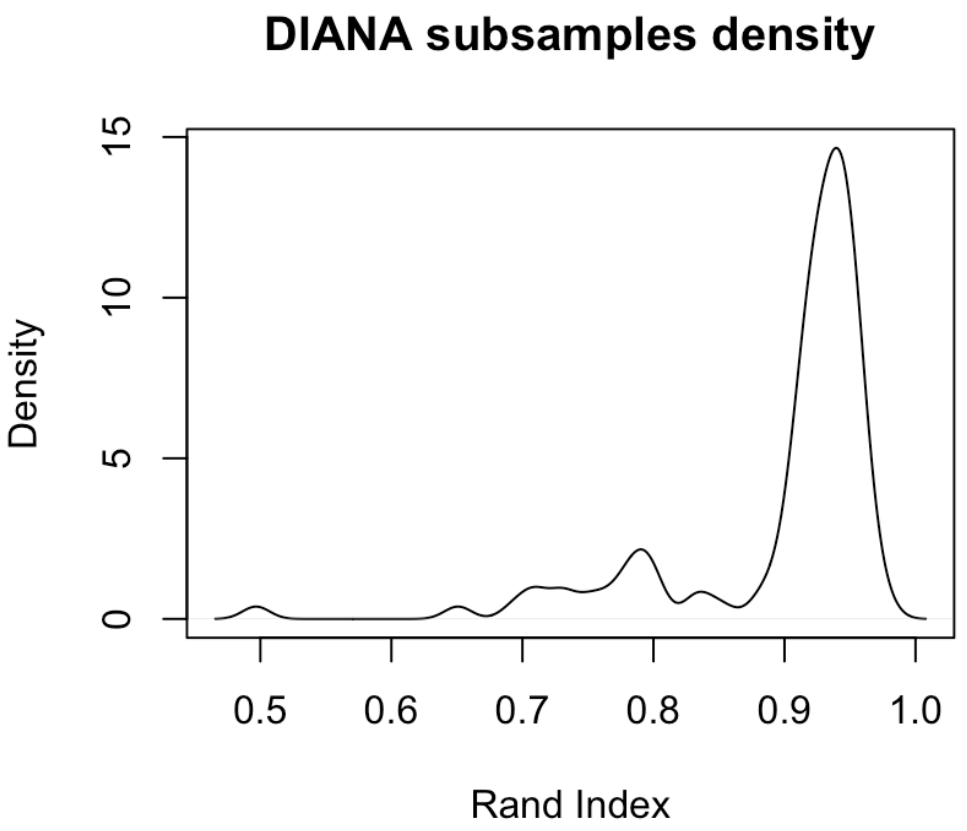
Устойчивость алгоритмов





Устойчивость алгоритмов





Спасибо за внимание!