НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Internet Studies Lab, Department of Applied Mathematics and Business Informatics

1. ЭЛЕМЕНТЫ МАТЕМАТИКИ В МАШИННОМ ОБУЧЕНИИ 2. КЛАСТЕРНЫЙ АНАЛИЗ

Анализ баз данных в публичном управлении

Кольцов С.Н.

Saint Petersburg, 07.09.2018

РАСПРЕДЕЛЕНИЕ РАЗДЕЛОВ МАТЕМАТИКИ В МАШИННОМ ОБУЧЕНИИ

- 35% – линейная алгебра;
- 25% – теория вероятности и математическая статистика;
- 15% – математический анализ;
- 15% – алгоритмы;
- 10% – подготовка данных.

ЭЛЕМЕНТЫ ЛИНЕЙНОЙ АЛГЕБРЫ

Матрицы Матрицей размера **mxn** называется прямоугольная таблица чисел, состоящая из m строк и n столбцов

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & a_{ij} & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

Числа a_{ij} – элементы матрицы:
i – номер строки
j – номер столбца.

Квадратная матрица (m=n)

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

ВИДЫ МАТРИЦ. ЕДИНИЧНАЯ И НУЛЕВАЯ МАТРИЦЫ

Единичная

$$E = \begin{pmatrix} 1 & 0 & .... & 0 \\ 0 & 1 & .... & 0 \\ .... & .... & ..... & ..... \\ 0 & 0 & ..... & 1 \end{pmatrix}$$

Нулевая

$$0 = \begin{pmatrix} 0 & .... & 0 \\ .... & .... & .... \\ 0 & .... & 0 \end{pmatrix}$$

Матрица-столбец (m×1)

$$A = \begin{pmatrix} a_1 \\ a_2 \\ .... \\ a_m \end{pmatrix}$$

Матрица-строка (1×n)

$$A = \begin{pmatrix} a_1 & a_2 & .... & a_n \end{pmatrix}$$

ДЕЙСТВИЯ С МАТРИЦАМИ

Сумма матриц

Суммой матриц A=(a_{ij}) и B=(b_{ij}) одинакового размера m×n называется матрица C=(c_{ij}) размера m×n, каждый элемент которой равен сумме соответствующих элементов матриц A и B

$$c_{ij} = a_{ij} + b_{ij}, \quad (i = \overline{1,m};\ j = \overline{1,n})$$

Пример.

$$A + B = \begin{pmatrix} 2 & 3 \\ -1 & 0 \end{pmatrix} + \begin{pmatrix} 4 & 5 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 2+4 & 3+5 \\ -1+2 & 0+3 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 1 & 3 \end{pmatrix}$$

Разность матриц

Разностью матриц A=(a_{ij}) и B=(b_{ij}) одинакового размера m×n называется матрица C=(c_{ij}) размера m×n, каждый элемент которой равен разности соответствующих элементов матриц A и B

$$c_{ij} = a_{ij} - b_{ij}, \quad (i = \overline{1,m};\ j = \overline{1,n})$$

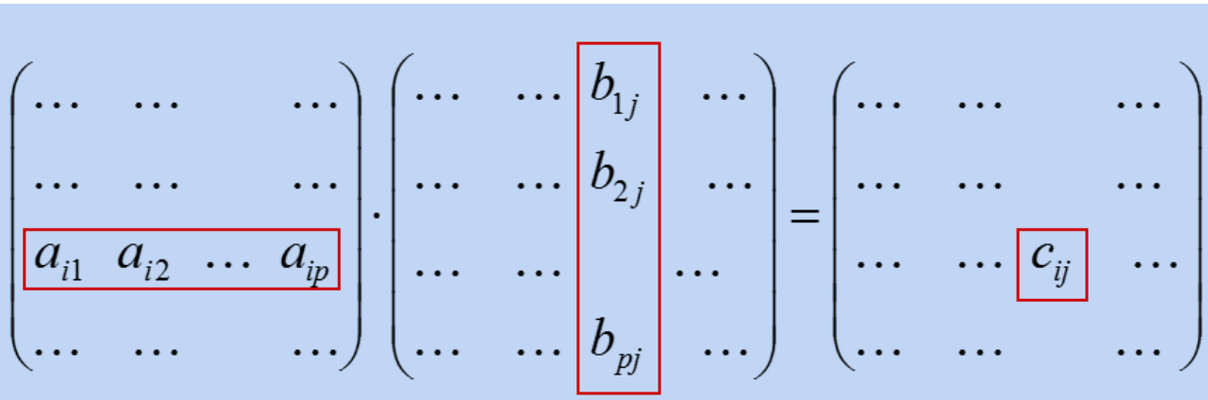
$$A = \begin{pmatrix} 4 & 3 \\ -1 & 0 \\ 1 & 2 \end{pmatrix} \text{ и } B = \begin{pmatrix} 0 & 2 \\ 3 & -1 \\ 2 & 2 \end{pmatrix}$$

$$A - B = \begin{pmatrix} 4 & 3 \\ -1 & 0 \\ 1 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 2 \\ 3 & -1 \\ 2 & 2 \end{pmatrix}$$

ДЕЙСТВИЯ С МАТРИЦАМИ

Умножение матриц

Произведением матрицы $A=(a_{ij})$ (размера $m\times p$) на матрицу $B=(b_{ij})$ (размера $p\times n$) называется матрица $C=(c_{ij})$ (размера $m\times n$), элементы которой вычисляются по формулам:



$$c_{ij} = \sum_{k=1}^{p} a_{ik} \cdot b_{kj} = a_{i1}b_{1j} + a_{i2} \cdot b_{2j} + \ldots + a_{ip} \cdot b_{pj}$$

$$i = \overline{1,m}; \quad j = \overline{1,n}.$$

$$A = (2, -1, 4) \quad B = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}$$

$$AB = 2 \cdot 1 - 1 \cdot 0 + 4 \cdot 3 = 14$$

ОБРАТНАЯ МАТРИЦА

Пусть дана невырожденная (det A≠0) квадратная матрица порядка n

$$A = \begin{pmatrix} a_{11} & a_{12} & .... & a_{1n} \\ a_{21} & a_{22} & .... & a_{2n} \\ .... & .... & ..... & ..... \\ a_{n1} & a_{n2} & ..... & a_{nn} \end{pmatrix}$$

Матрица A^{-1} называется обратной к матрице A, если выполняются равенства

$$A^{-1} \cdot A = A \cdot A^{-1} = E,$$

E – единичная матрица.

СИСТЕМЫ ЛИНЕЙНЫХ УРАВНЕНИЙ

$$a_1x_1 + a_2\ x_2 + \ldots + a_nx_n = b$$

где a_i, b – известные заданные числа

$x_1, x_2, \ldots, x_n$ - неизвестные уравнения.

a_i называются *коэффициентами уравнения*,
b называется *свободным членом*.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m. \end{cases}$$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} b_1 \\ b_2 \\ \ldots \\ b_m \end{pmatrix}$$

матричное уравнения

AX=B

ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТИ

Случайная величина — это величина, которая принимает в результате опыта одно из множества значений, причём появление того или иного значения этой величины до её измерения нельзя точно предсказать.

1. В частотном подходе (классический подход) предполагается, что случайность есть объективная неопределенность. Вероятность рассчитывается из серии экспериментов и является мерой случайности как эмпирической данности. Исторически частотный подход возник из практической задачи: анализа азартных игр — области, в которой понятие серии испытаний имеет простой и ясный смысл.

2. В байесовском подходе предполагается, что случайность характеризует наше незнания. Например, случайность при бросании кости связана с незнанием динамических характеристик игральной кости, сопротивления воздуха и так далее.

Многие задачи частотным методом решить невозможно (точнее, вероятность искомого события строго равна нулю). В то же время интерпретация вероятности как меры нашего незнания позволяет получить отличный от нуля осмысленный ответ.

Понятие условной вероятности

Условной вероятностью события **A** при условии, что произошло событие **B**, называется число **P(A|B)=P(B, A)/ P(B)**,

P(B, A) – произведение вероятностей,
P(B) – полная вероятность события **B.**

Например. В урне 3 белых и 3 черных шара. Из урны дважды вынимают по одному шару, не возвращая их обратно. Найти вероятность появления белого шара при втором испытании (**событие A**), если при первом испытании был извлечен черный шар (**событие B**).

Решение задачи:
Событие B – это вытаскивание первого шара (а именно черного). Вероятность события B=3/6=1/2 – вер. вытащить черный шар.
События A – это вытаскивание второго шара (а именно белого), так как в урне осталось 5 шаров, то вероятность этого события A=3/5
Таким образом, совместная вероятность событий A и B это произведение вероятностей этих событий P(B, A) =(3/6)*(3/5)=9/30
Полная вероятность события B=1/2

Итоговый результат: {3/6*3/5}/(1/2)=**3/5**

ДИФФЕРЕНЦИАЛЬНЫЕ УРАНЕНИЯ

Дифференциальным уравнением называют уравнение, содержащие производную или производные неизвестной функции.

Дифференциальное уравнение 1-го порядка: $f(y', y, x) = 0$

Задачей Коши для дифференциального уравнения 1-го порядка, разрешенного относительно производной, называют задачу об отыскании решения уравнения, удовлетворяющего начальному условию.

$$\begin{cases} y' = f(x, y) \\ y(x = x_0) = y_0 \end{cases}$$

Порядок старшей производной или старшего дифференциала искомой функции в уравнении называется **порядком уравнения**

СУЩЕСТВОВАНИЕ И ЕДИНСТВЕННОСТЬ РЕШЕНИЯ ЗАДАЧИ КОШИ

$$\begin{cases} y' = f(x, y) \\ y(x = x_0) = y_0 \end{cases}$$

Если правая часть уравнения непрерывна и имеет непрерывную частную производную

$\frac{df}{dx}$ в области D, то решение данного уравнения с заданными начальными условия существует и это решение единственно, то есть, через точку $y(x = x_0) = y_0$ проходит единственная интегральная кривая.

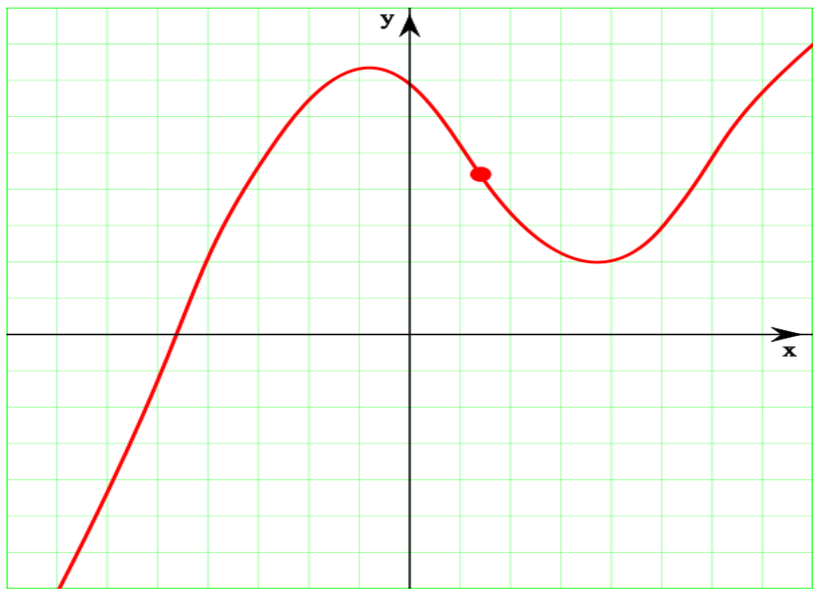


График решения уравнения называется интегральной кривой.

Важно также понимать, что теорема содержит только достаточные условия существования и единственности решения — при нарушении условий теоремы задача Коши может иметь или не иметь решений, может иметь несколько решений.

Понятие многомерного пространства

Векторное пространство называется n-мерным, если в нем можно найти n линейно независимых векторов, но больше, чем n линейно независимых векторов оно не содержит.

Размерность пространства – это максимальное число содержащихся в нем линейно независимых векторов.

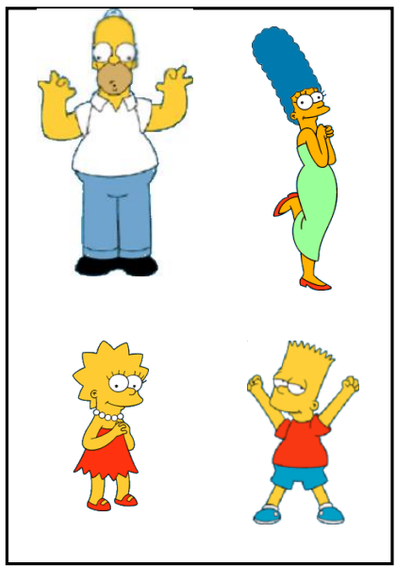
Система линейно независимых векторов векторного пространства называется базисом этого пространства, если любой вектор из может быть представлен в виде линейной комбинации векторов этой системы, т.е. для каждого вектора существуют вещественные числа такие, что имеет место равенство

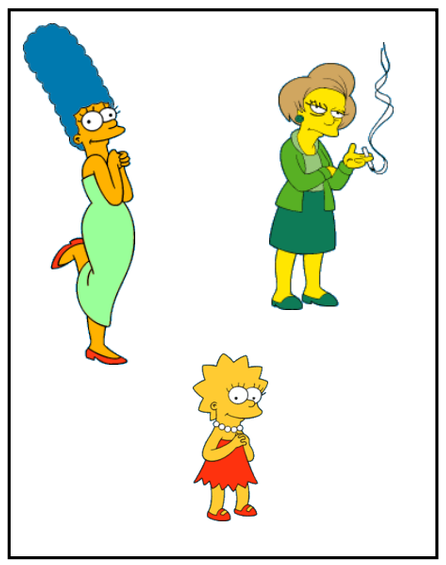
$$x = \alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_n e_n .$$

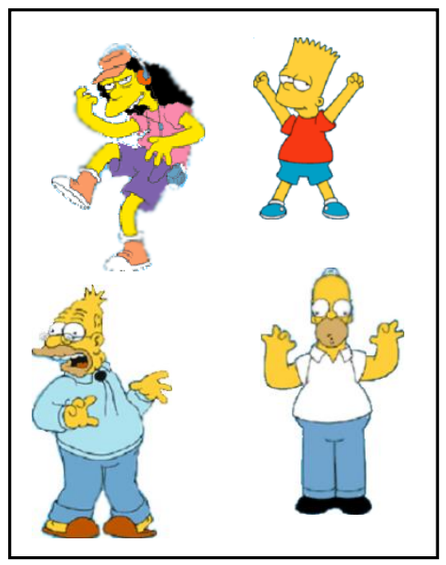
Это равенство называется разложением вектора **x** по базису **e**

Цели и задачи кластерного анализа

Кластеризация – это разбиение элементов некоторого множества на группы на основе их схожести. Задача кластеризации состоит в разбиении объектов из X на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров.







Семья

Сотрудники

Женщины

Мужчины

Лейбелинг групп – то что нужно найти

Цели и задачи кластерного анализа

Процедура кластеризации – зависит от меры сходства или не сходства. Такие меры выражаются виде функций расстояний, выраженных в виде той или иной функции.

Задача определения сходства является задачей Machine learning.

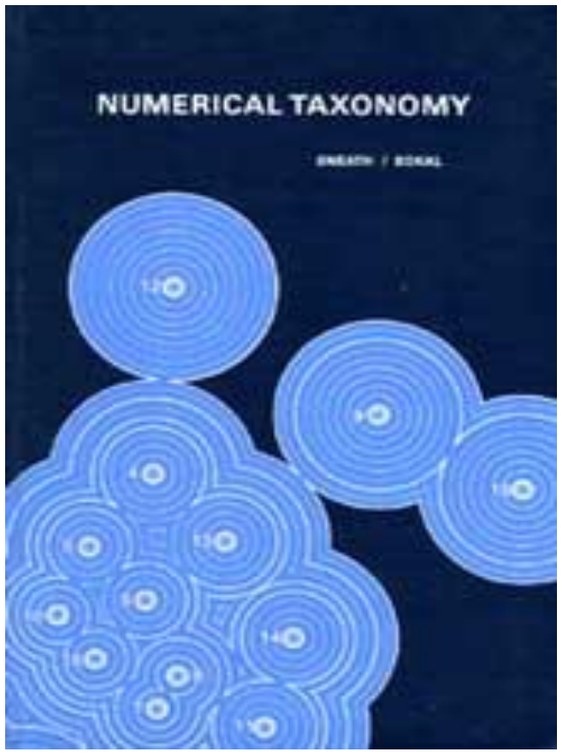
Сходство тяжело определить

“*We know it when we see it*”

Краткая история кластер анализа

Первые работы, описывающие методы кластерного анализа относятся к концу 30-х годов. Считается, что термин «кластерный анализ» первым в употребление ввёл американский психолог из университета Беркли Роберт Трайон (Robert C. Tryon) в 1939. Однако активный интерес к данной теме начался в 60 годы.

Импульсом для разработки многих кластерных методов послужила книга «**Начала численной таксономии**», опубликованная в 1963 г. двумя биологами — **Робертом Сокэлом** и **Петером Снитом**



Современные исследователи:

1. Миркин Б.Г. МЕТОДЫ КЛАСТЕР-АНАЛИЗА ДЛЯ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ (2011).
2. Everitt B.S., Landau S., Leese M. et al. (2011) Cluster Analysis. 5th ed. Wiley, 2011
3. García-Escudero L., Gordaliza A., Matrán C. et al. (2010) A review of robust clustering methods, Advances in Data Analysis and Classification, 4, 2–3: 89–109.

Направления в кластер - анализе

Partitioning approach: плоская кластеризация - предполагает разделение объектов на кластеры сразу, причем один объект относится только к одному кластеру.

Typical methods**: K-means**, k-medoids, CLARANS

Fuzzy approach: Метод нечеткой кластеризации позволяет разбить имеющееся множество объектов p на заданное число нечетких множеств, то есть один и тот же объект может принадлежать разным классам. Принадлежность характеризуется степенью принадлежности, например вероятностью.

Typical methods:**C-means** (С-средних)

Hierarchical approach:

Восходящая/нисходящая кластеризации: Иерархическая кластеризация (восходящая) - допускаем наличие подкластеров, осуществляется в несколько приемов, в результате образуется в иерархическое дерево (дендрограмму).

Typical methods: **Hierarchical**, Diana, Agnes, BIRCH, ROCK, CAMELEON

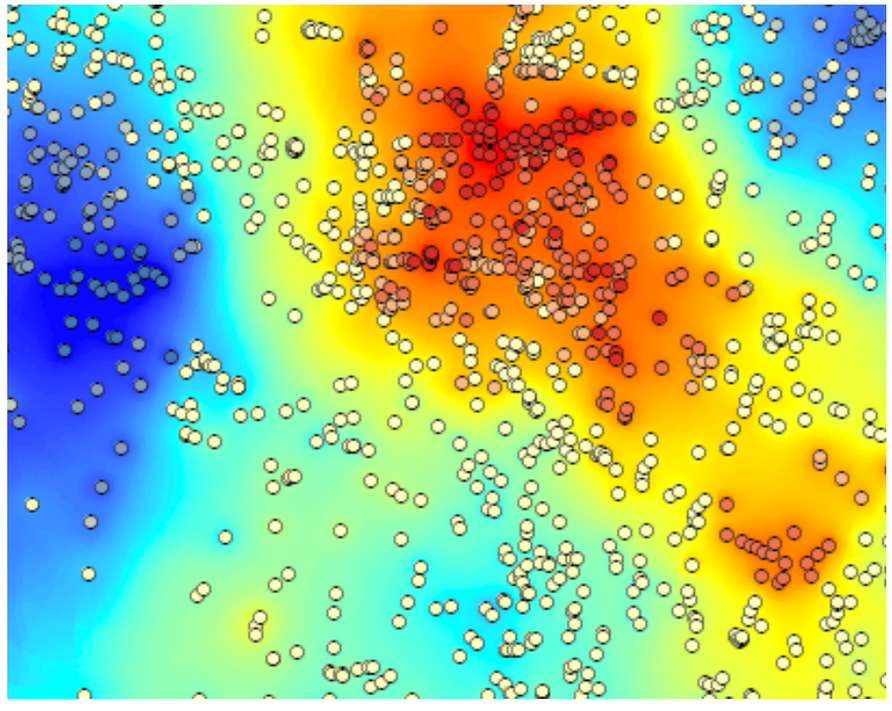
Density-based approach:

Based on connectivity and density functions

Typical methods: **DBSACN**, OPTICS, DenClue

ПРИМЕНЕНИЕ КЛАСТЕР - АНАЛИЗА

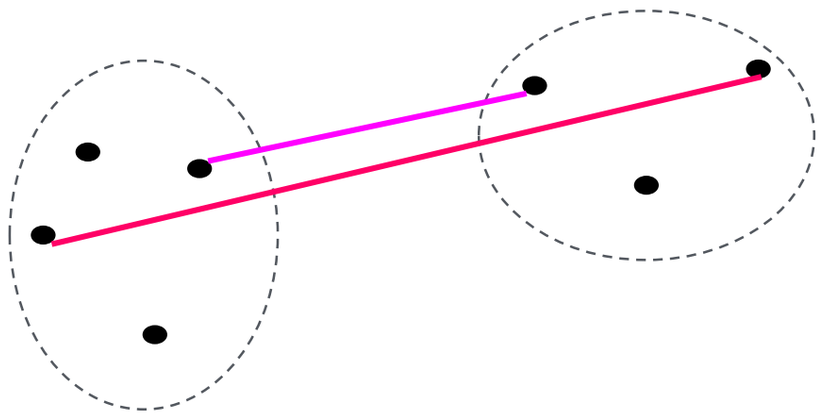
1. Статистика
2. Распознавание образов
3. Финансовая математика
4. Автоматическая классификация в различных областях науки (например, в археологи, биологии (кластеризация видов животных и растений))
5. Маркетинг. Маркетологи выделяют группы с целью оптимизации рекламной деятельности, оптимизации логистической деятельности.
6. Исследование свойств ДНК
7. Страхование (цель выделения групп населения и соотнесение групп с геогрф. расположением, заработком, семейным статусом и другой..)
8. Городское планирование.
9. Финансовое планирование города, района….
10. Социологические исследования.



Меры близости

Евклидово расстояние - наиболее общий тип расстояния. Является геометрическим расстоянием между точками в многомерном пространстве:

$$\rho_{ij} = \left[\sum_k (x_{ik} - x_{jk})^2 \right]^{1/2}$$



где: $\mathbf{X_i}$ **,** $\mathbf{X_j}$ **-** координаты **i**-го и **j**-го объектов в k-мерном пространстве;

$\mathbf{x_{il}}$ **-** $\mathbf{x_{jl}}$ **-** величина **l**-той компоненты у **i**-го (**j**-го) объекта (**l=1,2,...,k; i,j=1,2,...,n**).

Квадрат евклидова расстояния - используется, чтобы придать большие веса более отдаленным друг от друга объектам:

$$\rho_{ij} = \left[\sum_k (x_{ik} - x_{jk})^2 \right]$$

Меры близости

Расстояние city-block (городских кварталов) или манхэттенское расстояние - по сравнением с евклидовым расстоянием влияние отдельных больших разностей (выбросов) уменьшается, так как они не возводятся в квадрат:

$$\rho_{ij} = \sum_k | x_{ik} - x_{jk} |$$

где: $\mathbf{X_i}$, $\mathbf{X_j}$ - координаты **i**-го и **j**-го объектов в k-мерном пространстве;

$\mathbf{x_{il}}$ - $\mathbf{x_{jl}}$ - величина **l**-той компоненты у **i**-го (**j**-го) объекта (**l=1,2,...,k; i,j=1,2,...,n**).

Расстояние Минковского (Minkowski Metric)

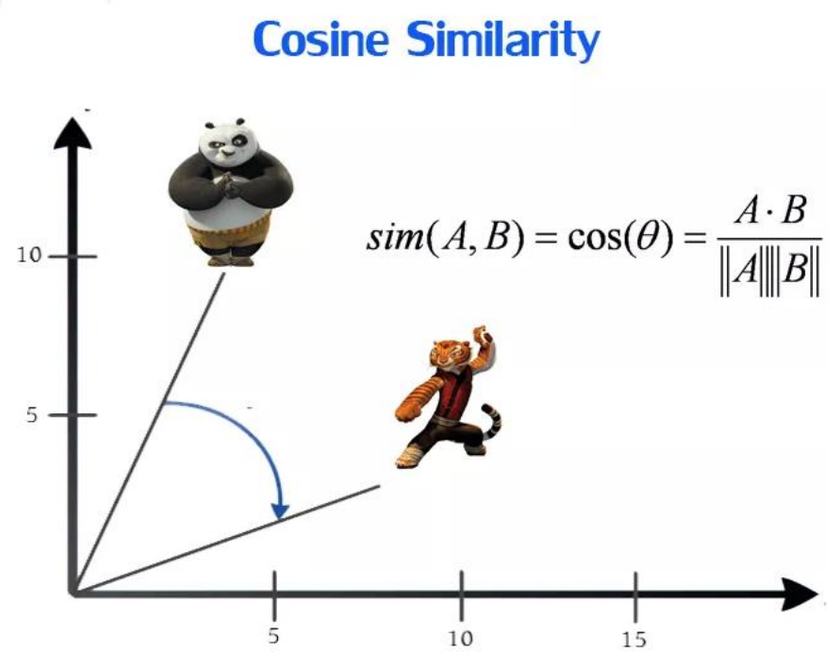
$$\rho_{ij} = \left[\sum_k \left| x_{ik} - x_{jk} \right|^p \right]$$

Если P=1 – расстояние городских кварталов,
Если P=2 – Евклидово расстояние

Меры близости

Cosine similarity: косинус угла между двумя векторами.
Скалярное произведение векторов

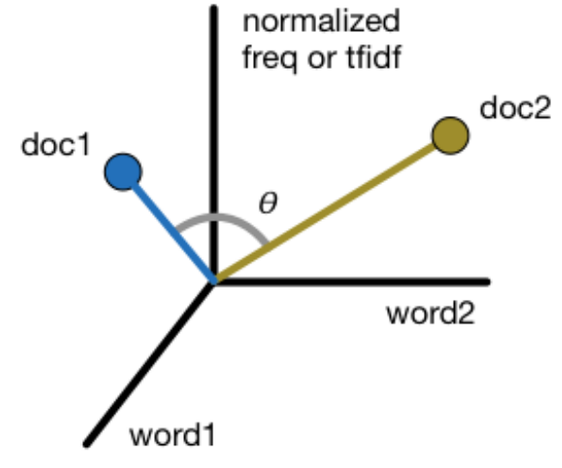
$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$



На основании скалярного произведения векторов, косинус угла можно bзразить следующим образом:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

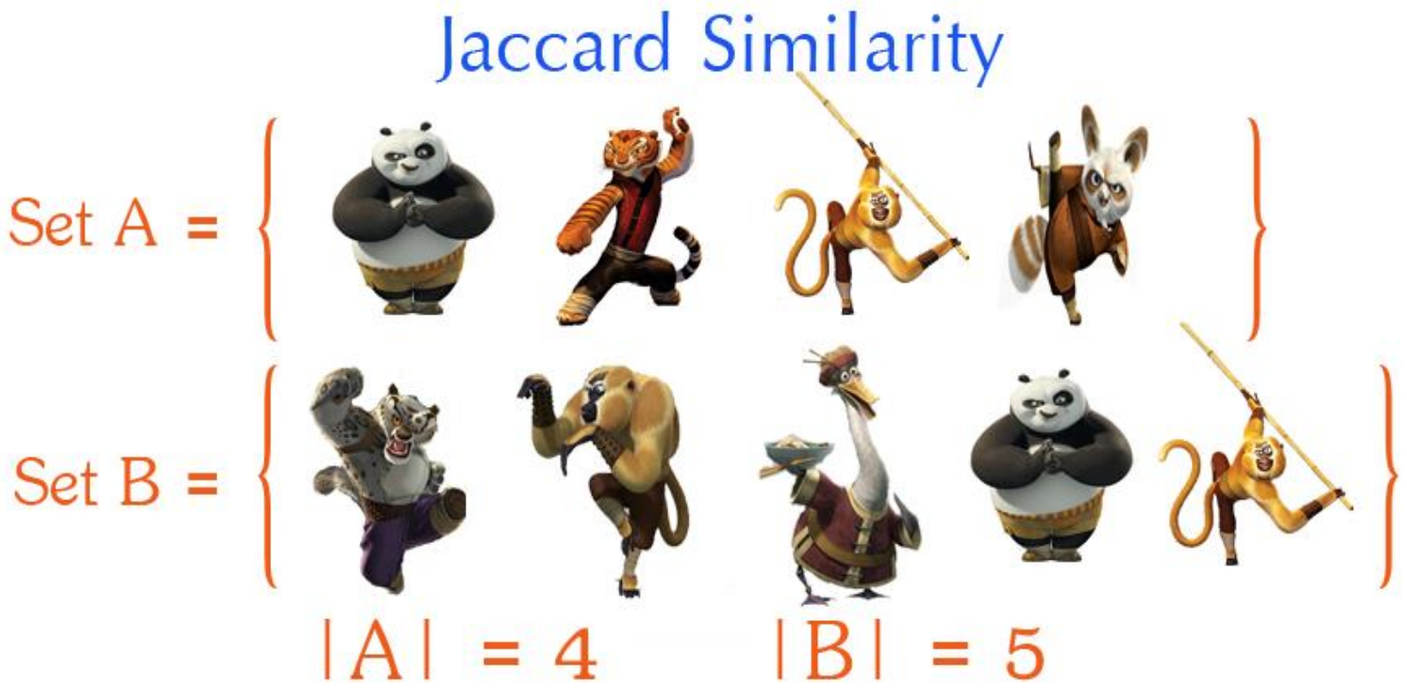
где $\mathbf{A_i}$ и $\mathbf{B_i}$ компоненты векторов в евклидовом пространстве.
Пропущенные значения компонент заполняются средними значениями, рассчитанными по существующим данным.



Например, для текстовых коллекций, атрибутами векторов являются tf-idf.

Меры близости

Jaccard distance:



Пусть у нас есть два множества X и Y. Тогда можно рассчитать следующие параметры:

A – множество величин в X, отсутствующих в Y;

B – множество величин Y, отсутствующих в X;

C – множество величин, общих для X и Y;

Тогда коэффициентом Жаккара называется следующая комбинация:

$$Jk=|c/(a+b-c)|.$$

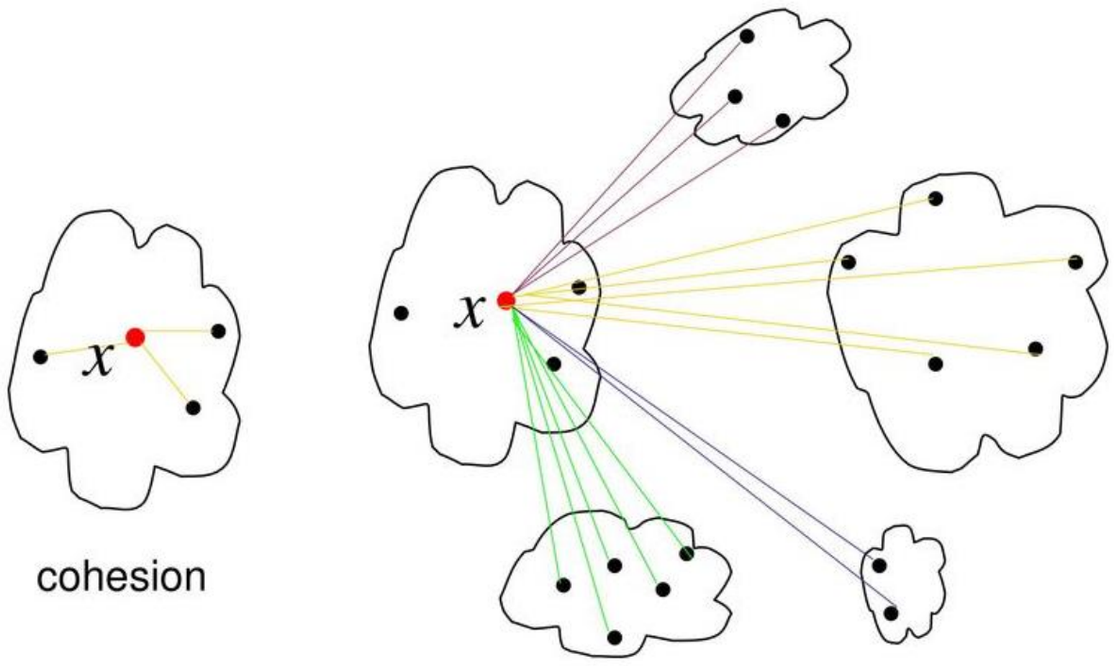
Коэффициент равен 1 в случае полного совпадения двух множеств и равны 0, если множества совершенно различны.

Например, в качестве двух множеств можно использовать наборы слова или разные фичи.

https://docs.orange.biolab.si/3/data-mining-library/reference/distance.html

Меры близости

Silhouette coefficient (SC)



$a(x)$: average distance in the cluster

b(x): average distances to others clusters, find minimal

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

$s(x)$ = [-1, +1]: -1=bad, 0=indifferent, 1=good

Silhouette coefficient (SC):

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(x)$$

https://docs.orange.biolab.si/3/data-mining-library/reference/distance.html

Алгоритм K means

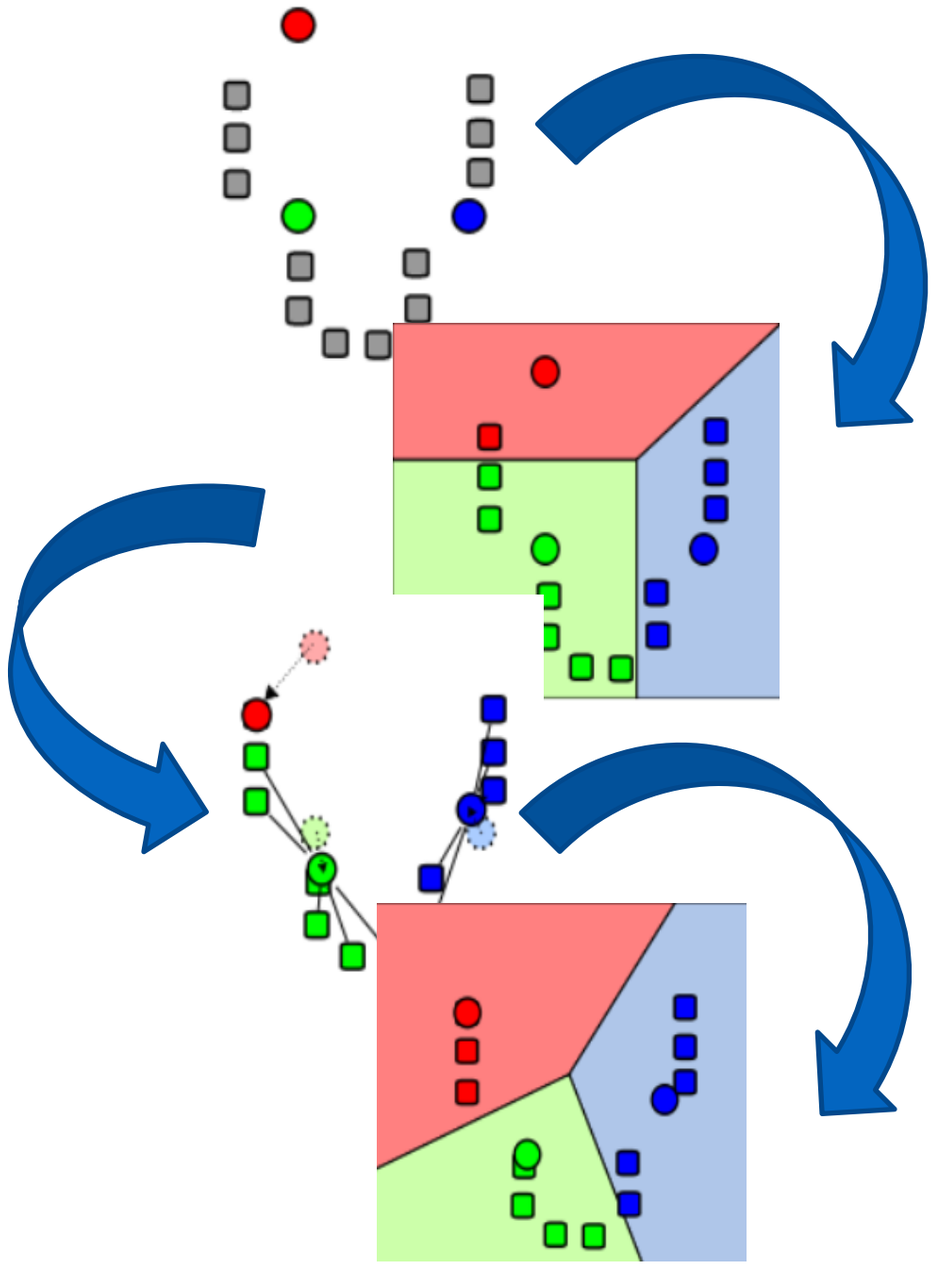
Основная суть работы алгоритма заключается в следующем: Пусть у нас есть совокупность объектов.

1. Выбираем начальные точки для кластеров.

2. Привязать ближайшие точки к центрам кластеров.

3. Пересчитать центры кластеров, исходя из того, что в кластер были добавлены новые объекты.

4. После того как нашли новые центры кластеризации, снова перераспределяем ближайшие точки по кластерам.

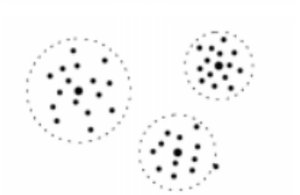


Алгоритм работает до тех пор пока центры кластеров не перестанут менять свое положение.

ВИДЫ КЛАСТЕРОВ

внутрикластерные расстояния, как правило, меньше межкластерных

ленточные кластеры



кластеры с центром

Разные виды кластеров ведут к проблеме выбора оптимального числа кластеров.

кластеры могут соединяться перемычками

кластеры могут накладываться на разрежен-ный фон из редко расположенных объектов

кластеры могут перекрываться

ПРОБЛЕМЫ КЛАСТЕРНОГО АНАЛИЗА

Задачу кластеризации можно ставить как задачу дискретной оптимизации: необходимо так приписать номера кластеров Y_i объектам X_i, чтобы значение выбранного функционала качества приняло наилучшее значение. Существует много разновидностей функционалов качества кластеризации, но нет «самого правильного» функционала.

Среднее внутрикластерное расстояние должно быть как можно меньше:

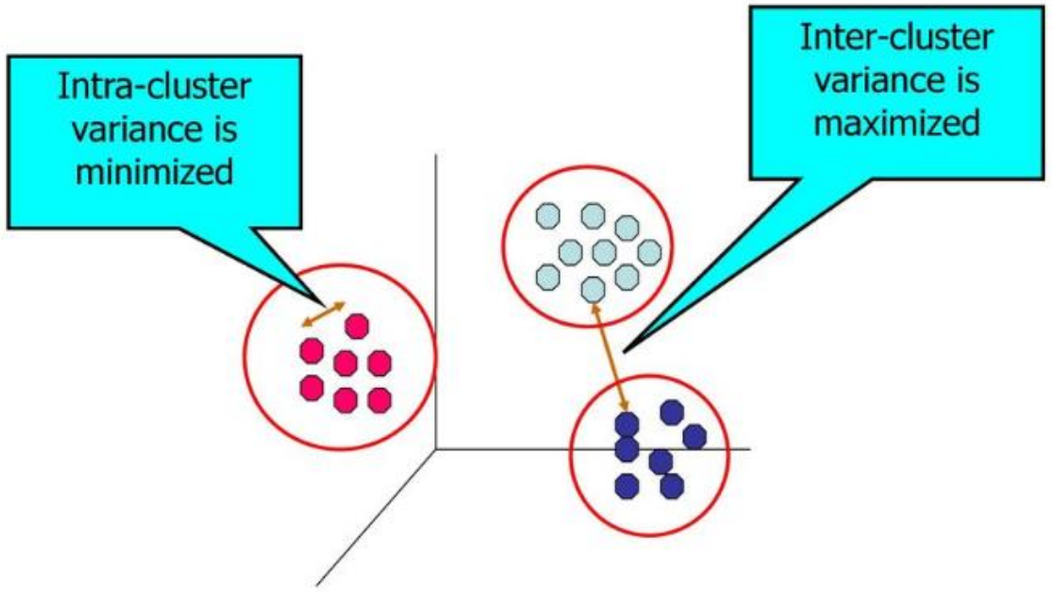
$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to \min.$$

Среднее межкластерное расстояние должно быть как можно больше:

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

Отношение пары функционалов: $F_0/F_1 \to \min.$

- Minimize within cluster variance (TSE)
- Maximize between cluster variance



ПРОБЛЕМЫ КЛАСТЕРНОГО АНАЛИЗА

1. Не гарантируется достижение глобального минимума суммарного квадратичного отклонения *V*, а только одного из локальных минимумов.
2. Результат зависит от выбора исходных центров кластеров, их оптимальный выбор заранее неизвестен.
3. Число кластеров надо знать заранее.

Как можно преодолеть эти проблемы?

1. Запускать алгоритм много раз (разными центрами кластеров), после чего выбрать результат с минимальной величиной ошибки.

2. Использовать дополнительные модели дл оценки количества кластеров.

ТЕОРИЯ СКАЧКОВ - АЛГОРИТМ SUGAR AND JAMES

1. Определение минимального искажения (distortion). В качестве минимального искажения берется минимальное значение внутрикластерной дисперсии, встречающееся в данном кластерном решении. Это значит следующее: для заданного кластерного решения (например, для 5 кластеров) рассчитываются дисперсии внутри каждого кластера (среднее внутрикластерное расстояние). Из этого множества чисел (5 штук в 5-кластерном решении) выбирается минимальное значение d.

2. Коэффициент трансформации. Согласно разработчикам метода, в качестве коэффициента трансформации можно взять величину Y=P/2, где P – размерность векторного пространства. В качестве коэффициента также можно взять величину 1/K, где K – число кластеров.

3. Transformed distortion. Данная величина рассчитывается следующим образом:

$$D_t(K) = d^Y(K)$$

4. Расчет скачков (‘Jumps’). Оценка поведения функции transformed distortion основана на оценке скачков функции, которые происходят при изменении числа кластеров. Скачок рассчитывается следующим образом:

$$J_K = d_K^{-Y} - d_{K-1}^{-Y}$$

ТЕОРИЯ СКАЧКОВ - АЛГОРИТМ SUGAR AND JAMES

Алгоритм расчета:

Процедура определения количества кластеров состоит из следующих шагов:

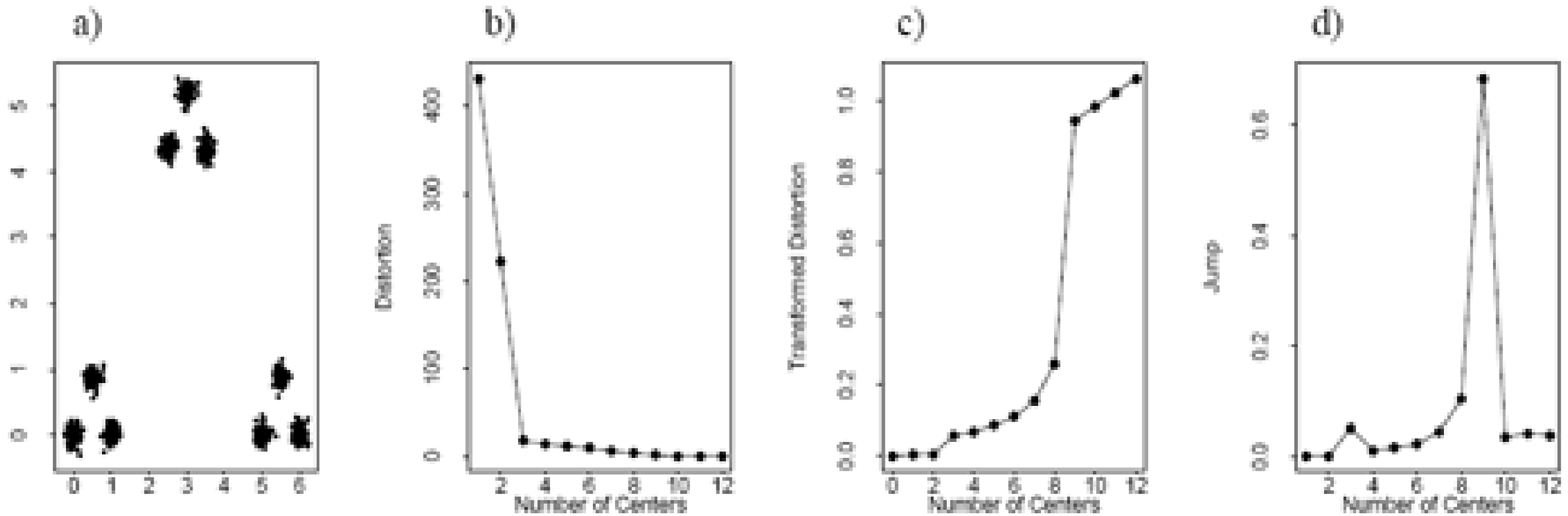
1. Запускается k-means алгоритм для K кластеров и определяется соответствующее "искажение" $\widehat{d}_k$. Для различных значений K строится набор $\widehat{d}_k$.

2. Выбирается степень трансформации $Y > 0$ (обычно принимается $Y = p/2$).

3. Вычисляются скачки по формуле $J_k = \widehat{d}_k^{-Y} - \widehat{d}_{k-1}^{-Y}$.

4. За итоговое количество кластеров выбирается то, которое соответствует наибольшему скачку $K^* = \arg\max_k J_k$.

ТЕОРИЯ СКАЧКОВ - АЛГОРИТМ SUGAR AND JAMES



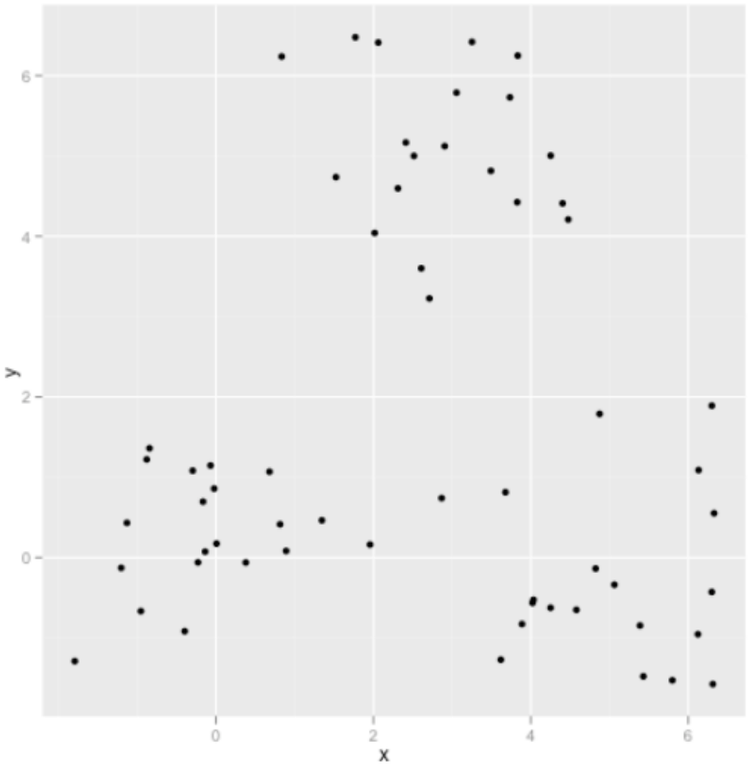
Catherine A. Sugar; Gareth M. James (2003). "Finding the number of clusters in a data set: An information-theoretic approach". Journal of the American Statistical Association. 98 (January): 750–763.

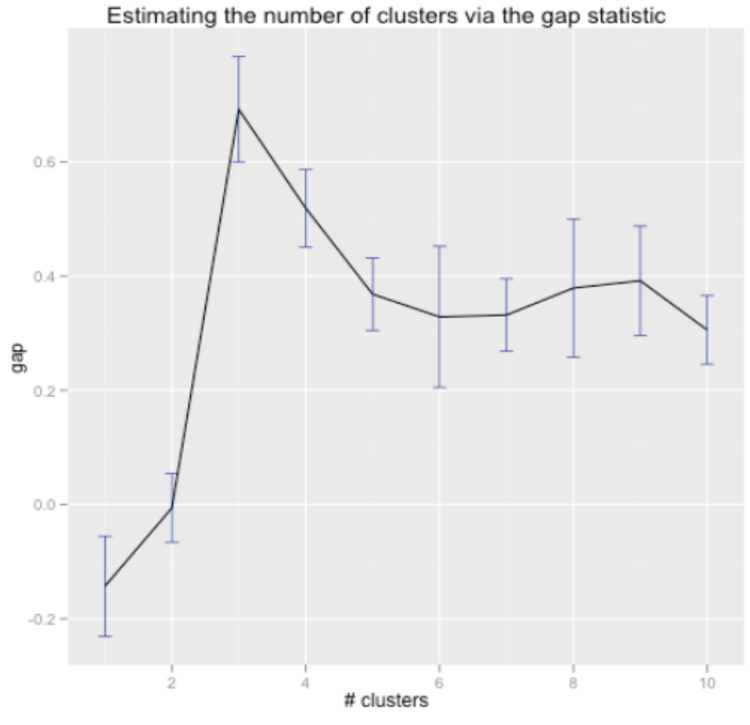
GAP STATISTIC

Ключевой идей gap statistic является измерение разницы между null reference distribution и распределением, которое получается в ходе кластеризации. Данная разница измеряется при различном числе кластеров. В рамках данного подхода авторы предполагают, что оптимальное число кластеров соответствует ситуации, когда логарифм от среднего внутри - кластерного расстояния падает ниже чем аналогичный логарифм, рассчитанный по null reference distribution:

$$Gap(k)_n = E_n\{\log(W_k)\} - \log(W_k),$$

где W_k - среднее внутри - кластерное расстояние, E_n – означает процедуру усреднения (то есть мат. ожидание) по сэмплам из null reference distribution. Соответственно разница между логарифмами дает максимальное значение.
В качестве null reference distribution используется набор объектов, которые выбираются случайным образом из исходного набора данных.





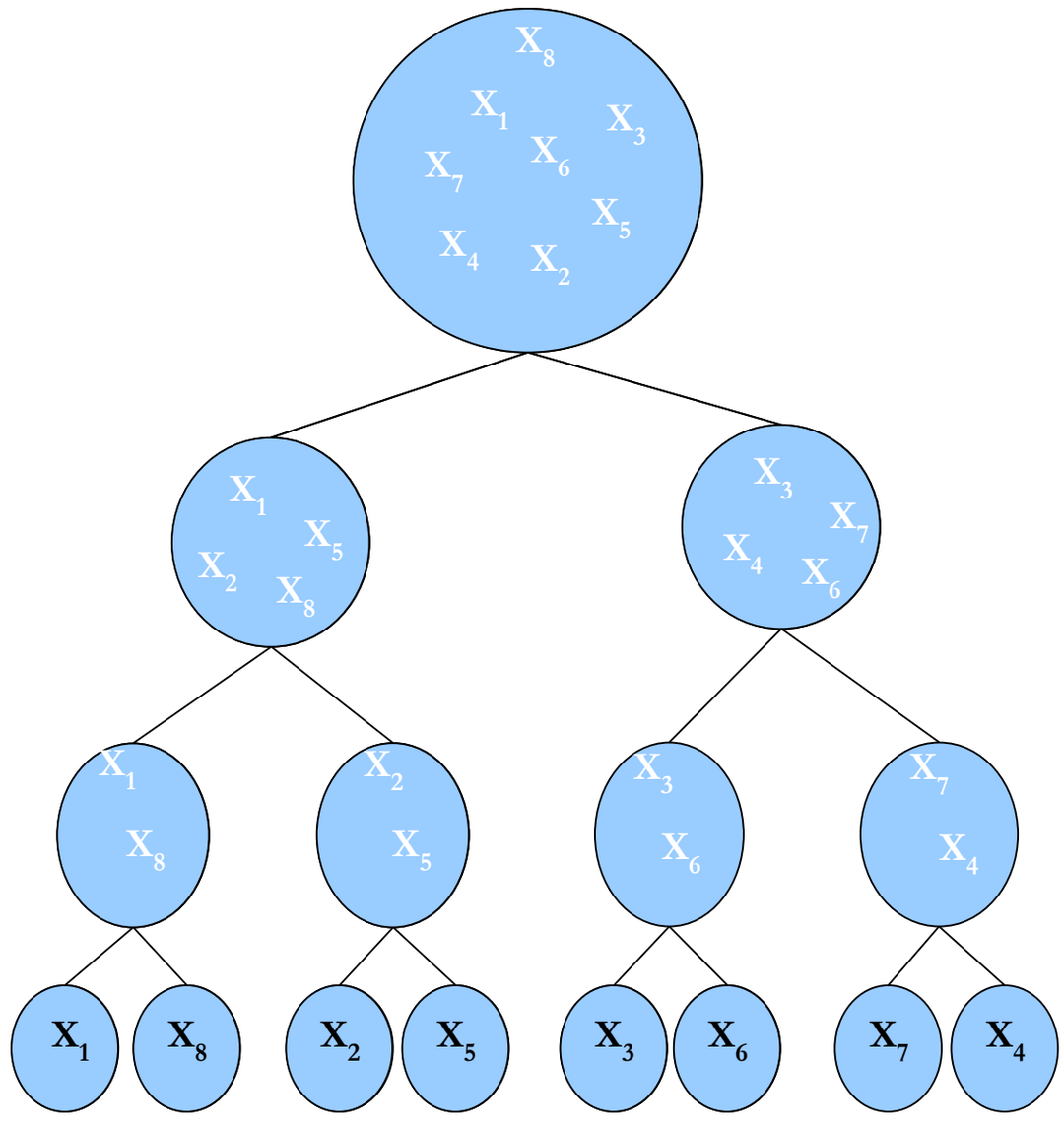
ПОЛЕЗНЫЕ СТАТЬИ ИЗ КЛАСТЕРНОГО АНАЛИЗА

1. Milligan, G.W.; Cooper, M.C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **1985**, *50*, 159–179.
2. Tibshirani, R., Walther, G., Hastie, T. (2002). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 411-423.
3. Catherine A. Sugar; Gareth M. James (2003). "Finding the number of clusters in a data set: An information-theoretic approach". Journal of the American Statistical Association. 98 (January): 750–763.
4. Fujita, André, Daniel Y. Takahashi and Alexandre G. Patriota. “A non-parametric method to estimate the number of clusters.” *Computational Statistics & Data Analysis*73 (2014): 27-39.
5. Aldana-Bobadilla, E.; Kuri-Morales, A. A Clustering Method Based on the Maximum Entropy Principle. *Entropy* **2015**, *17*, 151-180
6. Rose K, E. Gurewitz, and G. Fox. Statistical mechanics and phase transitions in clustering. Physical Review Letters, 65(8):945-948, 1990.
7. Basu S, I. Davidson, Wagstaff K. Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall, 2008.

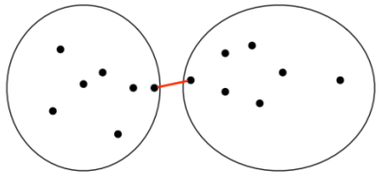
HIERARCHICAL APPROACH (ВОСХОДЯЩАЯ / НИСХОДЯЩАЯ КЛАСТЕРИЗАЦИИ)

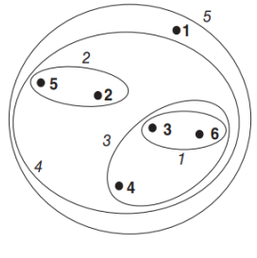
Восходящая кластеризация (agglomerative): В рамках данного алгоритма предполагается что каждый элемент нашего множества является отдельным кластером. Процесс образования новых кластеров заключается в объединение некоторых кластеров в один новый кластер. Объединение осуществляется на основе заданного расстояния между кластерами. Производя такое итеративное объединение мы получаем дерево кластеров, которое в итоге сходится к одному кластеру.

Нисходящая кластеризация (divisive): Данный вид кластеризация заключатся в следующем. Мы предполагаем, что все объекты принадлежат одному кластеру. В ходе итеративного процесса мы разделяем кластеры на несколько разных кластеров. Соответственно при этом, получаем дерево кластеров (дендрограмма).



HIERARCHICAL APPROACH



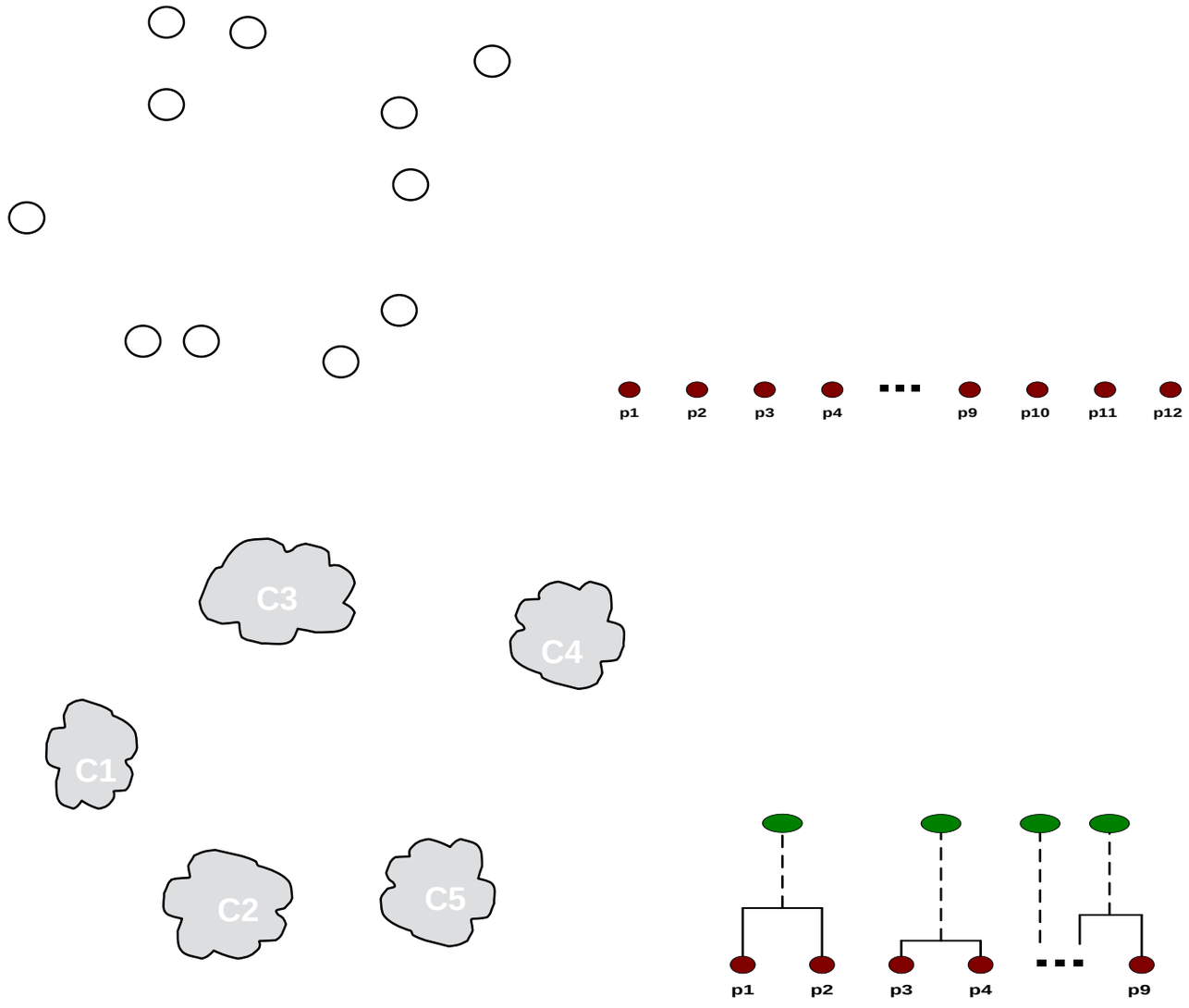


В основе иерархической кластеризации лежит использование двух вещей:
1. Расстояние.
'euclidean', 'cityblock', 'cosine'.. и другие
2. Алгоритм кластеризации на основе выбранного расстояния.
' Single-linkage clustering‘ - алгоритм в котором объединение кластеров происходит на основе расчета минимального расстояния между элементами двух кластеров.

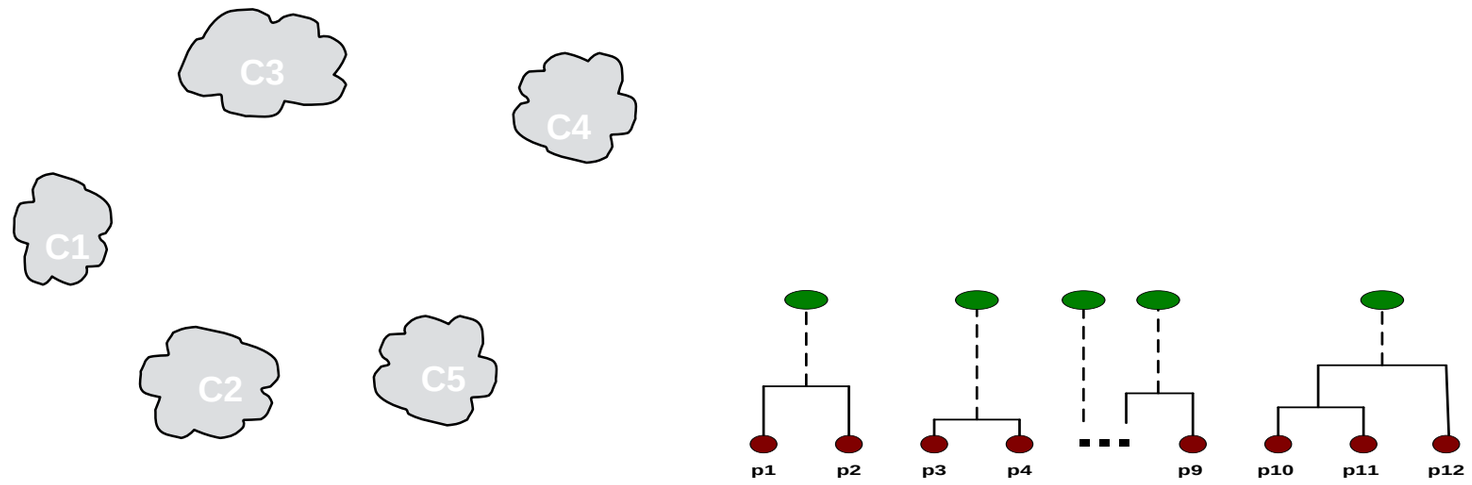
' Complete-linkage clustering‘ - Исходно каждый элемент выборки считается отдельным кластером. Кластеры последовательно объединяются, пока все элементы не попадут в один кластер. На каждом шаге алгоритма объединяются два кластера, расстояние между которыми минимальное. «Минимальное расстояние» определяется как максимум из множества расстояний между элементом первого кластера и элементом второго кластера.

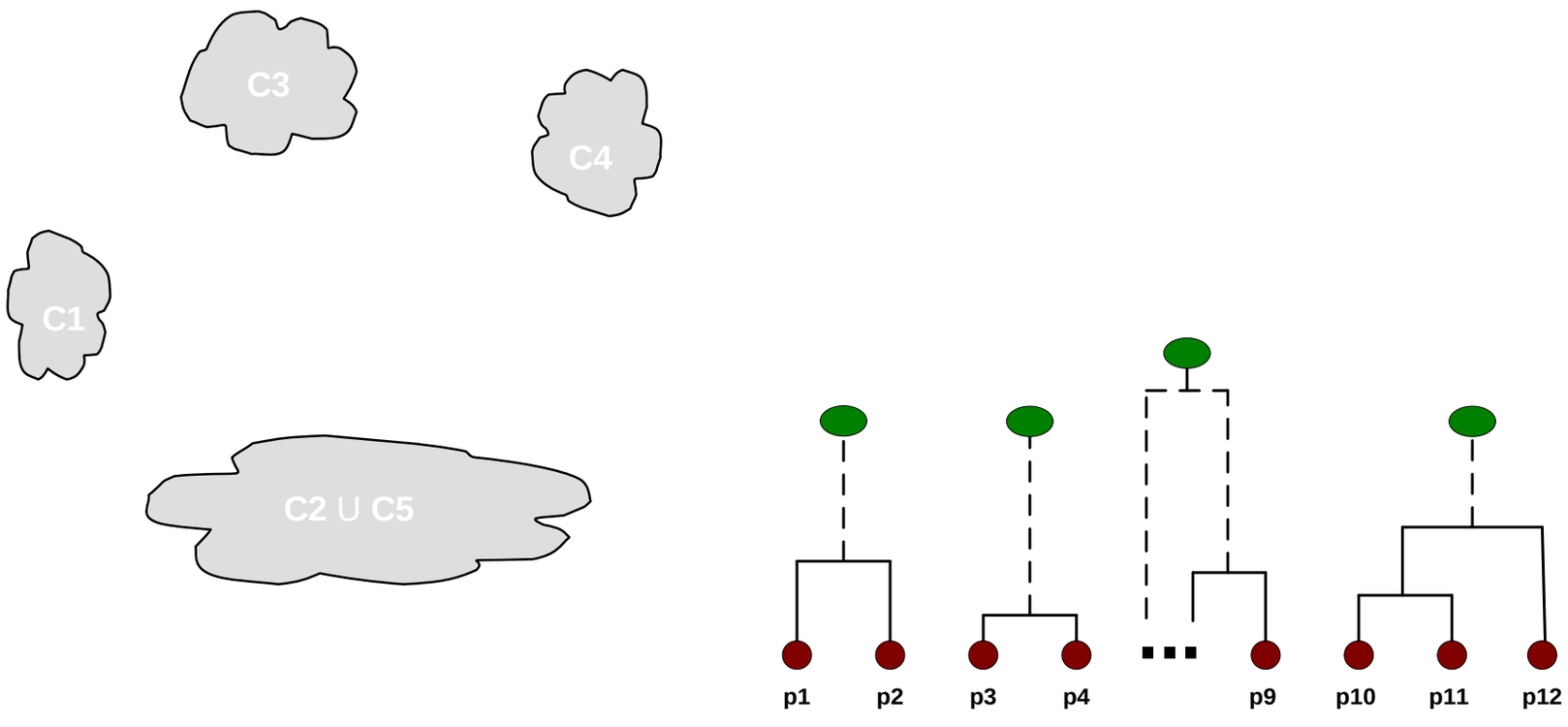
‘Average‘ (UPGMA (Unweighted Pair Group Method with Arithmetic Mean)) – алгоритм, в рамках которого происходит объединение кластеров с учетом усредненного расстояние между кластерами (усреднение по всем парам между двумя кластерами).

HIERARCHICAL APPROACH

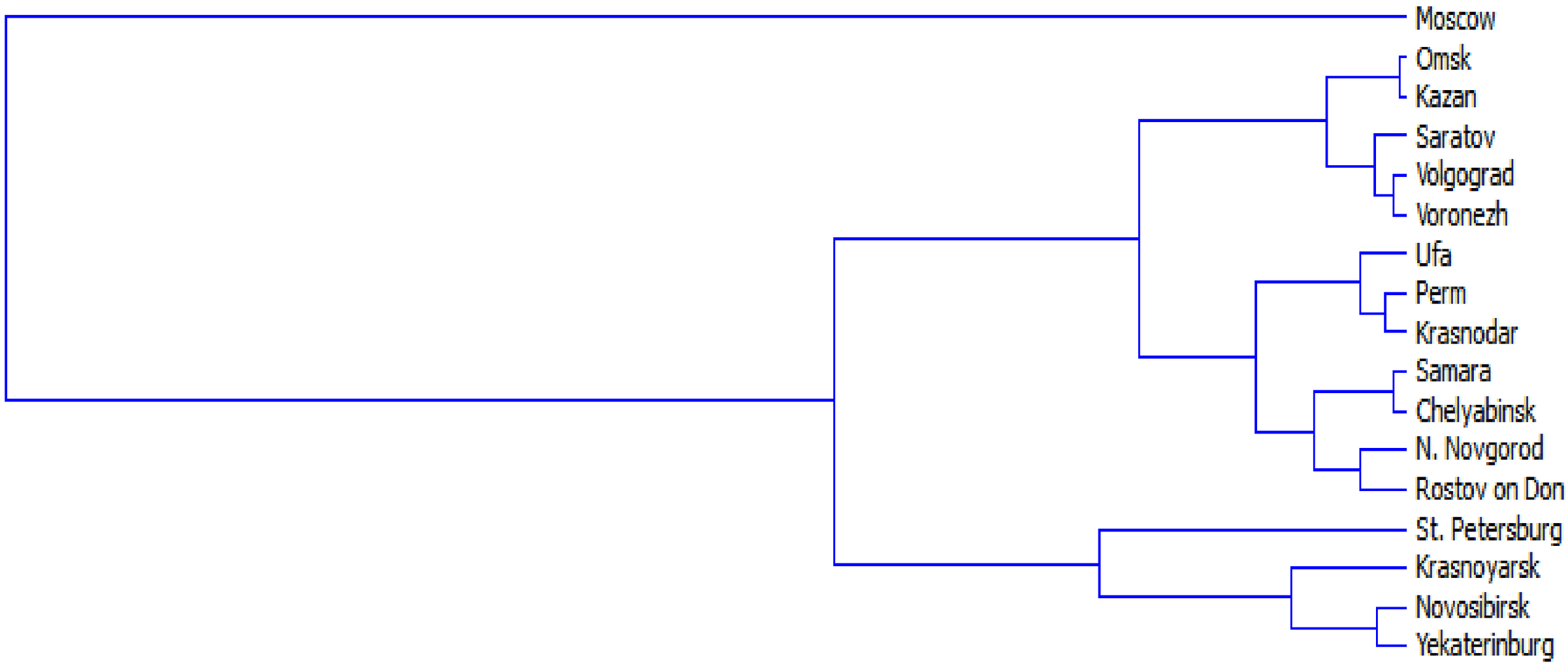


HIERARCHICAL APPROACH





HIERARCHICAL APPROACH

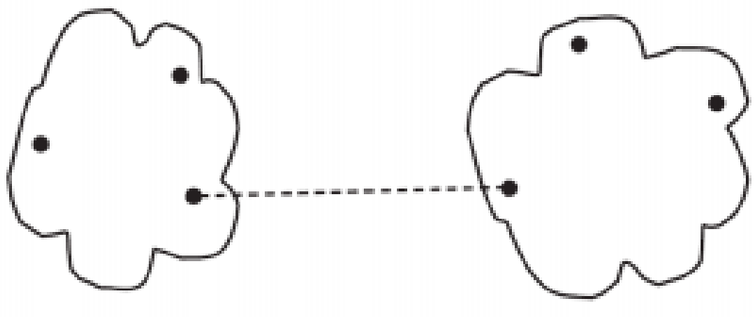


HIERARCHICAL APPROACH

1. Расстояние ближнего соседа:

$$R^{б}(W,S) = \min_{w\in W, s\in S} \rho(w,s);$$

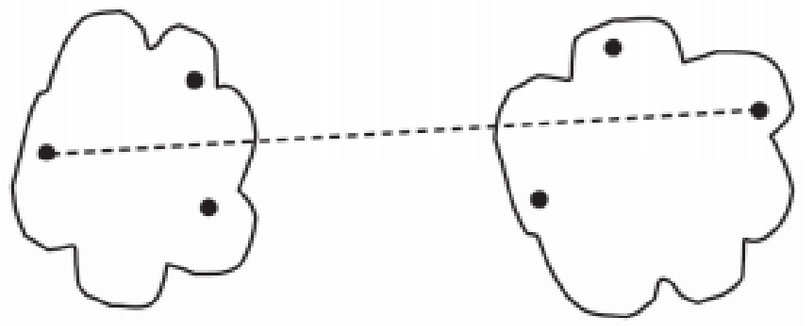
$$\alpha_U = \alpha_V = \frac{1}{2},\ \beta = 0,\ \gamma = -\frac{1}{2}.$$



2. Расстояние дальнего соседа:

$$R^{д}(W,S) = \max_{w\in W, s\in S} \rho(w,s);$$

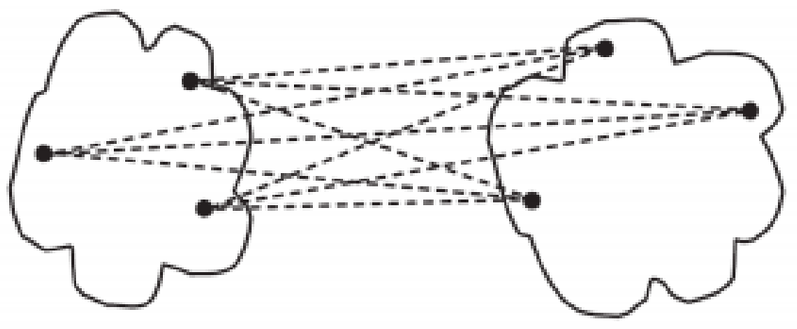
$$\alpha_U = \alpha_V = \frac{1}{2},\ \beta = 0,\ \gamma = \frac{1}{2}.$$



3. Групповое среднее расстояние:

$$R^{г}(W,S) = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s);$$

$$\alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = \gamma = 0.$$



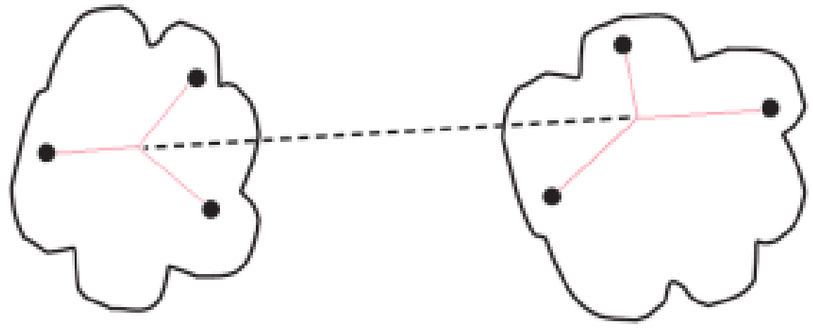
HIERARCHICAL APPROACH

4. Расстояние между центрами:

$$R^{ц}(W,S) = \rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|U|}{|W|},\ \ \alpha_V = \frac{|V|}{|W|},$$

$$\beta = -\alpha_U\alpha_V,\ \ \gamma = 0.$$



5. Расстояние Уорда:

$$R^{y}(W,S) = \frac{|S||W|}{|S|+|W|}\,\rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \ \beta = \frac{-|S|}{|S|+|W|},\ \ \gamma = 0.$$

Проблема выбора

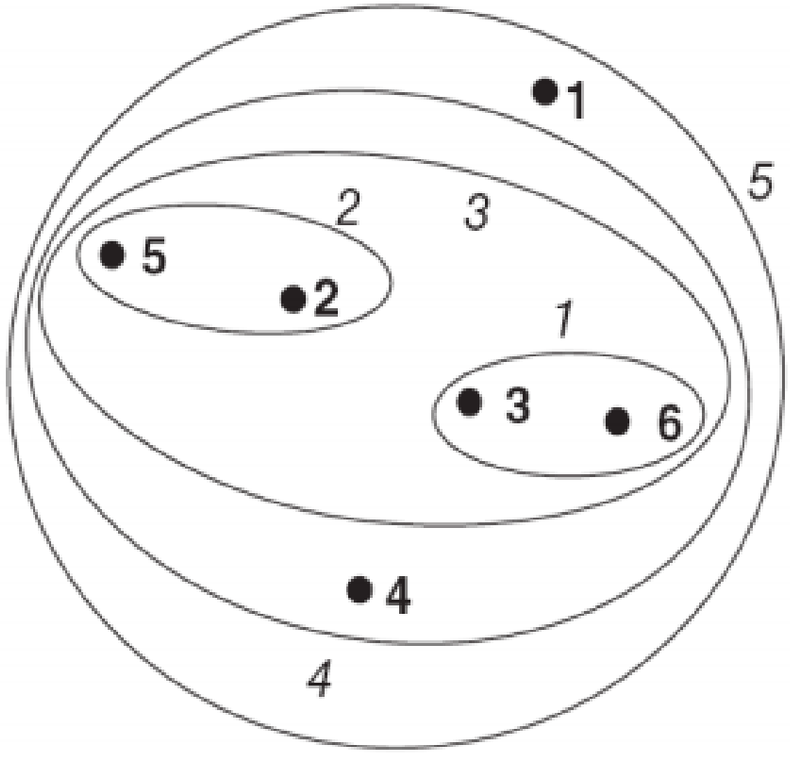
Какой тип расстояния лучше?

Правило Уорда (Варда). В этом методе в качестве целевой функции применяют внупригрупповую сумму квадратов отклонений, которая есть не что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект

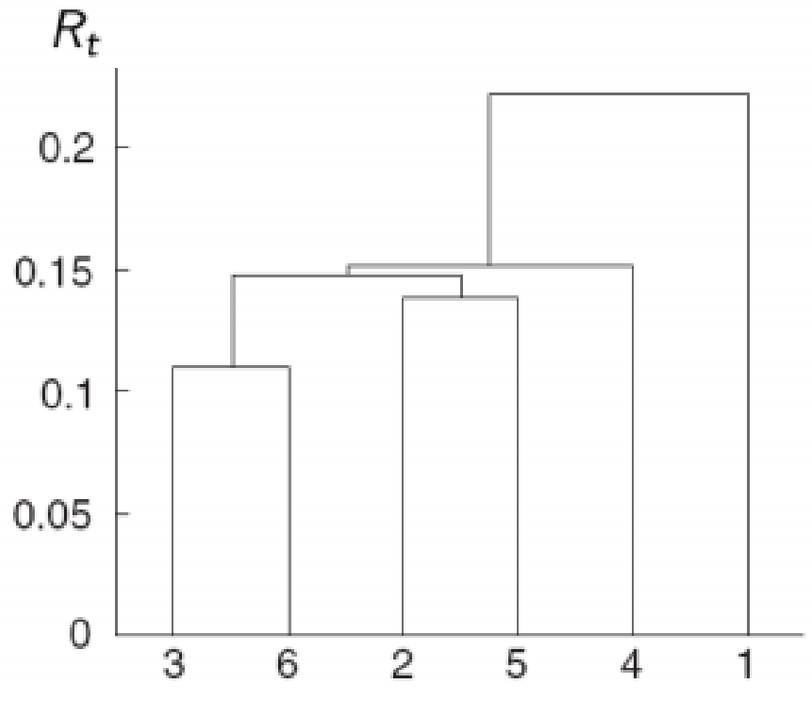
HIERARCHICAL APPROACH - ВИЗУАЛИЗАЦИЯ КЛАСТЕРНОЙ СТРУКТУРЫ

1. Расстояние ближнего соседа:

Диаграмма вложения



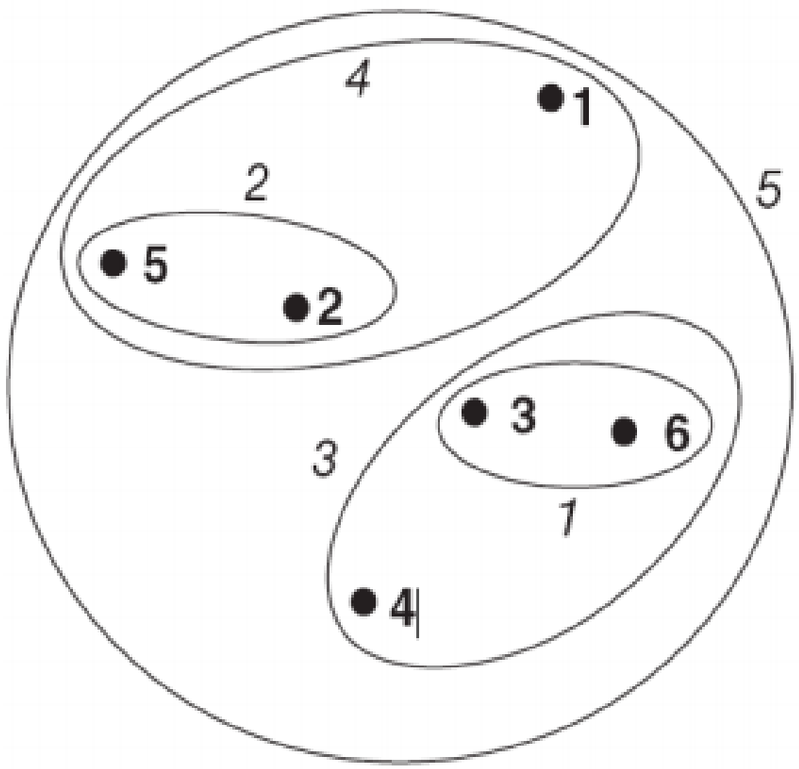
Дендрограмма



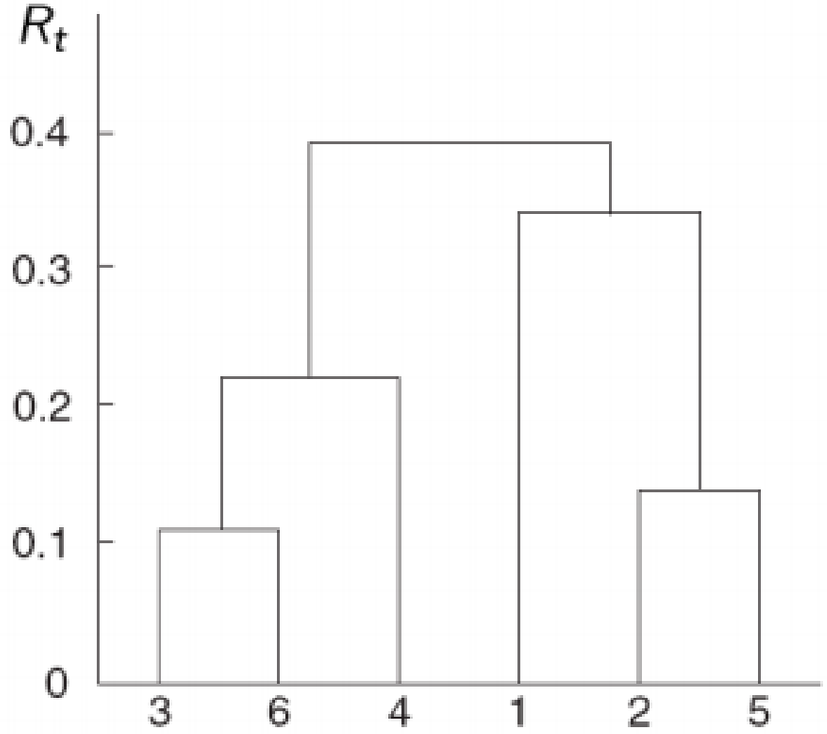
HIERARCHICAL APPROACH

2. Расстояние дальнего соседа:

Диаграмма вложения



Дендрограмма



HIERARCHICAL APPROACH – ПРОБЛЕМЫ

1. Первая проблема иерархических методов заключается в сложности определения условия остановки таким образом, чтобы выделить «естественные» кластеры и в то же время не допустить их разбиения.

2. Вторая проблема иерархических методов кластеризации заключается в выборе точки разделения или слияния кластеров. Этот выбор критичен, поскольку после разделения или слияния кластеров на каждом последующем шаге метод будет оперировать только вновь образованными кластерами, поэтому неверный выбор точки слияния или разделения на каком-либо шаге может привести к некачественной кластеризации.

3. Кроме того, иерархические методы не могут быть применены к большим наборам данных, потому как решение о разделении или слиянии кластеров требует анализа большого количества объектов и кластеров, что ведёт к большой вычислительной сложности метода.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru