РТУ МИРЭА

КЛАСТЕРИЗАЦИЯ. МЕТОДЫ K-MEANS, FUZZY C-MEANS

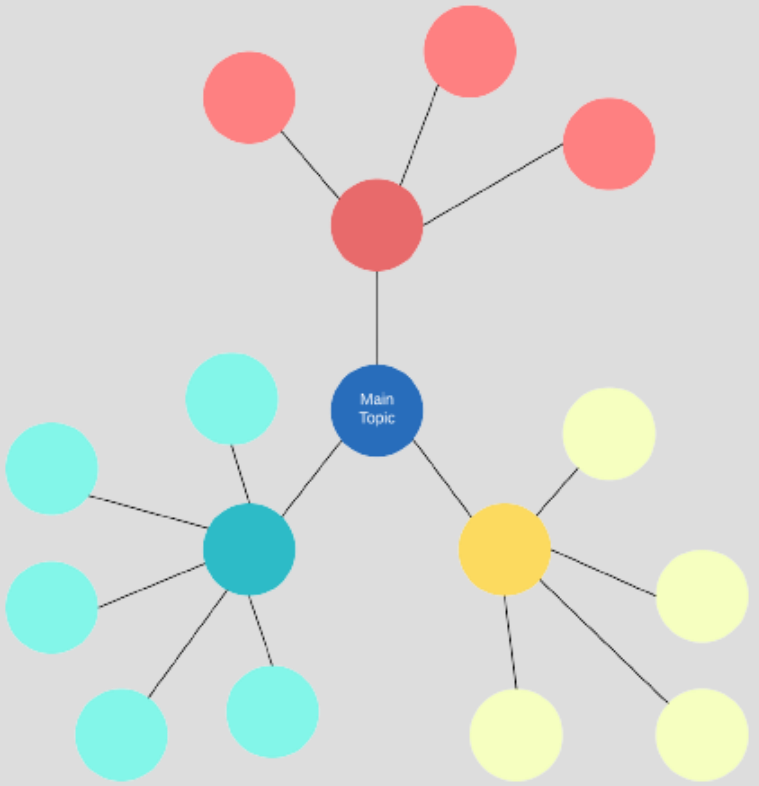
Москва, 2021

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Что такое кластеризация?

Кластер — группа элементов, характеризуемых общим свойством, главная цель кластерного анализа — нахождение групп схожих объектов в выборке.

Кластерный анализ — задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.



В чем отличие кластеризации от классификации?

Кластеризация

- Неконтролируемое обучение
- Обучение без учителя
- Метки класса обучающего множества неизвестны
- Дано множество данных с целью установления существования классов или кластеров данных

Классификация

- Контролируемое обучение
- Обучение с учителем
- Обучающее множество сопровождается меткой, указывающей класс, к которому относится наблюдение
- Новые данные классифицируются на основании обучающего множества

ЦЕЛИ КЛАСТЕРИЗАЦИИ

Три основных направления кластеризации

Понимание данных

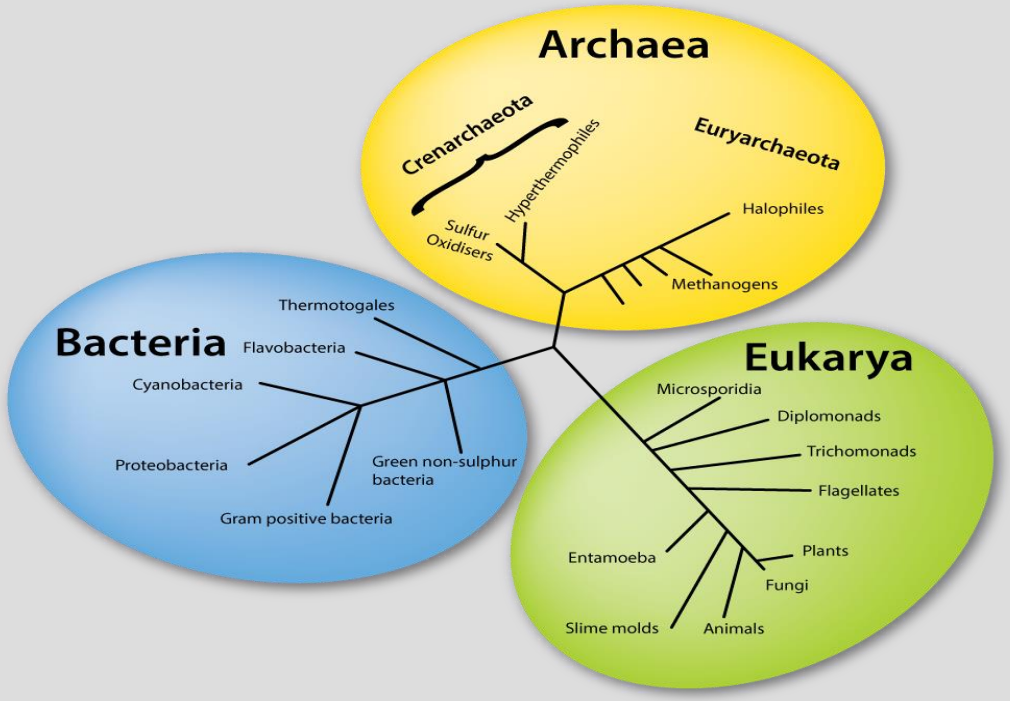
Предполагает разбиение выборки на группы схожих объектов, что позволяет упростить дальнейшую обработку данных и осуществлять принятие решений.

Обнаружение новизны

Нацелено на обнаружение новых или редких объектов в некоторых выборках.

Сжатие данных

Служит для уменьшения объемов данных с минимальными потерями информации.



ЗАДАЧИ КЛАСТЕРИЗАЦИИ

Задачи кластерного анализа

- Разработка типологии или классификации.
- Исследование полезных концептуальных схем группирования объектов.
- Порождение гипотез на основе исследования данных.
- Проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

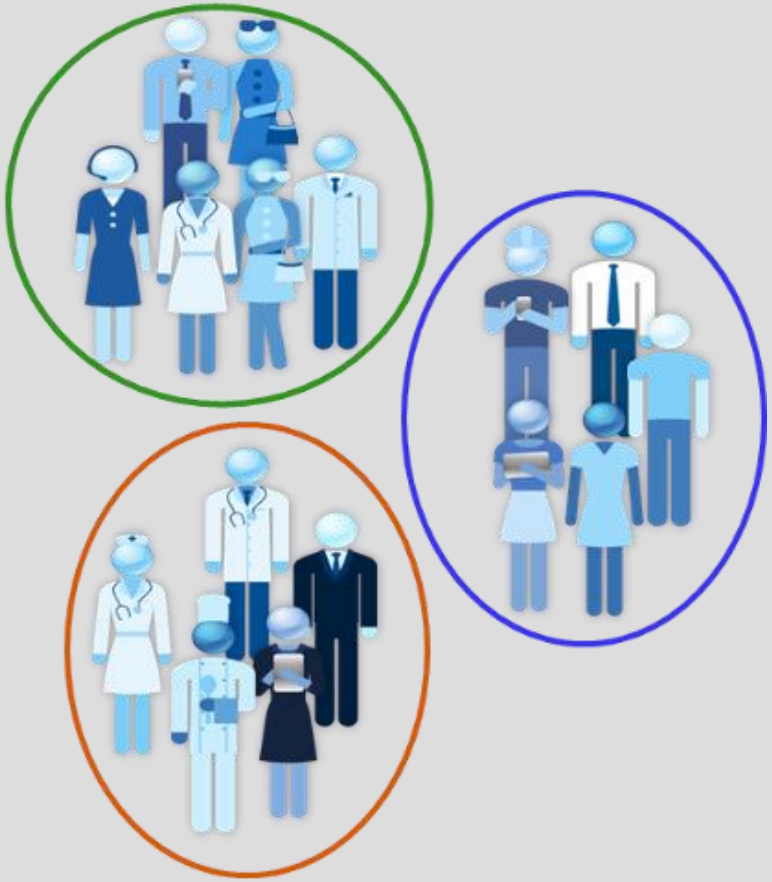
Формальная постановка задачи кластеризации описывается следующим образом:

«Требуется разбить множество X на непересекающиеся подмножества – кластеры – так, чтобы каждый кластер состоял из объектов схожих между собой, а объекты различных кластеров существенно отличались по своим свойствам.»

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА

Области в которых используется кластерный анализ

- Биология
- Биоинформатика
- Медицина
- Информатика
- Экономика
- Маркетинг
- Лингвистика
- Астрономия



ПРЕИМУЩЕСТВА КЛАСТЕРИЗАЦИИ

Сравнение с другими методами классификации данных

- Он позволяет производить разбиение объектов не по одному, а по целому набору признаков. Причем, влияние каждого из параметров может быть достаточно просто усилено или ослаблено путем внесения в математические формулы соответствующих коэффициентов.
- Кластерный анализ не накладывает ограничений на вид группируемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы.
- Многие алгоритмы кластеризации способны самостоятельно определить число кластеров, на которое следует разбить данные, а так же выделить характеристики этих кластеров без участия эксперта только при помощи используемого алгоритма.

ЭТАПЫ КЛАСТЕРИЗАЦИИ

Порядок действий при кластеризации

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:

1. Отбор выборки для кластеризации.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке.
3. Вычисление значений той или иной меры сходства между объектами.
4. Применение метода кластерного анализа для создания групп сходных объектов.
5. Проверка достоверности результатов кластерного решения.

ТРЕБОВАНИЯ ПРИ ПРОВЕДЕНИИ КЛАСТЕРИЗАЦИИ

Требования к данным

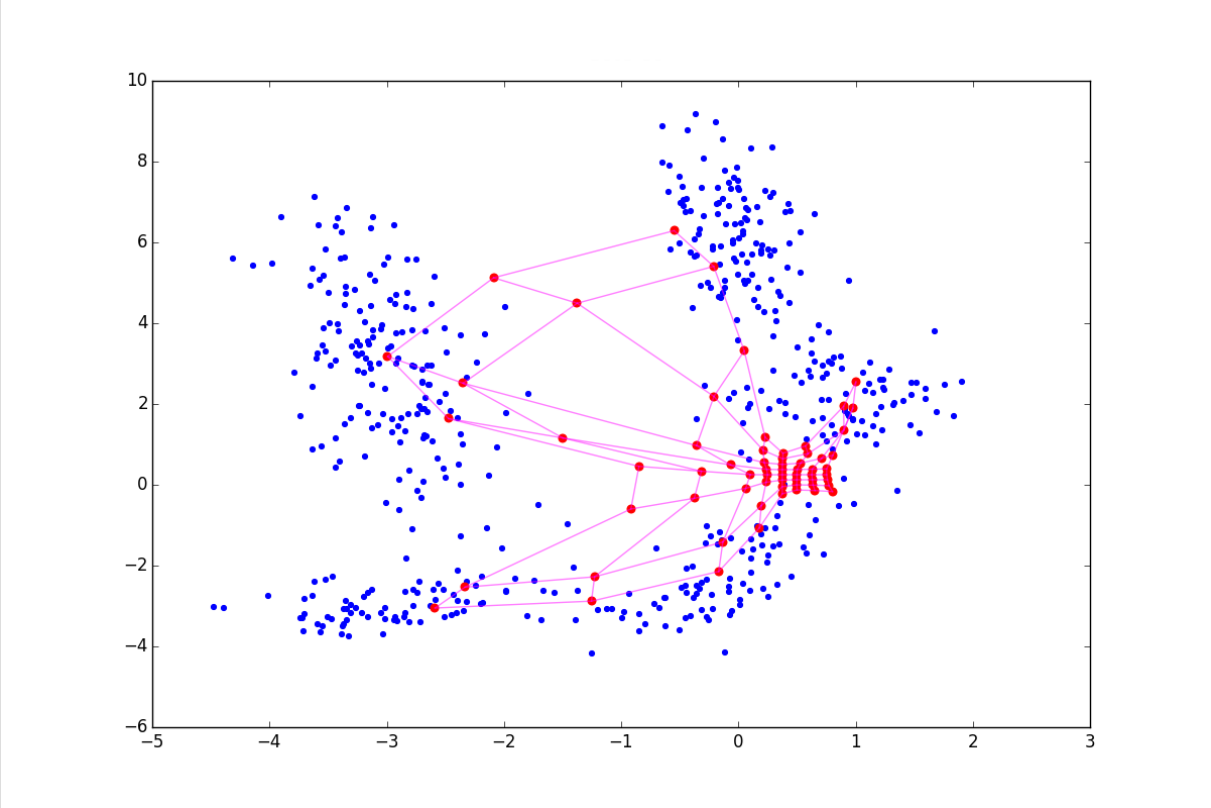
- Показатели не должны коррелировать между собой.
- Показатели должны быть безразмерными.
- Их распределение должно быть близко к нормальному.
- Показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов.
- Выборка должна быть однородна, не содержать «выбросов».

МЕРЫ СХОДСТВА

Основные меры сходства

Для вычисления расстояния между объектами используются различные меры сходства (меры подобия), называемые также метриками или функциями расстояний. Выделяют следующие меры сходства:

- Коэффициент корреляции.
- Мера расстояния.
- Коэффициенты ассоциативности.
- Вероятностные коэффициенты сходства.



Коэффициент корреляции

Коэффициент корреляции — показатель характера взаимного влияния изменения двух случайных величин.

Коэффициент корреляции обозначается латинской буквой K и может принимать значения от -1 до +1. Если значение по модулю находится ближе к 1, то это означает наличие сильной связи, а если ближе к 0 – связь отсутствует или является существенно нелинейной. При коэффициенте корреляции равном по модулю единице говорят о функциональной связи (а именно линейной зависимости), то есть изменения двух величин можно описать линейной функцией.

Мера расстояния

Мера расстояния устанавливает сходство или различие между объектами.

Для каждого типа данных существует несколько способов измерения расстояния или определения меры сходства объектов. Наиболее используемыми для интервальных данных являются:

- Евклидово расстояние (Euclidian Distance) = $\left\{\Sigma_i(x_i - y_i)^2\right\}^{1/2}$

Заметим, что евклидово расстояние (и его квадрат) вычисляется по исходным, а не по стандартизованным данным. Это обычный способ его вычисления, который имеет определенные преимущества (например, расстояние между двумя объектами не изменяется при введении в анализ нового объекта, который может оказаться выбросом).

- Квадрат Евклидова расстояния (Squared Euclidian distance) = $\{\Sigma_i(x_i - y_i)^2\}$

Иногда может возникнуть желание возвести в квадрат стандартное евклидово расстояние, чтобы придать большие веса более отдаленным друг от друга объектам. Это расстояние вычисляется следующим образом

Мера расстояния

- Расстояние городских кварталов (манхэттенское расстояние) = $\boldsymbol{\Sigma_i |x_i - y_i|}$

Это расстояние является просто средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако отметим, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

- Расстояние Чебышева = $\boldsymbol{max |x_i - y_i|}$

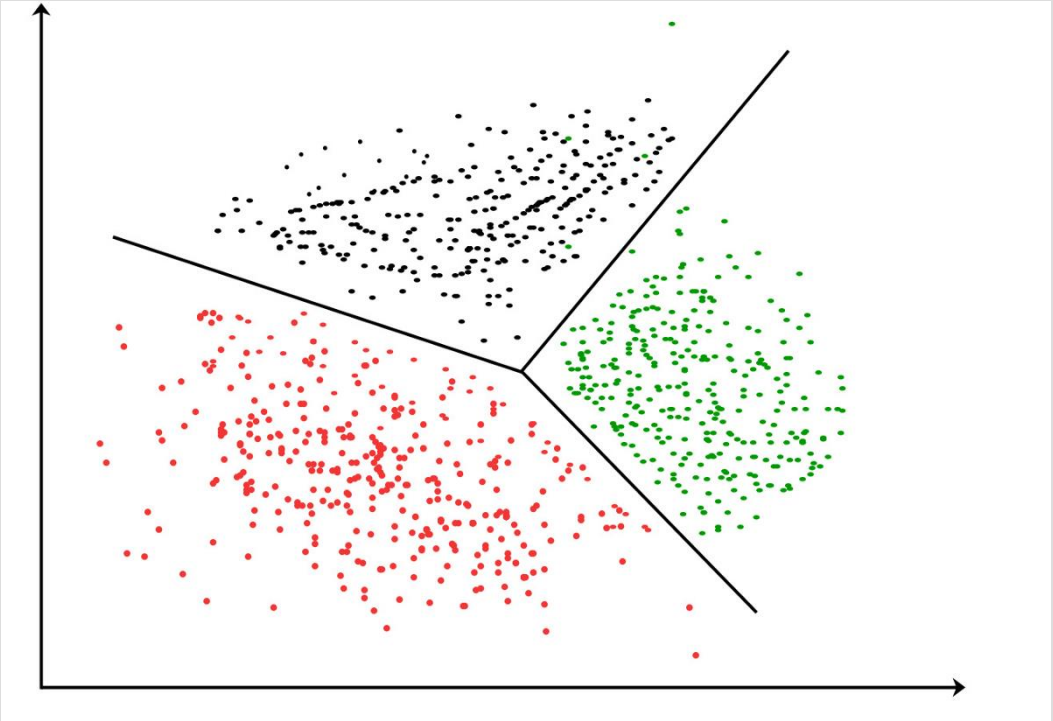
Это расстояние может оказаться полезным, когда желают определить два объекта как "различные", если они различаются по какой-либо одной координате (каким-либо одним измерением).

Коэффициенты ассоциативности

Применяются, когда необходимо установить сходство между объектами, описываемыми бинарными переменными, причем 1 указывает на наличие переменной, а 0 – на ее отсутствие.

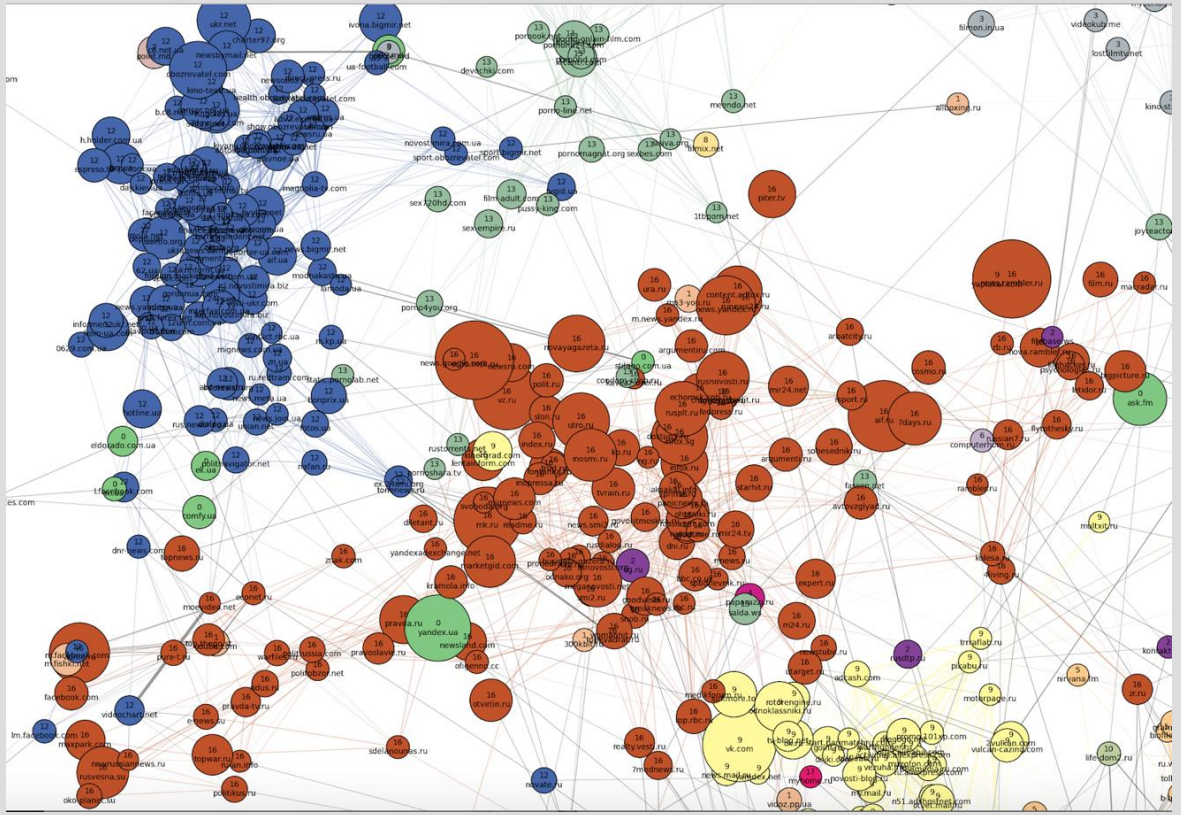
Простой коэффициент ассоциативности имеет вид: $S = \frac{(a+d)}{(a+b+c+d)}$

Коэффициент Жаккара: $S = \frac{a}{(a+b+c)}$



Вероятностные коэффициенты сходства

Вероятностные коэффициенты сходства – при образовании кластеров по этим мерам вычисляется информационный выигрыш от объединения двух объектов, а затем объекты с минимальным выигрышем рассматриваются как один



КЛАССИФИКАЦИЯ КЛАСТЕРНЫХ МЕТОДОВ

Методы кластеризации

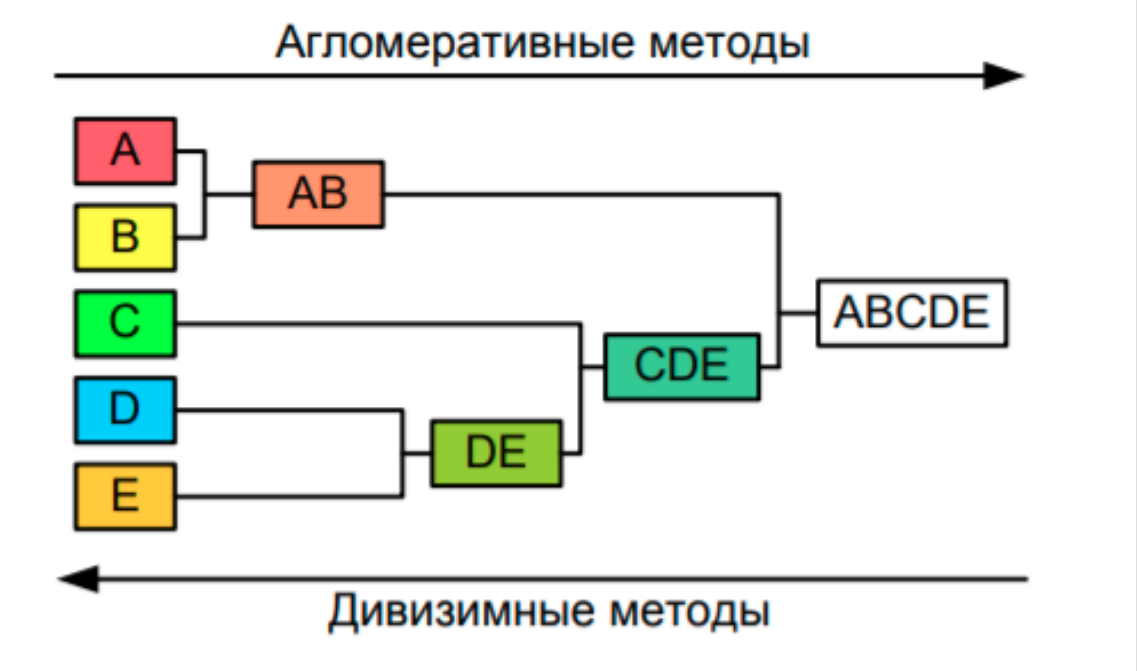
Единой системы классификации кластерных процедур на сегодняшний день *не существует*. Зачастую такие системы создаются отдельно в каждой отрасли применения кластерного анализа.

На рисунке приведена *обобщенная классификация кластерных методов*, характерная для большинства задач.

Иерархические методы кластеризации

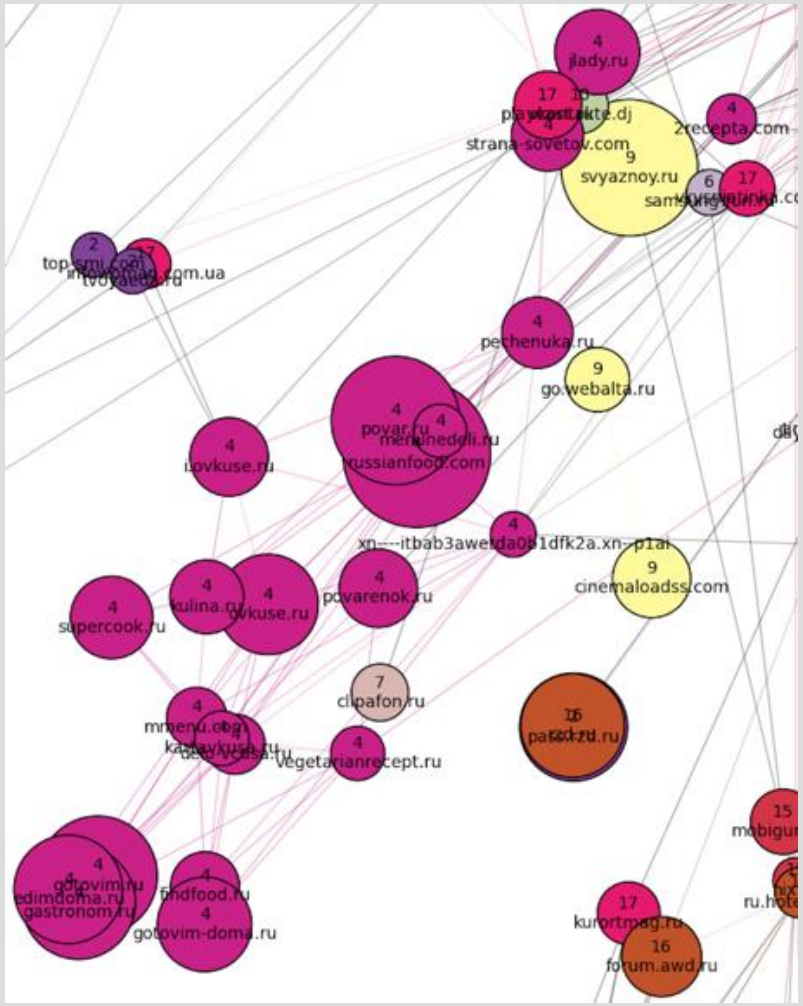
Большинство современных кластерных методов относятся к семейству ***иерархических.***

Главной отличительной особенностью таких методов является то, что процесс объединения объектов при их использовании имеет иерархический характер и может быть представлен *виде* ***дендрограммы*** (древовидной диаграммы), где каждый уровень соответствует одному шагу алгоритма. При этом на каждом шаге количество кластеров изменяется в сторону увеличения или уменьшения.



Неиерархические методы кластеризации

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют ***неиерархические метод***ы, основанные на разделении, которые представляют собой ***итеративные методы дробления исходной совокупности***. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

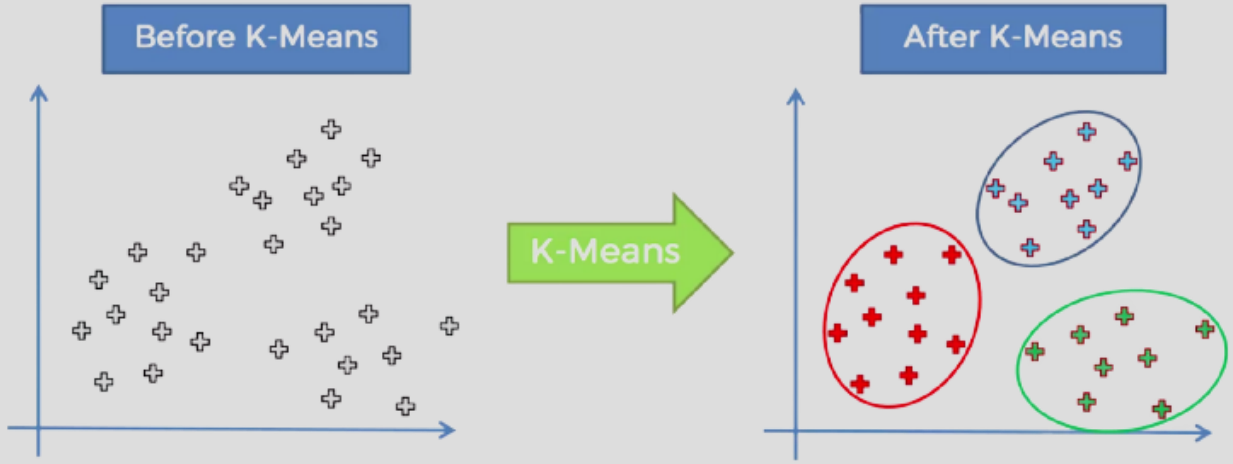


МЕТОД КЛАСТЕРИЗАЦИИ K-MEANS

Основная идея метода

Метод k-means так же известен как метод k-средних, hard-c-means.

Основная идея заключается в том, что на каждой итерации перечисляется **центр масс** для каждого кластера, полученного на предыдущем шаге, затем объекты снова разбиваются на кластеры в соответствии с тем, какой из новых центров оказался **ближе по выбранной метрике**. Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния.



Базовые определения алгоритма k-means

Данный алгоритм является прообразом практически всех алгоритмов нечеткой кластеризации, и его формализация поможет лучшему понимаю принципов, заложенных в более сложные алгоритмы.

Базовые определения и понятия в рамках данного алгоритма имеют вид:

- обучающее множество $M = \{m_j\}_{j=1}^{d}$, d— количество точек (векторов) данных;
- О вектор центров кластеров $C = \{c^{(i)}\}_{i=1}^{c}$, где $c^{(i)} = \frac{\sum_{j=1}^{d} (u_{ij})^w \cdot m_j}{\sum_{j=1}^{d} (u_{ij})^w}$, $1 \le i \le c$
- матрица разбиения $U = \{u_{ij}\}$, где $u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_A^2(m_J, c^{(i)})}{d_A^2(m_J, c^{(k)})} \right)^{\frac{1}{w-1}}}$

Базовые определения алгоритма k-means

- целевая функция: $J(M,U,C) = \sum_{i=1}^{c} \cdot \sum_{j=1}^{d} u_{ij}^{w} d_A^2\left(m_j, c^{(i)}\right)$,

где $w \in (1,\infty)$ — показатель нечеткости (взвешивающий коэффициент), регулирующий нечеткость разбиения. Обычно используется $w = 2$;

- Набор ограничений: $u_{ij} \in [0,1]; \sum_{i=1}^{c} u_{ij} = 1; 0 < \sum_{j=1}^{d} u_{ij} < d$,

который определяет, что каждый вектор данных может принадлежать различным кластерам с разной степенью принадлежности, сумма принадлежностей элемента данных всем кластерам пространства разбиения равна единице.

Общее описание алгоритма k-means

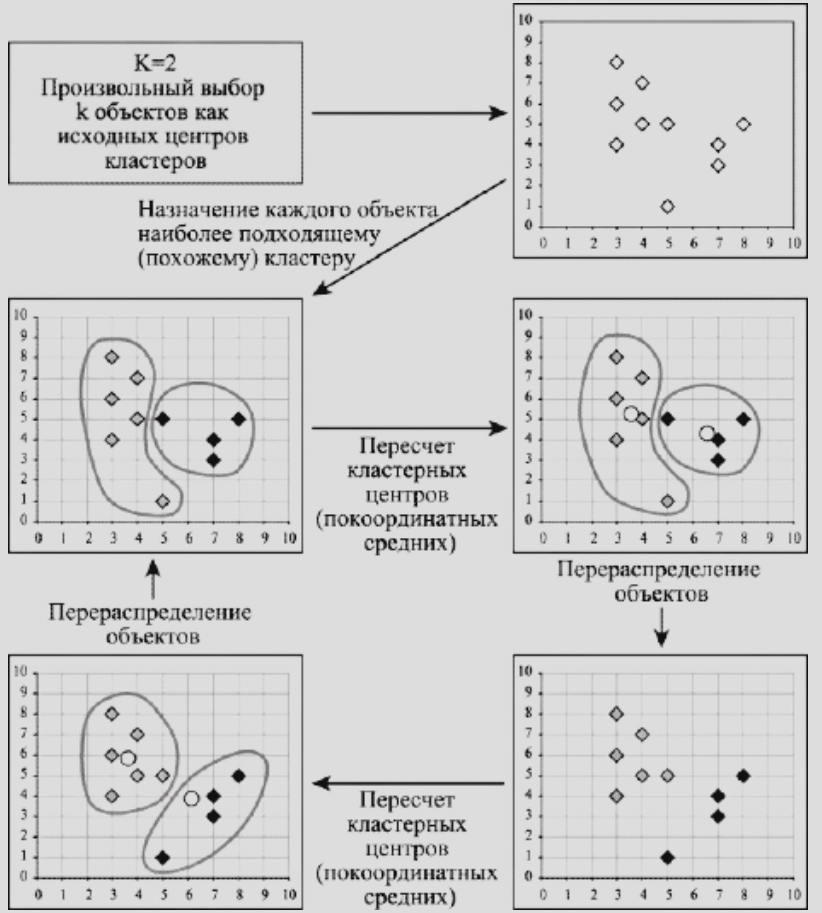
- Шаг 1. Проинициализировать начальное разбиение, выбрать точность, при достижении которой алгоритм завершится.
- Шаг 2. Определить центры кластеров:

$$C_j^{(i)} = \frac{\sum_{j=1}^{d} u_{ij}^{(l-1)} \cdot m_j}{\sum_{j=1}^{d} u_{ij}^{(l-1)}}, 1 \le \mathrm{i} \le C$$

- Шаг 3. Обновить разбиение.
- Шаг 4. Проверить условие завершения алгоритма:

$$\left\| U^{(l)} - U^{(l-1)} \right\| < \delta$$

Если не выполняется, то перейти к шагу 2.



Пример работы алгоритма

Рассмотрим алгоритм на примере набора данных, описывающих ирисы разных классов:

№	Длина чашелистника	Ширина чашелистника	Длина лепестка	Ширина лепестка	Класс
1	5,1	3,5	1,4	0,2	Iris setosa
2	4,9	3,0	1,4	0,2	Iris setosa
3	4,7	3,2	1,3	0,2	Iris setosa
4	4,6	3,1	1,5	0,2	Iris setosa
5	5,0	3,6	1,4	0,2	Iris setosa
51	7,0	3,2	4,7	1,4	Iris versicolor
52	6,4	3,2	4,5	1,5	Iris versicolor
53	6,9	3,1	4,9	1,5	Iris versicolor
54	5,5	2,3	4,0	1,3	Iris versicolor
55	6,5	2,8	4,6	1,5	Iris versicolor
101	6,3	3,3	6,0	2,5	Iris virginica
102	5,8	2,7	5,1	1,9	Iris virginica
103	7,1	3,0	5,9	2,1	Iris virginica
104	6,3	2,9	5,6	1,8	Iris virginica
105	6,5	3,0	5,8	2,2	Iris virginica

Будем делать вывод о принадлежности классу на основании длины b ширины чашелистника.

- Выберем k произвольных центров классов из исходного множества.
- Это могут быть любые элементы множества. Пусть центрами кластеров будут точки 1, 2 и 3. В качестве метрики близости выберем евклидово расстояние.
- Набор исходных данных в графическом представлении и разбиение на кластеры после первой итерации:

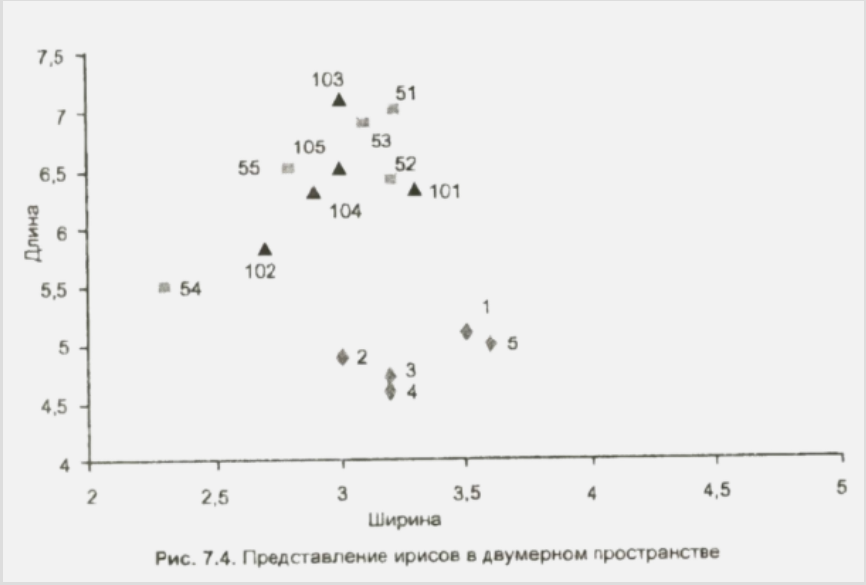


Рис. 7.4. Представление ирисов в двумерном пространстве

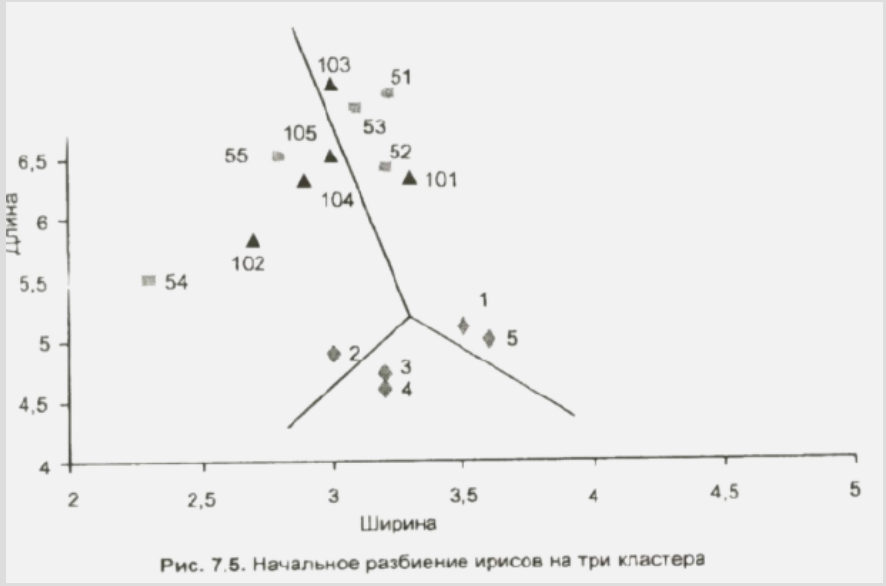


Рис. 7.5. Начальное разбиение ирисов на три кластера

Далее необходимо найти новые центры кластеров. Новый центр для каждого кластера – точка, координаты которой представляют собой среднее арифметическое для всех элементов в данном кластере. После этого необходимо повторить процедуру разбиения на кластеры.

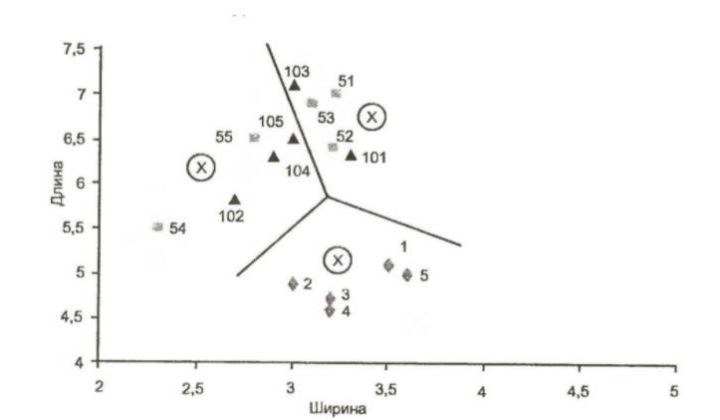


Рис. 7.6. Второе разбиение множества ирисов на кластеры

№	Длина чашелистника	Ширина чашелистника	Длина лепестка	Ширина лепестка	Класс
1	5,1	3,5	1,4	0,2	Iris setosa
2	4,9	3,0	1,4	0,2	Iris setosa
3	4,7	3,2	1,3	0,2	Iris setosa
4	4,6	3,1	1,5	0,2	Iris setosa
5	5,0	3,6	1,4	0,2	Iris setosa
51	7,0	3,2	4,7	1,4	Iris versicolor
52	6,4	3,2	4,5	1,5	Iris versicolor
53	6,9	3,1	4,9	1,5	Iris versicolor
54	5,5	2,3	4,0	1,3	Iris versicolor
55	6,5	2,8	4,6	1,5	Iris versicolor
101	6,3	3,3	6,0	2,5	Iris virginica
102	5,8	2,7	5,1	1,9	Iris virginica
103	7,1	3,0	5,9	2,1	Iris virginica
104	6,3	2,9	5,6	1,8	Iris virginica
105	6,5	3,0	5,8	2,2	Iris virginica

Второе разбиение множества ирисов на кластеры

После определятся новые центры кластеров. Процедура распределения по кластерам и переопределения центров кластеров будут повторяться до тех пор, пока значения центров перестанет меняться. Тогда можно будет сделать вывод, что кластеры внутри исходного определены.

Недостатки и достоинства метода

Достоинства

- Отлично работает, если данные по своей природе делятся на компактные, примерно сферические группы
- Простота использования
- Быстрота использования
- Понятность и прозрачность алгоритма

Недостатки

- Слишком чувствителен к выбросам, которые могут искажать среднее
- Медленная работа на больших базах данных
- Необходимо задавать количество кластеров

МЕТОД КЛАСТЕРИЗАЦИИ FUZZY C-MEANS

Базовые определения алгоритма soft k-means

Soft k-means, метод нечеткой кластеризации с-средних.

Данный алгоритм – обобщение алгоритма k-means, рассмотренного ранее, но в отличие от алгоритма k-средних кластеры здесь ***представлены в виде нечетких множеств***, каждая точка принадлежит различны кластерам с различной степенью принадлежности.

Точка считается принадлежащей данному кластеру по критерию максимума принадлежности данному кластеру (каждый из объектов не входит однозначно в какой-либо кластер, а принадлежит всем кластерам с различными степенями принадлежности).

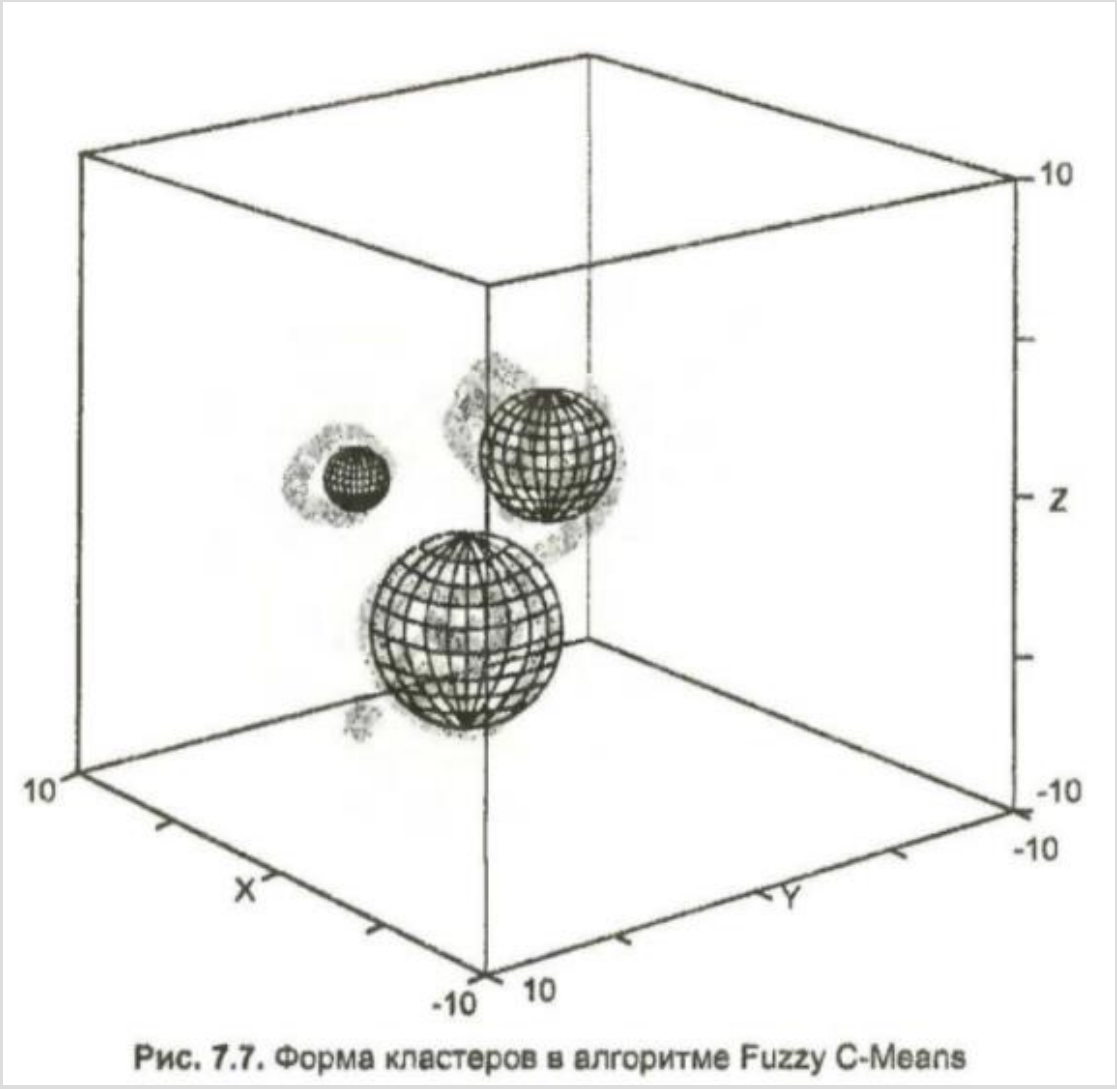
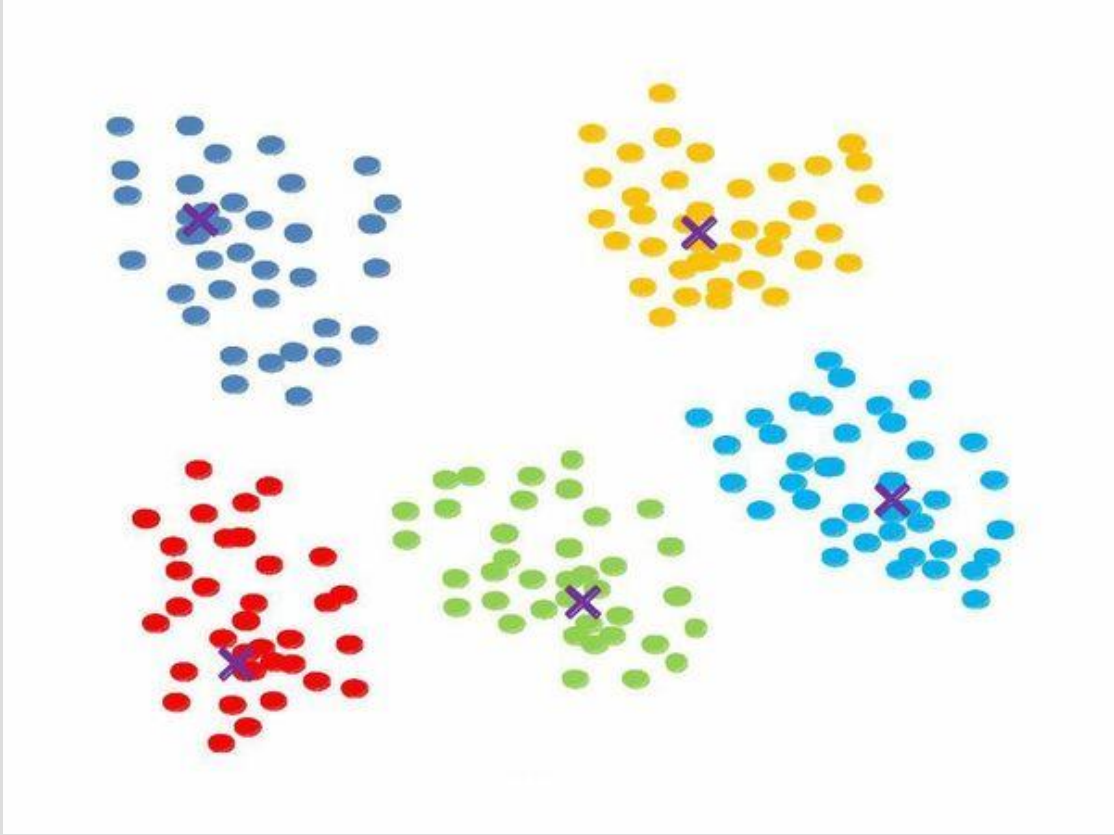


Рис. 7.7. Форма кластеров в алгоритме Fuzzy C-Means

Общее описание алгоритма

Алгоритм включает в себя три основных этапа:

1. вычисление центров кластеров,
2. вычисление расстояний между центрами кластеров и точками данных (включающее в себя макрооперации вычитания векторов и вычисления их норм)
3. пересчёт матрицы принадлежности.



Общее описание алгоритма

Шаг 1. Выбрать количество кластеров 2 ≤ c≤ d

Шаг 2. Выбрать скалярную метрику для отображения данных.

Шаг 3. Выбрать параметр остановки.

Шаг 4. Выбрать коэффициент нечеткости w ∈ (1,∞).

Шаг 5. Проинициализировать начальное разбиение.

Шаг 6. Вычислить центры кластеров:

$$C_l^{(i)} = \frac{\sum_{j=1}^{d} \left(u_{ij}^{(l-1)}\right)^n \cdot m_j}{\sum_{j=1}^{d} \left(n_{ij}^{(l-1)}\right)^w}, 1 \le \mathrm{i} \le c$$

Шаг 7. Для всех элементов вычислить квадраты расстояний до всех центров кластеров:

$$d_A^2\left(m_j, c_l^{(i)}\right) = (c_l^{(i)} - m_j)^t A \left(c_l^{(i)} - m_j\right)$$

Шаг 8. Обновить разбиение.

Шаг 9. Проверить условие завершения алгоритма (параметр остановки). Если не выполняется, то прейти к шагу 7.

Недостатки и достоинства метода

Достоинства

- лучше сходимость (по сравнению с алгоритмом k-средних)
- простота реализации
- интуитивная понятность

Недостатки

- вычислительная сложность
- необходимо знать количество кластеров
- чувствительность к начальному разбиению

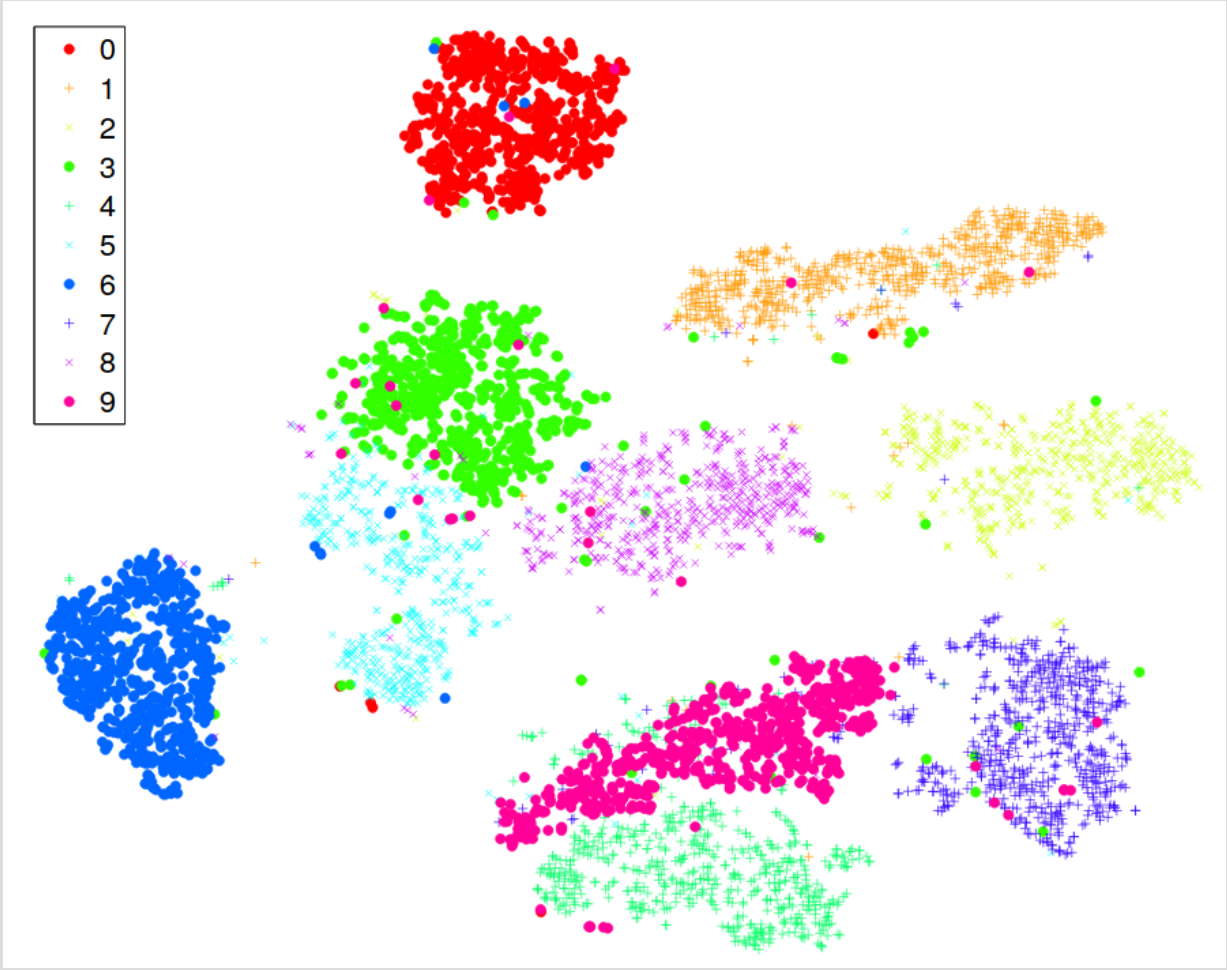
АЛГОРИТМ МАШИННОГО ОБУЧЕНИЯ T-SNE

Базовые определения алгоритма t-SNE

t-SNE (t-distributed stochastic neighbor embedding, стохастическое вложение соседей с распределением Стьюдента) - алгоритм уменьшения размерности.

Проще говоря, t-SNE дает вам представление о том, как данные расположены в многомерном пространстве.

Разработанный Лоренсом ван дер Маатеном и Джеффри Хинтоном в 2008 году, он был успешно применен ко многим реальным наборам данных.



Общее описание алгоритма t-SNE

Алгоритм t-SNE вычисляет меру сходства между парами экземпляров в пространстве с высокой размерностью и в пространстве с низкой размерностью. Затем он пытается оптимизировать эти два показателя сходства. Давайте разберем это на 3 основных шага.

Шаг 1 Для всех точек рассчитывается многомерное евклидово расстояние

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\,\sigma_i^2\right)}{\Sigma_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

Эта формула показывает близость точек x_i и x_j при гауссовом распределении вокруг x_i с заданным отклонением σ, вычисляемым для каждой точки отдельно таким образом, чтобы точки в областях с большей плотностью имели меньшую дисперсию.

Для этого используется оценка перплексии:

$$p_{erp}(p_i) = 2^{H(p_i)} = 2^{-\sum p_{j|i} \log_2(pj|i)}$$

где $H(p_i)$ — энтропия по Шеннону. На практике перплексия задается в качестве параметра метода.

Общее описание алгоритма t-SNE

Шаг 2 Для двумерных или трехмерных соседей пары x_i и x_j , назовем их y_i и y_j , не представляет труда оценить условную вероятность, приняв стандартное отклонение равным $\frac{1}{\sqrt{2}}$:

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\Sigma_{k \neq i} \exp(-\|y_i - y_k\|^2)}$$

Если точки отображения y_i и y_j корректно моделируют сходство между исходными точками высокой размерности x_i и x_j , то соответствующие условные вероятности $p_{j|i}$ и $q_{j|i}$ будут эквивалентны.

Шаг 3 В качестве оценки качества в классическом SNE используется расстояние Кульбака-Лейблера. SNE минимизирует сумму таких расстояний для всех точек отображения при помощи градиентного спуска.

Общее описание алгоритма t-SNE

Шаг 3 При реализации метода t-SNE в качестве альтернативы минимизации суммы дивергенций Кульбака-Лейблера между условными вероятностями $p_{j|i}$ и $q_{j|i}$ минимизирует одиночную дивергенцию между совместной вероятностью P в многомерном пространстве и совместной вероятностью Q в пространстве отображения:

$$C_{ost} = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

где p_{ij} и q_{ij} = 0, $p_{ij} = p_{ji}$, $q_{ij} = q_{ji}$ для любых i и j,

а p_{ij} определяется по формуле:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n},$$

где n — количество точек в наборе данных

Градиент для симметричного SNE получается существенно проще, чем для классического:

$$\frac{\partial C_{ost}}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)$$

Общее описание алгоритма t-SNE

Для того, чтобы избежать скученности точек, в t-SNE используется t-распределение с одной степенью свободы. Совместная вероятность для пространства отображения в этом случае будет определяться следующей формулой:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\Sigma_{k=1}(1 + \|y_k - y_t\|^2)^{-1}}$$

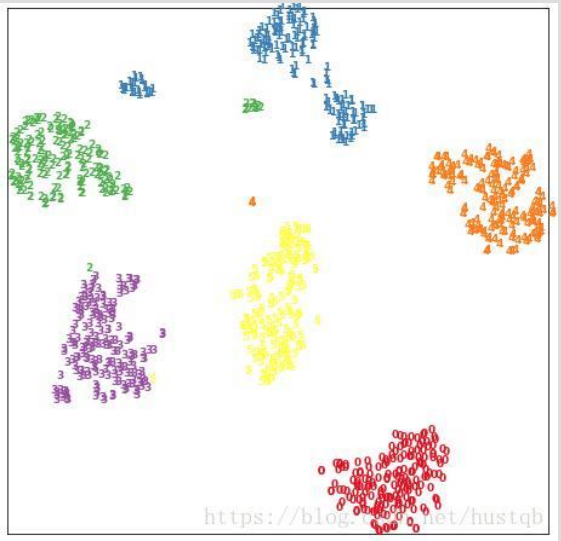
А соответствующий градиент:

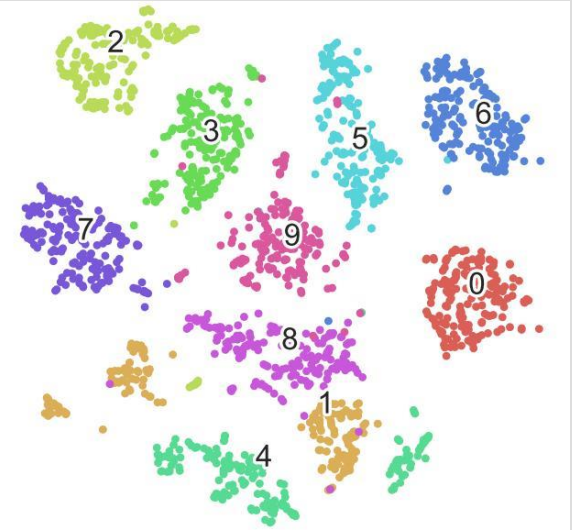
$$\frac{\partial Cost}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1}$$

Пример реализации алгоритма t-SNE на python 3

T-SNE - это алгоритм уменьшения размерности данных. Основанием для его создания является предположение, что, хотя многие наборы данных в реальном мире встроены в многомерное пространство, все они имеют очень низкие внутренние измерения.

Другими словами, после уменьшения размерности, многомерные данные могут показать свои **существенные** характеристики в низкоразмерном состоянии. Эта основная идея также известна как **Нелинейное уменьшение размерности.**





Выводы

Кластеризация и ее методы активно используются в современном мире в совершенно разных предметных областях, что позволяет более эффективно достигать поставленные цели.

Достоинства и недостатки метода определяют его спектр задач, выполняемых в определенных условиях. Несмотря на это, есть концепции при которых использование кластеризации невозможно или несущественно, но при должном развитии технологий и повышении уровня запросов в предметных областях, решение задач кластеризации имеет все шансы стать более приоритетной методологией в сравнении с другими вариантами в вопросах статистического анализа или при использовании машинного обучения.