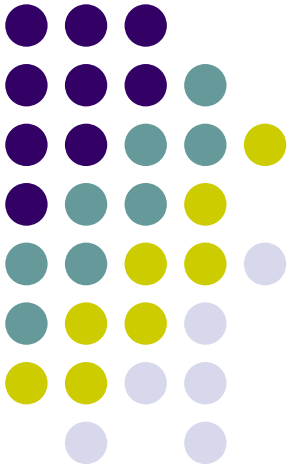
Кластер-анализ

Корлякова М.О.

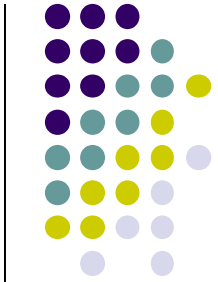
2018



Обучение

- Без учителя (выделение классов)
- С учителем (отнесение к классу)

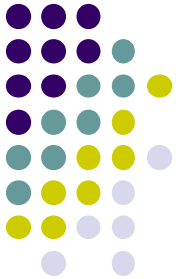
Задача обобщения



- n – число координат
- Dn – пространство в координатах <x1,x2, …,xn>
- xi – номинальный, дискретный упорядоче. или непрерывн.

- X – объект из Dn
- X = (a1, a2, …, an) , ai – значение xi для X

Задача обобщения



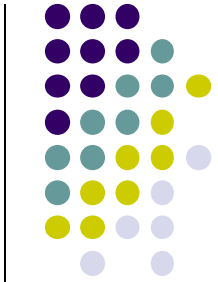
- O – множество объектов известных в Dn
- V – общее множество объектов в Dn

Задача обобщения

- Необходимо в Dn определить группы Ci, которые сформированы на основании близости по мере m(Xj, Ci).

- P(Xj, Ci) = 1 для m(Xj, Ci)<g
- P(Xj, Ci) = 0 для m(Xj, Ci)>g

Кластеризация



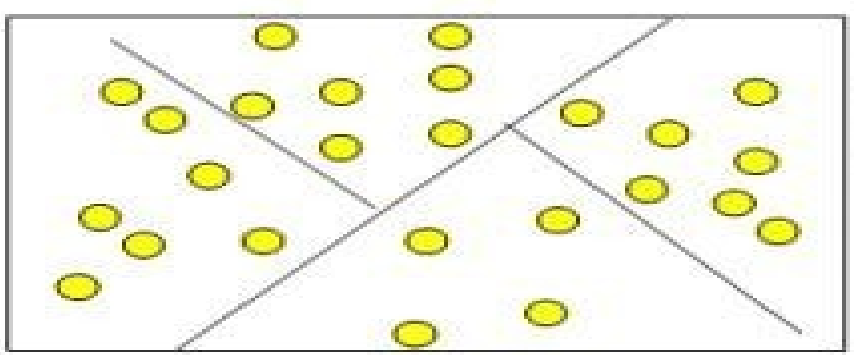
- T={(Xi)}

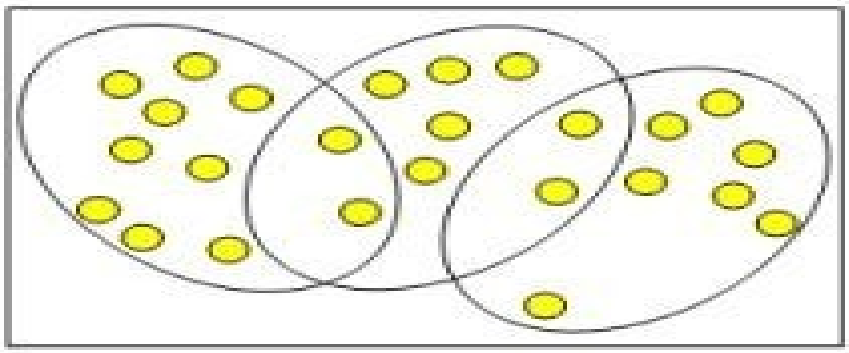
- Цель: Найти классы

Типы кластерного анализа

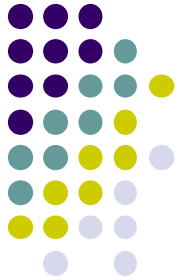
- Одноуровневый
  - Фиксация числа кластеров
  - Фиксация размеров кластера
- Иерархический

- Кластеры без пересечения и с пересечением



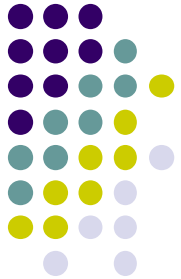


Расстояния между объектами



- Метрики : Минковский
- Меры: Хемминг
- И МНОГО ДРУГИХ МЕТОДОВ!!!!

Расстояние между множествами

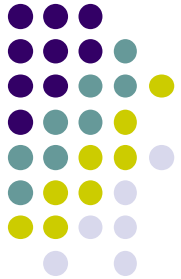


- Ближний сосед
- Средний
- Дальний сосед

- Метрика Хаусдорфа

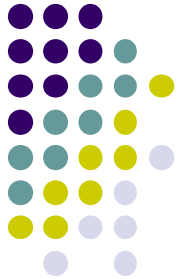
- И МНОГО ДРУГИХ МЕТОДОВ!!!!!

Методы одноуровневого анализа



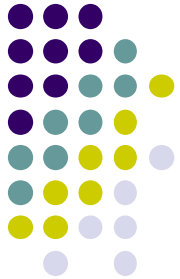
- Фиксация числа кластеров (k - средних)
  - Минимизация различий в примерах кластера
- Фиксация размера - FOREL
  - Поиски сгущений в пространстве
- Комбинации

Порядок решения задачи кластеризации



- Определить Dn – пространство признаков и T={(Xi)} – примеры
- Выбрать способ вычисления расстояния между объектами
- Определить отношение эквивалентности
- Определить тип алгоритма кластеризации.

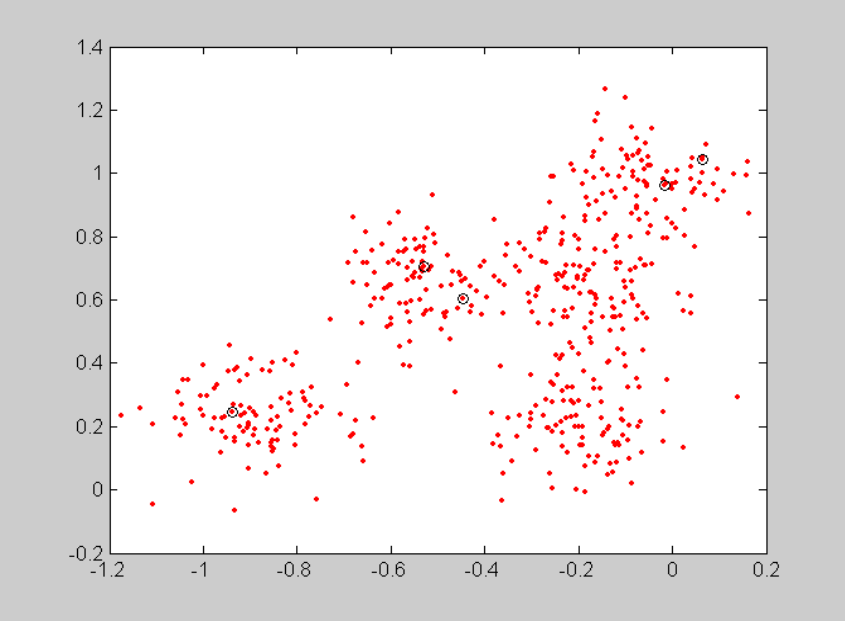
Алгоритм кластеризации по максимальному расстоянию

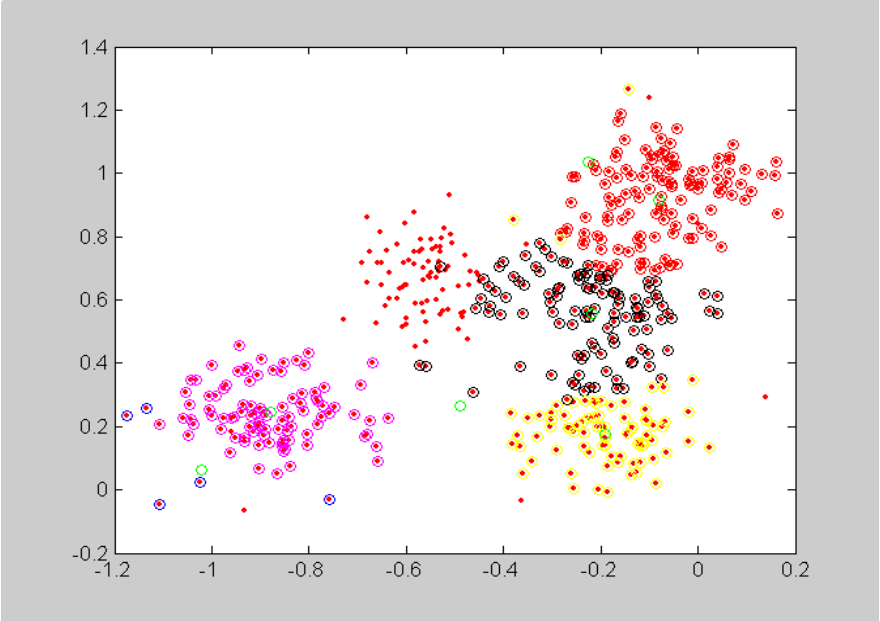


- Фиксирует размер кластера
  1. Зафиксировать диаметр кластера dmax
  2. Установить центр кластера в свободный (вне других кластеров) объект Xi из обучающей выборки.
  3. Добавить к кластеру объект Xj такой, что для всех Xi из текущего кластера.
  4. Продолжать процесс расширения кластера пока не исчерпана выборка примеров.
  5. Вычислить центр кластера.
  6. Если остались свободные примеры, то перейти к процедуре формирования кластеров (п.2.), иначе остановить процесс.

Пример

Исходное множество Результат

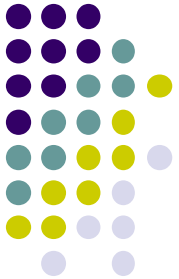




Алгоритм к-средних

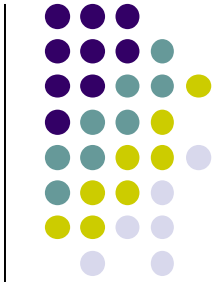
- Фиксирует число классов
  1. Номер итерации s=0
  2. Связать с каждым кластером Kj объект Xi из обучающей выборки (случайно).
  3. если число кластеров меньше N, то перейти к процедуре формирования кластеров (п.4.).
  4. Вычислить расстояния от всех объектов до всех центров кластеров.
  5. присоединить объект Xi к кластеру Ck, если $Ck = \min_{j=1.M} d(Xi, Cj)$
  6. повторить для всех объектов выборки.
  7. вычислить новое положение центров кластеров $\mathbf{centr}_b = \frac{1}{N_b} \sum_{i=1}^{N_b} \mathbf{x}_i$
  8. где Nb – число примеров множества Cb.
  9. Повторять от п.4. пока кластер смещается более чем на ε (задано пользователем), иначе остановить процесс.

Достоинства алгоритма k-средних:



- простота использования;
- быстрота использования;
- понятность и прозрачность алгоритма.

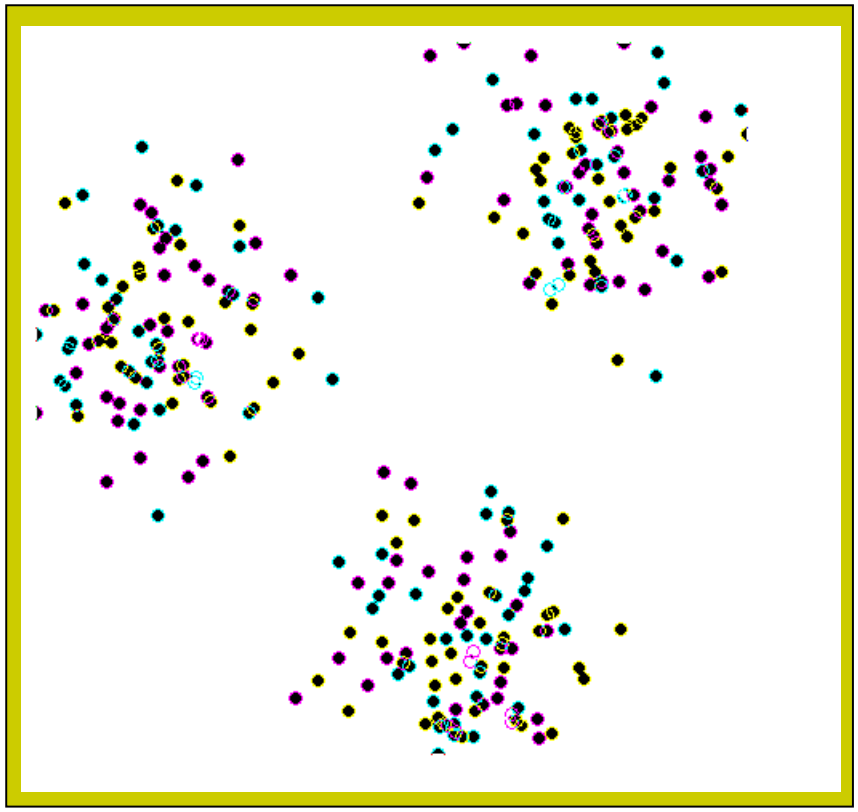
Недостатки алгоритма k-средних:

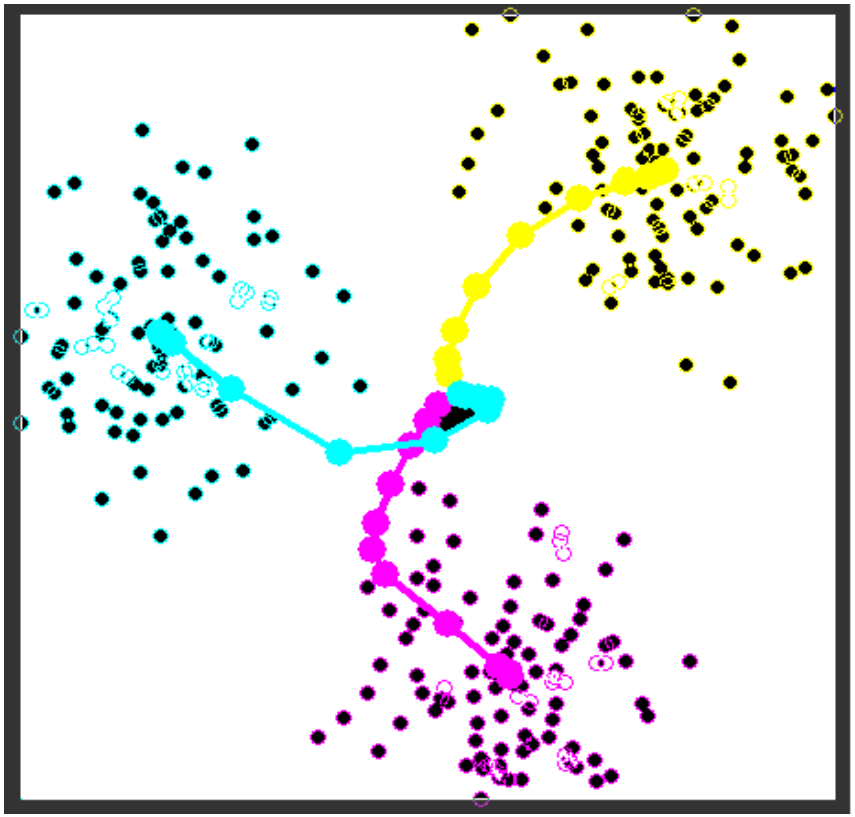


- Чувствителен к выбросам, которые могут искажать среднее;
- Может медленно работать на больших базах данных.

Пример

Исходное множество Результат



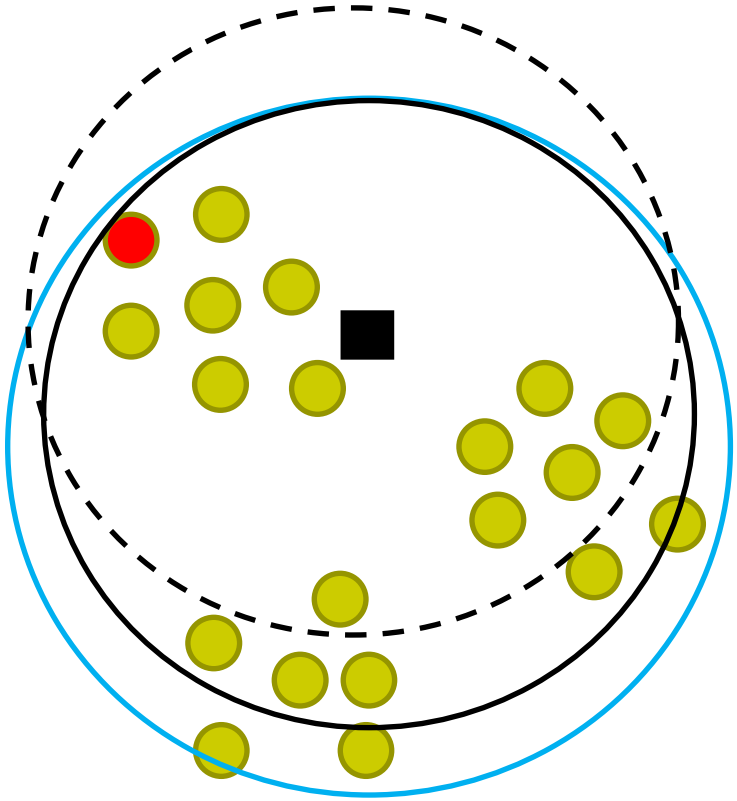


FOREL

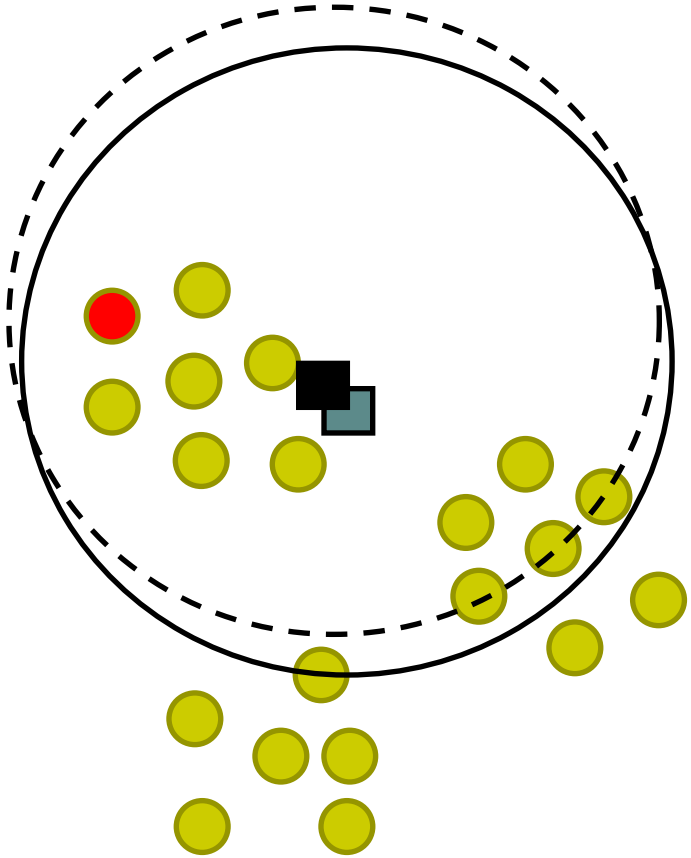
1. T={Xi}, i=1,N
2. R0=max(dist(Xi,Xj)) Xi,Xj из T – радиус кластера
3. Ck=Xi – центр кластера
4. Xj из Ck, если dist(Ck,Xj)<R0*0.9
5. Новый Ck*=1/|Ck| ∑Xj
6. Если | Ck* -Ck|>a, то к 4, иначе 7
7. Точки Ck – исключаем из T
8. Если |T|>0 , то k=k+1 и к 1, иначе конец.

Критерий min ∑dist(Ck,Xi), Xi из Ck

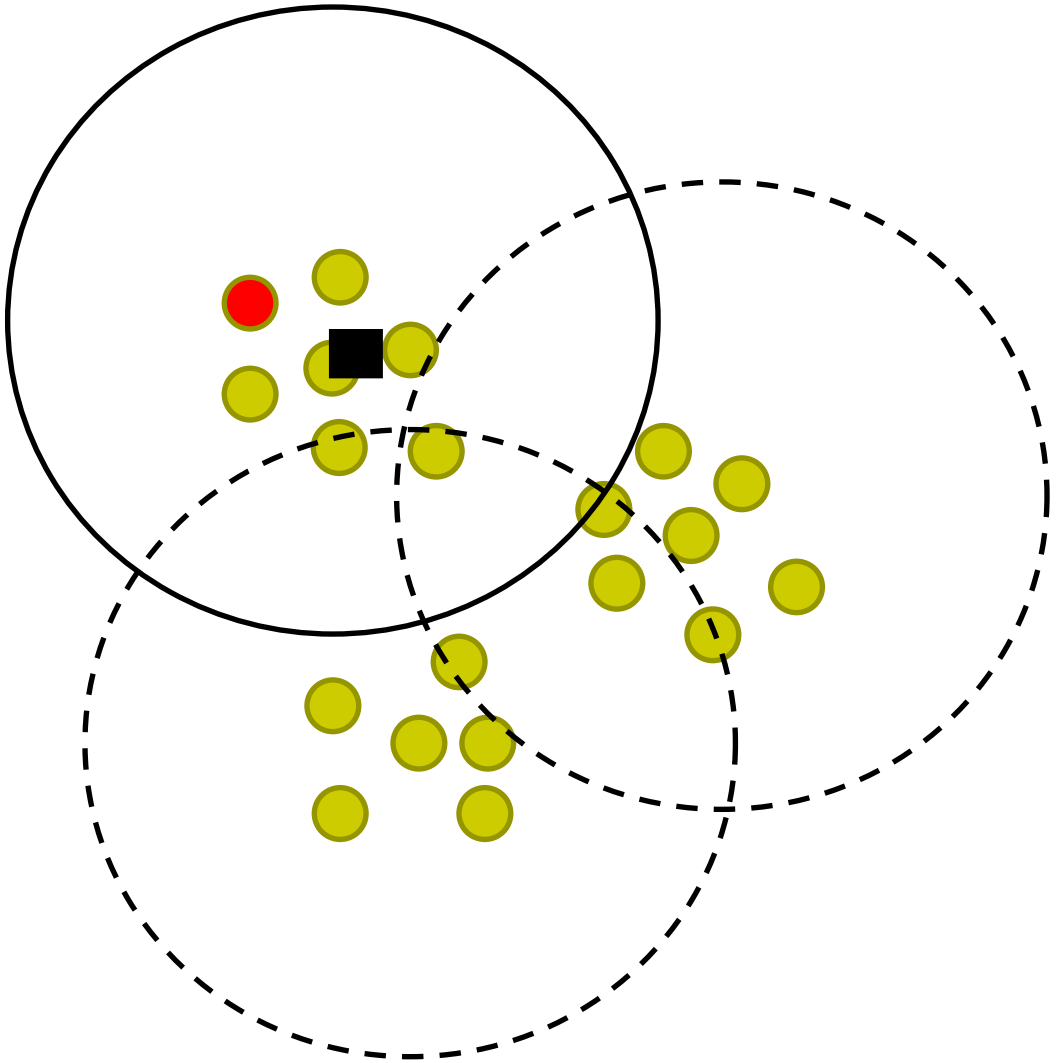
FOREL



FOREL



FOREL

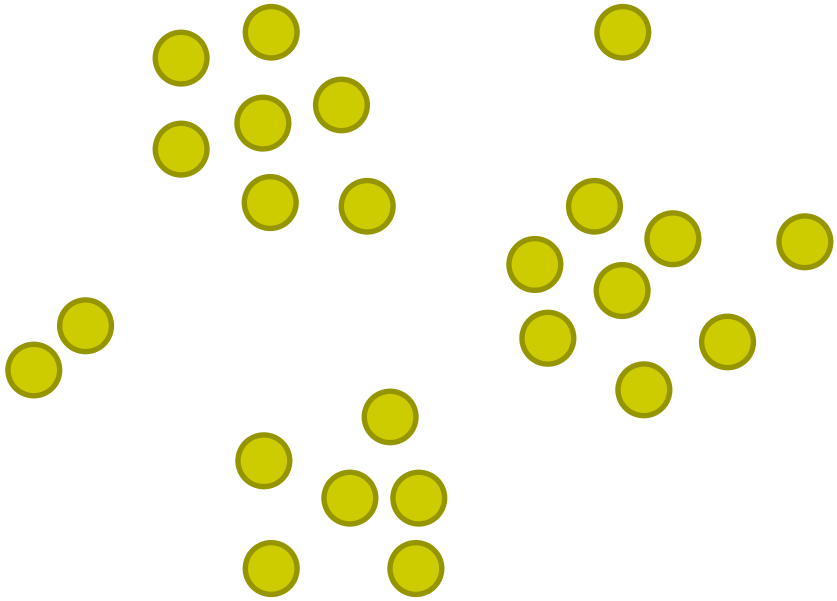


KOLAPS

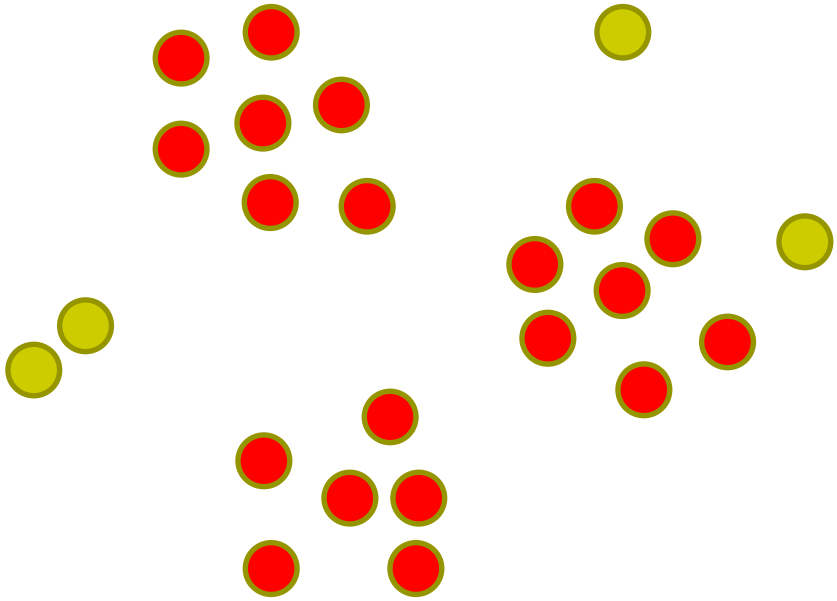
- T={Xi},i=1,N, d- мощность кластера
- Проводим генерацию по FOREL – {Ck}, k=1,m
- Находим все |Ck|>d и заносим их в список L – кандидатов
- Кластеры кандидаты Lm^{Ck*}:
  - Уменьшаем радиус от R до Rmin с шагом dR
  - Если число теряемых точек за шаг увеличилось, то остановить сжатие

Критерий max∑|Lm|

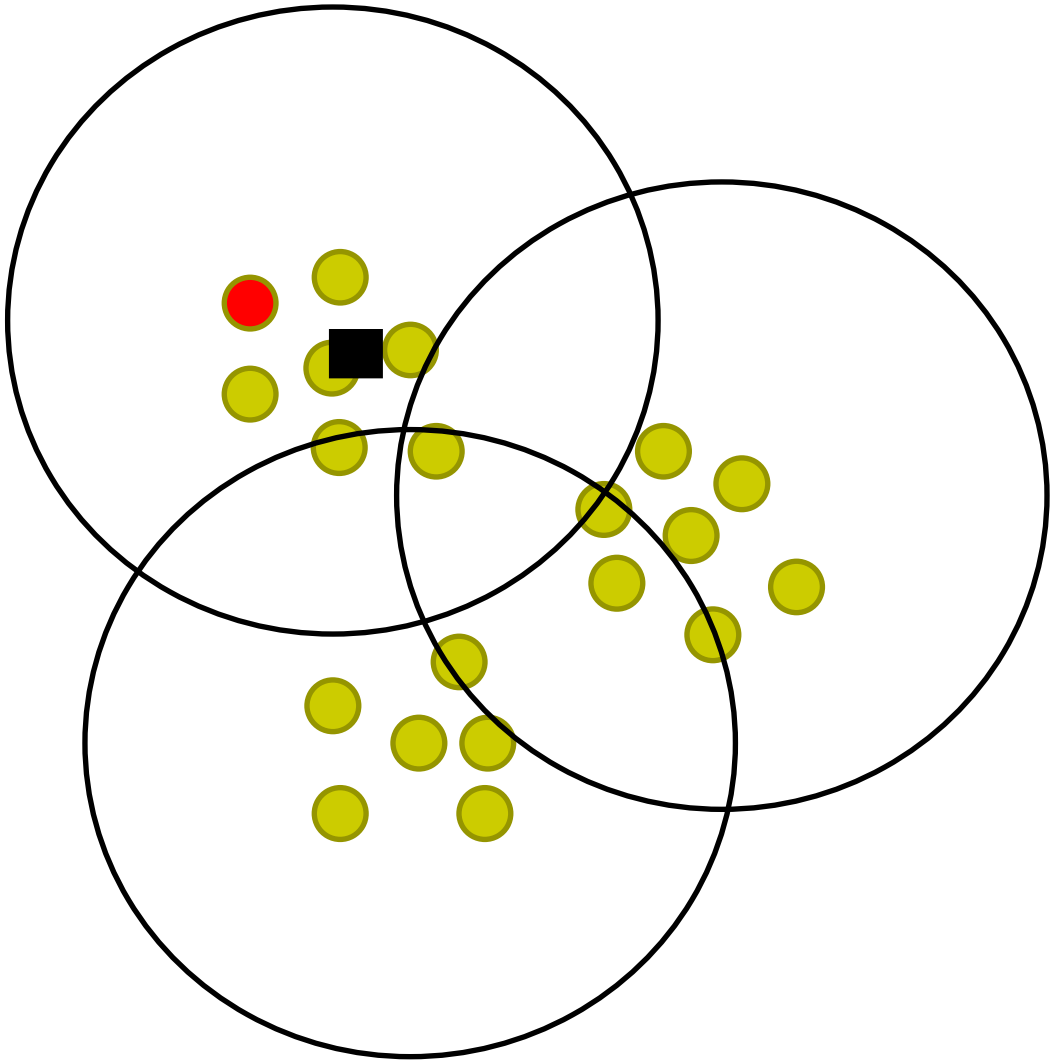
KOLAPS



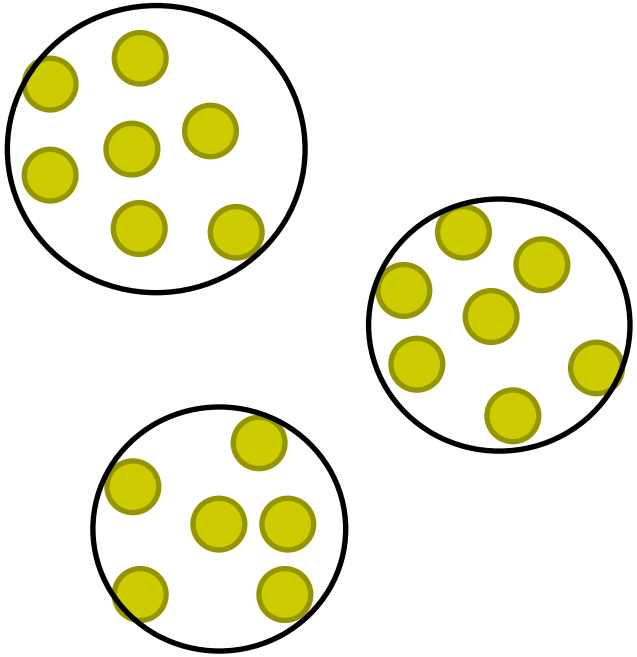
KOLAPS



KOLAPS

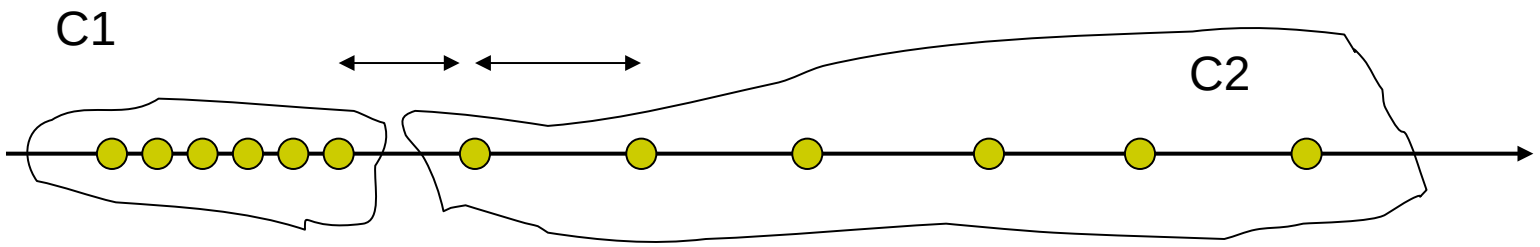


KOLAPS



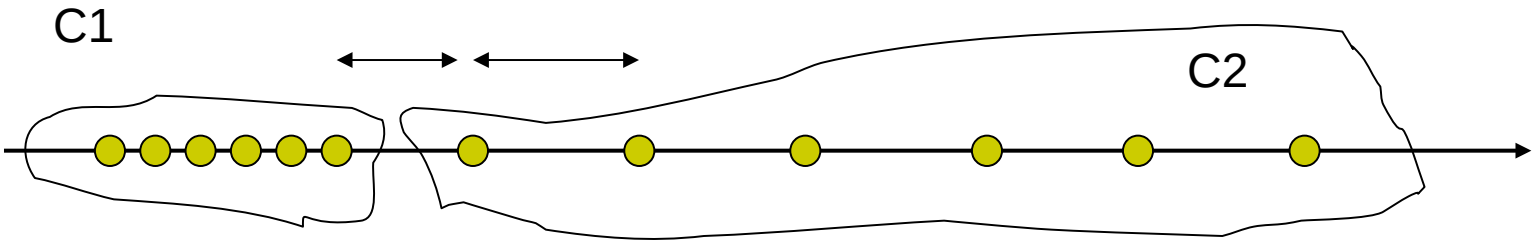
Гипотеза компактности

- Гипотеза λ-компактности
- Расстояние мало, но есть неоднородность.
- G – полный граф для T={Xi}
- A(a,b) – расстояние от точки a к b – длина ребра
- D=max(A(a,b))



Гипотеза компактности

- d=A(a,b)/D
- Для ребер смежных с (a,b) определим Bmin=min(A(a,b))
- r*=A(a,b)/Bmin
- Rmax=max(r*)
- R=r*/Rmax
- $\lambda=f(R,d)=R^2\ d$



λ-KRAB

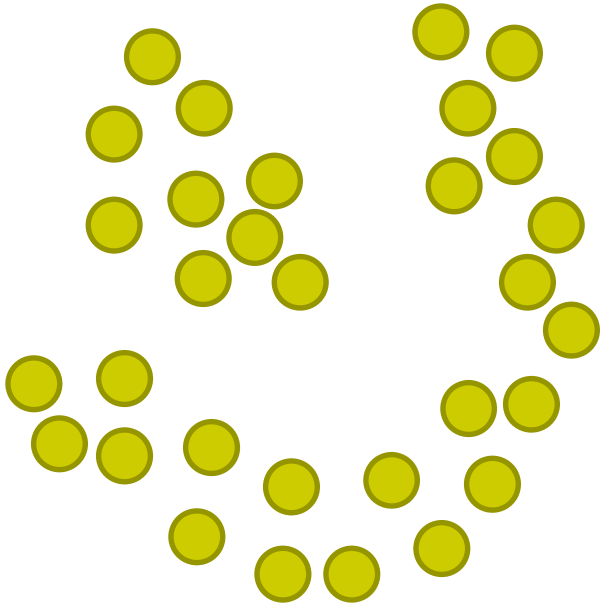
- Критерий равномерности таксонов (число объектов в кластерах приблизительно равно)
- $H=_{i=1,k}$ П (mi/M) ,i=1,k , mi – число объектов в кластере, M – общее число объектов

- Критерий алгоритма max ($H^q R^s d^v$) - q, s, v – параметры модели – степень влиятельности H, R и d,
- Экспериментально - max ($H^4 R^2 d$)

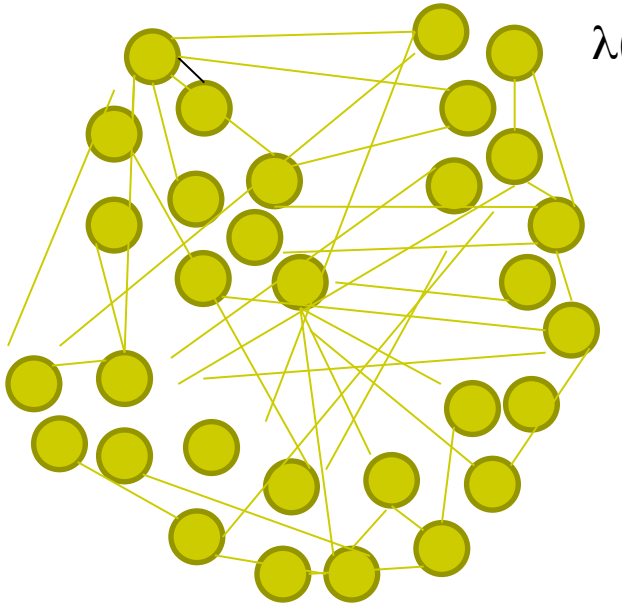
λ-KRAB – 2 кластера

1. Для всех Xi, Xj min(λ(Xi,Xj))
2. Строим ребро a(Xi, Xj) графа G.
3. Исключаем Xi, Xj из числа свободных
4. Если остались точки вне графа, то к 1, иначе к 5
5. G – кратчайший незамкнутый путь
6. Рассмотрели графы G1(a),G2(a) с разрывом по ребру a и определили для каждого случая $f(a)= H(a)^4 R(a)^2 d(a)$ по 2-м кластерам.
7. Делаем разрыв в a=arg max f(a)

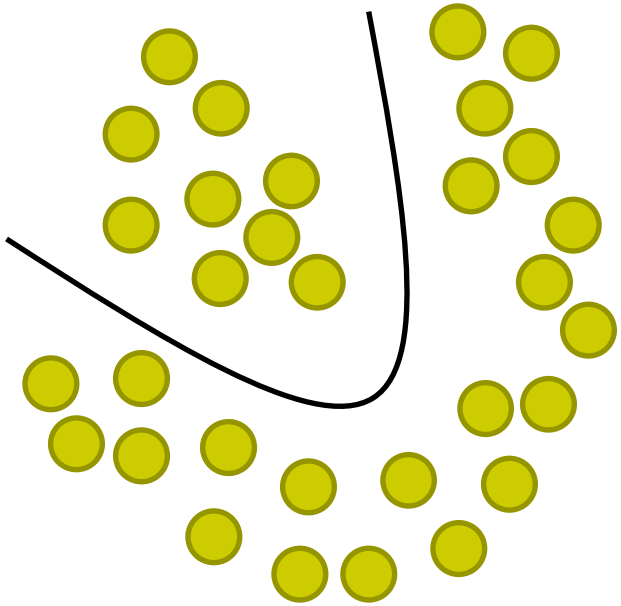
λ-KRAB



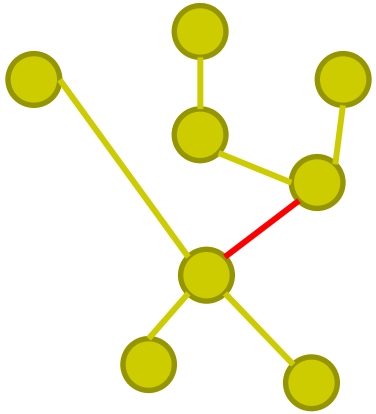
λ-KRAB



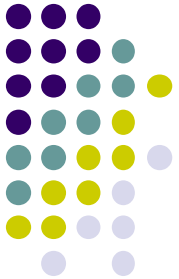
λ-KRAB



max f(a)



Иерархическая кластеризация



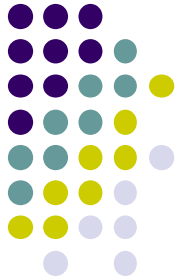
- Отношение эквивалентности

$$R_{Э}(X_i, X_j) = \begin{cases} 0, \rho(X_i, X_j) < \xi \\ 1, \rho(X_i, X_j) \geq \xi \end{cases}$$

$\rho(X_i, X_j)$ - *мера близости*

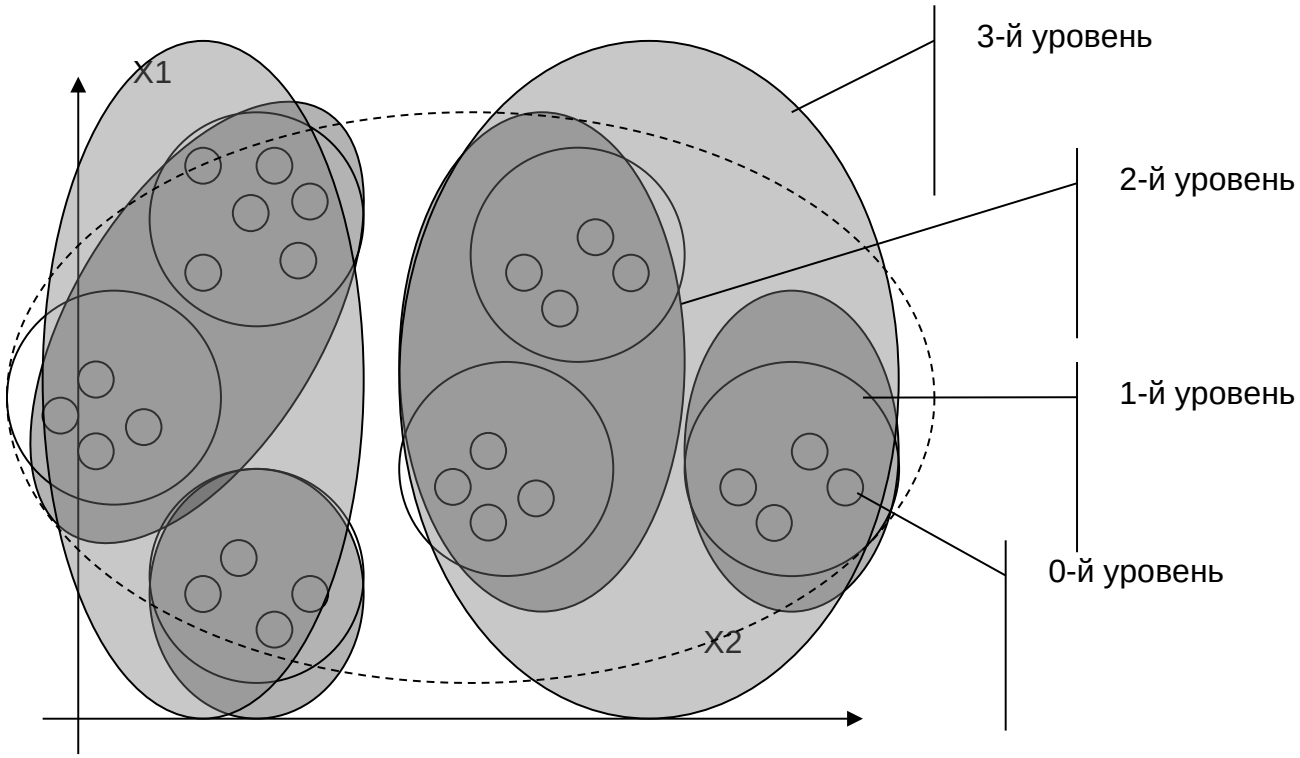
- ξ=*0.*
- *Когда* X_i, X_j *идентичны, то* $R_э(X_i, X_j) = 0$
- Работает эффективно для небольших объемов данных

Иерархическая кластеризация



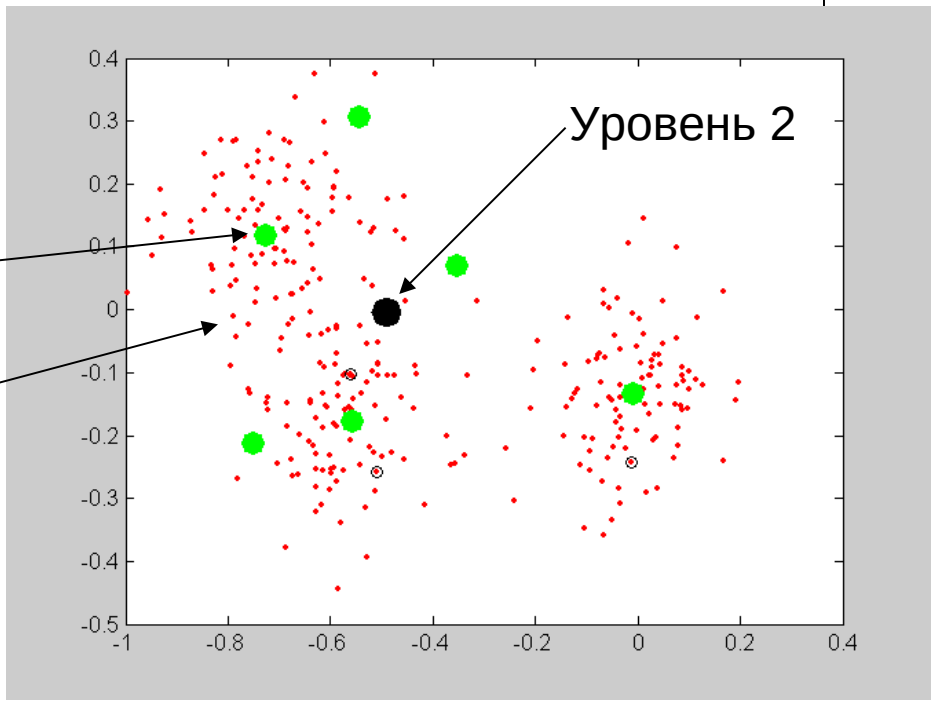
- Каждому кластеру соответствует один уникальный объект.
- Для формирования следующего уровня – вычисляем центр кластеров текущего уровня и рассматривать эти центры в качестве входной информации следующего уровня. Для перехода на этот уровень следует увеличить порог $\xi + \delta$
- Продолжать процесс пока не останется один кластер.

Иерархическая кластеризация



Пример

N=3;
n=100;



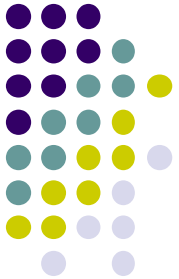
P1 =

-0.0110 -0.3528 -0.5584 -0.7277 -0.5441 -0.7518
-0.1335 0.0694 -0.1763 0.1191 0.3077 -0.2120

P2 =

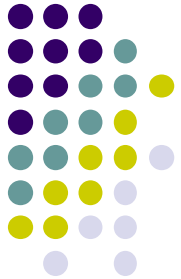
-0.4910
-0.0043

Смысловые цели кластеризации



- Минимизировать изменчивость внутри кластеров,
- Максимизировать изменчивость между кластерами.

Новые методы

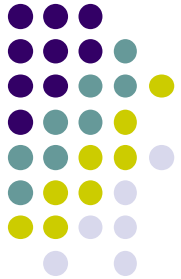


- Обработка сверхбольших объемов БД
- Требования:
  - Масштабируемость.
  - Работа в рамках оперативной памяти.
- методы кластеризации:
  - BIRCH,
  - CURE,
  - CHAMELEON,
  - ROCK

BIRCH

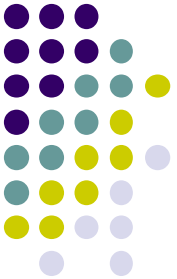
- Balanced Iterative Reducing and Clustering using Hierarchies - Тьян Зангом
- Алгоритм:
  - формируется предварительный набор кластеров.
  - к выявленным кластерам применяются другие алгоритмы кластеризации - пригодные для работы в оперативной памяти.

WaveCluster



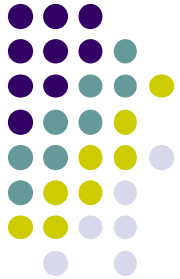
- Основан на волновых преобразованиях.
- Алгоритм:
  - данные обобщаются путем наложения на пространство данных многомерной решетки.
  - анализируются не отдельные точки, а обобщенные характеристики точек, попавших в одну ячейку решетки.
  - На последующих шагах для определения кластеров алгоритм применяет волновое преобразование к обобщенным данным.

особенности WaveCluster:



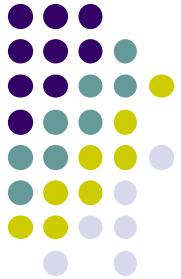
- сложность реализации;
- может обнаруживать кластеры произвольных форм;
- не чувствителен к шумам;
- применим только к данным низкой размерности.

Анализ результатов кластеризации.



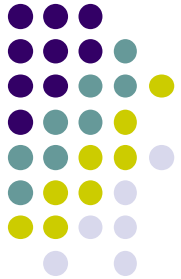
- не является ли полученное разбиение на кластеры случайным;
- является ли разбиение надежным и стабильным на подвыборках данных;
- существует ли взаимосвязь между результатами кластеризации и переменными, которые не участвовали в процессе кластеризации;
- можно ли интерпретировать полученные результаты кластеризации.

Процедуры проверки качества кластеризации:



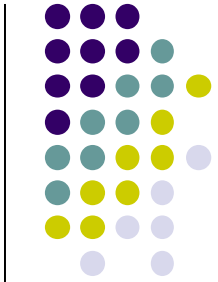
- анализ результатов кластеризации, полученных на определенных выборках;
- кросс-проверка;
- проведение кластеризации при изменении порядка наблюдений в наборе данных;
- проведение кластеризации при удалении некоторых наблюдений;
- проведение кластеризации на небольших выборках.

Использование нескольких методов



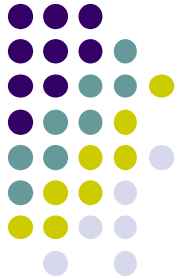
- Отсутствие подобия не будет означать некорректность результатов,
- Присутствие похожих групп считается признаком качественной кластеризации.

Как сделать кластер анализ быстрее



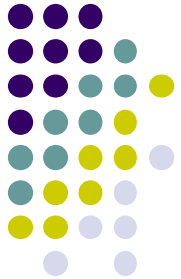
- Провести предобработку данных
  - Правильный выбор координат (оценка информативности)
  - Удаление выбросов (статистика и нормализация модели)
  - Редукция размерности
    - Факторный анализ (МЕТОД ГЛАВНЫХ КОМПОНЕНТ)
    - Многомерное шкалирование

Многомерное шкалирование



- целенаправленном преобразовании матриц сходства D, заранее сформированных на исходном множестве показателей
- Отображаем многомерные данные в пространство 2-х координат наибольшей значимости

Этапы кластер анализа

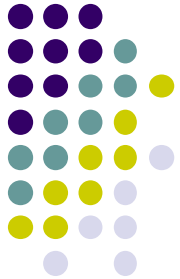


- Использовать ли все наблюдения либо же исключить некоторые данные или выборки из набора данных.
- Выбор метрики и метода стандартизации исходных данных.
- Определение количества кластеров (для итеративного кластерного анализа).
- Определение метода кластеризации (правила объединения или связи)

Этапы кластер анализа

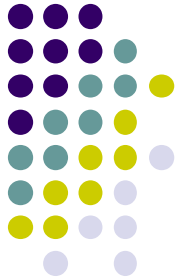
- Проведение кластеризации

Этапы кластер анализа



- Анализ результатов кластеризации.
  - не является ли полученное разбиение на кластеры случайным;
  - является ли разбиение надежным и стабильным на подвыборках данных;
  - существует ли взаимосвязь между результатами кластеризации и переменными, которые не участвовали в процессе кластеризации;
  - можно ли интерпретировать полученные результаты кластеризации

Этапы кластер анализа



- Оценка качества кластеризации и(возможно) возврат к предшествующим этапам
  - анализ результатов кластеризации, полученных на определенных выборках набора данных;
  - кросс-проверка;
  - проведение кластеризации при изменении порядка наблюдений в наборе данных;
  - проведение кластеризации при удалении некоторых наблюдений;
  - проведение кластеризации на небольших выборках.

Таблица 5.2. Сравнение классификации и кластерзации

Характеристика	Классификация	Кластеризация
Контролируемость обучения	Контролируемое обучение	Неконтролируемое обучение
Стратегия	Обучение с учителем	Обучение без учителя
Наличие метки класса	Обучающее множество сопровождается меткой, указывающей класс, к которому относится наблюдение	Метки класса обучающего множества неизвестны
Основание для классификации	Новые данные классифицируются на основании обучающего множества	Дано множество данных с целью установления существования классов или кластеров данных

Этапы классификации

Кластеризация
изучение исходных данных на предмет наличия в них групп, классов и определение признаков, которые за это отвечают

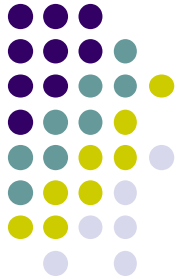
↓

Построение модели
нахождение зависимости между значениями признаков объектов и принадлежность их к определенной группе

↓

Классификация новых образцов
отождествление неизвестных образцов с одним из известных классов

Возможные ситуации



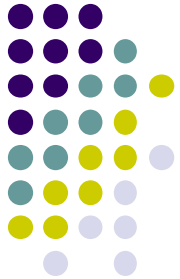
1. **В начале ни одного класса не определено**

   первым шагом в этом случае является предварительный анализ данных на предмет обнаружения потенциальных групп. В зависимости от результата возможны варианты:

   - **Имеется одна ярко выраженная группа**
   - **Имеется несколько ярко выраженных групп**

   Эти же варианты могут быть известны априори

Возможные ситуации



2. **Имеется одна ярко выраженная группа**

   В этом случае основная задача классификации найти и выделить типичную зависимость в данных для объектов, принадлежащих к одной группе и использовать ее для классификации новых объектов

3. **Имеется несколько ярко выраженных групп**

   Необходимо использовать методы распознавания образов для выяснения принадлежности новых объектов к тому или иному классу. Задачу можно свести к предыдущей ситуации.