Национальный исследовательский
Томский государственный университет

world skills Russia

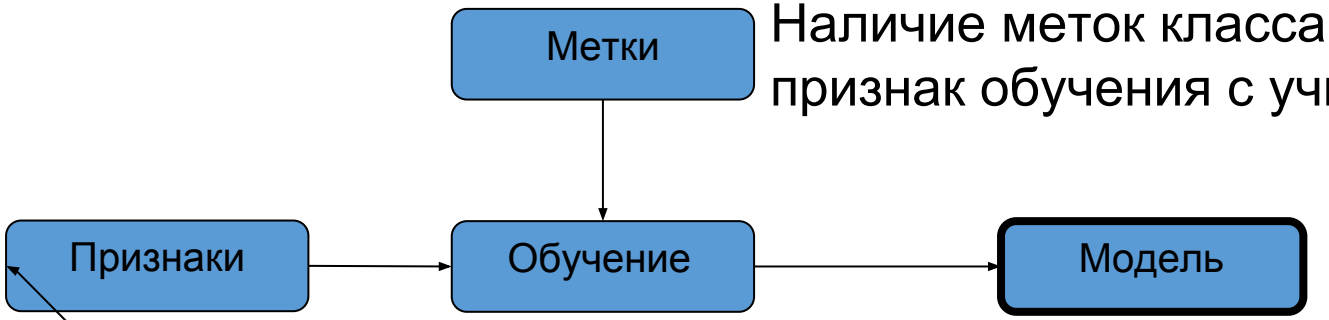
Основные концепции

Сергей Аксёнов, к.т.н., доцент кафедры
Теоретических основ информатики ТГУ

Обучение с учителем

Проектирование модели

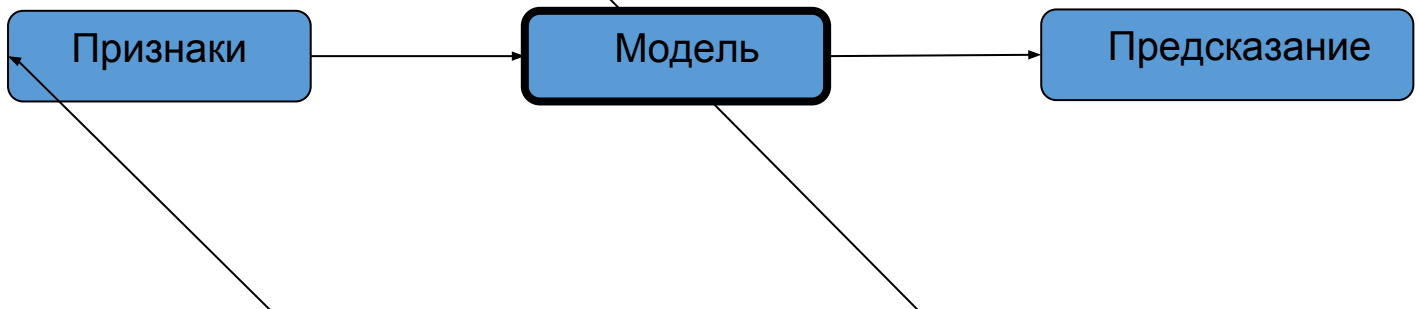
Выборка



Наличие меток класса – главный признак обучения с учителем

Тестирование / Использование

Тестирующий пример



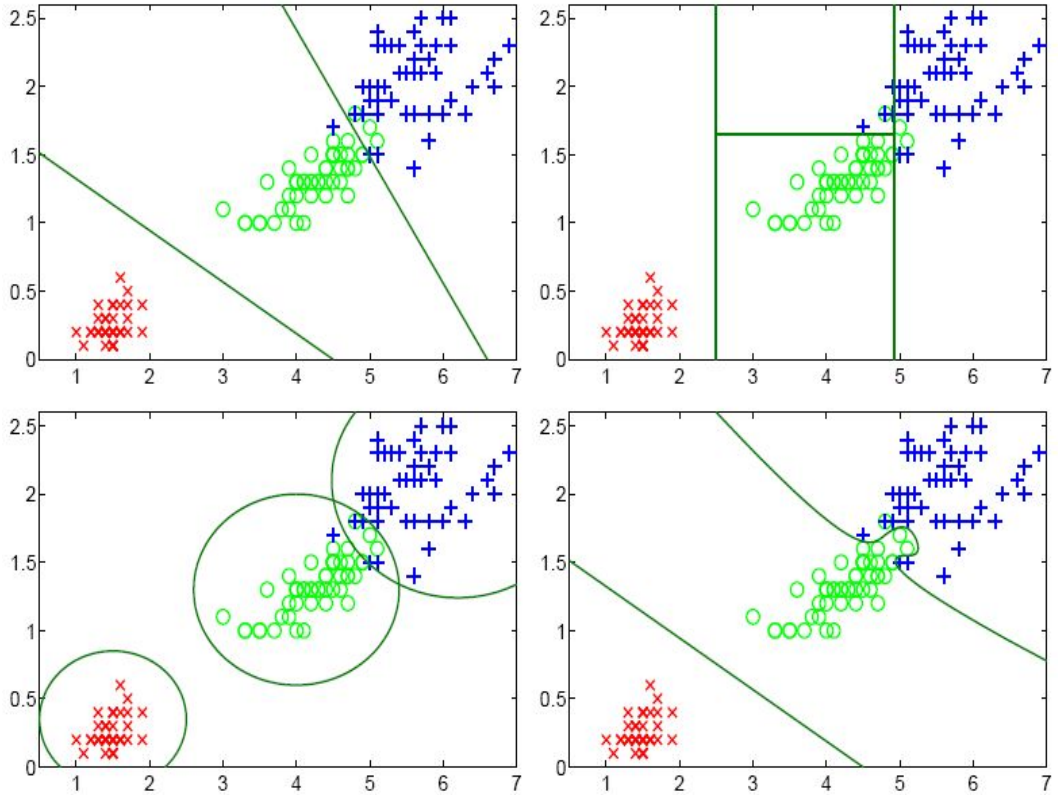
Обучение с учителем. Классификация. Пример

Цель обучения: разделить пространство признаков на регионы, в которых располагаются объекты принадлежащие только одному классу.

Разделение пространства признаков (длина и ширина лепестка ириса) разными алгоритмами

Линейная модель

Дерево решений



Гауссовы смеси

Метод опорных векторов

Классы объектов

setosa

virginica

versicolor

Оценка качества моделей классификации

Основные критерии:

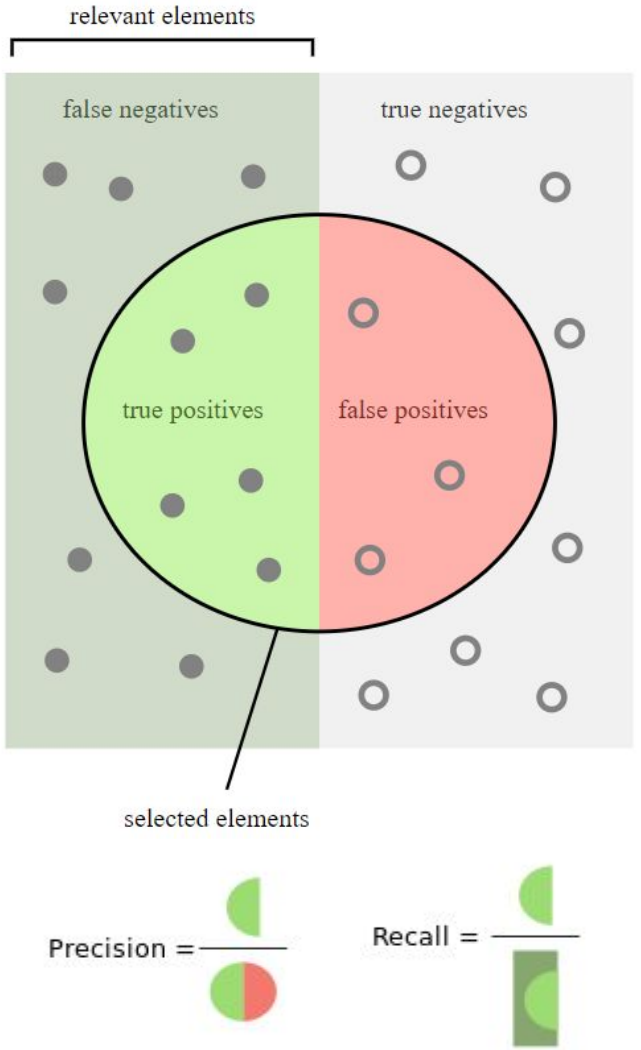
1. Матрица ошибок/несоответствий (Confusion matrix)

		Истинная метка	
		Положит. класс	Отрицат. класс
Предсказанная метка	Положительный класс	TP	FP
	Отрицательный класс	FN	TN

2. Верность: $ACC = \frac{TP+TN}{TP+FP+TN+FN}$

3. Точность: $Precision = \frac{TP}{TP+FP}$

4. Полнота: $Recall = \frac{TP+TN}{TP+FN}$



Пример оценки модели класстфикации

accuracy: 96.00%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	25	0	0	100.00%
pred. Iris-versicolor	0	23	1	95.83%
pred. Iris-virginica	0	2	24	92.31%
class recall	100.00%	92.00%	96.00%	

Матрица несоответствий для задачи с ирисами

Если классов больше чем два для получения точности и полноты применяется методика OvR (One versus Rest).

Для случаев трех классов: 1-й класс(+) против 2-й и 3-й классы(-),
2-й(+) против 1-й и 3-й классы(-), 3-й класс(+) против 1-й и 2-й классы(-)

Обучение с учителем. Регрессия. Пример

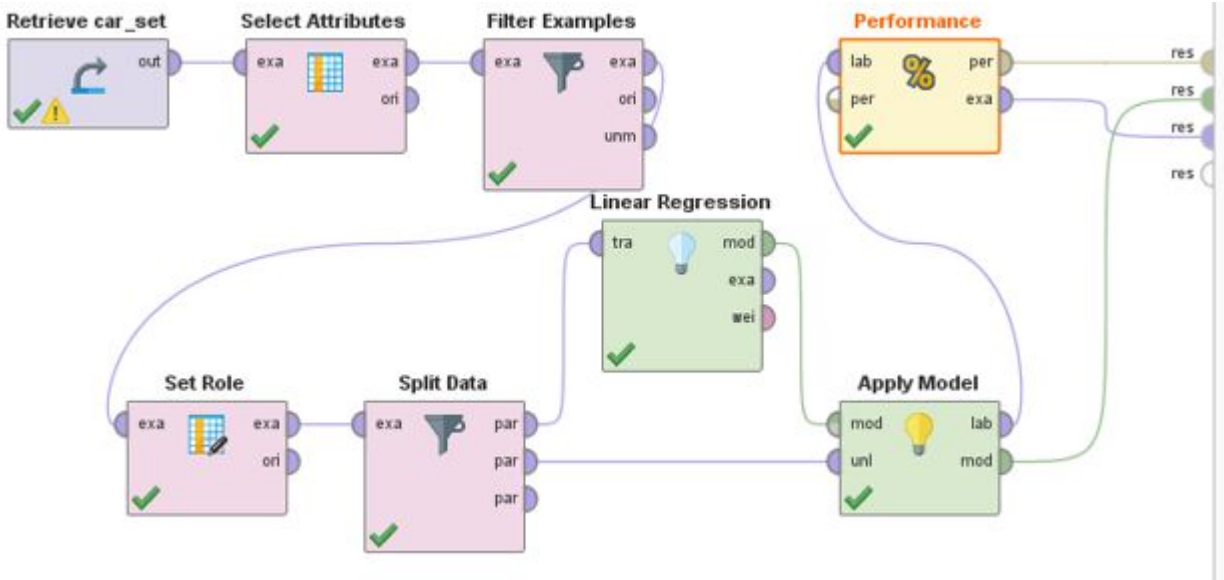
Цель обучения: получить выражение зависимости типа Y=f(X), где Y – целевая переменная, а X – входные признаки.

Пример из https://archive.ics.uci.edu/ml/datasets/Auto+MPG

Выборка моделей автомобилей. Задача построить модель позволяющую оценить показатель Mpg (сколько миль проезжает автомобиль на галлоне топлива), т.е. 1/расход топлива

Набор входных параметров:

1. cylinders: Кол-во цилиндров двигателя
2. displacement: Объём двигателя
3. horsepower: Мощность двигателя
4. weight: Масса автомобиля
5. acceleration: Ускорение
6. model year: Год выпуска
7. car name: Наименование модели



Построение модели в среде Rapid Miner Studio

Оценка качества модели регрессии. Пример

1.Среднеквадратичная ошибка

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

2.Средняя абсолютная ошибка

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i|$$

3.Коэффициент детерминации

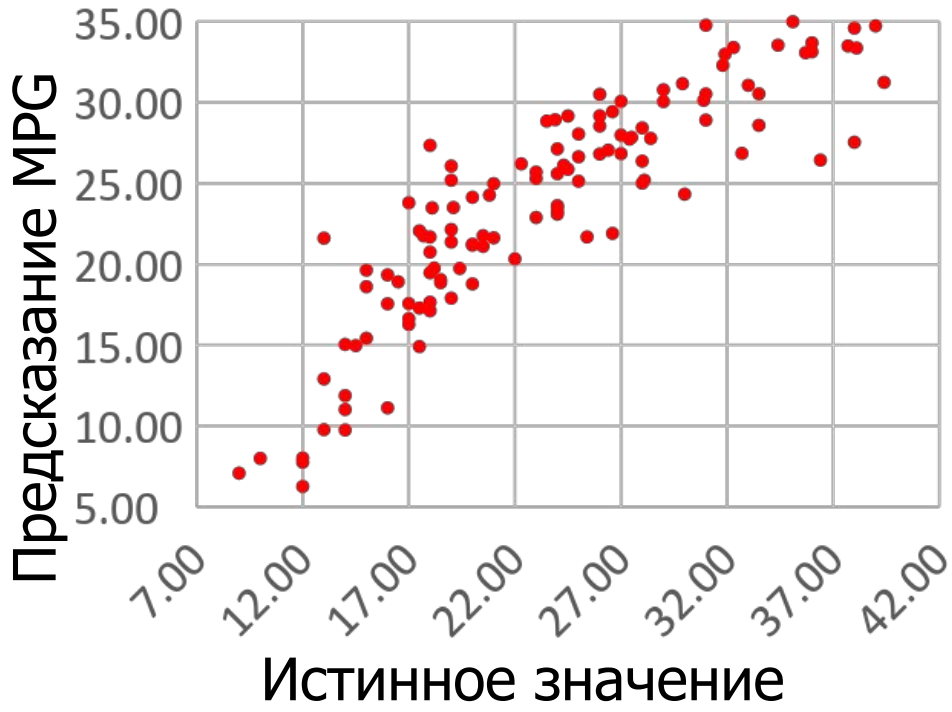
$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

y_i - Истинное значение

$\bar{y}$ - Среднее значение

$\tilde{y}_i$ - Предсказанное значение

Row No.	MPG	prediction(MPG)
1	15	15.419
2	14	15.038
3	24	23.522
4	22	20.327
5	18	20.750
6	24	23.098
7	21	21.635
8	10	7.990
9	9	7.077
10	28	25.012
11	17	17.565
12	14	9.749
13	14	11.874
14	12	6.254
15	19	17.899
16	23	25.303



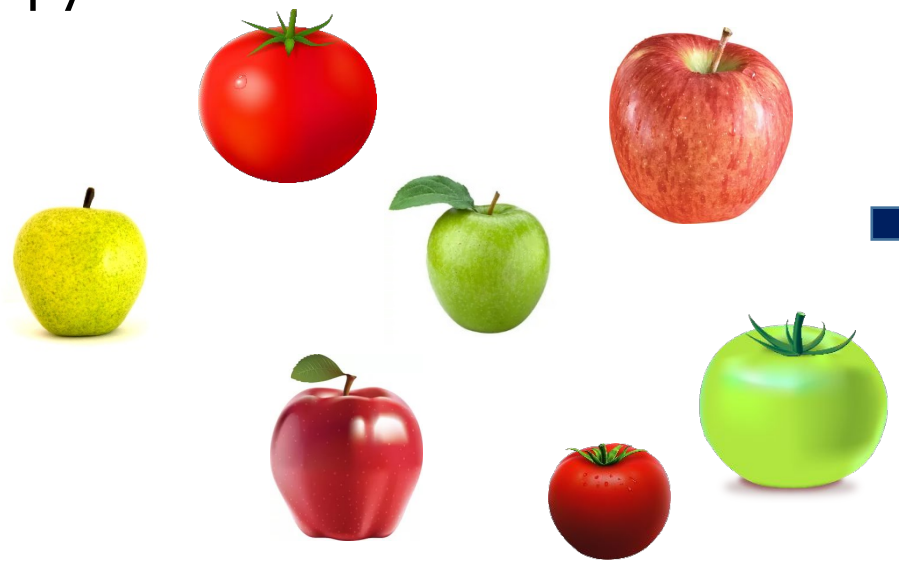
MSE = 3.48
R2 = 0.881

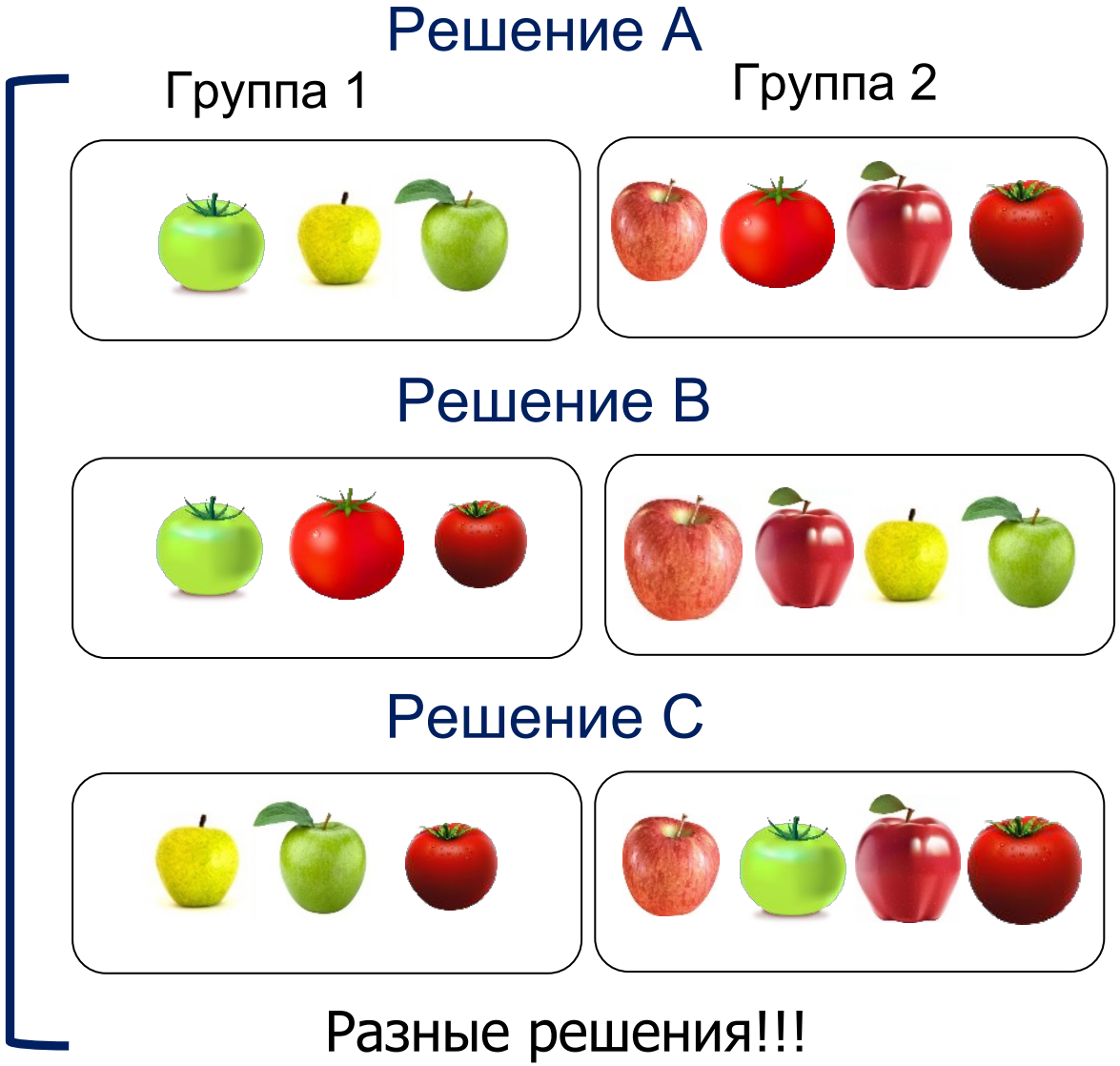
Обучение с учителем

Меток класса нет. Метод используется для изучения данных.

Особенность: Субъективность кластеризации.

Задача: Разложить объекты на две группы





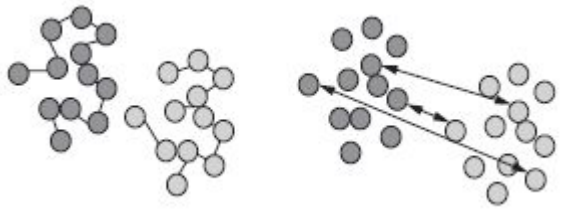
Оценка качества обучения с учителем

Индексы качества кластеризации.
Оценка производится методом сравнения нескольких структур

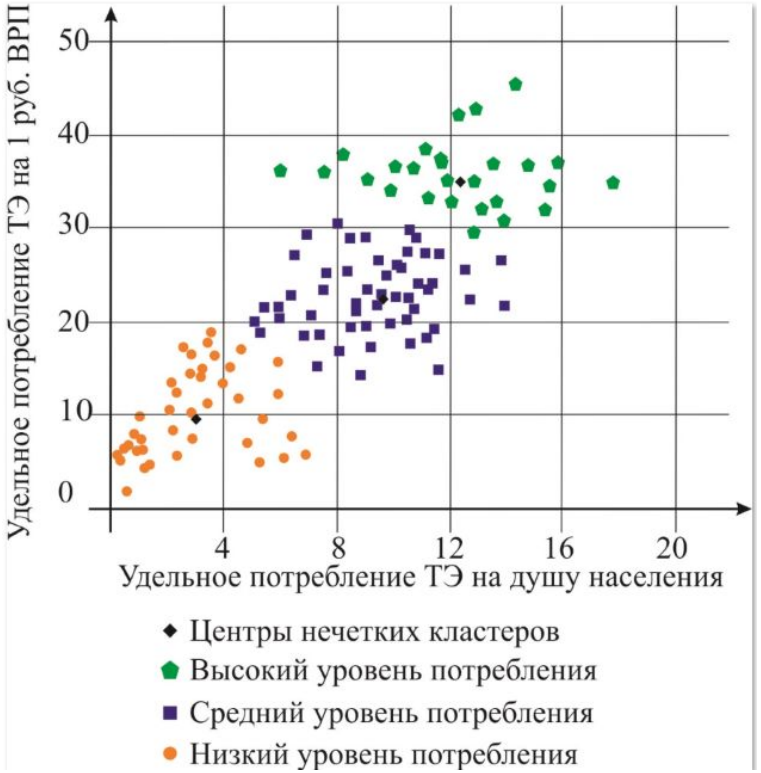
- Нескольких запусков одного и того же алгоритма
- Запуск алгоритма с разными параметрами
- Запуск разных алгоритмов

Критерии оценки качества:

Компактность - элементы из одного кластера должны быть как можно ближе друг к другу.
Отделимость – элементы из разных кластеров должны быть как можно дальше друг от друга.



Пример кластеризации



Примеры индексов

Индекс Данна

$$D = min_{i,j\in\{1\dots c\},i\neq j}\left\{\frac{d(c_i,c_j)}{\max_{k\in\{1\dots c\}} diam(c_k)}\right\}$$

$$d(c_i,c_j) = min_{x\in c_i, y\in c_j}\|x-y\|$$

$$diam(c_i) = \max_{x,y\in c_i}\|x-y\|$$

Индекс Дэвиса-Болдуина

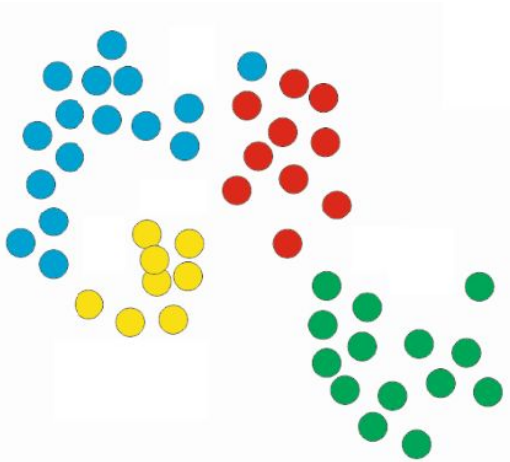
$$DB = \frac{1}{c}\sum_{i=1}^{c} R_i$$

$$R_i = \max_{i,j\in\{1\dots c\},i\neq j}(R_{ij}) \qquad R_{ij} = \frac{S_i + S_j}{d_{ij}}$$

$$S_i = \left\{\frac{1}{n_{c_i}}\sum_{x\in c_i}\|x - v_i\|^q\right\}^{\frac{1}{q}} \qquad d_{ij} = \left\{\sum_{k=1}^{dim}|v_i^k - v_j^k|^p\right\}^{\frac{1}{p}}$$

Обозначения

X	кластеризуемое множество
N	количество элементов в X
c	число кластеров
n_{c_i}	число элементов в кластере c_i
v_i	центр кластера c_i: $v_i = \frac{\sum_{x\in c_i} x}{n_{c_i}}$



Для примера