Оценка качества четкой кластеризации

Елена Сивоголовко

Санкт-Петербургский государственный университет
математико-механический факультет

План доклада

- **Кластеризация: основные понятия**
- Оценка качества кластеризации
  - Внешние метрики
  - Внутренние метрики
  - Относительные метрики
- Сравнение метрик оценки качества
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- Эксперименты
- Заключение

Кластер

Кластер (cluster) — объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица, обладающая определёнными свойствами.

- Информатика
- Астрономия
- Химия
- Экономика
- Лингвистика
- Музыка

Кластер: информатика

- единица хранения данных на диске;
- группа компьютеров, использующихся как единый ресурс;
- специализированный объект базы данных для физически совместного хранения одной или нескольких таблиц

Кластер: data mining

Кластер — объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица.
Кластеризация – задача разбиения множества на однородные группы, так, чтобы элементы в одной группе были максимально схожи друг с другом, а элементы из разных групп значительно отличались.
Data Mining
Machine learning
Unsupervised learning

Кластеризация: общая схема

1. Выделение значимых характеристик
2. Определение метрики схожести
3. Разбиение на группы
4. Оценка качества результатов
5. Представление и интерпретация результатов

План доклада

- Кластеризация: основные понятия
- **Оценка качества кластеризации**
  - Внешние метрики
  - Внутренние метрики
  - Относительные метрики
- Сравнение метрик оценки качества
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- Эксперименты
- Заключение

Качество кластеризации

- **Качество ПО стандарт ISO 9000**:
  "The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs"
- **Качество кластеризации (Cluster validity)**:
  "The adequacy of a clustering structure refers to the sense in which the clustering structure provides true information about the data, or the ability of recovered structure to reflect the intrinsic character of the data"

Оценка качества кластеризации

Методы (индексы) оценки качества (cluster validity methods) — инструментарий для количественной оценки результатов кластеризации.

Методы оценки качества

- Для четкой кластеризации: кластеры не пересекаются.
- для нечеткой кластеризации: допускается пересечение кластеров.

- Для иерархических структур.
- Для не иерархических структур.
- Для отдельных кластеров

- Внешние (external)
- Внутренние (internal)
- Относительные (relative)

План доклада

- Кластеризация: основные понятия
- Оценка качества кластеризации
  - **Внешние метрики**
  - Внутренние метрики
  - Относительные метрики
- Сравнение метрик оценки качества
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- Эксперименты
- Заключение

Внешние метрики оценки качества

Используют дополнительные знания о кластеризуемом множестве: распределение по кластерам, количество кластеров и т.д.
Хорошая структура кластеров: та же самая, что и предопределенная.
Методы:

- Rand statistic
- Jaccard index
- Folkes and Mallows index
- F1-measure

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: $v_i = \frac{\sum_{x \in c_i} x}{n_{c_i}}$
$\bar{X}$	центральный элемент множества $\bar{X} = \frac{1}{N} \sum_{j=1}^{N} x_j$
$\bar{v}$	центр центров $\bar{v} = \frac{1}{c} \sum_{i=1}^{c} v_i$
dim	размерность множества X

Rand, Jaccard, FM

Рассмотрим пары (x_i, x_j) из элемнтов X. Подсчитаем количество пар, в которых :

1. элементы принадлежат одному кластеру и одному классу: SS.
2. элементы принадлежат одному кластеру, но разным классам: SD
3. элементы принадлежат разным кластерам, но одному классу: DS
4. элементы принадлежат разным классам и разным кластерам: DD

Rand, Jaccard, FM

$$Rand = \frac{SS + DD}{SS + DS + SD + DD} \tag{1}$$

$$Jaccard = \frac{SS}{SS + SD + DS} \tag{2}$$

$$FM = \sqrt{\frac{SS}{SS + SD} * \frac{SS}{SS + DS}} \tag{3}$$

F1-measure

Для кластера c_i и класса g_j рассмотрим:

$$Precision(i,j) = \frac{n_{ij}}{n_i}$$

$$Recall(i,j) = \frac{n_{ij}}{n_j}$$

где n_{ij} — количество объектов $x_k : x_k \in c_i \& x_k \in g_j$, $n_i = |c_i|$ и $n_j = |g_j|$. F1-мера для c_i и g_j определяется как

$$F1(i,j) = \frac{2 * Precision(i,j) * Recall(i,j)}{Precision(i,j) + Recall(i,j)}$$

Общий показатель для структуры кластеров:

$$F1 = \sum_j \frac{n_j}{N} \max_i F1(i,j) \tag{4}$$

План доклада

- Кластеризация: основные понятия
- Оценка качества кластеризации
  - Внешние метрики
  - **Внутренние метрики**
  - Относительные метрики
- Сравнение метрик оценки качества
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- Эксперименты
- Заключение

Внутренние методы оценки качества

Оценивают качество структуры кластеров опираясь только на непосредственно на нее.
Методы:

- Cophenetic Correlation Coefficient
- Hubert Γ Statistics

Внутренние методы: статистика

Какова вероятность получить тоже самое значение индекса случайно?

Нулевые гипотезы:

- случайное положение
- случайные метки кластеров
- случайная матрица близости

Внутренние методы: статистика

$P(T \geq t_\alpha | H_0) = \alpha$

T — значение индекса оценки качества

t_α — пороговое значение

α — уровень значимости

- Monte-Carlo method
- Bootstrapping method

Cophenetic Correlation Coefficient

Cophenetic matrix P_c: $P_c(i,j)$ — уровень иерархии, на котором элементы x_i и x_j первый раз встречаются в одном кластере.
P — матрица близости

$$C = \frac{\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}c_{ij} - \mu_P\mu_C}{\sqrt{\left[\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 - \mu_P^2\right]\left[\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2 - \mu_C^2\right]}} \tag{5}$$

где d_{ij} — элемент P, c_{ij} — элемент P_c, $M = \frac{N(N-1)}{2}$ и

$$\mu_P = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij} \qquad \mu_C = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}$$

Hubert Γ Statistics

$$\Gamma = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} P(i,j) Y(i,j) \tag{6}$$

здесь $P(i,j)$ — матрица близости, а

$$Y(i,j) = \begin{cases} 1 & \text{, если } x_i \text{ и } x_j \text{ в одном кластере} \\ 0 & \text{, в другом случае.} \end{cases}$$

Обзначает сходство между матрицами X и Y. Чем больше, тем лучше.

План доклада

- Кластеризация: основные понятия
- Оценка качества кластеризации
  - Внешние метрики
  - Внутренние метрики
  - **Относительные метрики**
- Сравнение метрик оценки качества
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- Эксперименты
- Заключение

Относительные методы оценки качества

Оценка производится методом сравнения нескольких структур

- несколько 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 дальше друг от друга.

Модифицированная Hubert Γ Statistic

$$\Gamma = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} P(i,j)Q(i,j) \tag{7}$$

$Q(i,j)$: расстояние между центрами кластеров v_{c_i} и v_{c_j}, к которым принадлежат элементы x_i и x_j, соответственно.

Нормализованная Hubert Γ Statistic

$$\hat{\Gamma} = \frac{\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(P(i,j)-\mu_P)(Q(i,j)-\mu_Q)}{\sigma_P\sigma_Q} \tag{8}$$

здесь μ_P, μ_Q, σ_P и σ_Q — средние значения и среднеквадратичные отклонения для матриц P и Q соответственно.

$\hat{\Gamma} \in [-1, 1]$

Большие значения $\hat{\Gamma}$ подразумевают лучшую структуру кластеров.

Calinski-Harabasz индекс

$\bar{d}^2$ — средний квадрат расстояния между элементами в кластеризуемом множестве
$\bar{d}^2_{c_i}$ — средний квадрат расстояния между элементами в кластере c_i.
Сумма квадратов расстояний внутри групп:

$$WGSS = \frac{1}{2}\sum_{i=1}^{c}(n_{c_i} - 1)\bar{d}^2_{c_i}$$

Сумма квадратов расстояний между группами:

$$BGSS = \frac{1}{2}((c-1)\bar{d}^2 + (N-c)A_c)$$

где $A_c = \frac{1}{N-c}\sum_{i=1}^{c}(n_{c_i} - 1)(\bar{d}^2 - \bar{d}^2_{c_i})$

$$VRC = \frac{\frac{BGSS}{c-1}}{\frac{WGSS}{N-c}} = \frac{\bar{d}^2 + \frac{N-c}{c-1}A_c}{\bar{d}^2 - A_c} \quad (9)$$

Индекс Данна

$$D = min_{i,j\in\{1\dots c\},i\neq j}\left\{\frac{d(c_i,c_j)}{\max_{k\in\{1\cdots c\}} diam(c_k)}\right\} \qquad (10)$$

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

$diam(c_i)$ — диаметр кластера: $diam(c_i) = \max_{x,y\in c_i}\|x-y\|$.

Индекс Девиса-Болдуина

Пусть $S_i = \left\{ \frac{1}{n_{c_i}} \sum_{x \in c_i} \|x - v_i\|^q \right\}^{\frac{1}{q}}$ — мера разброса внутри c_i

$d_{ij} = \left\{ \sum_{k=1}^{dim} |v_i^k - v_j^k|^p \right\}^{\frac{1}{p}}$ — мера различия между c_i и c_j

Тогда R_{ij} мера схожести между c_i и c_j если:

1. $R_{ij} \geq 0$
2. $R_{ij} = R_{ji}$
3. При $S_i = 0$ и $S_j = 0$ $R_{ij} = 0$
4. При $S_j > S_k$ и $d_{ij} = d_{ik}$ $R_{ij} > R_{ik}$
5. При $S_j = S_k$ и $d_{ij} < d_{ik}$ $R_{ij} > R_{ik}$

Индекс Девида-Болдуина

$$R_{ij} = \frac{S_i + S_j}{d_{ij}}$$

Тогда сам индекс вычисляется по формуле:

$$DB = \frac{1}{c} \sum_{i=1}^{c} R_i \tag{11}$$

где $R_i = \max_{i,j \in \{1 \cdots c\}, i \neq j}(R_{ij})$.
DB индекс определяет среднюю схожесть между кластером c_i и наиболее близким к нему кластером.

CS индекс

$$CS = \frac{\sum_{i=1}^{c}\{\frac{1}{n_{c_i}}\sum_{x_j,x_k\in c_i} max(\|x_j - x_k\|)\}}{\sum_{i=1}^{c}\min_{j\neq i}(\|v_i - v_j\|)} \quad (12)$$

Измеряет отношение максимального расстояния между точками в кластере к минимальному межкластерному расстоянию.

Оптимальная структура характеризуется меньшим показателем CS.

SD (Scatter-Distance) индекс

Дисперсия на множестве:

$$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}(x_k^p - \bar{x}^p)^2$$

$$\sigma_X = \begin{bmatrix} \sigma_X^1 \\ \vdots \\ \sigma_X^{dim} \end{bmatrix}$$

Дисперсия внутри кластера:

$$\sigma_{v_i}^p = \frac{1}{n_{c_i}}\sum_{x_k \in c_i}(x_k^p - v_i^p)^2$$

$$\sigma_{v_i} = \begin{bmatrix} \sigma_{v_i}^1 \\ \vdots \\ \sigma_{v_i}^{dim} \end{bmatrix}$$

SD (Scatter-Distance) индекс

Средний разброс для кластеров определяется как:

$$Scatt = \frac{1}{c}\sum_{i=1}^{c}\frac{\|\sigma_{v_i}\|}{\|\sigma_x\|}$$

Отделимость кластеров измеряется как:

$$Dist = \frac{\max_{i,j\in\{1\cdots c\}}(\|v_j - v_i\|)}{\min_{i,j\in\{1\cdots c\}}(\|v_j - v_i\|)}\sum_{i=1}^{c}\left(\sum_{j=1,i\neq j}^{c}\|v_i - v_j\|\right)^{-1}$$

Собственно *SD* индекс:

$$SD = \alpha * Scatt + Dist \tag{13}$$

где α – взвешивающий коэффициент

S_Dbw (Scatter-Density) индекс

Плотность между кластерами:

$$Dens\_bw = \frac{1}{c(c-1)} \sum_{i=1}^{c} \left(\sum_{j=1, i \neq j}^{c} \frac{dens(u_{ij})}{\max(dens(v_i), dens(v_j))} \right)$$

Здесь u_{ij} – есть середина линии,соединяющей кластерные центры v_i и v_j.
Функция плотности: $dens(u_{ij}) = \sum_{x \in c_i \cup c_j} f(x, u_{ij})$, где $f(x, u_{ij})$ – окрестность точки u_{ij}: гиперсфера с центром в u_{ij} и радиусом равным $stdev = \frac{1}{c} \sqrt{\sum_{i=1}^{c} \|\sigma_{v_i}\|}$

$$f(x, u_{ij}) = \begin{cases} 0, & \text{если } \|x - u_{ij}\| > stdev \\ 1, & \text{в другом случае} \end{cases}$$

S_Dbw (Scatter-Density) индекс

$$Scatt = \frac{1}{c}\sum_{i=1}^{c}\frac{\|\sigma_{v_i}\|}{\|\sigma_x\|}$$

Общая формула индекса для индекса S_Dbw

$$S\_Dbw = Dens\_bw + Scatt \tag{14}$$

Значение индекса должно быть минимальным для получения наилучшего результата.

RMSSTD(root – mean – square standard deviation) индекс

$$RMSSTD = \left[\frac{pooled\ sum\ of\ squares\ for\ all\ variables}{pooled\ degrees\ of\ freedom\ for\ all\ variables} \right]^{\frac{1}{2}}$$

Итоговая формула:

$$RMSSTD = \left[\frac{\sum_{i=1}^{c} \sum_{x_j \in c_i} \|x_j - v_i\|^2}{dim * \sum_{i=1}^{c} (n_{c_i} - 1)} \right]^{\frac{1}{2}} \qquad (15)$$

RS (R Squared) Индекс

Пусть

1. SS_w сумма квадратов расстояний внутри кластера
2. SS_b сумма квадратов расстояний между кластерами
3. SS_t сумма квадратов расстояний по всему множеству, причем $SS_t = SS_w + SS_b$

Формула индекса RS:

$$RS = \frac{SS_b}{SS_t} = \frac{SS_t - SSw}{SS_t} \tag{16}$$

Индекс оценки силуэта(Silhouette index)

Пусть $a_{pj} = \frac{1}{n_{c_p}-1} \sum_{x_k \in c_p} \|x_j - x_k\|$ — среднее расстояние от $x_j \in c_p$ до других объектов из кластера c_p.
$d_{qj} = \frac{1}{n_{c_q}} \sum_{x_k \in c_q} \|x_j - x_k\|$ — среднее расстояние от x_j до объектов из другого кластера $c_q : q \neq p$.
Положим: $b_{pj} = \min_{q \neq p} d_{qj}$.

Тогда "силуэт " элемента x_j: $S_{x_j} = \frac{b_{pj} - a_{pj}}{\max(a_{pj}, b_{pj})}$

Оценка для всей кластерной структуры:

$$SWC = \frac{1}{N} \sum_{j=1}^{N} S_{x_j} \tag{17}$$

Индекс оценки силуэта(Silhouette index)

Упрощенный силуэт: a_{pj} и b_{pj} вычисляются через центры кластеров.

Альтернативный силуэт: $S_{x_j} = \frac{b_{pj}}{a_{pj} + \epsilon}$

индекс Maulik-Bandoypadhyay

$$MB = \left(\frac{1}{c} \frac{E_1}{E_c} D \right)^p \tag{18}$$

$E_c = \sum_{i=1}^{c} \sum_{x \in c_i} \|x - v_i\|$ — сумма внутрикластерных расстояний.

E_1 – сумма расстояний от центра множества до каждого элемента.

$D = \max_{i,j} \|v_i - v_j\|$ – максимальное расстояние между кластерами.

Константа p определяется произвольно, в оригинале: $p = 2$

Score function

Определим расстояние между кластерами как:

$$bcd = \frac{\sum_{i=1}^{c} \|v_i - \bar{v}\| * n_{c_i}}{N * c}$$

Стандартный подход для измерения близости точек внутри кластера:

$$wcd = \sum_{i=1}^{c} \left(\frac{1}{n_{c_i}} \sum_{x \in c_i} \|x - v_i\| \right)$$

Кластерная структура является хорошей, если bcd высокий, а wcd - низкий.

$$SF = 1 - \frac{1}{\exp^{\exp^{bcd - wcd}}} \tag{19}$$

Чем выше SF, тем лучше структура кластеров.

VNND индекс (индекс ближайших соседей)

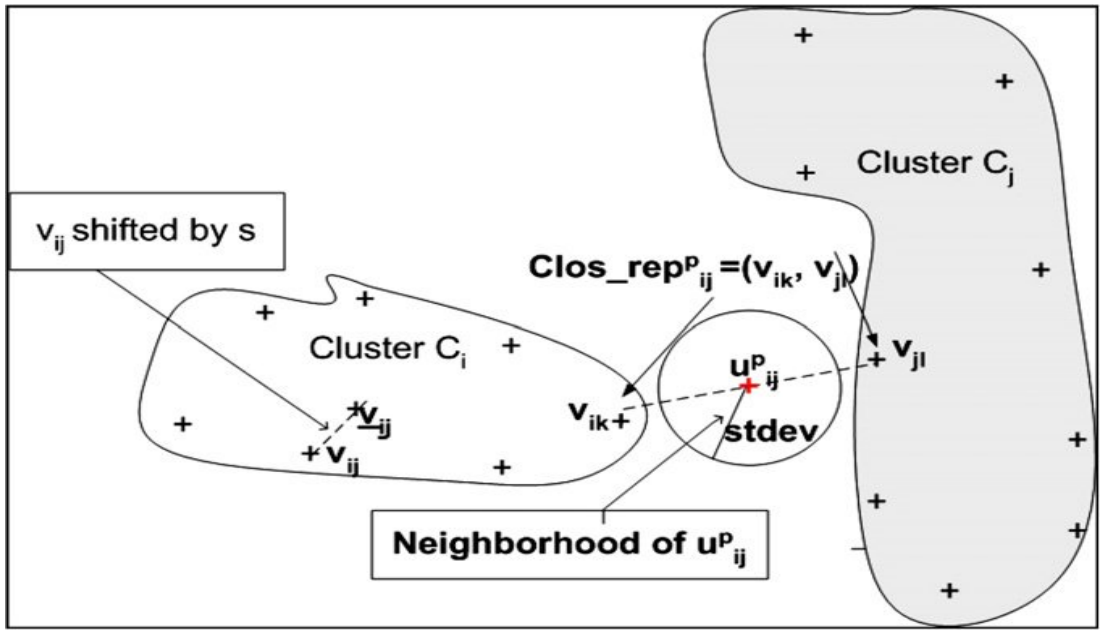
$d_{min}(x_j) = \min_{y \in c_i}(\|x_i - y\|)$ — расстояние от элемента x_j до его ближайшего соседа.
$\overline{d_{min}(c_i)} = \frac{1}{n_{c_i}}(\sum_{x_j \in c_i} d_{min}(x_j))$ — среднее расстояние между ближайшими соседями в кластере c_i.
Отклонение для расстояния между ближайшими соседями:

$$V(c_i) = \frac{1}{n_{c_i} - 1} \sum_{x_j \in c_i} (d_{min}(x_j) - \overline{d_{min}(c_i)})^2$$

Итоговое значение индекса определяется как

$$VNND = \sum_{i=1}^{c} V(c_i) \tag{20}$$

Индекс плотности CDbw



Индекс плотности CDbw

Множество представителей для кластера c_i: $V_{c_i} = \{v_1 \cdots v_r\}$, где r — произвольное число, $r \geq 10$.
$v_{ik} = closest\_rep^i(v_{jl})$: v_{ik} является ближайшим представителем кластера c_i по отношению к v_{jl}.
Множество соответствующих ближайших представителей: $RCR_{ij} = \{(v_{ik}, v_{jl}) | v_{ik} = closest\_rep^i(v_{jl}) \& v_{jl} = closest\_rep^j(v_{ik})\}$.

Индекс плотности CDbw: отделимость

Плотность между двумя кластерами в этом случае вычисляется как

$$Dens_{ij} = \frac{1}{|RCR_{ij}|} * \sum_{(v_k, v_l) \in RCR_{ij}}^{RCR_{ij}} \left(\frac{\|v_k - v_l\|}{2 * stdev} * cardinality(u_{kl}) \right)$$

здесь $stddev = \frac{1}{c}\sqrt{\sum_{i=1}^{c} \|\sigma_{v_i}\|}$, u_{kl} – средина отрезка, между точками v_l и v_k, и $cardinality(u_{kl}) = \frac{\sum_{x \in c_i \cup c_j}^{c_i \cup c_j} f(x, u_{kl})}{n_{c_i} + n_{c_j}}$, где

$$f(x,u) = \begin{cases} 0, & \text{если } \|x - u_{kl}\| > stdev \\ 1, & \text{в другом случае} \end{cases}$$

индекс плотности CDbw: отделимость

Межкластерная плотность для всей структуры:

$$Inter\_dens = \frac{1}{c}\sum_{i=1}^{c}\max_{i\neq j}(Dens_{ij})$$

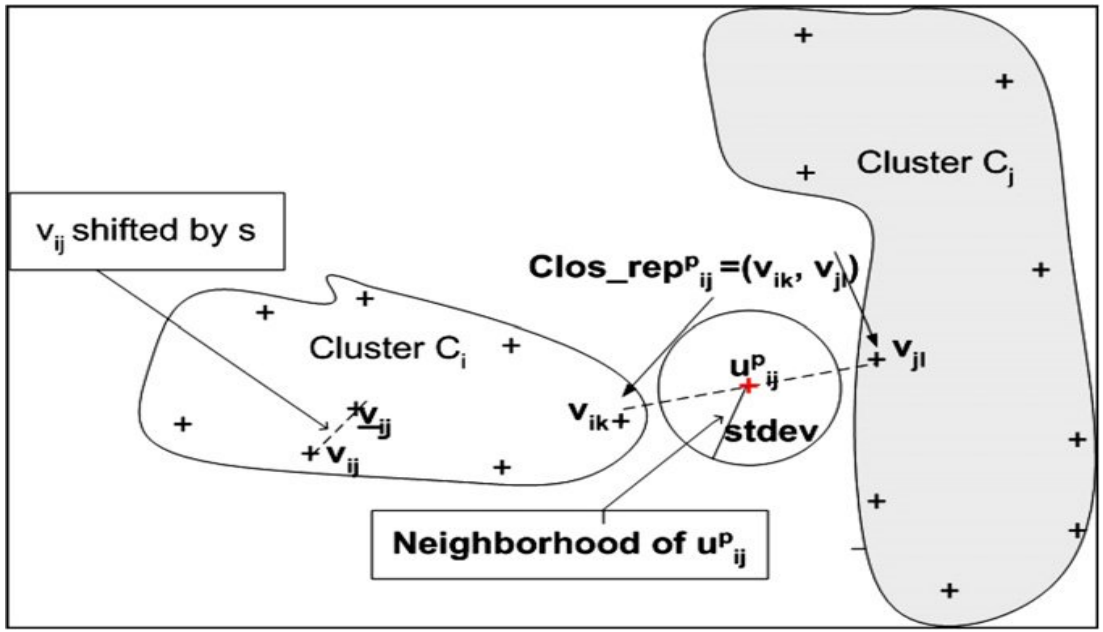
Расстояние между кластерами:

$$Dist_{ij} = \frac{1}{|RCR_{ij}|}\sum_{(v_k,v_l)\in RCR_{ij}}^{RCR_{ij}}\|v_k - v_l\|$$

Общая мера отделимости для всей структуры:

$$Sep = \frac{\frac{1}{c}\sum_{i=1}^{c}\min_{i\neq j} Dist_{ij}}{1 + Inter\_dens}$$

Индекс плотности CDbw



индекс плотности CDbw: компактность

Сдвинутые "представители"(shifted representatives): $v_k \in V_{c_i}$
$v_k^s = v_k + s * (center(c_i) - v_k)$, где $s \in [0, 1]$.
Возьмем $0.1 \leq s \leq 0.8$, $s_i = s_{i-1} + 0.1$. Обозначим число итераций за n_s.
Плотность структуры кластеров относительно s:

$$Intra\_dens(s) = \frac{\frac{1}{r}\sum_{i=1}^{c}\sum_{v_k \in V_{c_i}} cardinality(v_k^s)}{c * stdev}$$

индекс плотности CDbw: компактность

Компактность кластерной структуры:

$$Compactness = \frac{1}{n_s} \sum_{i=1}^{n_s} Intra\_dens(s_i)$$

Изменение плотность кластера в зависимости от s:

$$Intra\_change = \frac{\sum_{i=1}^{n_s} |Intra\_dens(s_i) - Intra\_dens(s_{i-1})|}{n_s - 1}$$

"Связанность" кластеров:

$$Cohension = \frac{Compactness}{1 + Intra\_change}$$

Общая формула индекса:

$$CDbw = Cohension * Sep * Compactness \qquad (21)$$

План доклада

- Кластеризация: основные понятия
- Оценка качества кластеризации
  - Внешние метрики
  - Внутренние метрики
  - Относительные метрики
- **Сравнение метрик оценки качества**
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- Эксперименты
- Заключение

Тестовый стенд

- тестовые множества
- алгоритмы кластеризации
- индексы оценки качества

Тестовые множества

- синтетические
- реальные

Тестовые множества

1. Равномерно распределенные данные
2. Кластеры неправильной формы и вложенные кластеры
3. Плохоотделимые кластеры
4. Кластеры с гауссовым распределением

Алгоритмы кластеризации

- разбивающие: K-средних
- плотностные: DBScan

Алгоритмы кластеризации: K-средних

Algorithm. 1 K-Means

Require: c — количество кластеров

Init: случайным образом, выбрать c точек, которые будут центрами кластеров на первой итерации.

repeat

Определить каждый элемент из множества в кластер, с ближайшим центром.

Пересчитать кластерные центры с учетом текущего распределения элементов.

until Пока не выполнится условие остановки

Алгоритмы кластеризации: DBScan

Algorithm. 2 DBScan

Require: ϵ, min_pts
  for all $x_j \in X$ **do**
    if $(\nexists k : x_j \in C_k)$ & $(x_j \notin Outliers)$ **then**
      $eps\_n = |\{y : y \in N_\epsilon(x_j)\}|$
      if $eps\_n < min\_pts$ **then**
        $Outliers = x_j \cup Outliers$
      else
        $x_j \in C_{k+1}$
        Для всех $y : y \in N_\epsilon(x_j)$ повторить оценку точек в $N_\epsilon(y)$.
      end if
    end if
  end for

Параметры запуска

K-Means: число кластеров — от 2 до 10
DBSCan: *min_pts* - менялось с 3 до 6;
ϵ - с 0.1 до 1 с шагом 0.1 (0.5 до 4 с шагом 0.5)

Индексы качества

1. Dunn,
2. DB,
3. SD,
4. S_Dbw,
5. индекс силуэта и упрощенный индекс силуэта,
6. CS,
7. VNND,
8. Score Function,
9. MB,
10. CDbw.

План доклада

- Кластеризация: основные понятия
- Оценка качества кластеризации
  - Внешние метрики
  - Внутренние метрики
  - Относительные метрики
- Сравнение метрик оценки качества
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- **Эксперименты**
- Заключение

Случайное распределение данных: K-Means

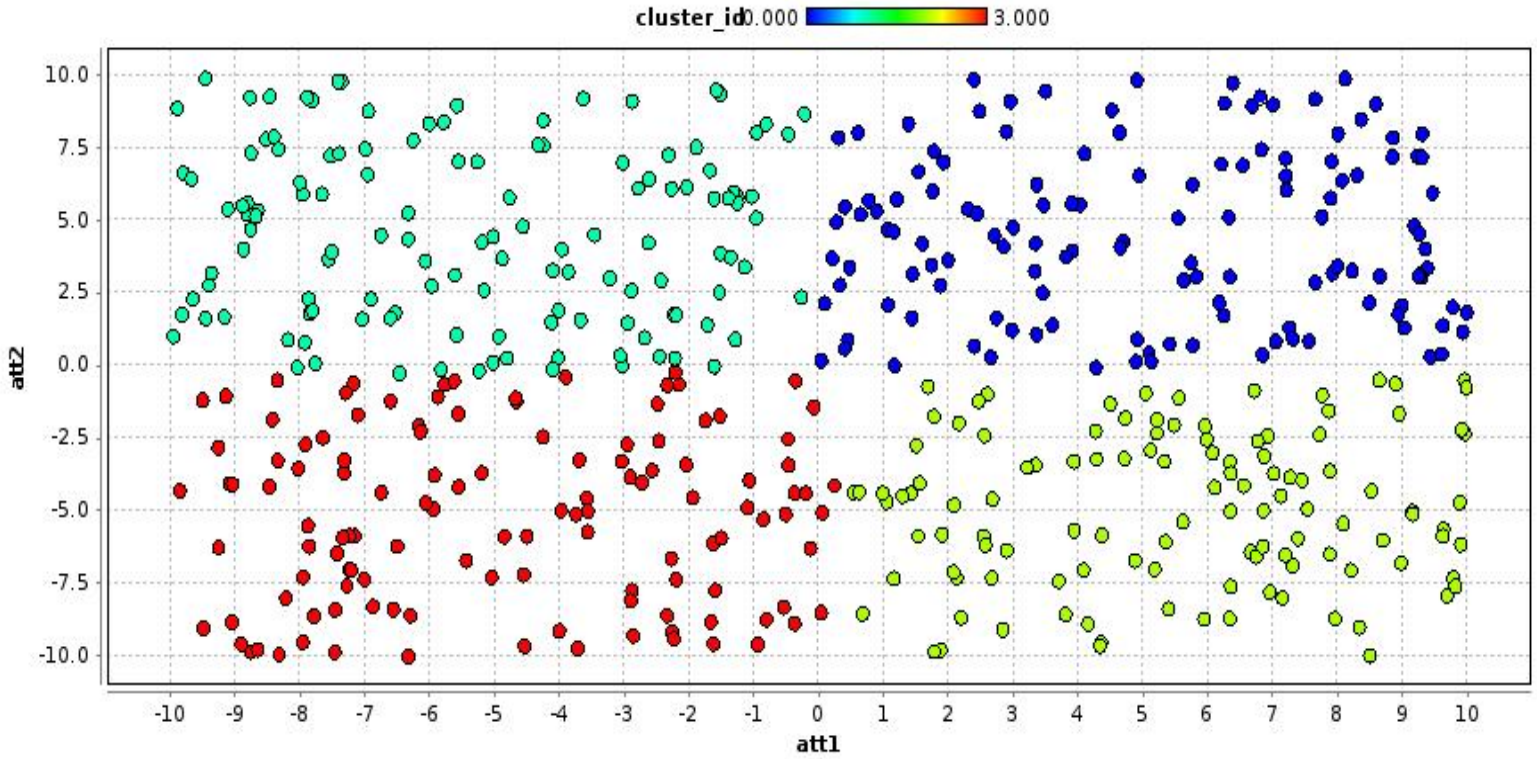


Рис.: Деление на четыре кластера при помощи K-Means.

Случайное распределение данных: K-Means

Индекс	Количество кластеров
Dunn	10
DB	9
SD	3
S_Dbw	9
Sil.	4
Simp. Sil.	4
CS	9
VNND	2
Score Func.	2
MB	4
CDbw	2

Таблица: Результаты оценки для K-Means

Случайное распределение данных: DBScan

Индекс	minpts	eps
Dunn	3	0.2
DB	3	0.2
SD	3	0.2
S_Dbw	3	0.2
Sil.	3	0.7
Simp. Sil.	3	0.7
CS	3	0.2
VNND	3	0.2
Score Func.	3	0.1
MB	3	0.2
CDbw	3	0.2

Таблица: Случайное распределенные данные: результаты для DBScan

Случайное распределение данных: DBScan

Индексы: Dunn, DB, SD, S_Dbw, CS, VNND, MB, CDbw.

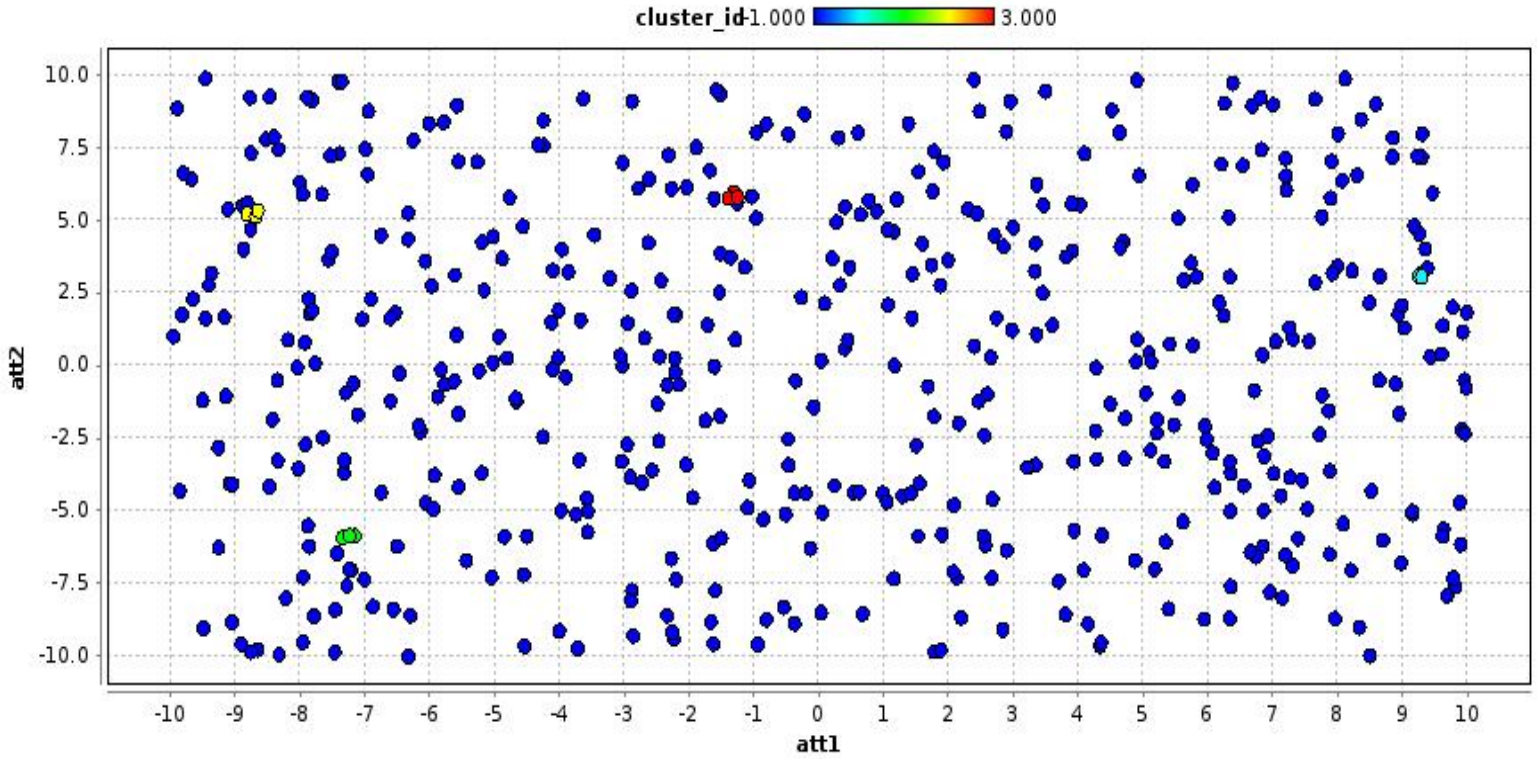


Рис.: Случайные данные: результаты для DBScan(3, 0.2)

Гауссово распределение данных: K-Means

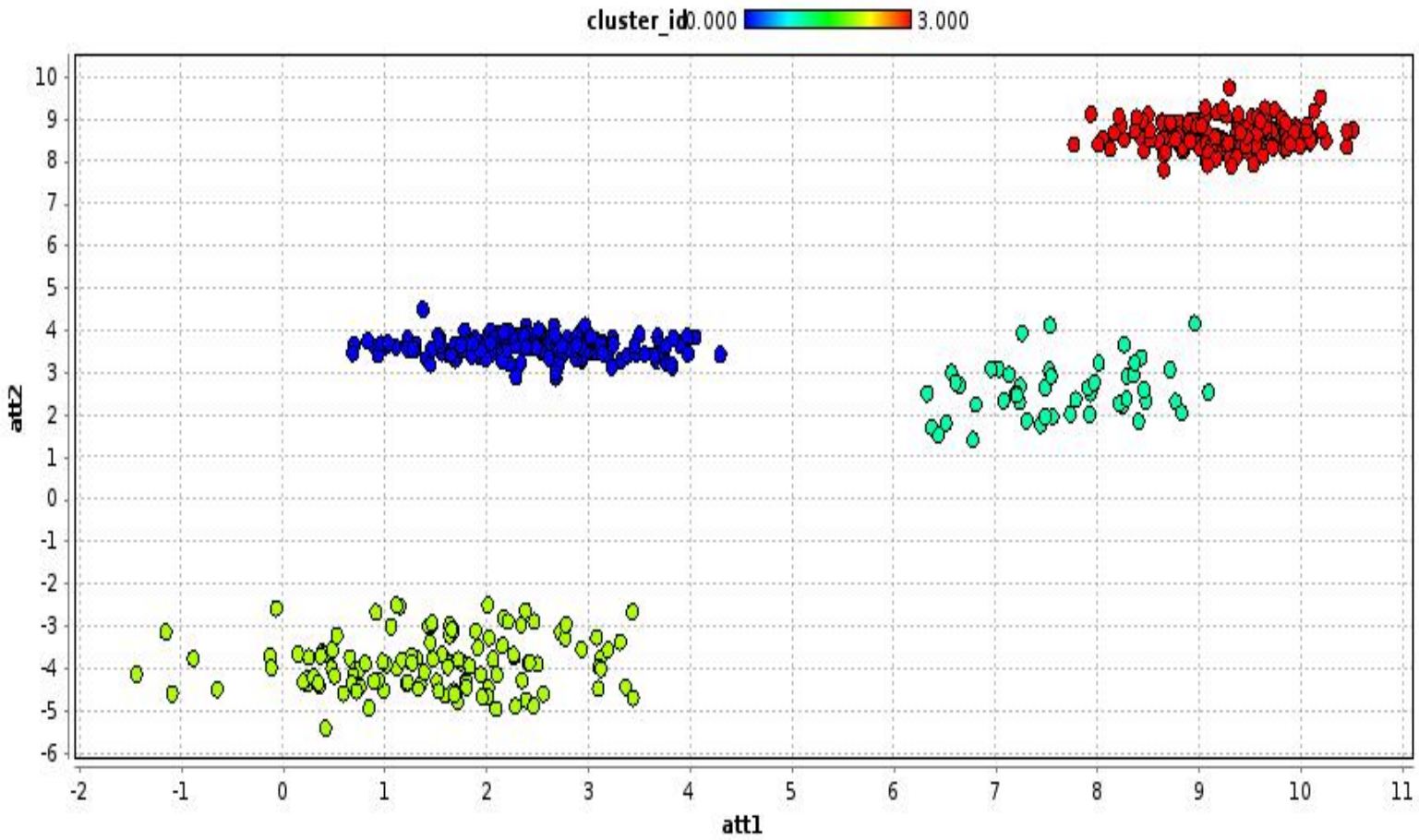


Рис.: Гауссово распределение данных: результаты для K-Means

Гауссово распределение данных: K-Means

Индекс	Количество кластеров
Dunn	4
DB	4
SD	2
S_Dbw	4
Sil.	4
Simp. Sil.	4
CS	4
VNND	2
Score Func.	4
MB	4
CDbw	4

Таблица: Гауссово распределение данных: результаты для K-Means

Гауссово распределение данных: DBScan

Индекс	minpts	eps
Dunn	5	0.5
DB	5	0.5
SD	5	0.5
S_Dbw	**6**	**0.1**
Sil.	5	0.9
Simp. Sil.	5	0.9
CS	6	0.2
VNND	**6**	**0.1**
Score Func.	**6**	**0.1**
MB	**6**	**0.1**
CDbw	6	0.6

Таблица: Гауссово распределение данных: результаты для DBScan

Гауссово распределение данных: DBScan

Индексы: S_Dbw, VNND, Score Func., MB

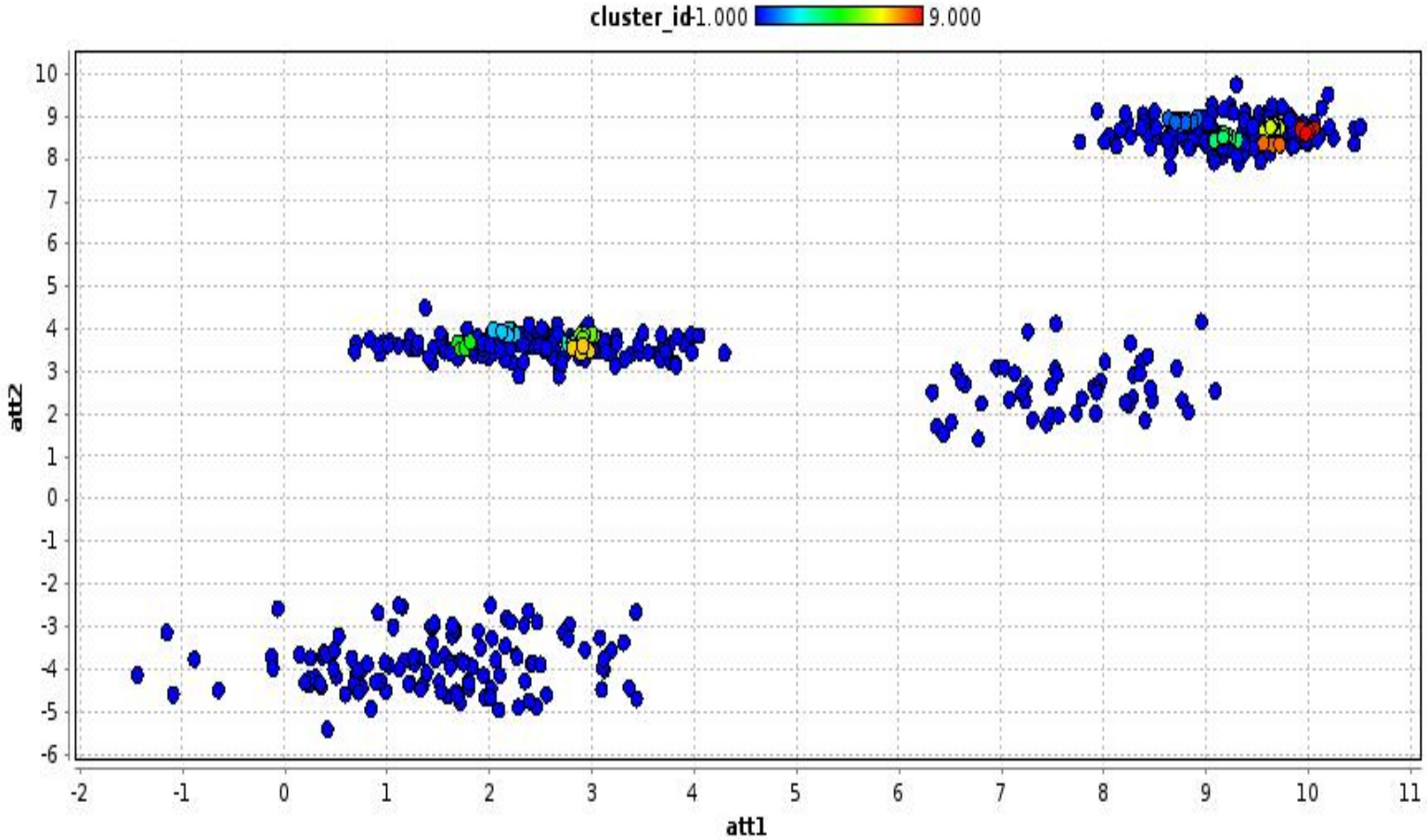


Рис.: Гауссово распределение данных: результаты для DBScan(6,0.1)

Гауссово распределение данных: DBScan

Индексы: Dunn, DB, SD

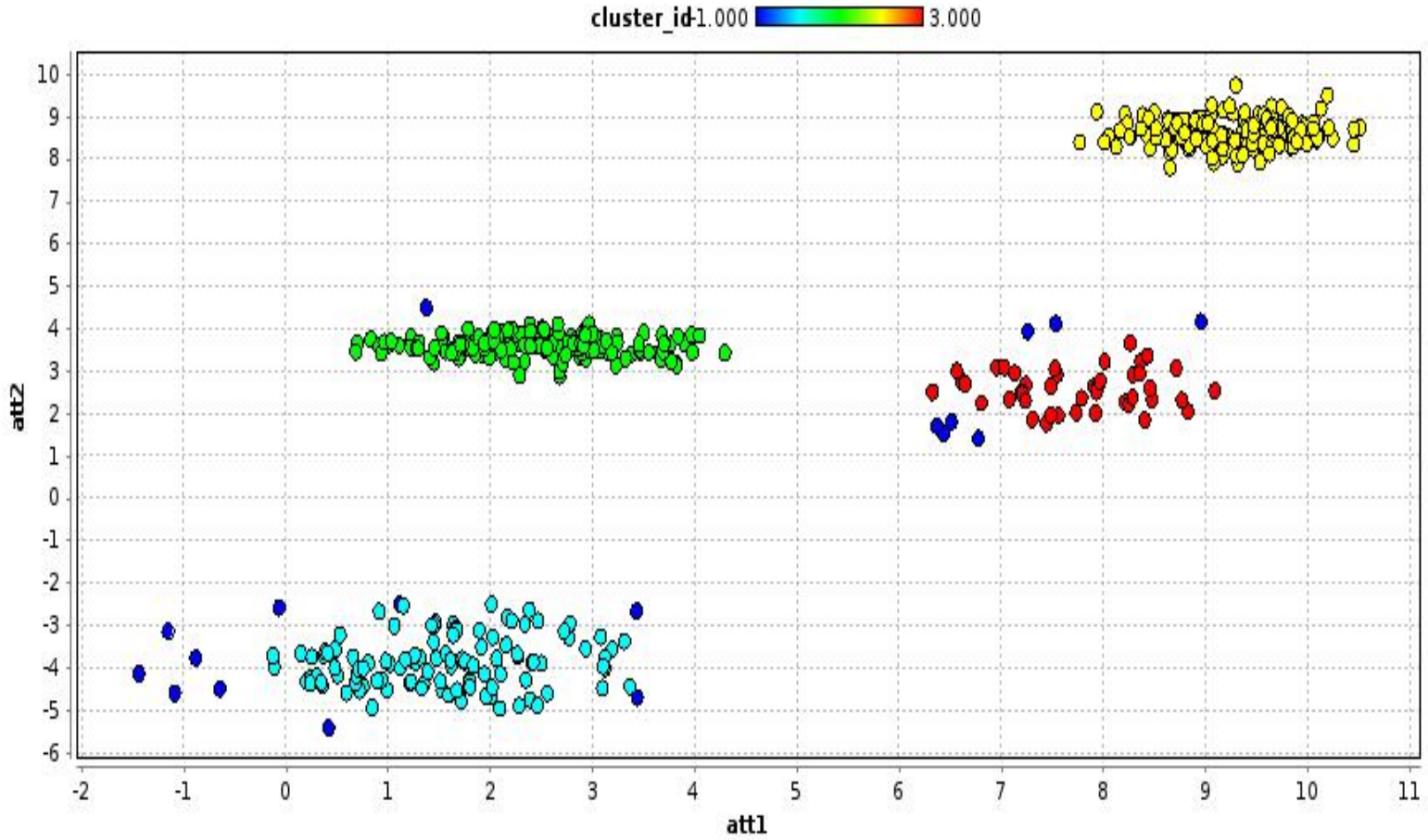


Рис.: Гауссово распределение данных: результаты для DBScan(5,0.5)

Кластеры произвольной формы: DBScan

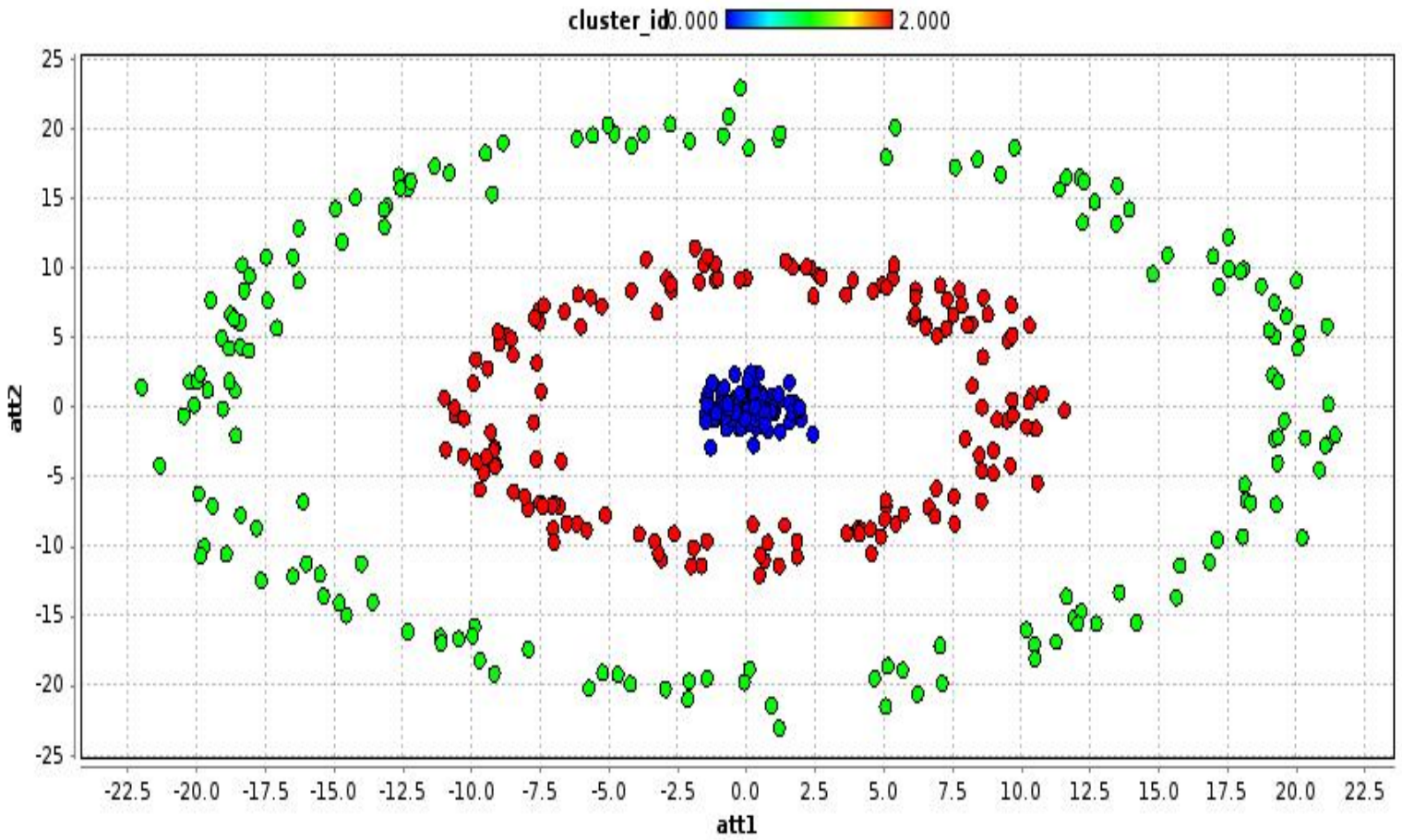


Рис.: Концентрические кластеры: результаты для DBScan(3,4.5)

Кластеры произвольной формы: DBScan

Индекс	minpts	eps
Dunn	**4**	**0.5**
DB	**4**	**0.5**
SD	**4**	**0.5**
S_Dbw	3	0.5
Sil.	4	1.5
Simp. Sil.	4	1.5
CS	**4**	**0.5**
VNND	6	0.5
Score Func.	6	0.5
MB	**4**	**0.5**
CDbw	3	4.5

Таблица: Концентрические кластеры: результаты для DBScan

Кластеры произвольной формы: DBScan

Индексы: Dunn, DB, SD, CS, MB

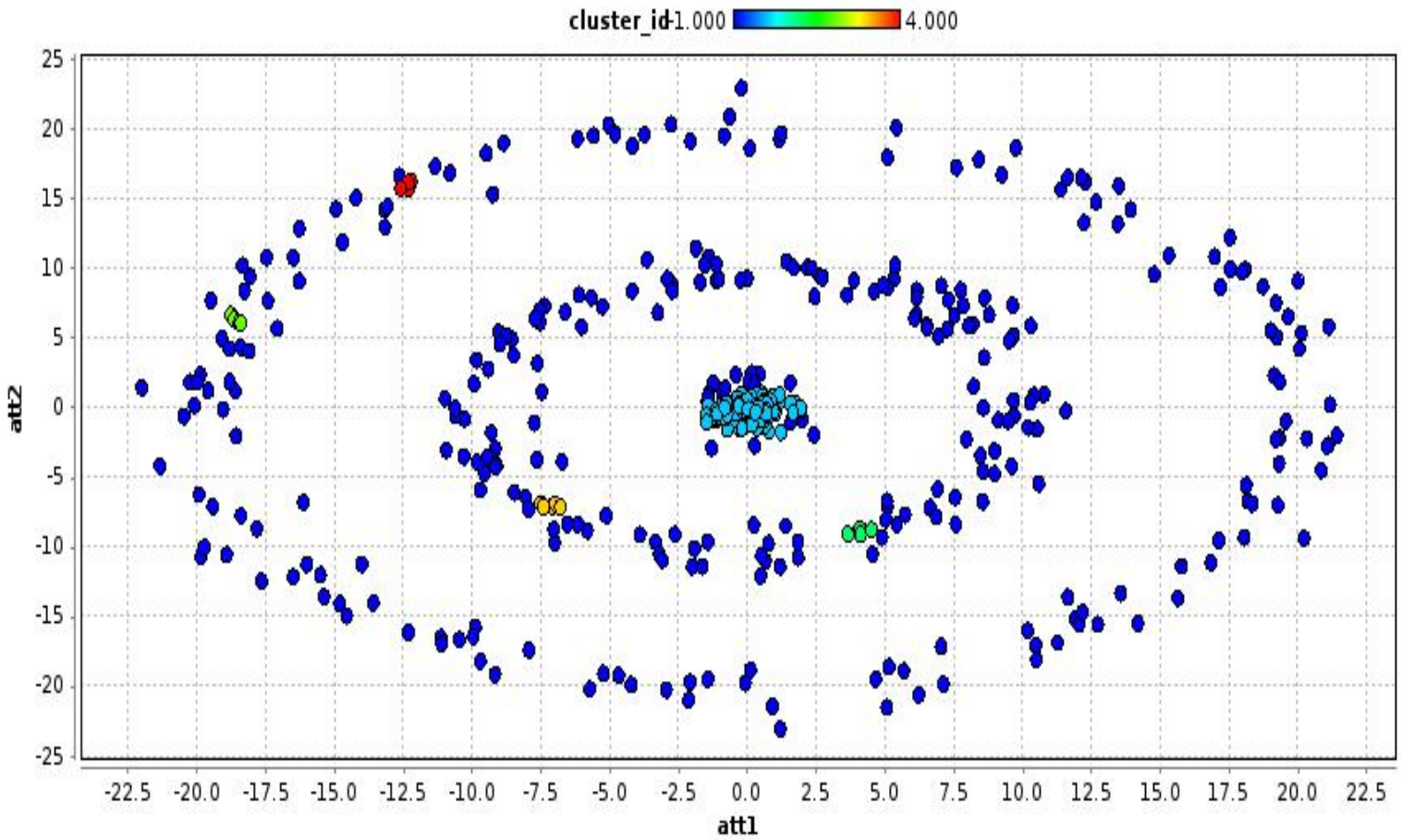
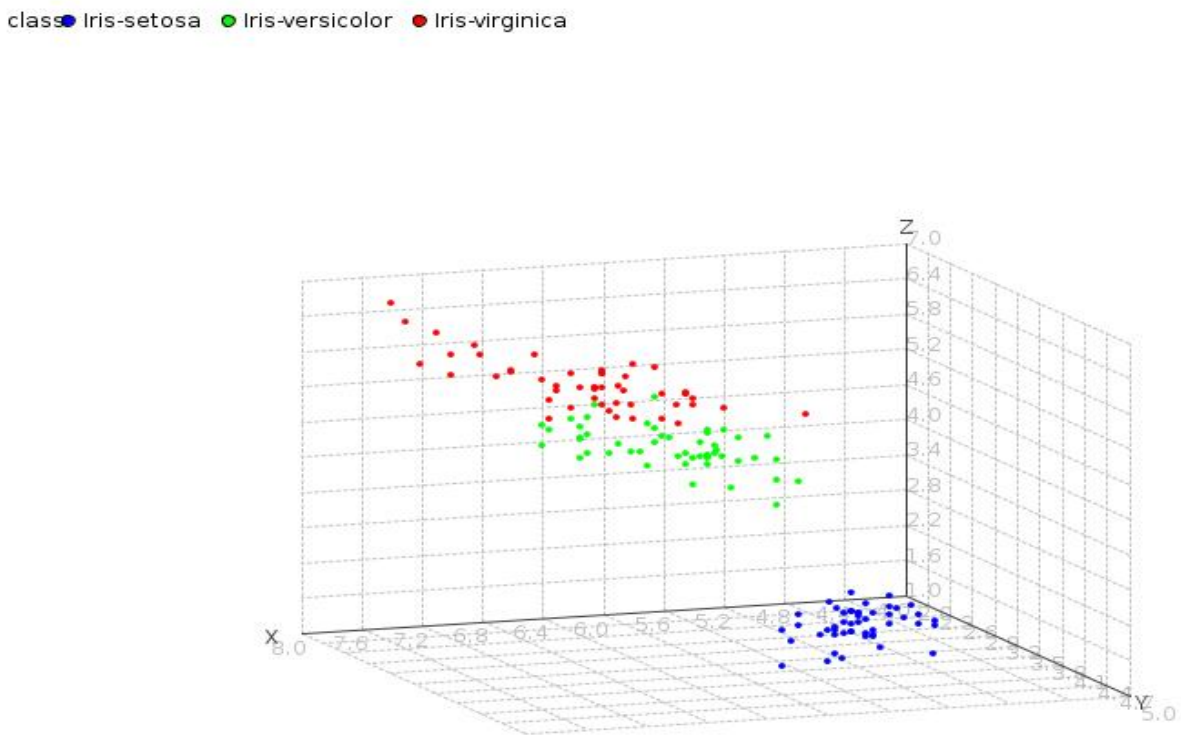


Рис.: Концентрические кластеры: результаты для DBScan(4,0.5)

IRIS

3 класса по 50 элементов.

4 аттрибута: длина чашелистика, ширина чашелистика, длина лепестка и ширина лепестка



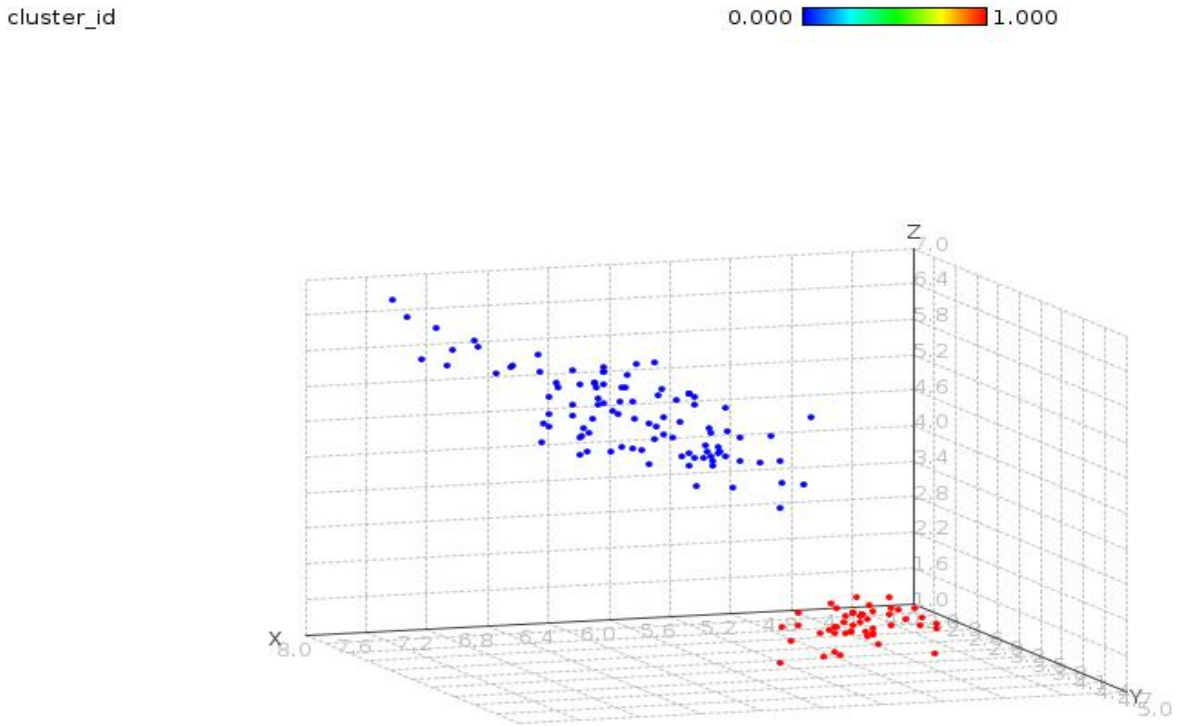
IRIS: K-Means

Индекс	Количество кластеров
Dunn	2
DB	2
SD	2
S_Dbw	10
Sil.	2
Simp. Sil.	2
CS	2
VNND	2
Score Func.	2
MB	3
CDbw	3

Таблица: IRIS: результаты для K-Means

IRIS: K-Means

Индексы: Dunn, DB, SD, CS, Sil, Simpl. Sil, CS, VNND, Score. Func



IRIS: K-Means

Индексы: MB, CDbw

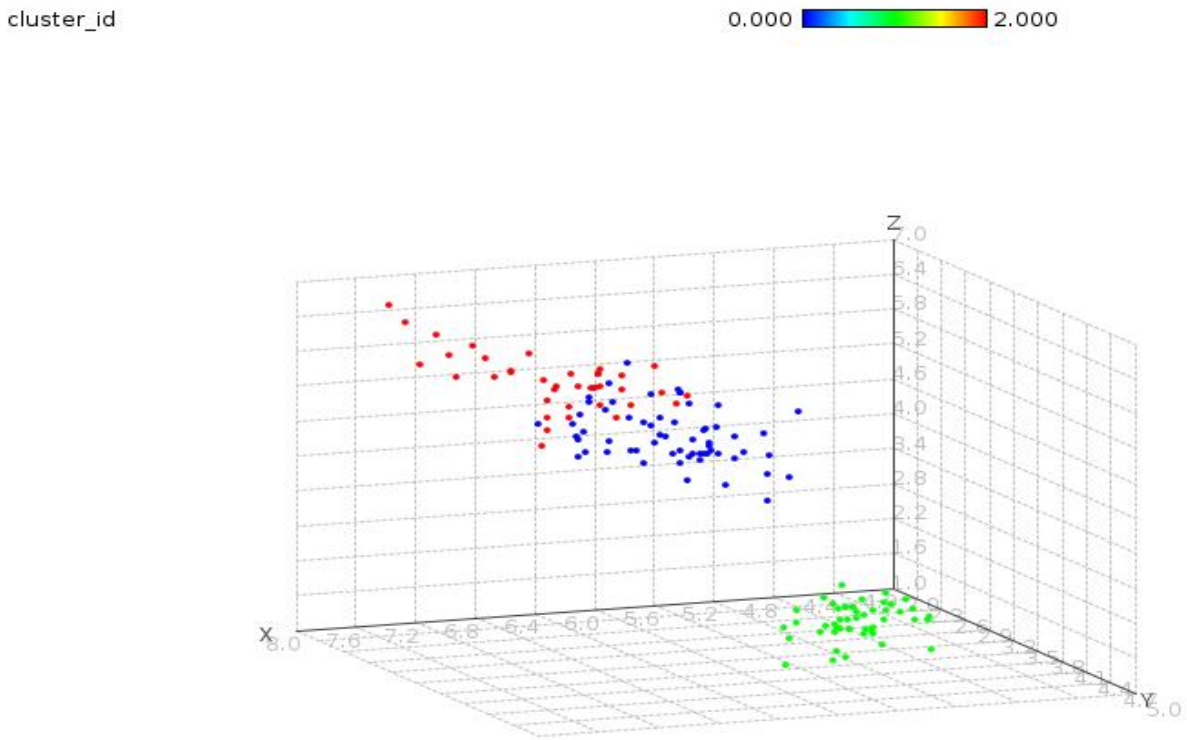


Рис.: IRIS: результаты для K-Means (3 кластера)

IRIS: DBScan

Индекс	minpts	eps
Dunn	**6**	**0.3**
DB	**6**	**0.3**
SD	5	0.6
S_Dbw	6	0.2
Sil.	3	1.5
Simp. Sil.	3	1.5
CS	4	0.2
VNND	**6**	**0.3**
Score Func.	3	0.1
MB	4	0.2
CDbw	5	0.6

Таблица: IRIS: результаты для DBScan

IRIS: DBScan

Индексы: Dunn, DB, VNND

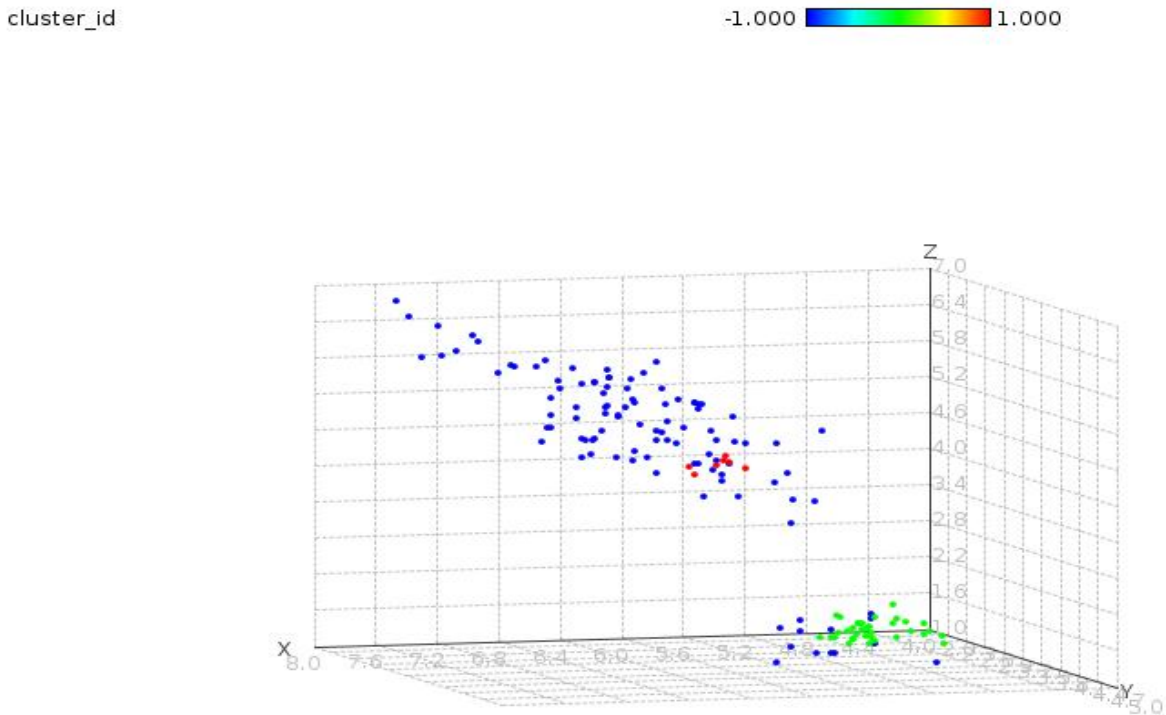


Рис.: IRIS: результаты для DBScan(6,0.3)

IRIS: DBScan

Индексы: SD, CDbw

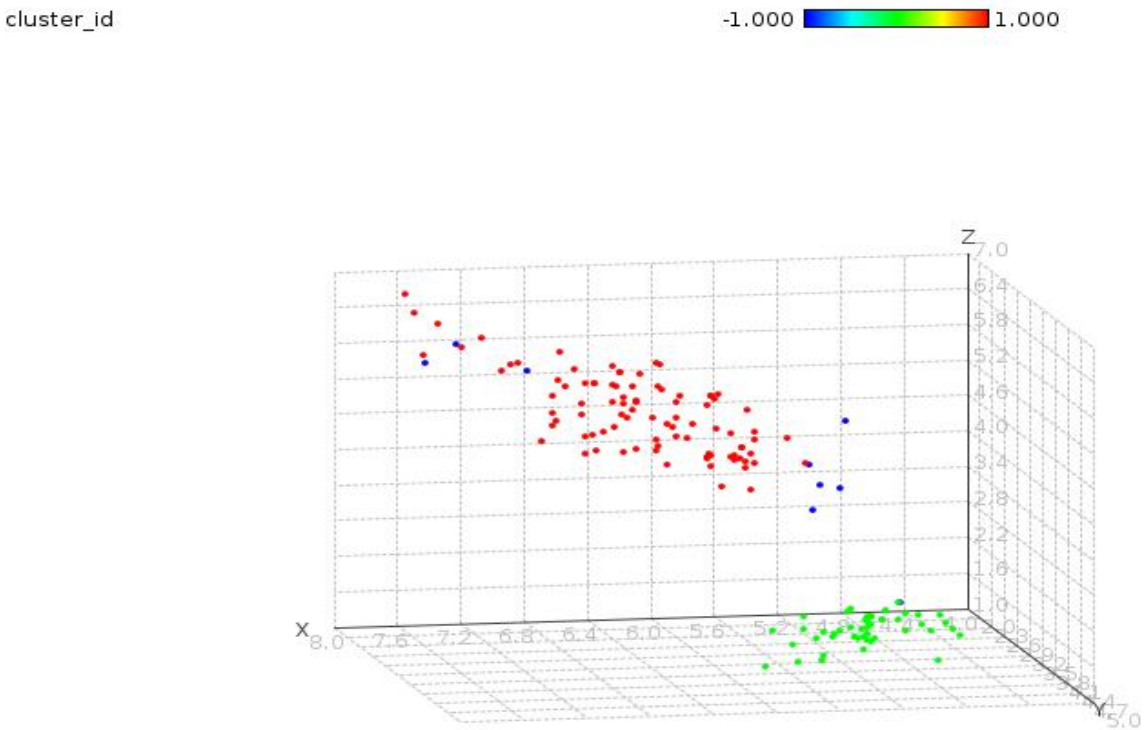


Рис.: IRIS: результаты для DBScan(5,0.6)

План доклада

- Кластеризация: основные понятия
- Оценка качества кластеризации
  - Внешние метрики
  - Внутренние метрики
  - Относительные метрики
- Сравнение метрик оценки качества
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- Эксперименты
- **Заключение**

Результаты проверки: сводная таблица

Индекс	Gaussian(4)		3 rings	IRIS	
	K-Means	DBScan	DBScan	K-Means	DBScan
Dunn	+	+	-	+	-
DB	+	+	-	+	-
SD	-	+	-	+	+
S_Dbw	+	-	-	-	-
Sil.	+	+	-	+	+
Simp. Sil.	+	+	-	+	+
CS	+	-	-	+	-
VNND	-	-	-	+	-
Score Func.	+	-	-	+	-
MB	+	-	-	+	-
CDbw	+	+	+	+	+

Результаты проверки: выводы

1. Ни один индекс не выдаст вам правильной оценки, если во множестве отсутствует структура кластеров
2. Ни один индекс не сможет корректно оценить структуру из одного кластера.
3. Не стоит использовать индексы, которые учитывают только компактность, а отделимость не учитывают.
4. Для повышения эффективности в оценке качества и получения объективного результата лучше пользоваться не одним каким-то индексом, а их совокупностью.
5. Лучшие индексы в порядке уменьшения точности оценки: 1) CDbw 2)индексы силуэта 3)Dunn и DB

Результаты проверки: рекомендации

K-Means: нет особой разницы, какой из описанных индексов брать для оценки качества – они все справятся с задачей.

DBScan: лучше использовать индексы, учитывающие геометрическую структуру кластеров (CDbw) и измеряющие компактность и отделимость в терминах средних расстояний между элементами кластеров (Silhouette).

Заключение

- библиотека индексов
- сравнительный анализ
- рекомендации

Future work

- алгоритмы других типов
- разные виды данных
- семантика