УДК 004.051

Сивоголовко Елена Владимировна

МЕТОДЫ ОЦЕНКИ КАЧЕСТВА ЧЁТКОЙ КЛАСТЕРИЗАЦИИ*

Аннотация

В данной работе производится сравнительный анализ наиболее известных индексов оценки качества кластеризации. Рассматривается эффективность индексов качества для множеств с различной структурой и делаются выводы о том, индексы с какими свойствами наиболее точны при оценке разбивающих или плотностых алгоритмов кластеризации.

Ключевые слова: кластеризация, оценка качества кластеризации, методы оценки качества кластеризации.

1. ВВЕДЕНИЕ

Кластеризация является одним из мощнейших вспомогательных средств в сфере анализа данных. В настоящее время множество различных алгоритмов кластеризации с успехом используются в самых разных научно-технических областях от физики до лингвистики и психологии. Кластеризацию используют и как самостоятельный инструмент анализа данных, и как предварительный этап для других методов анализа, таких как, например, классификация или деревья решений. Однако при всем этом оценка качества кластеризации является мало разработанной областью, и зачастую вопрос о том, насколько хороша или плоха структура кластеров, приходится решать «вручную», вне зависимости от того, являются ли данные точками на двумерной плоскости или же кусками генетической последовательности, лежащими в пространстве высокой размерности. Как известно, общая схема кластеризации включает в себя четыре пункта:

1. Выделение характеристик.
2. Определение метрики.
3. Разбиение объектов на группы.
4. Представление результатов.

В сущности, на каждом из первых трех этапов мы можем допустить существенные ошибки, которые могут серьезно исказить итоговый результат. Мы можем неправильно выделить интересующие нас характеристики, что, например, часто бывает при создании векторного представления для текстов. Мы можем неправильно подобрать метрики сравнения: так, в случае с кластеризацией изображений евклидова метрика не дает адекватного представления о сходстве двух рисунков. И также алгоритм разбиения на группы может допустить какие-

* Работа была частично поддержана грантом РФФИ №10-07-00156.

то неточности в процессе работы. Исходя из этого, кажется странным, что оценке качества кластеризации уделяется так мало внимания, поскольку каждый раз, получая кластерную структуру, мы хотим быть уверенными в том, что эта структура корректна, отвечает нашим требованиям и годится для дальнейшего использования.

Основная задача кластеризации обычно формулируется так: разделить объекты на группы таким образом, чтобы сходство между объектами одной группы было велико, а сходство между объектами разных групп – мало. Исходя из этого, понятие качества кластеризации состоит из следующих пунктов

Компактность: элементы кластера должны быть как можно ближе друг к другу. Это свойство можно выразить через расстояния между элементами в кластере, плотностью внутри кластера или же объемом, занимаемым кластером в многомерном пространстве.

Отделимость: расстояние между различными кластерами должно быть как можно больше. Это правило применяется как для четкой так и для нечеткой кластеризации. В первом случае расстояние между кластерами обычно измеряется одним из трех следующих способов: 1) как расстояние между ближайшими элементами кластеров, 2) как расстояние между наиболее удаленными друг от друга элементами кластеров и 3) как расстояние между кластерными центрами. В случае нечеткой кластеризации для измерения межкластерного расстояния обычно используется матрица принадлежности.

[*Концентрация*: элементы кластера должны быть сконцентрированы вокруг центра кластера. Этот пункт используется гораздо реже, чем первые два, потому что далеко не во всех алгоритмах кластеризации используется понятие центра кластера].

Для самих метрик качества обычно вводят следующую классификацию:

Внутренние – к ним относятся метрики, которые при оценке качества используют какую-либо уже известную информацию о структуре кластеров, существующей в рассматриваемом множестве. Как правило, такие метрики применяются при оценке эффективности работы алгоритма кластеризации, когда в качестве тестового множества используется какое-либо множество данных с известной структурой классов.

Внешние – к ним относятся метрики, которые не имеют априори знаний о структуре классов и при оценке опираются только на ту информацию, которую можно получить опираясь на множество данных.

Относительные – оценивают качество, сравнивая несколько кластерных структур между собой, не имея априорной информации и принимая в расчет только сведения о кластерной структуре и кластеризуемом множестве.

Это деление является несколько нечетким в том смысле, что «внешняя» метрика, применяемая для разных структур, может быть рассмотрена как относительная, и «относительная» метрика, основанная на сравнении показателей, получаемых для каждой структуры по отдельности, может быть использована как внешняя. В случае с внешними метриками не происходит сравнения с показателями для других структур: для признания структуры качественной можно использовать либо метод Монте-Карло, либо просто какой-то определенный порог для значений индекса качества.

В данной статье мы будем рассматривать относительные метрики, предназначенные для оценки качества алгоритмов четкой кластеризации, то есть той в которой не допускается пересечение кластеров, также мы проведем некоторый сравнительный анализ по использованию тех или иных метрик для различных классов алгоритмов кластеризации.

2. ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ

Используемые обозначения:
X – кластеризуемое множество,
N – количество элементов во множестве X,

c – число кластеров,

n_{c_i} – число элементов в кластере c_i,

v_i – центр кластера c_i: $v_i = \dfrac{\sum_{x \in c_i} x}{n_{c_i}}$,

$\bar{X}$ – центральный элемент множества $\bar{X} = \dfrac{1}{N}\sum_{j=1}^{N} x_j$,

$\bar{v}$ – центр центров $\bar{v} = \dfrac{1}{c}\sum_{i=1}^{c} v_i$,

d – размерность множества X,

$\|\cdot\|$ – евклидова метрика в X.

Как уже говорилось ранее, оценка качества кластеризации является менее популярной областью научного исследования, чем сам кластерный анализ как таковой, поэтому индексов оценки качества в принципе существует не очень много. Основные из них приведены в следующем списке.

1. ***Модифицированная Hubert*** Γ ***Statistic*** [1]

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

здесь $M = N * \dfrac{(N-1)}{2}$, P – матрица близости *(proximity matrix)* для кластеризуемого множества, Q – матрица $N \times N$, в которой элемент (i,j) соответствует расстоянию между центрами кластеров (v_{c_i}, v_{c_j}), к которым принадлежат элементы – x_i и x_j соответственно. Чем выше значение статистики, тем лучше структура кластеров.

Так может использоваться нормализованная версия статистики

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

Здесь μ_P, μ_Q, σ_P и σ_Q – суть средние значения и среднеквадратичные отклонения для матриц P и Q соответственно. $\hat{\Gamma}$ принимает значения в промежутке [–1, 1]. Точно так же, как и для не нормализованного варианта, большие значения $\hat{\Gamma}$ подразумевают лучшую структуру кластеров.

2. ***Calinski-Harabasz индекс*** [2]. Пусть $\bar{d}^2$ – средний квадрат расстояния между элементами в кластеризуемом множестве и $\bar{d}_{c_i}^2$ – средний квадрат расстояния между элементами в кластере c_i. Тогда сумма расстояний внутри групп

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

и сумма расстояний между группами

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

где $A_c = \dfrac{1}{N-c}\sum_{i=1}^{c}(n_{c_i}-1)(\bar{d}^2 - \bar{d}_{c_i}^2)$ – взвешенная средняя разница расстояний между центрами кластеров и общим центром множества. Тогда определим формулу индекса как

$$VRC = \frac{\frac{BGSS}{c-1}}{\frac{WGSS}{N-c}} = \frac{\bar{d}^2 + \frac{N-c}{c-1}A_c}{\bar{d}^2 - A_c} = \frac{1+\frac{N-c}{c-1}a_c}{1-a_c},$$

где $a_c = \dfrac{A_c}{\bar{d}^2}$. Легко видеть, что если все расстояния между точками одинаковы, то $a_c = 0$ и

$VRC = 1$. $a_c = 1$ исключительно для «идеальной» кластеризации, когда в кластерах нет никакого отклонения. При нормальном распределении данных a_c медленно, но постоянно возрастает при увеличении c. Однако, VRC убывает при постоянном a_c и возрастающем c, что немножко балансирует рост a_c в случае нормального распределения. Максимальное значение индекса VRC соответствует оптимальной структуре кластеров.

3. ***Индекс Данна*** [3]

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

В этом случае d – расстояние между кластерами c_i и c_j, определяющееся, например как расстояние между двумя их ближайшими элементами $d(c_i,c_j) = \min_{x\in c_i, y\in c_j} \| x-y \|$, $diam(c_i)$ – диаметр кластера – может быть рассчитан как максимальное расстояние между элементами одного кластера: $diam(c_i) = \max_{x,y\in c_i} \| x-y \|$. Таким образом, индекс Данна сравнивает межкластерное расстояние с диаметром кластера. Считается, что если диаметр кластера мал по сравнению с межкластерным расстоянием, то кластеры полученной структуры достаточно компактны и отделимы. Следовательно, чем больше значение индекса, тем лучше кластеризация. Следует заметить, что сам по себе индекс Данна чувствителен к шуму и выбросам в данных. Чтобы устранить эту погрешность, были созданы некоторые модификации индекса, изменения в которых касались измерения межкластерного расстояния. Так, например, одним из вариантов подсчета является использование минимального остовного дерева. В этом случае для каждого кластера c_i строится граф G_i, в вершинах которого находятся элементы кластера, а вес для каждого ребра e_{jk} определяется как расстояние между x_j и x_k соответственно: $w(e_{jk}) = \| x_j - x_k \|$. Затем для графа G_i строится минимальное остовное дерево MST_{G_i}, и диаметр кластера принимается равным максимальному весу ребра в MST_{G_i}: $diam(c_i) = \min_{e\in E(MST_{G_i})} w(e)$. Также вместо минимального остовного дерева можно использовать относительный граф ближайших соседей (relative neighbourhood graph) и граф Габриэля (Gabriel graph) [4].

4. ***Индекс Девида-Болдуина*** [5]

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 кластера c_i и $d_{ij} = \left\{\sum_{k=1}^{d}(v_i^k - v_j^k)^p\right\}^{\frac{1}{p}}$ – мера различия между кластерами (dissimilarity measure), тогда мерой схожести между кластерами c_i и c_j может являться любая функция R_{ij}, удовлетворяющая следующим условиям:

a) $R_{ij} \geq 0$,
b) $R_{ij} = R_{ji}$,
c) при $S_i = 0$ и $S_j = 0$ $R_{ij} = 0$,
d) при $S_j > S_k$ и $d_{ij} = d_{ik}$ $R_{ij} > R_{ik}$,
e) при $S_j = S_k$ и $d_{ij} < d_{ik}$ $R_{ij} > R_{ik}$.

Сами авторы предлагают следующий вариант подсчета схожести:

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

Тогда сам индекс вычисляется по формуле:

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

где $R_i = \max_{i,\, j\in\{1\ldots c\}, i\neq j}(R_{ij})$. Как видно из определения, DB индекс определяет среднюю схожесть между кластером c_i и наиболее близким к нему кластером. Поскольку подразумевается, что кластеры в структуре значительно отличаются друг от друга, наилучшей будет структура с минимальным DB. Существуют также другие варианты подсчета разброса внутри кластера, в частности, с применением такой же методики на графах, как и для индекса Данна [4].

5. ***CS validity index*** [6]

Создан как сборка из индексов Данна и Деви-Болдуина

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

Измеряет отношение максимального расстояния между точками в одном кластере к минимальному межкластерному расстоянию. Оптимальная структура характеризуется меньшим показателем *CS*.

6. ***PS validity index*** [7]

$$PS = \frac{1}{c}\sum_{i=1}^{c}\left[\frac{1}{n_{c_i}}\sum_{x_j\in c_i}\frac{d_s(x_j, v_i)\,\| x_j - v_i \|}{\min_{m\neq n}\| v_m - v_n \|}\right],$$

где $d_s(x_j, v_i)$ – расстояние симметрии (point symmetry distance) для точки x_i, вычисляемое по формуле

$$d_s(x_j, v_i) = \min_{x_k\in c_i, k\neq j}\left\{\frac{\| (x_j - v_i) + (x_k - v_i) \|}{\| x_j - v_i \| + \| x_k - v_i \|}\right\}.$$

Чем меньше показатель *PS*, тем лучше структура кластеров.

7. ***SD индекс*** [8]

Определим дисперсию на множестве и дисперсию внутри кластера следующим образом:

Дисперсия на множестве:

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

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

Дисперсия внутри кластера:

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

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

Средний разброс для кластеров определяется так:

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

Несложно понять, что разброс в кластере служит мерой его компактности. Чем меньше значение *Scatt*, тем компактнее кластер.

Отделимость кластеров измеряется так:

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

Собственно SD индекс:

$$SD = \alpha * Scatt + Dist,$$

где α – взвешивающий параметр (равен $Dist(c_{\max})$, если идет проверка нескольких кластерных структур на более вероятное число кластеров). Появление α объясняется тем, что в общем случае значение $Dist$ возрастает с количеством кластеров, и весовой коэффициент необходим, для того чтобы сбалансировать $Dist$ и $Scatt$.

Низкое значение индекса SD означает лучшее разбиение, так как в этом случае кластеры являются компактными ($Scatt$ параметер) и отделимыми ($Dist$ параметер).

8. ***S_Dwb индекс*** [9] использует для определения качества кластеризации два понятия: дисперсия по кластеру и плотность между кластерами. Дисперсия определяется точно так же, как и в SD индексе:

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

Плотность между кластерами определяется следующим образом:

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

Здесь u_{ij} – есть середина линии, соединяющей кластерные центры v_i и v_j. В свою очередь, функция плотности $dens(u_{ij}) = \sum_{x \in c_i \cup c_j} f(x, u_{ij})$. Функция $f(x, u_{ij})$ определяет окрестность точки u_{ij} как гиперсферу с центром в u_{ij} и радиусом, равным среднему стандартному отклонению по всем кластерам $stdev = \frac{1}{c}\sqrt{\sum_{i=1}^{c}\| \sigma_{v_i} \|}$:

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

Таким образом, $Dens\_bw$ рассматривает среднее количество элементов между кластерами как меру отделимости кластеров друг от друга. Чем лучше кластеры отделены – тем ниже значение $Dens\_bw$. Общая формула индекса для индекса S_Dwb:

$$S\_Dwb = Dens\_bw + Scatt.$$

Поскольку малые значения $Dens\_bw$ и $Scatt$ характеризуют лучшую структуру кластеров, то и все значение индекса должно быть минимальным для получения наилучшего результата.

9. ***RMSSTD и RS индексы*** [10]

$RMSSTD$ (root – mean – square standard deviation) индекс вычисляется через сумму дисперсий (pooled sample variance) по всем аттрибутам данных, используемых в процессе кластеризации. Индекс измеряет неоднородность полученных кластеров на каждом шаге алгоритма(изначально он использовался для иерархических алгоритмов кластеризации). Поскольку задача кластеризации – это выделение однородных групп данных, $RMSSTD$ индекс должен быть как можно ниже, если его значение повышается, значит полученная на следующем шаге кластерная структура будет хуже уже существующей.

$$RMSSTD = \sqrt{pooled\ \text{var}iance} = \left[\frac{pooled\ sum\ of\ squares\ for\ all\ variables}{pooled\ degrees\ of\ freedoms\ for\ all\ variables}\right]^{1/2}.$$

Итоговая формула:

$$RMSSTD = \left[\frac{\sum_{i\leq c,\, p\leq d}\sum_{j=1}^{n_{c_i p}}(x_j^p - \bar{v}_i^p)^2}{\sum_{i\leq c,\, p\leq d}(n_{c_i p} - 1)}\right]^{1/2} = \left[\frac{\sum_{i\leq c}\sum_{p\leq d}\sum_{j=1}^{n_{c_i p}}(x_j^p - \bar{v}_i^p)^2}{d * \sum_{i\leq c}(n_{c_i} - 1)}\right]^{1/2}.$$

Здесь d – размерность пространства, x_j^p – компонента p вектора x_j.

Индекс *RS* (R Squared) измеряет несхожесть между кластерами, его значение варьируется от 0 до 1, где значение 0 свидетельствует о том, что различие между кластерами очень невелико, то есть полученная структура неинформативна. Он состоит из трех основных частей:

a) SS_w – сумма квадратов расстояний внутри кластера,

b) SS_b – сумма квадратов расстояний между кластерами,

c) SS_t – сумма квадратов расстояний по всему множеству, причем $SS_t = SS_w + SS_b$.

Формула индекса *RS*:

$$RS = \frac{SS_b}{SS_t} = \frac{SS_t - SS_w}{SS_t},$$

где $SS_t = \sum_{p=1}^{d}\sum_{j=1}^{N}(x_j^p - \bar{x}^p)^2$, $SS_w = \sum_{p=1}^{d}\sum_{i=1}^{c}\sum_{j=1}^{n_{ci}}(x_j^p - v_i^p)^2$.

Изначально *RMSSDT*, *SPR*, *RS*, *CD* индексы вводились как индексы для иерархических алгоритмов: они использовались, для того чтобы понять на какой итерации алгоритма следует остановиться, однако не имеется никаких противопоказаний, к тому чтобы использовать их и для оценки качества кластеризации в целом. Однако, в случае не-иерархических алгоритмов наилучшая структура выбирается не по предельному минимальному и максимальному значению, поскольку, вообще говоря, и *RMSSTD* и *RS* монотонно возрастают или убывают, соответственно, в зависимости от числа кластеров, а по «пику» («knee») на графике индекса [11], что несколько затрудняет оценку структуры с их помощью, потому что «пик» не всегда легко определить.

10. ***Индекс оценки силуэта*** (Silhouette index) [12]

«Силуэт» каждого кластера определяется следующим образом: допустим элемент x_j принадлежит кластеру c_p. Обозначим среднее расстояние от этого объекта до других объектов из того же кластера c_p через a_{pj}. Теперь обозначим среднее расстояние от x_j до объектов из другого кластера c_q, $q \neq p$ через $d_{q,j}$. Положим $b_{pj} = \min_{q \neq p} d_{qj}$. Смысл этой величины можно определить как меру несхожести отдельного элемента с элементами ближайшего кластера. Таким образом, «силуэт» каждого отдельного элемента определяется как

$$S_{x_j} = \frac{b_{pj} - a_{pj}}{\max(a_{pj},\, b_{pj})}.$$

Знаменатель введен в целях нормализации. Очевидно, что высокое значение показателя S_{x_j} характеризует собой «лучшую» принадлежность элемента x_j к кластеру p. Оценка для всей кластерной структуры достигается усреднением показателя по элементам:

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

Лучшее разбиение характеризуется максимальным SWC, что достигается когда расстояние внутри кластера a_{pj} мало, а расстояние между элементами соседних кластеров b_{pj} велико. Также на практике используются вариации индекса силуэта: упрощенный силуэт (Simplified Silhouette) и альтернативный силуэт (Alternative Silhouette) [13]. В первом случае при определении a_{pj} и b_{pj} используются расстояния не между одним элементом и всеми остальными элементами кластеров, а между элементом и центроидом соответствующего кластера. Во втором случае $S_{x_j} = \frac{b_{pj}}{a_{pj} + \varepsilon}$. ε – малая константа, введенная для того чтобы избежать деления на 0, если $a_{pj} = 0$.

11. ***Индекс Maulik-Bandoypadhyay*** [14, 15]

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

Здесь $E_c = \sum_{i=1}^{c} \sum_{x \in c_i} \| x - v_i \|$ – измерение внутрикластерного расстояния (inter-cluster distance). По аналогии с этим E_1 – сумма расстояний от центра множества до каждого элемента. $D = \max_{i,j=1} \| v_i - v_j \|$ – максимальное расстояние между кластерами. Верное количество кластеров определяется высоким показателем MB. Константа p определяется произвольно, сами авторы используют $p = 2$.

12. ***Score function*** [16].

Определим расстояние между кластерами так:

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

Размер кластера n_{c_i} позволяет ограничить чувствительность к выбросам. Деление на N уменьшает влияние общего числа элементов, и c используется, для того чтобы учесть «пенальти» от добавления нового кластера и избежать тупиковых ситуаций вида: один элемент – один кластер.

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

Стандартный подход для измерения близости точек внутри кластера. Кластерная структура является хорошей, если bcd высокий, а wcd – низкий.

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

Чем выше SF, тем лучше структура кластеров.

13. ***Индекс плотности CDbw*** [17]

Учитывает геометрическую структуру кластеров с помощью выборки «представителей» (representatives) $V_{c_i} = \{v_1 ... v_r\}$, где r – произвольное число, но, как утверждают авторы, более информативно брать $r \geq 10$. Список представителей составляется итеративно: сначала берется наиболее удаленный от центра кластера, а потом – наиболее удаленные от уже добавленных в список. Для расчета отделимости и плотности между кластерами c_i и c_j используется множество соответствующих ближайших представителей:

$$RCR_{ij} = \{(v_{ik}, v_{jl}) \mid v_{ik} = closest\_rep^i(v_{jl}) \,\&\, v_{jl} = closest\_rep^j(v_{ik})\}.$$

Обозначение $v_{ik} = closest\_rep^i(v_{jl})$ означает, что представитель v_{ik} является ближайшим представителем кластера c_i по отношению к v_{jl}. Таким образом, RCR_{ij} представляет собой как бы «границу» из представителей между двумя кластерами. 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 $stdev = \frac{1}{c} \sqrt{\sum_{i=1}^{c} \| \sigma_{v_i} \|}$, u_{kl} – середина отрезка между точками v_l и v_k, и

$$cardinality(u_{kl}) = \frac{\sum_{x \in c_i \cup c_j} f(x, u_{kl})}{n_{c_i} + n_{c_j}}, \text{ где } f(x,u) = \begin{cases} 0, \text{ если } \| x - u_{kl} \| > stdev, \\ 1, \text{ в другом случае.} \end{cases}$$

Межкластерная плотность для всей структуры определяется как

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

В свою очередь, расстояние между кластерами вычисляется как

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

и общая мера отделимости для всей структуры:

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

Компактность кластера измеряется с помощью сдвинутых «представителей» (shifted representatives): для $v_k \in V_{c_i}$

$$v_k^s = v_k + s*(center(c_i) - v_k),$$

где $s \in [0,1]$. Необходимо провести несколько итераций сдвига. Допустим, $0.1 \le s \le 0.8$, $s_i = s_{i-1} + 0.1$. Обозначим число таких итераций за n_s. Тогда плотность структуры кластеров относительно s будет равна

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

и компактность кластерной структуры определяется как

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

Также посмотрим, как меняется плотность кластера в зависимости от s:

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

и определим «связанность» кластера как

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

Тогда общая формула индекса будет

$$CDbw = Cohension * Sep * Compactness.$$

Оптимальным разбиением будет то, у которого максимальный *CDbw*. Следует заметить, что значение индекса не определено при $c = 1$, и, по утверждению авторов, индекс реально рассчитать за $O(n)$.

14. ***VNND индекс*** [18].

В этом случае авторы решили измерять компактность кластера, исходя из ближайших соседей для каждого кластерного элемента.

Введем несколько определений. Пусть $d_{\min}(x_j) = \min_{y\in c_i}(\| x_i - y \|)$ – расстояние от элемента x_j до его ближайшего соседа. Тогда $\overline{d_{\min}(c_i)} = \frac{1}{n_{c_i}}\left(\sum_{x_j\in c_i} d_{\min}(x_j)\right)$ – среднее расстояние между ближайшими соседями в кластере c_i. Отклонение для расстояния между ближайшими соседями будет

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

Итоговое значение индекса определяется как $VNND = \sum_{i=1}^{c} V(c_i)$.

Индекс измеряет однородность кластеров. Чем ниже его значение, тем больше однородность и, соответственно, лучше структура кластеров. Однако индекс совершенно не учитывает отделимость и потому не сможет опознать случай, когда два компактных, хорошо отделимых кластера оказались объединены в один.

3. СРАВНЕНИЕ ИНДЕКСОВ

3.1. ИНСТРУМЕНТЫ ТЕСТИРОВАНИЯ

3.1.1. Тестовые множества

Процесс тестирования кластерных индексов не задокументирован, однако традиционно проверка проводится в два этапа: 1) на синтетически сгенерированных и 2) на реально существующих данных. Синтетические данные обычно создаются, для того чтобы, во-первых, смоделировать «идеальное» для кластеризации множество: с плотными, отделимыми кластерами, элементы в которых, как правило, имеют гауссово распределение, и, во-вторых, смоделировать всевозможные ситуации, «мешающие» алгоритму кластеризации: кластеры с «шумом», нечетко отделяемые кластеры, вложенные кластеры, кластеры произвольной формы и т. д. Мы используем следующие тестовые множества:

1. *Равномерно распределенные данные*: своего рода «ночной кошмар» для алгоритмов кластеризации, в сущности, их бессмысленно как-либо кластеризовать, однако представляет интерес, как поведут себя кластерные индексы при оценке заведомо «плохих» кластерных структур.

2. *Кластеры неправильной формы и вложенные кластеры*: множества такого вида традиционно представляют сложность для центроориентированных алгоритмов кластеризации (например семейства к-средних), в которых центры кластеров играют важную роль при построении кластерной структуры. Мы используем этот тестовый случай для проверки того, насколько хорошо справятся с оценкой данного разбиения различные, в том числе и центроориентированные, индексы кластеризации.

3. *Плохоотделимые кластеры*: представляют известную трудность и для алгоритмов кластеризации и для индексов оценки качества. Отделимость кластеров является признаком хорошей структуры кластеров, однако что же делать в том случае, когда кластеры действительно расположены очень близко друг к другу?

4. *Кластеры с гауссовым распределением*: наиболее удобный для кластеризации вариант, с которым должны справиться как алгоритмы, так и индексы.

Все описанные множества, кроме третьего, были сгенерированы синтетически с помощью инструмента Rapid-Miner [19]. Каждое из них состоит из 500 элементов в двумерном пространстве. Третий вариант – плохоотделимые кластеры – представлено множеством реальных данных IRIS [20], являющимся классическим тестовым множеством для многих задач интеллектуального анализа данных.

Также следует упомянуть об одном граничном значении, которое редко рассматривается как при проверке алгоритмов кластеризации, так и при проверке индексов оценки качества: это множество из одного кластера. Большинство как алгоритмов, так и индексов предполагают, что в множестве, по меньшей мере, два различных кластера. Понятие отделимости для случая одного кластера как минимум не определено, однако же только для индекса *CDbw* авторы заранее предупреждают о том, что значение индекса при $c = 1$ является неопределенным. Таким образом, на данный момент, даже если решение с однокластерной структурой действительно является оптимальным и отражает существующие на множестве данных взаимодействия, большинство индексов потеряют его.

Алгоритм 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 структура не стабилизируется или не выполнится условие остановки (например, максимальное число итераций)

3.1.2. Алгоритмы кластеризации

В данной работе мы проверяем методы оценки качества всего на двух классах алгоритмов кластеризации: центроориентированных и плотностных, поскольку именно алгоритмы этих классов в настоящее время используются наиболее широко. Принцип работы у алгоритмов этих классов существенно различается. Грубо говоря, центроориентированные алгоритмы считают кластером область «близкую» к центру кластера, а плотностные – область, где плотность элементов из кластеризуемого множества выше какого-то порогового значения. При такой разнице в подходах к кластеризации, логично ожидать, что к вопросам оценки их результатов следует подходить по-разному. В качестве представителя разбивающих алгоритмов было решено использовать K-Means [21, 22] (алгоритм 1), а в качестве представителя плотностных алгоритмов – DBScan [23] (алгоритм 2). Этот выбор мотивирован тем, что оба эти алгоритма уже давно используются для кластеризации в любых областях для любых типов данных и заслуженно считаются эффективными. Параметры для кластеризации выбирались следующим образом: число кластеров в K-Means варьировалось от 2 до 10, число соседей – minpts – в DBSCan менялось с 3 до 6, а окрестность точки – eps – с 0.1 до 1 с шагом 0.1 и для некоторых экспериментов от 0.5 до 4 с шагом 0.5.

3.1.3. Индексы качества

В качестве индексов для тестирования были выбраны следующие: *Dunn*, *DB*, *SD*, *S_Dwb*, индекс силуэта, упрощенный индекс силуэта, *CS*, *VNND*, *Score Function*, *MB*, *CDbw*. (Следует заметить, что в индексе

Алгоритм 2. DBScan

Require: *eps* – радиус *eps*-окрестности, *min_pts* – минимальное число элементов, которые должны попасть в *eps*-окрестность каждого элемента ядра.
for all элементов кластеризуемого множества **do**
if элемент x_j еще не отнесен к кластеру или к выбросам
eps_n = число элементов, в *eps*-окрестности x_j
if $eps\_n < min\_pts$ **then**
пометить элемент как выброс
else
Поместить x_j в кластер со следующим порядковым номером (x_j – первый элемент нового кластера).
Повторить процедуру оценки точек в *eps*-окрестности рекурсивно для всех элементов в *eps*-окрестности x_j.
end if
end if
end for

Табл. 1. Случайно распределенные данные: результаты для K-Means

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

Табл. 2. Случайно распределенные данные: результаты для DBScan

Индекс	minpts	eps
Dunn	3	0.2
DB	3	0.2
SD	3	0.2
S_Dwb	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

CDbw расстояние от точки до множества при нахождении «представителей» кластера может вычисляться разными способами, при проведении данных экспериментов для измерения этого расстояния использовалась сумма расстояний от уже существующих «представителей» до каждого элемента кластера, элемент, на котором достигался максимум, и выбирался в качестве очередного представителя). Мы не взяли в сравнение модифицированную *Hubert* Γ *Statistic*, поскольку ее область применения ограничена небольшими объемами данных из-за операций с матрицами смежности, и индексы *RMSSTD* и *RS*, поскольку они предназначены для иерархических алгоритмов кластеризации, которые в данной работе не рассматриваются. Тем не менее, мы сочли нужным упомянуть об этих метриках в обзоре для создания более полной картины описываемой области.

3.2. ЭКСПЕРИМЕНТЫ

1. ***Множество со случайным распределением.*** Если во множестве данных нет кластерной структуры, то сами по себе метрики качества не смогут это понять. В данном случае необходимо использовать методы для оценки возможности кластеризации. При использовании алгоритма K-Means (табл. 1) оба разбиения (рис. 1 и 2), которые были выбраны большинством индексов в качестве оптимальных, можно назвать кластерной структурой лишь условно.

Поскольку по принцип работы DBScan основан не на разбиении всего множества на части, а на отыскании в нем участков определенной плотности, при выборе из всех кластерных структур, полученных с помощью DBScan, индексы более единодушны (табл. 2). Наилучшей признается структура, где есть хотя бы какие-то, хоть сколь угодно

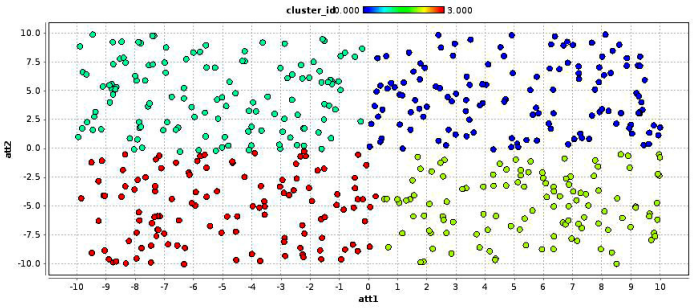

Рис. 1. Случайные данные: деление на четыре кластера при помощи K-Means

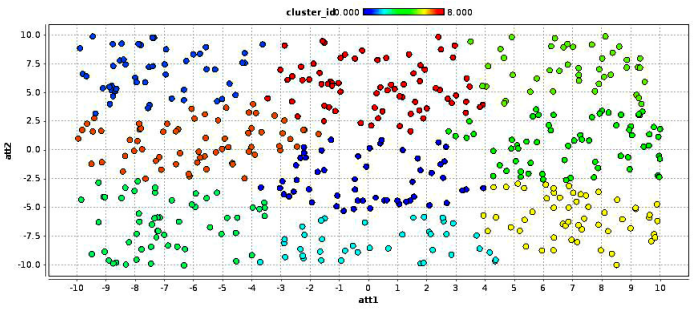

Рис. 2. Случайные данные: деление на девять кластеров при помощи K-Means

Табл. 3. Гауссово распределение данных: результаты для K-Means

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

Табл. 4. Гауссово распределение данных: результаты для DBScan

Индекс	minpts	eps
Dunn	5	0.5
DB	5	0.5
SD	5	0.5
S_Dwb	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

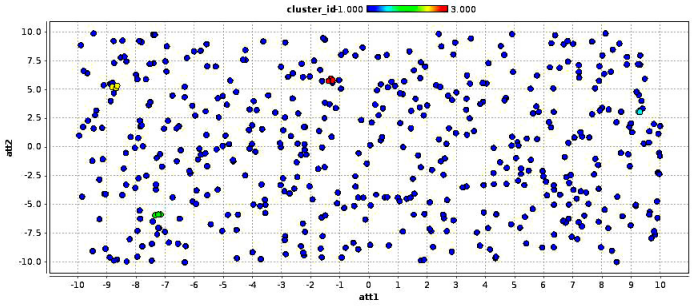

Рис. 3. Случайные данные: результаты для DBScan (3, 0.2)

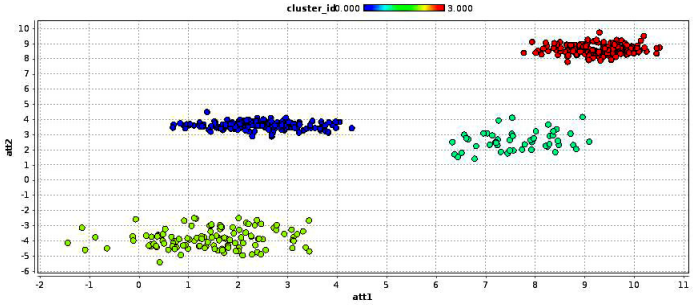

Рис. 4. Гауссово распределение данных: результаты для K-Means

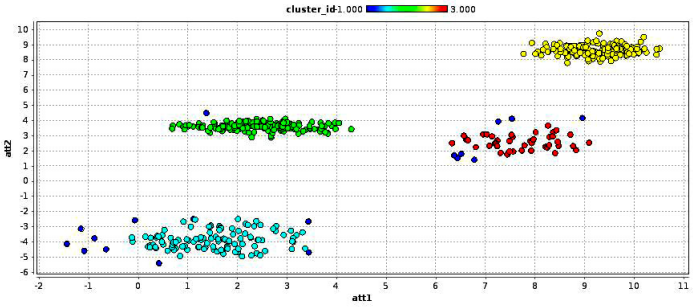

Рис. 5. Гауссово распределение данных: результаты для DBScan (5,0.5)

малые кластеры, но подходящие под понятие компактных и отделимых (рис. 3). Кластер с номером –1 в случае с DBScan означает «кластер» выбросов.

Таким образом, следует признать, что там, где нет структуры кластеров, нет и понятия оптимальной структуры кластеров. При использовании алгоритмов семейства к-средних то, что индексы оценки качества указывают на разное число кластеров в качестве оптимального, скорей всего свидетельствует о том, что алгоритм семейства к-средних (K-Means) не справился с кластеризацией данного множества. В свою очередь, среди структур, построенных с помощью DBScan, выбирается та, где основное множество элементов определено как шум, и лишь незначительная часть распределена по очень маленьким кластерам. Такое поведение индексов и алгоритма должно натолкнуть аналитика на мысль, что данные в принципе плохо кластеризуемы, и ему следует либо сменить представление данных, либо произвести дополнительную обработку, либо понизить размерность данных до подпространства, где кластеризация более возможна.

2. ***Гауссово распределение данных.*** Большинство алгорит-

мов предполагают, что данные имеют гауссово распределение, и это же неявно предполагается в критериях качества кластеризации. Результаты оценки для K-Means и DBScan можно видеть в табл. 3 и 4 соответственно. Если в случае K-Means практически все индексы выбирают правильный вариант рис. 4, то среди кластеризаций DBScan имеются два варианта: рис. 5 и 6, соответственно. Первый из них, полученный с параметрами (5, 0.5), практически совпадает с делением на четыре кластера при помощи K-Means (а разбиение (5, 0.9), выбранное индексами силуэта, полностью идентично разбиению от к-средних), второй вариант – (6, 0.1) и (6, 0.2) ошибочен, потому что выбравшие его индексы стараются добиться от всех кластеров одинаковой максимальной плотности, и из-за этого в качестве «шума» отбрасывается значительная часть данных. В частности, ошибки связаны с попыткой некоторых индексов измерить отделимость как некоторое предельное расстояние между кластерными центрами (*MB* и *CS*) или же представить компактность как сумму расстояний от центра кластера до всех его элементов (*MB* и *Score*).

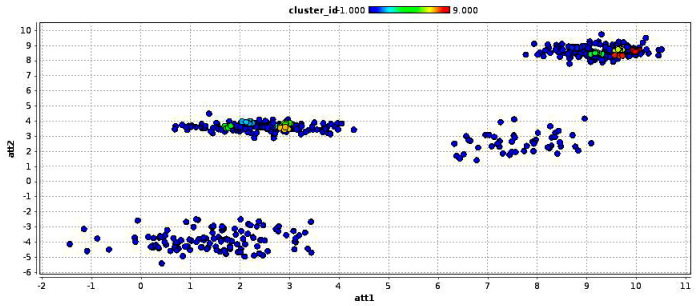

Рис. 6. Гауссово распределение данных: результаты для DBScan (6,0.1)

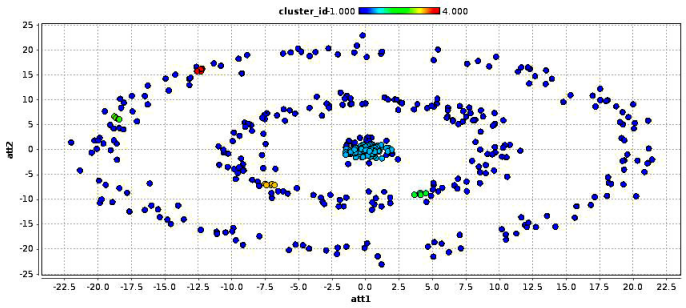

Рис. 7. Концентрические кластеры: результаты для DBScan (4,0.5)

3. ***Кластеры произвольной формы.*** Несмотря на то, что гауссово распределение данных встречается достаточно часто, нельзя ожидать, что оно будет вообще во всех множествах. Для кластеризации в этой категории был использован только DBScan, потому что K-Means в принципе не приспособлен для выделения кластеров такого типа. В качестве тестового множества использовались три концентрических, вложенных друг в друга кластера. В этом случае (табл. 5) большинство индексов исходят из того же принципа, что и при оценке случайно распределенных данных: лучшей структурой признается структура с маленькими и очень компактными кластерами (рис. 7), а не та, которую ожидает пользователь (рис. 8). Это некоторым образом логично, поскольку концентрические кластеры сами по себе не подходят под то понятие о компактности и отделимости, которое используется в большинстве метрик качества: в этом случае их нельзя корректно измерять, используя центры кластеров. Верно определить оптимальную структуру смог только один индекс, CDbw, поскольку он принимает в расчет геометрическую форму кластеров.

Табл. 5. Концентрические кластеры: результаты для DBScan

Индекс	minpts	eps
Dunn	4	0.5
DB	4	0.5
SD	4	0.5
S_Dwb	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

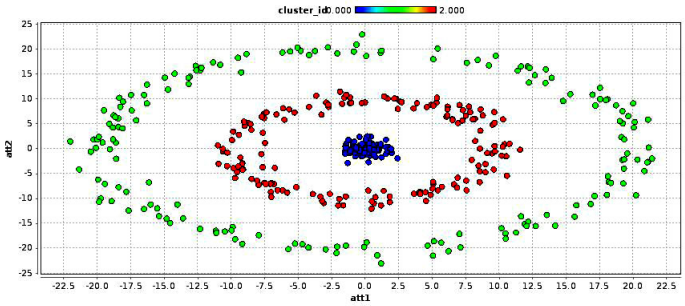

Рис. 8. Концентрические кластеры: результаты для DBScan (4,4.5)

4. ***IRIS***. Множество данных IRIS очень часто используется при тестировании различных инструментов интеллектуального анализа данных. Оно состоит из трех классов по 50 элементов, каждый класс соответствует некоторому типу ирисов. Каждый элемент представляет собой описание цветка и имеет 4 атрибута (длина чашелистика, ширина чашелистика, длина лепестка и ширина лепестка). Один класс хорошо отделим от других, но оставшиеся два расположены очень близко друг к другу (рис. 9).

При кластеризации с помощью K-Means (табл. 6) большинство индексов выбирают деление на два кластера (рис. 10) как оптимальный вариант кластеризации, и только два из них – деление на три (рис. 11). Следует заметить, что даже деление на три кластера в данном случае не совпадает с исходным делением по классам.

Выбор среди структур, полученных с помощью DBScan (табл. 7), сильно разнится от индекса к индексу. Так, *Dunn*, *DB* и *S_Dwb* индексы выбирают структуру с маленькими, но как можно более компактны-

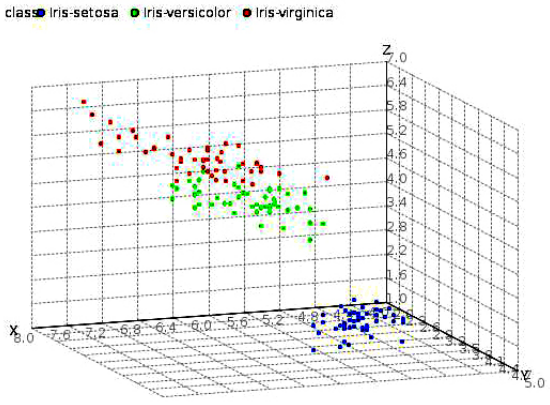

Рис. 9. Множество данных IRIS

Табл. 6. IRIS: результаты для K-Means

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

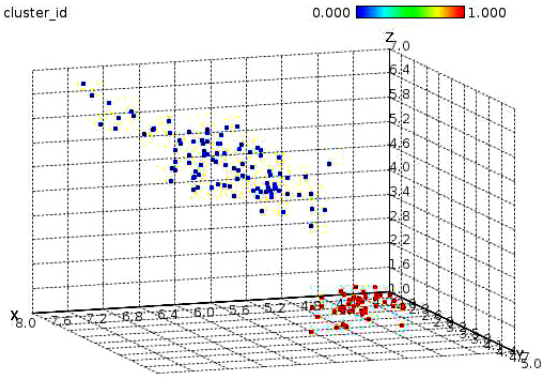

Рис. 10. IRIS: результаты для K-Means (2 кластера)

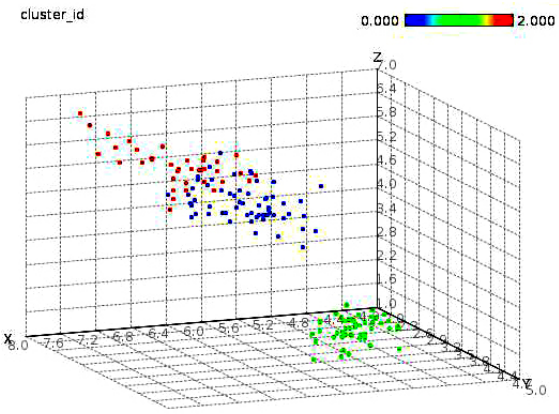

Рис. 11. IRIS: результаты для K-Means (3 кластера)

Табл. 7. IRIS: результаты для DBScan

Индекс	minpts	eps
Dunn	6	0.3
DB	6	0.3
SD	5	0.6
S_Dwb	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

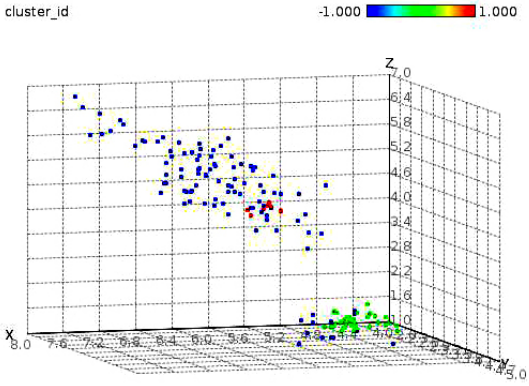

Рис. 12. IRIS: результаты для DBScan (6,0.3)

ми и плотными кластерами, отбрасывая большую часть данных как шум (рис. 12). Выбор индексов *CDbw* и *SD* (рис. 13) более похож на ожидаемый – в качестве шума отброшена лишь незначительная часть данных, и индексы силуэта выбрали структуру, состоящую из двух кластеров, аналогичную разбиению к-средних.

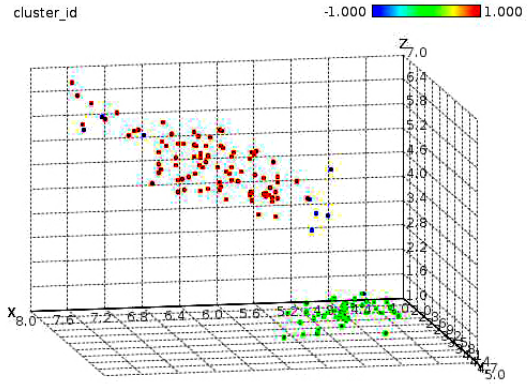

Рис. 13. IRIS: результаты для DBScan (5,0.6)

4. ЗАКЛЮЧЕНИЕ

Результаты проверки индексов на различных множествах даны в табл. 8. Знаком «+» обозначен верный выбор структуры кластеров конкретным индексом. (Правильность выбора в данном случае определялась визуально: если в исходной структуре два кластера, а индекс выбирает четыре – это неверно, но если индекс выбирает структуру из двух почти идентичных оригинальным кластеров и шума – то это верно).

Исходя из представленной таблицы и экспериментов, можно сделать следующие выводы:

1. Ни один индекс не выдаст правильной оценки, если во множестве отсутствует структура кластеров, однако по поведению

Табл. 8. Результаты проверки индексов

Индекс	Gaussian (2)		Gaussian (4)		3 rings	IRIS	
	K-Means	DBScan	K-Means	DBScan	DBScan	K-Means	DBScan
Dunn	+	+	+	+	–	+	–
DB	+	+	+	+	–	+	–
SD	+	+	–	+	–	+	+
S_Dwb	+	+	+	–	–	–	–
Sil.	+	+	+	+	–	+	+
Simp. Sil.	+	+	+	+	–	+	+
CS	+	+	+	–	–	+	–
VNND	+	–	–	–	–	+	–
Score Func.	+	+	+	–	–	+	–
MB	+	+	+	–	–	+	–
CDbw	+	+	+	+	+	+	+

группы индексов можно сделать определенные выводы о том, что используемое множество данных плохо поддается кластеризации.

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. Это заметно по индексу VNND.

4. Если вы используете для кластеризации разбивающие алгоритмы, такие как K-Means, то нет особой разницы, какой из описанных индексов (не включая входящие в предыдущий пункт) брать для оценки качества – они все справятся с задачей.

5. Если вы используете для кластеризации DBScan, то лучше использовать индексы, учитывающие геометрическую структуру кластеров (*CDbw*) и измеряющие компактность и отделимость в терминах средних расстояний между элементами кластеров (Silhouette). Попытка измерить компактность и отделимость как предельное расстояние между элементами (как в случае с индексом *Dunn*) или в зависимости от центра кластера (как у *MB* или *Score*) может привести к тому, что в качестве наилучшей будет выбрана структура с максимальной в терминах алгоритма DBScan плотностью (то есть с максимальным числом соседей и минимальным радиусом окрестности), и при этом значительная часть данных может быть отсеяна как выбросы.

6. Для повышения эффективности в оценке качества и получения объективного результата лучше пользоваться не одним каким-то индексом, а их совокупностью.

7) Лучшие индексы в порядке уменьшения точности оценки: 1) *CDbw*, 2) индексы силуэта, 3) *Dunn* и *DB*.

Следует заметить, что оценка качества кластеризации заслуженно считается сложной областью. Проблему качества сложно выразить семантически и также сложно подогнать ее под математическую модель. Как правило, для любого индекса существует такое множество, на котором его оценка неверна. Даже индекс *CDbw*, показавший хорошие результаты в процессе этого исследования, «ломается», например, на множестве из двух вложенных друг в друга спиралей. В целом, складывается впечатление, что универсального решения в этой области не существует, и потому лучше пользоваться комбинированным подходом. Отказ от использования индексов оценки качества все-таки не представляется целесообразным, потому что, даже если *M* индексов выбрали в качестве оптимальной структуры *M* разных структур, зачастую, проверить вручную их проще и быстрее, чем проверять все полученные кластеризации по отдельности.

Литература

1. *S. Theodoridis and K. Koutroumbas*.Pattern recognition. Academic Press, San Diego, CA, 1999.

2. *R.B. Calinski and J. Harabasz.* A dendrite method for cluster analysis // Comm. in Statistics, 3:1–27, 1974.

3. *J.C. Dunn.* Well separated clusters and optimal fuzzy-partitions // Journal of Cybernetics, 4:95–104, 1974.

4. *N.R. Pal and J.Biswas.* Cluster validation using graph theoretic concepts // Pattern Recognition, 30(6), 1997.

5. *D.L. Davies and D.W. Bouldin.* A cluster separation measure // IEEE Transactions on Pattern Analysis and Machine Intelligence, 1, 1979.

6. *C.H. Chou, M.C. Su, and E.Lai.* A new cluster validity measure and its application to image compression // Pattern Analysis and Applications, 7:205–220, 2004.

7. *C.H. Chow, M.C. Su, and Lai Eugene.* Symmetry as a new measure for cluster validity // 2nd WSEAS International Conference of scientific on Computation and Soft Computing. P. 209–213, 2002.

8. *M. Halkidi, M. Vazirgiannis, and I. Batistakis.* Quality scheme assessment in the clustering process // Proceedings of PKDD, 2000.

9. *M. Halkidi and M. Vazirgiannis.* Clustering validity assessment: Finding the optimal partitioning of a data set // Proc. of ICDM, 2001.

10. *Subhash Sharma.* Applied multivariate techniques. John Wiley and Sons, Inc., 1996.

11. *M. Halkidi, Y. Batistakis, and M. Vazirgiannis.* On clustering validation techniques // Intelligent Information Systems Journal, 17:107–145, 2001.

12. *L. Kaufman and P. Rousseeuw.* Finding Groups in Data. An Introduction to Cluster Analysis. Wiley, 2005.

13. *E.R. Hruschka, L. Vendramin, and R.J. G.B. Campello.* On the comparison of relative clustering validity criteria // 2009 SIAM International Conference on Data Mining (SDM 09), 1: 733–744, 2009.

14. *U. Maulik and S. Bandyopadhyay.* Performance evaluation of some clustering algorithms and validity indices // IEEE Transactions Pattern Analysis Machine Intelligence, 24(12): 1650–1654, 2002.

15. *M.K. Pakhira, S. Bandyopadhyay, and U. Maulik.* Validity index for crisp and fuzzy clusters // Pattern Recognition, 37(3): 487–501, 2004.

16. *S. Saitta, B. Raphael, and I. F.C. Smit.* A bounded index for cluster validity // 5th International Conference on Machine Learning and Data Mining in Pattern Recognition, 2007.

17. *M. Halkidi and M. Vazirgiannis.* A density-based cluster validity approach using multi-representatives // Pattern Recognition Letters, 29(6): 773–786, 2008.

18. *F. Kovacs and R. Ivancsy.* Cluster validity measurement for arbitrary shaped clusters // 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and DB, 2008.

19. *I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz and T. Euler.* Yale: Rapid prototyping for complex data mining tasks // 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

20. *R.A. Fisher.* The use of multiple measurements in taxonomic problems // Annals of Eugenics, 7: 179–188, 1936.

21. *J. Hartigan.* Clustering Algorithms. John Wiley & Sons, 1975.

22. *J. Hartigan and M. Wong.* Algorithm as 136: A k-means clustering algorithm // Applied Statistics, 28: 100–108, 1979.

23. M. Ester, H-P Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise // In Proc. of the 2nd ACM SIGKDD,1996. P. 226–231.

Abstract

In this paper the most commonly used cluster validity indexes are compared. Theirs features, efficiency and applicability to different context are discussed. Some recommendations about usage of validity indexes with different clustering algorithms are made.

Keywords: clustering, cluster validity, cluster validity indexes.

Сивоголовко Елена Владимировна,
аспирант математико-механического факультета СПбГУ,
efecca@gmail.com