Саитта С., Рафаэль Б. и Смит И.Ф.К., "Ограниченный индекс валидности кластеров", Машинное обучение и добыча данных в распознавании образов, LNAI 4571, Springer, Heidelberg, pp. 174-187, 2007. Окончательная публикация доступна в Springer по адресу WOS:000248523200013.

# Ограниченный индекс валидности кластеров

Сандро Саитта, Бенни Рафаэль и Ян Ф.К. Смит

Федеральная политехническая школа Лозанны (EPFL)
Станция 18, 1015 Лозанна, Швейцария
sandro.saitta@epfl.ch,bdgbr@nus.edu.sg,ian.smith@epfl.ch

Аннотация. Кластеризация - один из наиболее известных типов несамостоятельного обучения. Оценка качества результатов и определение количества кластеров в данных является важным вопросом. Большинство существующих индексов достоверности охватывают лишь подмножество важных аспектов кластеров. Более того, эти индексы актуальны только для наборов данных, содержащих не менее двух кластеров. В данной работе представлен новый ограниченный индекс достоверности кластеров, названный функцией оценки (SF). Функция оценки основана на стандартных свойствах кластеров. Для оценки эффективности функции оценки используется несколько искусственных и реальных наборов данных. Функция оценки тестируется в сравнении с четырьмя существующими индексами достоверности. Оказалось, что индекс, предложенный в данной работе, всегда не хуже или лучше этих индексов в случае гиперсфероидальных кластеров. Показано, что он хорошо работает на многомерных наборах данных и способен учитывать уникальные и субкластерные случаи.

Ключевые слова: кластеризация, валидность кластеров, индекс валидности, k-средние

## 1 Введение

Цель кластеризации [1, 2] - сгруппировать точки данных, которые похожи друг на друга в соответствии с выбранной метрикой сходства (обычно используется евклидово расстояние). Методы кластеризации применяются в таких областях, как поиск текстов [3], обнаружение вторжений [4] и распознавание объектов [5]. В этих областях, как и во многих других, количество кластеров обычно не известно заранее.

В литературе можно найти несколько методов кластеризации. Обычно они относятся к одной из следующих категорий [6]: условная кластеризация, иерархическая кластеризация, кластеризация на основе плотности и кластеризация на основе сетки. Дополнительной категорией является подход на основе смеси Гаусса. Поскольку его вычислительная сложность высока, он вряд ли будет использоваться на практике. У всех этих категорий есть недостатки. Например, иерархическая кластеризация имеет более высокую сложность. Алгоритмы кластеризации на основе плотности часто требуют настройки неинтуитивных параметров. Наконец, алгоритмы кластеризации на основе плотности не всегда выявляют кластеры хорошего качества. Наиболее широко используется алгоритм K-means [1], являющийся частью условной кластеризации. К преимуществам K-means относятся вычислительная эффективность, быстрота реализации и простота математического обоснования. Однако у K-means есть и ограничения. К ним относятся случайный выбор местоположения центроидов в начале процедуры, обработка категориальных переменных и неизвестное

количество кластеров.

*k*. Что касается первого ограничения, то решением может стать многократный прогон. В работе [7] приводится возможное решение второго ограничения за счет использования меры соответствия несходства для обработки категориальных параметров. Наконец, третья проблема связана с количеством кластеров и, следовательно, с достоверностью кластеров.

Кластеризация по определению является субъективной задачей, и именно это делает ее сложной [8]. Примеры проблем в кластеризации включают i) количество кластеров, присутствующих в данных, и ii) качество кластеризации [9]. Элементы ответов на эти два вопроса можно найти в области валидации кластеров. В кластеризации важны и другие проблемы, такие как начальные условия и наборы данных высокой размерности. Целью методов валидации кластеров является оценка результатов кластеризации [6, 8, 10]. Эта оценка может быть использована для определения количества кластеров в наборе данных. В современной литературе приводится несколько примеров индексов валидности [9, 11-13]. Недавно была проведена работа по их оценке [14].

Индекс Данна [11] объединяет несходство между кластерами и их диаметры для оценки наиболее надежного количества кластеров. Как отмечается в [6], индекс Данна требует больших вычислительных затрат и чувствителен к зашумленным данным. Понятия дисперсии кластера и несходства между кластерами используются для вычисления индекса Дэвиса-Болдина [12]. Индекс Дэвиса-Болдина был признан одним из лучших индексов [14]. Индекс Silhouette [13] использует среднее несходство между точками для определения структуры данных и выделения возможных кластеров. Индекс Silhouette подходит только для оценки первого выбора или наилучшего разбиения [15]. Наконец, индекс Maulik-Bandyopadhyay [9] связан с индексом Данна и предполагает настройку параметра.

Все эти индексы требуют указания как минимум двух кластеров. Как отмечается в [16], случай с одним кластером важен и, скорее всего, встречается на практике. В качестве предварительного условия для идентификации одного кластера важно определение того, что такое кластер. Среди существующих в литературе определений одно из возможных дано в [17]. Вкратце оно гласит, что кластер считается "настоящим", если он значительно компактен и изолирован. Понятия компактности и изолированности основаны на двух параметрах, определяющих внутренние свойства кластера. Хотя это определение и является точным, оно часто оказывается слишком ограничительным, поскольку лишь немногие наборы данных удовлетворяют таким критериям. Более подробно об одиночных кластерных тестах можно прочитать в [16]. В литературе существуют и другие индексы достоверности. Некоторые из них требуют больших вычислительных затрат [6], а другие не могут определить реальное количество кластеров во всех наборах данных [14]. В данной статье предлагается новый индекс достоверности, который помогает преодолеть эти ограничения. Данная статья организована следующим образом. В разделе 2 описаны существующие в литературе показатели валидности. В разделе 3 предлагается новый показатель валидности, названный функцией оценки. В разделе 4 описывается работа функции оценки. Сайт
В последнем разделе представлены выводы и направления дальнейшей работы.

## 2 Существующие индексы

В этом разделе описаны четыре показателя валидности, подходящие для жесткой условной кластеризации. Эти индексы служат основой для оценки результатов функции оценки на эталонных наборах данных. Обозначения для этих индексов были адаптированы

чтобы обеспечить согласованную основу. Для нормализованных данных используется метрика стандартное евклидово расстояние, определяемое как $||x - y|| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$, где $x$ и $y$ - точки данных, а $d$ - количество измерений.

Индекс Данна: Один из самых цитируемых индексов предложен в работе [11]. Индекс Данна (DU) определяет кластеры, которые хорошо разделены и компактны. Таким образом, цель состоит в том, чтобы максимизировать межкластерное расстояние при минимизации внутрикластерного расстояния. Индекс Данна для $k$ кластеров определяется уравнением 1:

$$DU_k = \min_{i=1,\ldots,k} \; \text{мин}_{j=1+1,\ldots,k} \; \frac{diss(c_i, c)_j}{\max_{m=1,\ldots,k} diam(c_m)} \tag{1}$$

где $diss(c_i, c_j) = \min_{x \in c\, i, y \in cj} ||x - y||$ - несходство между кластерами $c_i$ и $c_j$ , а $diam(C) = \max_{x,y \in C} ||x - y||$ - внутрикластерная функция (или диаметр) кластера. Если индекс Данна велик, это означает, что существуют компактные и хорошо разделенные кластеры. Поэтому максимум наблюдается для $k$, равного наиболее вероятному числу кластеров в наборе данных.

Индекс Дэвиса-Болдина: Как и индекс Данна, индекс Дэвиса-Болдина [12] определяет кластеры, которые находятся далеко друг от друга и компактны. Индекс Дэвиса-Болдина (DB) определяется в соответствии с уравнением 2:

$$DB_k = \frac{1}{k} \sum_{i=1}^{k} \text{макс}_{j=1,\ldots,k, i \neq j} \; \frac{diam(c_i) + diam(c_j)}{||c_i - c||_j} \tag{2}$$

где в данном случае диаметр кластера определяется как:

$$diam(c_i) = \frac{1}{n_i} \sum_{x \in c_i} ||x - z||_i^{1/2} \tag{3}$$

где $n_i$ - количество точек, а $z_i$ - центроид кластера $c_i$ . Поскольку цель состоит в получении кластеров с минимальными внутрикластерными расстояниями, интересны малые значения DB. Поэтому при поиске оптимального числа кластеров этот показатель минимизируется.

Индекс силуэта: Статистика силуэтов [13] - еще один известный способ оценки количества групп в наборе данных. Силуэтный индекс (SI) рассчитывает для каждой точки ширину силуэта, зависящую от ее принадлежности к какому-либо кластеру. Эта ширина силуэта является средним значением по всем наблюдениям. Это приводит к уравнению 4:

$$SI_k = \frac{1}{n} \sum_{i=1}^{n} \frac{(b_i - a_i)}{\max(a_i, b)_i} \tag{4}$$

где $n$ - общее количество точек, $a_i$ - среднее расстояние между точкой $i$ и всеми другими точками в ее собственном кластере, а $b_i$ - минимальное из средних несходств между $i$ и точками в других кластерах. Наконец, оптимальным считается разбиение с наибольшим значением SI.

Индекс Маулика-Бандиопадьяя: Недавно разработанный индекс был назван индексом *I* [9]. Для согласованности с другими индексами он переименован в MB. Этот индекс, представляющий собой комбинацию трех терминов, определяется с помощью уравнения 5:

$$MB_k = \left( \frac{1}{k} \cdot \frac{E_1}{E_k} \cdot D_k \right)^p \quad (5)$$

где внутрикластерное расстояние определяется $E_k = \sum_{i=1}^{\Sigma k} \sum_{x \in c_i} ||x - z_i||$ и межкластерное расстояние по $D_k = \max_{i,j=1}^{k} ||z_i - z_j||$. Как и ранее, $z_i$ - это центр кластера $c_i$ . Правильное количество кластеров оценивается путем максимизации уравнения 5. В данной работе $p$ выбрано равным двум.

Обсуждение: Хотя все эти индексы полезны в определенных ситуациях, они не являются универсальными. Например, индекс Данна требует больших вычислительных затрат и с трудом справляется с зашумленными данными. Он полезен для выявления чистых кластеров в наборах данных, содержащих не более сотни точек. Индекс Дэвиса-Болдина дает хорошие результаты при выделении отдельных групп. Однако он не рассчитан на работу с перекрывающимися кластерами. Индекс Silhouette способен определить только первый вариант, поэтому его не следует применять к наборам данных с подкластерами. Особенностью индекса Маулика-Бандиопадьяя является то, что он зависит от заданного пользователем параметра.

## 3 Функция оценки

В этой статье мы предлагаем функцию для оценки количества кластеров в наборе данных. Предлагаемый показатель, а именно функция оценки (SF), основан на межкластерных и внутрикластерных расстояниях. Функция оценки используется для двух целей:
i) для оценки количества кластеров и ii) для оценки качества результатов кластеризации. Функция оценки представляет собой функцию, объединяющую два понятия: расстояние между кластерами и расстояние внутри кластера. Первое понятие определяется как "межкластерное расстояние" (*bcd*), а второе - как "внутрикластерное расстояние" (*wcd*).

Для измерения расстояния между двумя кластерами существуют три распространенных подхода: одиночная связь, полная связь и сравнение центроидов. Данное предложение основано на третьей концепции, поскольку первые две имеют слишком высокие вычислительные затраты [6]. В данной работе *bcd* определяется уравнением 6:

$$bcd = \frac{\sum_{i=1}^{\Sigma k} ||z_i - z_{tot}|| \cdot n_i}{n - k} \quad (6)$$

где $k$ - количество кластеров, $z_i$ - центроид текущего кластера, а $z_{tot}$ - центроид всех кластеров. Размер кластера, $n_i$ , определяется количеством точек внутри него. Наиболее важной величиной в *bcd* является расстояние между $z_i$ и $z_{tot}$ . Чтобы ограничить влияние промахов, каждое расстояние взвешивается по размеру кластера. Это позволяет снизить чувствительность к шуму. Благодаря $n$ удается избежать чувствительности *bcd* к общему количеству точек. Наконец, значения $k$ используются для штрафа за добавление нового кластера. Таким образом, удается избежать ограничения в одну точку на кластер. Значение *wcd приведено* в уравнении 7:

$$wcd = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{x \in ci} ||x - z_i|| \quad (7)$$

Вычисление значений для *wcd* включает определение расстояния между каждой точкой и центроидом ее кластера. Это расстояние суммируется по всем *k* кластерам. Обратите внимание, что $||z_i - x||$ уже учитывает размер соответствующего кластера. Как и в уравнении *bcd* (уравнение 6), размер кластера в знаменателе позволяет избежать чувствительности к общему количеству точек. Благодаря уравнениям 6 и 7, *bcd* и *wcd* не зависят от количества точек данных.

Чтобы функция оценки была эффективной, она должна i) максимизировать *bcd*, ii) минимизировать *wcd* и iii) быть ограниченной. Максимизация уравнения 8 удовлетворяет вышеуказанным условиям:

$$SF = 1 - \frac{1}{e^{bcd-wcd}} \quad (8)$$

Чем выше значение *SF*, тем более подходящее количество кластеров. Таким образом, с помощью предложенного SF теперь можно оценить количество кластеров для заданного набора моделей. Такие трудности, как идеальные кластеры (*wcd* = 0) и уникальные кластеры (*bcd* = 0), преодолены. Более того, предложенная функция оценки ограничена пределами ]0, 1[. Верхняя граница позволяет определить, насколько текущий набор данных близок к случаю идеального кластера. Таким образом, мы стремимся максимизировать уравнение 8, чтобы получить наиболее достоверное количество кластеров. Как видно из уравнений 6 и 7, вычислительная сложность линейна. Если *n* - количество точек данных, то предложенная функция оценки имеет сложность O(*n*). В следующем разделе функция оценки тестируется на нескольких эталонных задачах и сравнивается с существующими индексами.

## 4 Результаты

В этом разделе сравниваются показатели валидности. Для этого используется стандартный алгоритм K-means. K-means - это процедура, которая перебирает *k* кластеров с целью минимизации их внутрикластерных расстояний. Процедура K-means выглядит следующим образом. Сначала случайным образом выбираются *k* центроидов среди всех точек. Затем набор данных разбивается на группы в соответствии с минимальным квадратичным расстоянием. Новые позиции центроидов рассчитываются в соответствии с точками внутри кластеров. Процесс разбиения и обновления повторяется до тех пор, пока не будет достигнут критерий остановки. Это происходит, когда центры кластеров или внутрикластерные расстояния существенно не меняются в течение двух последовательных итераций.

Чтобы контролировать случайность K-средних, они запускаются 10 раз с $k_{min}$ по $k_{max}$ кластеров. Оптимум - минимальный или максимальный, в зависимости от индекса

- выбрано как наиболее подходящее количество кластеров. Индексы для сравнения были выбраны в соответствии с их производительностью и использованием в литературе (см. раздел 1). Выбраны индексы Данна (DU), Дэвиса-Болдина (DB), Силуэта (SI) и Маулика-Бандиопадьяя (MB). Они сравниваются с

Функция оценки (SF). Далее показаны результаты в зависимости от количества кластеров, выявленных для предложенных эталонов. Затем проверяются особенности функции оценки, такие как идеальные и уникальные кластеры, а также иерархия кластеров. Наконец, приводятся примеры ограничений, связанных с функцией оценки.

## 4.1 Количество кластеров

Функция оценки была протестирована на эталонных наборах данных, и результаты сравниваются с другими индексами. $k_{min}$ и $k_{max}$ принимаются равными соответственно 2 и
10. Искусственные наборы данных, используемые в данном разделе, состоят из 1000 точек в двух измерениях.

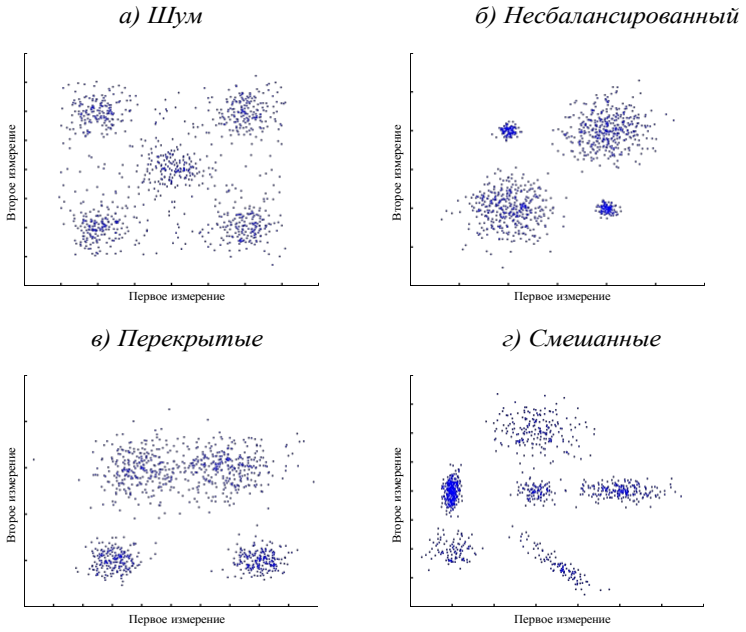


Рис. 1. Четыре искусственных набора данных: *шумный*, *несбалансированный*, *перекрытый* и *смешанный*. Все эти наборы данных содержат 1000 точек в двумерном пространстве.

*Пример 1*: В первом наборе данных, *Noisy*, присутствуют пять кластеров в шумной среде (см. рис. 1a). Маловероятно, чтобы набор данных не содержал шума. Поэтому кластеры часто окружены шумом. Из таблицы 1 видно, что, в отличие от других индексов, индекс Данна не способен правильно оценить количество кластеров (пять). Этот результат подтверждает идею о том, что индекс Данна чувствителен к шуму [6].

*Пример 2*: Второй набор данных, *Unbalanced*, состоит из четырех кластеров (см. рис. 1b). Эти кластеры имеют разный размер и плотность. Согласно [18],

| k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| DU | 0.018 | 0.016 | 0.019 | 0.019 | 0.032 | **0.035** | 0.027 | 0.028 | 0.023 |
| DB | 1.060 | 0.636 | 0.532 | **0.440** | 0.564 | 0.645 | 0.665 | 0.713 | 0.729 |
| SI | 0.534 | 0.573 | 0.719 | **0.821** | 0.785 | 0.768 | 0.733 | 0.706 | 0.669 |
| MB | 1.314 | 2.509 | 3.353 | **5.037** | 4.167 | 3.323 | 2.898 | 2.515 | 2.261 |
| SF | 0.424 | 0.489 | 0.553 | **0.592** | 0.584 | 0.578 | 0.575 | 0.573 | 0.572 |

Таблица 1. Результаты пяти индексов достоверности на наборе данных *"Шум"* (пример 1). Набор данных показан на рисунке 1a. Жирные цифры показывают максимальные значения для всех индексов, кроме DB, где желательны минимальные значения. Это обозначение используется для таблиц 1-6. Правильное число кластеров равно пяти.

кластеры с различной плотностью имеют важное значение. В таблице 2 приведены результаты для этого набора данных. В то время как DU недооценивает количество кластеров, MB переоценивает его. Это не относится к DB, SI и SF, которые правильно определяют четыре кластера.

| k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| DU | **0.154** | 0.066 | 0.025 | 0.024 | 0.016 | 0.018 | 0.014 | 0.012 | 0.016 |
| DB | 0.739 | 0.522 | **0.347** | 0.552 | 0.633 | 0.712 | 0.713 | 0.722 | 0.733 |
| SI | 0.709 | 0.688 | **0.803** | 0.689 | 0.704 | 0.701 | 0.679 | 0.683 | 0.590 |
| MB | 3.900 | 3.686 | 4.795 | 4.751 | **4.941** | 4.844 | 4.540 | 3.575 | 3.794 |
| SF | 0.549 | 0.563 | **0.601** | 0.593 | 0.591 | 0.589 | 0.589 | 0.588 | 0.589 |

Таблица 2. Результаты пяти индексов валидности на наборе *несбалансированных* данных (пример 2). Набор данных показан на рисунке 1b. Правильное количество кластеров - четыре.

*Пример 3*: Этот набор данных, названный *Overlapped*, содержит четыре кластера, два из которых перекрываются. Это видно на рисунке 1c. В реальных наборах данных два кластера, скорее всего, будут перекрываться. Поэтому способность справляться с перекрывающимися кластерами - один из лучших способов сравнения индексов [19]. В таблице 3 приведены результаты для этого набора данных. Видно, что DU и DB недооценивают правильное количество кластеров. Только SI, MB и SF способны определить четыре кластера.

| k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| DU | **0.030** | 0.025 | 0.013 | 0.013 | 0.012 | 0.019 | 0.021 | 0.012 | 0.012 |
| DB | 0.925 | **0.451** | 0.482 | 0.556 | 0.701 | 0.753 | 0.743 | 0.774 | 0.761 |
| SI | 0.635 | 0.740 | **0.818** | 0.728 | 0.713 | 0.669 | 0.683 | 0.669 | 0.656 |
| MB | 1.909 | 3.322 | **5.755** | 5.068 | 4.217 | 3.730 | 3.527 | 3.150 | 3.009 |
| SF | 0.452 | 0.555 | **0.610** | 0.601 | 0.593 | 0.589 | 0.588 | 0.585 | 0.584 |

Таблица 3. Результаты пяти индексов достоверности на наборе данных *Overlapped* (пример 3). Набор данных показан на рисунке 1c. Правильное количество кластеров - четыре.

*Пример 4*: Следующий набор данных, названный *Mixed*, содержит шесть кластеров. Они имеют разный размер, компактность и форму. Набор данных показан на рисунке 1d. В таблице 4 представлены результаты. Во-первых, видно, что DU максимален для двух последовательных значений (хотя и не правильных). MB - единственный показатель, который переоценивает правильное количество кластеров. Наконец, только DB, SI и SF способны правильно определить шесть кластеров.

| k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| DU | 0.015 | 0.041 | 0.041 | 0.027 | 0.018 | 0.020 | 0.014 | 0.018 | 0.017 |
| DB | 1.110 | 0.751 | 0.630 | 0.575 | 0.504 | 0.554 | 0.596 | 0.641 | 0.662 |
| SI | 0.578 | 0.616 | 0.696 | 0.705 | 0.766 | 0.744 | 0.758 | 0.730 | 0.687 |
| MB | 1.523 | 1.574 | 2.379 | 2.813 | 3.389 | 3.661 | 3.857 | 3.490 | 3.236 |
| SF | 0.442 | 0.492 | 0.540 | 0.559 | 0.583 | 0.579 | 0.577 | 0.576 | 0.579 |

Таблица 4. Результаты пяти индексов валидности на наборе данных *Mixed* (пример 4). Набор данных показан на рисунке 1d. Правильное количество кластеров - шесть.

*Пример 5*: Набор данных, используемый в этом примере, *Iris* - один из наиболее часто используемых наборов реальных данных в сообществах машинного обучения и интеллектуального анализа данных [20]. Он состоит из 150 точек в четырех измерениях. *Iris* содержит три кластера (два из них не являются линейно разделяемыми). Это хороший пример случая, когда размерность больше двух и кластеры перекрываются. В таблице 5 приведены значения индексов для этого набора данных. В этом случае только SF может правильно определить три кластера. Перекрытие слишком сильное, чтобы другие испытанные индексы смогли перечислить кластеры.

| k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| DU | 0.267 | 0.053 | 0.070 | 0.087 | 0.095 | 0.090 | 0.111 | 0.091 | 0.119 |
| DB | 0.687 | 0.716 | 0.739 | 0.744 | 0.772 | 0.791 | 0.833 | 0.752 | 0.778 |
| SI | 0.771 | 0.673 | 0.597 | 0.588 | 0.569 | 0.561 | 0.570 | 0.535 | 0.580 |
| MB | 8.605 | 8.038 | 6.473 | 6.696 | 5.815 | 5.453 | 4.489 | 4.011 | 4.068 |
| SF | 0.517 | 0.521 | 0.506 | 0.507 | 0.503 | 0.503 | 0.497 | 0.510 | 0.513 |

Таблица 5. Результаты пяти индексов достоверности на наборе данных *Iris* (пример 5). Набор данных состоит из 150 точек в четырехмерном пространстве. Правильное количество кластеров - три (два из них пересекаются).

*Пример 6*: Следующий набор данных, названный *Wine*, также является реальным набором данных [20]. Он содержит 178 точек в 13 измерениях. Набор данных *Wine* содержит три кластера. Результаты пяти индексов приведены в таблице 6. Если DU переоценивает правильное количество кластеров, то MB недооценивает его. DB, SI и SF способны обнаружить три кластера.

| k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| DU | 0.160 | 0.232 | 0.210 | 0.201 | 0.202 | 0.208 | 0.235 | 0.206 | 0.214 |
| DB | 1.505 | 1.257 | 1.499 | 1.491 | 1.315 | 1.545 | 1.498 | 1.490 | 1.403 |
| SI | 0.426 | 0.451 | 0.416 | 0.394 | 0.387 | 0.347 | 0.324 | 0.340 | 0.288 |
| MB | 5.689 | 5.391 | 3.548 | 2.612 | 2.302 | 2.124 | 1.729 | 1.563 | 1.387 |
| SF | 0.131 | 0.161 | 0.151 | 0.146 | 0.143 | 0.145 | 0.147 | 0.149 | 0.150 |

Таблица 6. Результаты пяти индексов валидности на наборе данных *Wine* (пример 6). Набор данных состоит из 178 точек в 13-мерном пространстве. Правильное количество кластеров - три.

В таблице 7 приведены результаты применения пяти индексов к четырем искусственным и двум реальным наборам данных. Среди пяти протестированных индексов наилучшие результаты показал SF. SF правильно определил количество кластеров во всех шести наборах данных. SF успешно обрабатывает стандартный случай с кластерами и шумом (*Noisy*), кластерами разного размера и компактности (*Unbalanced* ), перекрывающимися кластерами (*Overlapped* ), кластерами нескольких видов (*Mixed* ) и многомерными данными (*Iris* и *Wine*).

| Наборы данных | DU | DB | SI | MB | SF |
|---|---|---|---|---|---|
| *Шумный* | 7(X) | 5(O) | 5(O) | 5(O) | 5(O) |
| *Небалансный* | 2(X) | 4(O) | 4(O) | 6(X) | 4(O) |
| *Перекрытые* | 2(X) | 3(X) | 4(O) | 4(O) | 4(O) |
| *Смешанные* | 3/4(X) | 6(O) | 6(O) | 8(X) | 6(O) |
| *Айрис* | 2(X) | 2(X) | 2(X) | 2(X) | 3(O) |
| *Вино* | 8(X) | 3(O) | 3(O) | 2(X) | 3(O) |

Таблица 7. Расчетное количество кластеров для шести наборов данных и пяти индексов валидности кластеров. Условные обозначения указывают, когда найдено правильное количество кластеров (O) или нет (X).

## 4.2 Идеальные кластеры

Поскольку функция оценки ограничена, ее верхний предел (1,0) можно использовать для оценки близости наборов данных к идеальным кластерам. Следующие два набора данных используются для проверки того, как SF справляется с идеальными кластерами. Наборы данных *Perfect3* и *Perfect5 состоят* из 1000 точек в 2D и содержат три и пять кластеров соответственно, которые близки к идеальным (т. е. имеют очень высокую компактность). Хотя количество кластеров определено правильно, интересно отметить, что максимальное значение SF в обоих случаях различно. В случае с тремя кластерами максимум (0,795) выше, чем во втором (0,722). Это связано с зависимостью SF от количества кластеров *k*. Это видно из знаменателя уравнения 6. Тем не менее, SF дает представление о том, насколько хороши кластеры, по близости значения индекса к его верхней границе, равной единице.

### 4.3 Уникальный кластер

Цель SF - учесть уникальный случай кластера. Этот случай обычно не рассматривается другими. В этом подразделе $k_{min}$ и $k_{max}$ принимаются равными 1 и 8 соответственно. Когда SF строится в зависимости от количества кластеров, могут возникнуть две ситуации. Либо количество кластеров имеет локальный максимум (рис. 2, слева), либо SF монотонно растет между $k_{min}$ и $k_{max}$ (рис. 2, справа).

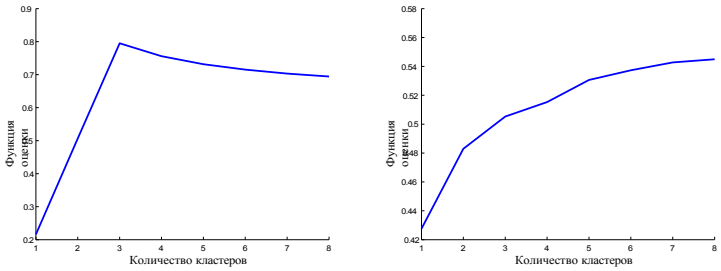


Рис. 2. Разница тренда SF с набором данных, содержащим три кластера (слева) и один кластер (справа).

Поскольку в первой ситуации количество кластеров можно определить, сложность заключается во второй ситуации. Здесь возможны три варианта. Это: i) отсутствие структуры в данных, ii) данные, образующие один кластер, и iii) правильное число кластеров больше, чем $k_{max}$ . Первая ситуация выходит за рамки данной статьи. Более подробную информацию о том, структурированы данные или нет, называемую кластерной тенденцией, можно найти в [1]. В последних двух ситуациях SF монотонно растет с увеличением числа кластеров.

Были замечены два наблюдения. Во-первых, в случаях с уникальными кластерами значение SF при $k = 2$, обозначаемое как $SF_2$ , ближе к значению для $k = 1$ ($SF_1$ ), чем в других наборах данных. Во-вторых, SF зависит от размерности набора данных. Поэтому в качестве индикатора используется наклон между $SF_2$ и $SF_1$ , взвешенный на размерность набора данных. Для проверки случая уникального кластера вводятся два новых набора данных: *UniqueN* - уникальный кластер с добавленным шумом и *Unique30* - уникальный кластер в 30-мерном пространстве. Результаты этого показателя для всех наборов данных приведены в таблице 8.

Согласно таблице 8, эмпирически установлено, что набор данных, скорее всего, содержит более одного кластера, если выполняется уравнение 9.

$$(SF_2 - SF_1) - d > 0{,}2 \tag{9}$$

где $d$ - размерность данных, $SF_2$ и $SF_1$ - соответственно значение SF при $k = 2$ и $k = 1$. Только два набора данных, содержащих уникальные

| Наборы данных | Индикатор | Наборы данных | Индикатор |
|---|---|---|---|
| *Шумный* | 0.37 | *UniqueN* | 0.11 |
| *Небалансный* | 0.65 | *Unique30* | 0.10 |
| *Перекрытые* | 0.45 | *Айрис* | 1.49 |
| *Смешанные* | 0.41 | *Вино* | 1.31 |

Таблица 8. Результаты показателя ($SF_2$ - $SF_1$) - *d* для восьми эталонных наборов данных.

кластеры не удовлетворяют условию уравнения 9. Таким образом, индекс SF является единственным из всех протестированных индексов, способным идентифицировать уникальную кластерную ситуацию.

## 4.4 Подкластеры

Еще одно интересное исследование касается случая с субкластерами. Эта ситуация возникает, когда существующие кластеры можно рассматривать как кластерную иерархию. Если эта иерархия может быть отражена индексом валидности, то пользователю может быть предоставлено больше информации о структуре данных. Набор данных *Sub-cluster* на рисунке 3 является примером такой ситуации. Индекс SF сравнивается с ранее упомянутыми индексами по этой теме. На рисунке 3 показана эволюция каждого индекса достоверности в зависимости от количества кластеров.

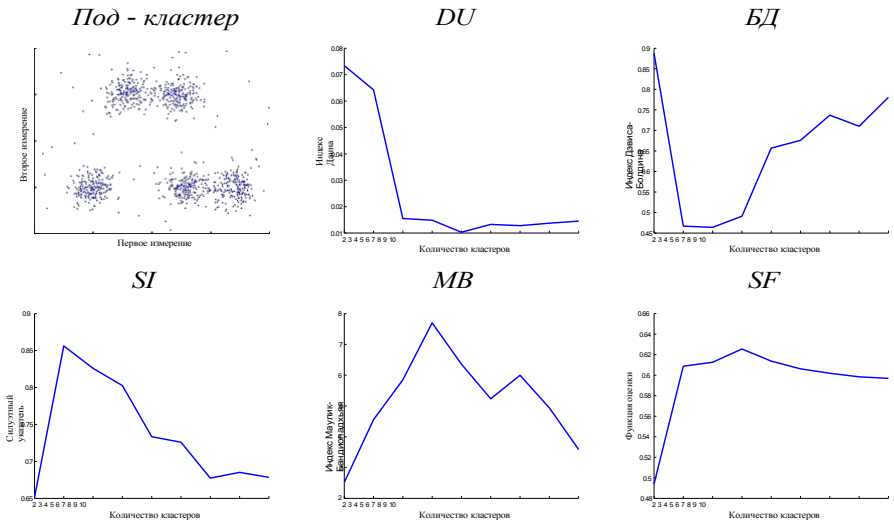


Рис. 3. Сравнение DU, DB, SI, MB и SF для случая субкластера. DB должен быть минимизирован.

DU не может найти правильное количество кластеров (ни подкластеров, ни общих кластеров). Хотя MB находит подкластеры, информация об иерархии не видна. В случае DB, даже если ему не удается найти пять кластеров (он находит четыре), иерархия подкластеров видна, поскольку значение индекса быстро падает при трех кластерах. Индекс SI не способен восстановить правильное количество кластеров (т. е. подкластеров), хотя и находит три общих кластера. Наконец, единственным индексом, способным дать правильные пять кластеров, а также указание на три общих кластера, является SF.

## 4.5 Ограничения

В предыдущих подразделах наборы данных, использованные для тестирования различных индексов, содержат гигантские сфероидальные кластеры. В этом подразделе мы кратко рассмотрим кластеры произвольной формы на примере двух новых наборов данных. *Pattern* - это набор данных, содержащий 258 точек в 2D. Он содержит три кластера с определенным рисунком и разными формами. *Rectangle cocmoum* из 1000 точек в 2D, которые представляют собой три прямоугольных кластера. Эти наборы данных показаны на рисунке 4.

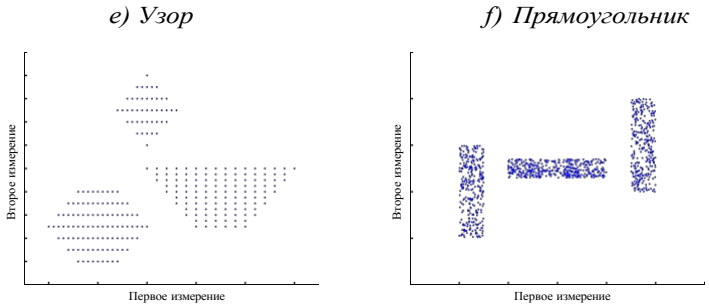


Рис. 4. Два новых искусственных набора данных. *Pattern* и *Rectangle* содержат соответственно 258 и 1000 точек в 2D.

Что касается набора данных *Pattern, mo* все индексы способны найти правильное количество кластеров (3). Предложенные фигуры и узор не выявили слабых мест ни в одном индексе. Что касается набора данных *"Прямоугольник",* то здесь результаты иные. Предложенная функция оценки не может обнаружить три кластера. Все остальные протестированные индексы также не справляются. Все индексы завышают правильное количество кластеров: DU (9), DB (7), SI (8), MB (8) и SF (10). Вероятное объяснение заключается в том, что кластеры имеют далеко не гиперсфероидальную форму. Таким образом, ограничением функции оценки, как и других протестированных индексов, является их ограничение наборами данных, содержащими гиперсфероидальные кластеры.

## 5 Выводы

Хотя в литературе существует несколько предложений по индексам валидности, большинство из них успешно работают только в определенных ситуациях. В данной работе представлен и исследован новый индекс для жесткой кластеризации - функция баллов (SF). Предлагаемый индекс основан на комбинации внутриклассовых и межклассовых расстояний. Он может учитывать особые случаи, такие как случай уникального кластера и случай идеального кластера. SF способен правильно оценить количество кластеров в нескольких искусственных и реальных наборах данных. SF успешно оценивает количество кластеров в наборах данных, содержащих несбалансированные, перекрывающиеся и зашумленные кластеры. Кроме того, SF был успешно протестирован на многомерных наборах реальных данных. Ни один другой индекс не показал столь же высоких результатов на всех наборах данных. Наконец, в случае субкластерных иерархий только SF смог оценить пять кластеров и в целом три группы. Таким образом, индекс SF превосходит четыре других индекса достоверности (Dunn, Davies-Bouldin, Silhouette и Maulik-Bandyopadhyay) для алгоритма k-means на гиперсфероидальных кластерах. Предложенный индекс также может учитывать идеальные и уникальные кластеры. Для определения единственного случая кластера было сформулировано эмпирическое условие. Наконец, определение значений для индекса является вычислительно эффективным.

В настоящее время ведется работа по расширению данной работы. Например, изучается теоретическое обоснование условия уникальности кластера (уравнение 9). Необходимо провести более обширное тестирование на кластерах произвольной формы. Наконец, проводятся исследования других алгоритмов кластеризации.

## Благодарности

Данное исследование финансируется Швейцарским национальным научным фондом (грант № 200020-109257). Авторы выражают признательность доктору Флере за плодотворные обсуждения и двум анонимным рецензентам за полезные комментарии.

## Ссылки

1. Jain, A., Dubes, R.: Algorithms for Clustering Data. Prentice Hall (1988)
2. Вебб, А.: Статистическое распознавание образов. Wiley (2002)
3. SanJuan, E., Ibekwe-SanJuan, F.: Text mining without document context. Инф. Process. Manage. 42(6) (2006) 1532-1552
4. Пердиши, Р., Джачинто, Г., Роли, Ф.: Кластеризация сигналов тревоги для систем обнаружения вторжений в компьютерных сетях. Инженерные приложения искусственного интеллекта 19(4) (2006) 429-438
5. Jaenichen, S., Perner, P.: Приобретение концептуальных описаний путем концептуальной кластеризации. In Perner, P., Amiya, A., eds: MLDM 2005. LNAI 3587, Springer-Verlag Berlin Heidelberg (2005) 153-162
6. Халкиди, М., Батистакис, Й., Вазиргианнис, М.: О методах проверки кластеризации. Журнал интеллектуальных информационных систем 17(2-3) (2001) 107-145
7. Хуанг, З.: Расширения алгоритма k-means для кластеризации больших наборов данных с категориальными значениями. Data Mining and Knowledge Discovery 2(3) (1998) 283-304

8. Jain, A.K., Murty, M.N., Flynn, P.J.: Кластеризация данных: обзор. ACM Computing Surveys 31(3) (1999) 264-323
9. Maulik, U., Bandyopadhyay, S.: Оценка производительности некоторых алгоритмов кластеризации и индексов валидности. IEEE Transactions Pattern Analysis Machine Intelli- gence 24(12) (2002) 1650-1654
10. Бездэк, Й., Пал, Н.: Некоторые новые показатели достоверности кластеров. IEEE Transactions on Systems, Man and Cybernetics 28(3) (1998) 301-315
11. Данн, Дж.: Хорошо разделенные кластеры и оптимальные нечеткие разбиения. Journal of Cyber- netics 4 (1974) 95-104
12. Дэвис Д., Болдин В.: Мера разделения кластеров. IEEE PAMI 1 (1979) 224-227
13. Kaufman, L., Rousseeuw, P.: Поиск групп в данных: введение в кластерный анализ. John Wiley & Sons (1990)
14. Ким, М., Рамакришна, Р.: Новые индексы для оценки валидности кластеров. Pattern Recognition Letters 26(15) (2005) 2353-2363
15. Большакова, Н., Азуахе, Ф.: Методы проверки кластеров для данных об экспрессии генома. Обработка сигналов 83(4) (2003) 825-833
16. Гордон А.: Валидация кластеров. In: Data science, classification and related methods (eds. Hayashi, C. and Yajima, K. and Bock H.H. and Ohsumi, N. and Tanaka, Y. and Baba, Y.). Springer (1996) 22-39
17. Линг, Р.: О теории и построении k-кластеров. Компьютерный журнал 15 (1972) 326-332
18. Чоу, К., Су, М., Лай, Е.: Новая мера достоверности кластеров и ее применение для сжатия изображений. Pattern Analysis Applications 7(2) (2004) 205-220
19. Бугесса, М., Ванг, С., Сун, Х.: Объективный подход к проверке кластеров. Pattern Recognition Letters 27(13) (2006) 1419-1430
20. Мерц, С., Мерфи, Р.: *UCI* машина обучение репозиторий (1996) http://www.ics.uci.edu/~mlearn/MLSummary.html.