УДК 004.855.5

# ВЫБОР ЧИСЛА КЛАСТЕРОВ В ЗАДАЧАХ КЛАСТЕРИЗАЦИИ С ИСПОЛЬЗОВАНИЕМ МЕТОДА СИЛУЭТОВ

***В.П. Корячко***
*Заведующий кафедрой систем автоматизированного проектирования РГРТУ им. В.Ф. Уткина, доктор технических наук, профессор*
*koryachko.v.p@rsreu.ru*

***В.И. Орешков***
*Доцент кафедры систем автоматизированного проектированияя РГРТУ им. В.Ф. Уткина, кандидат технических наук, доцент*
*vyacheslav.oreshkov@yandex.ru*

***В.П. Корячко***

*Окончил Рязанский радиотехнический институт. Область научных интересов связана с применением технологий машинного обучения в системах автоматизированного проектирования, разработкой методов представления знаний с использованием нечёткой логики и мягких вычислений.*

***В.И.Орешков***

*Окончил Рязанскую государственную радиотехническую академию. Область научных интересов связана с использованием технологий искусственного интеллекта в системах автоматизированного проектирования, машинным обучением, проектированием интеллектуальных информационных систем.*

**Аннотация.** Рассмотрена проблема выбора числа кластеров в задачах кластеризации. Предложена методика выбора числа кластеров на основе метода кластерных силуэтов. В результате эксперимента на реальных данных показано, что лучшее число кластеров с точки зрения минимума внутрикластерных и максимума междукластерных расстояний не всегда соответствует числу естественных групп данных.

**Ключевые слова:** интеллектуальный анализ данных, машинное обучение, обучение без учителя, обучающие данные, кластер, кластеризация, коэффициент силуэта.

**Введение.**

Одной из основных задач интеллектуального анализа данных (Data Mining), который называют ядром технологий Big Data [1], является кластеризация - обнаружение в структурированных данных групп наблюдений, близких по своим признакам, и называемых кластерами (от англ. cluster - гроздь, сгусток) [2]. В целом, задача кластеризация похожа на классификацию: в обоих задачах имеет место группировка объектов по близости значений их признаков. Но если классы заранее задаются, т.е. метки классов примеров в обучающем наборе данных должны быть известны (обучение с учителем), то в кластеризации кластеры заранее не определяются, а формируются в процессе построения модели исключительно на основе информации о расстоянии (обычно, евклидовом) между векторами наблюдений в пространстве признаков (обучение без учителя).

Формально, кластер можно определить, как группу объектов такую, что расстояние между любыми двумя объектами в группе будет меньше, чем расстояние между любым объектом группы и любым объектом из другого кластера. Кластеризация является одним из наиболее популярных методов анализа данных. Это связано с тем, что она лучше подходит для обнаружения новизны. Действительно, если в классификации классы задаются заранее, то любой новый объект, предъявленный классификатору, всегда будет отнесён к одному из имеющихся классов, что не позволит обнаружить появление объектов с принципиально новыми, ранее не наблюдавшимися свойствами. В случае кластеризации для объектов с нестандартными свойствами будет создан новый кластер, что позволит их обнаружить.

Практическое применение кластерных аналитических моделей заключается в содержательной интерпретации кластеров. Если свойства объектов внутри кластера известны, то они могут быть обобщены на любой новый объект, распределённой моделью в этот кластер. Таким образом, кластерная

модель любому предъявленному ей объекту должна присвоить номер или иную метку кластера. Например, если при кластеризации клиентов банка в некотором кластере окажутся клиенты с хорошей кредитной историей, то можно предположить, что любой новый клиент, распределённый в этот же кластер, тоже будет иметь хорошую кредитную историю.

Успех решения задачи кластеризации зависит от правильного выбора алгоритма кластеризации и его параметров, особенно числа кластеров. Все алгоритмы кластеризации можно разделить на три группы:

- машинного обучения (сети Кохонена, *k*-средних, CLOPE, FOREL);
- вероятностные, в которых наблюдениям присваивается не метки кластеров, а вероятность принадлежность объекта кластеру (*EM*-алгоритм);
- иерархические - формируют кластеры путём объединения (агломеративные) малых групп в большие или наоборот (дивизимные);
- графовые.

Выбор алгоритма не очевиден и может потребовать экспериментов или привлечения сведений о решении аналогичных задач. Тем не менее выбором алгоритма проблемы кластеризации не ограничиваются. Даже если алгоритм подходит, остаётся ещё задача выбора числа кластеров.

На первый взгляд здесь всё очевидно: число кластеров модели должно соответствовать числу групп наблюдений в исходном наборе данных. Целевой функцией, минимизируемой при обучении модели может быть средний квадрат расстояния от объектов, попавших в кластер, до его центра, что показывает насколько похожи объекты внутри кластера. Другой вариант - максимизировать расстояние от объектов одного кластера до центров других кластеров, что показывает степень различия объектов в разных кластерах. Действительно, чем сильнее объекты похожи внутри кластеров и чем выше междукластерные отличия, тем лучше кластеризация, и тем проще интерпретировать модель.

Однако, часто встречается ситуация, когда кластерная модель, оптимальная с точки зрения формальных критериев, оказывается совершенно бессмысленной с точки зрения целей и логики анализа. Причиной этого может быть разная кластеризующая способность признаков. Например, пусть требуется выполнить кластеризацию клиентов банка по трём признаками - доход, стаж работы и возраст. Если кластеризующая сила двух первых признаков значительно лучше, чем возраста, то мы получим кластеры, где по уровню дохода и стажу клиенты сгруппированы хорошо, а по возрасту - нет. Противоречие здесь в том, что по идее, чем выше возраст, тем выше стаж, и, в общем случае, доход. Однако при неудачной кластеризации данная тенденция окажется не выявлена, что может привести к неправильным выводам по результатам анализа. Сделать модель более интерпретируемой и повысить эффективность принятых с её помощью решений можно, подбирая число кластеров таким образом, чтобы оно отвечало, как формальным критериям точности, так и логике задачи, что требует поиска компромисса. Поэтому и в настоящее время число формируемых кластеров, как параметр алгоритма кластеризации, является дискуссионной темой в сообществе аналитиков, которое непрерывно генерирует какие-то новые идеи и решения в контексте данной проблемы.

Наиболее простым решением можно считать эвристическое правило: "не менее пяти, но не более десяти". Оно означает, что число кластеров менее пяти, скорее всего будет плохо отражать разнообразие свойств объектов, а больше десяти - будет очень сложным для интерпретации. Данное правило является, конечно, слишком общим и не учитывает особенностей конкретной задачи анализа, но позволяет получить приемлемый результат во многих практически значимых случаях. Тем не менее, в сообществе аналитиков бытует мнение, что однозначно лучшей кластеризации не бывает, поэтому получение хорошей кластерной модели может потребовать множества итераций выбора алгоритма кластеризации и его параметров, и особенной числа кластеров.

Кроме приведённого общего правила, в литературе можно встретить множество других критериев [3]:

- метод отношения дисперсий (индекс Калински-Карабаш);
- метод статистики разрыва;
- информационные критерии (Акаике и байесовский);
- метод *ISODATA*;
- метод Девиса-Болдуина;
- метод локтя и др.

Следует отметить, что большинство перечисленных методов не является универсальными, и каждый из них способен привести к разным кластерным решениям в рамках одной и той же задачи. В

связи с этим возникает проблема разработки подхода, который опирается не на какой-то формальный критерий или алгоритм выбора числа кластеров, а предоставляет аналитику возможность оперативного подбора данного параметра с хорошим визуальным представлением кластерной структуры, которая позволила бы оперативно подобрать приемлемое кластерное решение. Особенно данный подход представляет интерес в рамках разведочного анализа Тьюки, когда полученные решения впоследствии могут уточняться с помощью более строгих подходов.

В данной работе авторы рассматривают в качестве подхода к выбору числа кластеров использование кластерных силуэтов. Изначально метод разрабатывался для оценки качества кластерных моделей, но может быть использован и для оценки числа кластеров.

**Теоретическая часть.**

Впервые метод силуэтов в задаче кластеризации был предложен бельгийским статистиком Питером Руссо (Peter Rousseeuw) в 1987 г [4]. Изначально метод силуэта использовался для оценки качества кластеризации, а именно того, несколько результаты кластеризации соответствуют исходным данным. Метод предлагает компактное графическое представление результатов кластеризации. При этом для каждого объекта из обучающего набора данных вычисляется коэффициент силуэта, который отражает степень того, насколько данный объект похож на другие, попавшие в этот же кластер, и не похож на объекты из других кластеров. Значение коэффициента варьируется от -1 до1, при этом чем выше значение, тем лучше объект "подходит" своему кластеру, и менее "подходит" другим кластерам.

Таким образом, если в результате кластеризации окажется, что большинство объектов имеет высокое значение коэффициента силуэта (0,5 – 1), то и кластерная структура достаточно выражена и соответствует групповой структуре данных. Если же, напротив, большинство объектов имеет низкое значение коэффициента (около 0) или отрицательное, то результаты кластеризации плохие [5].

Метод силуэта можно использовать для большинства алгоритмов кластеризации и с любыми мерами расстояния, используемыми в них (евклидово, степенное, манхэттенское и др.).

Пусть имеются результаты кластеризации набора данных, при этом для каждой точки данных $i \in C_I$, может быть вычислено значение

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d_{ij},$$

где $C_I$ - кластер с номером *I*, $|C_I|$ - мощность кластера, $d_{ij}$ – расстояние между разными объектами *i* и *j* в кластере. Тогда $a_i$ можно проинтерпретировать как среднее расстояние между объектом *i* и всеми другими объектами этого же кластера. Эта величина показывает, насколько объект *i* "подходит" своему кластеру (чем меньше расстояние, тем лучше).

Теперь следует определить, насколько "плохо" объект *i* подходит к другим кластерам, для чего необходимо вычислить среднее расстояние от этого объекта до всех объектов некоторого кластера $C_J$, где $C_J \neq C_I$. После того, как среднее расстояние будет вычислено для всех кластеров, следует выбрать тот их них, для которого оно будет наименьшим, т.е.

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d_{ij}.$$

Тогда можно сказать, что кластер *J* является ближайшим к кластеру *I*, поэтому будет рассматриваться как следующий кандидат чтобы стать "подходящим" для объекта *i*. Теперь можно вычислить собственно коэффициент силуэта:

$$s(i) = \frac{b(i) - a(i)}{max\big(a(i), b(i)\big)},$$

если $C_I > 1$. Если $C_I = 1$, то $s(i) = 0, -1 \leq s(i) \leq 1$.

Следует отметить, что для кластеров, которые содержат только один элемент $a(i)$ не определено. В этом случае устанавливаем $s(i) = 0$, что соответствует середине интервала изменения величины.

Чтобы $s(i)$ оказалось близким к 1, необходимо чтобы выполнялось $a(i) \ll b(i)$. Поскольку $a(i)$ показывает, насколько объект $i$ в среднем отличается от других объектов в том же кластере, малое значение этой величины показывает хорошее соответствие объекта кластеру. Большое значение $b(i)$ указывает на плохое соответствие объекта $i$ соседнему кластеру. Таким образом, значение коэффициента силуэта $s(i)$ близкое к 1 указывает на высокую степень соответствия объекта $i$ своему кластеру, и низкую - соседнему, что и является признаком хорошей кластеризации. Значение $s(i)$ близкое к нулю означает, что объект расположен на границе двух естественных кластеров.

Среднее значение $s(i)$ по всем объектам кластера является мерой того, насколько плотно сгруппированы объекты кластере. Таким образом, среднее значение $s(i)$ по всем объектам исходного набора данных показывает, насколько группировка данных в кластеры соответствует их естественной группировке. Если число кластеров больше или меньше числа естественных групп, что имеет место при неудачном выборе числа кластеров в модели, то из-за низких значений коэффициента силуэта, силуэт всего "плохого" кластера будет уже, чем у остальных. Это происходит, когда один естественный кластер оказывается разбит на два или более результирующих кластера, либо два или более естественных кластера окажутся объединены в один. В любом случае, количество естественных и результирующих кластеров оказывается несоответствующим. Таким образом, кластерные силуэты удобно использовать для выбора числа кластеров в модели, соответствующему числу естественных групп а данных.

Для этого используются специального вида диаграммы, которые так и называют - диаграммы силуэтов [5]. На диаграмме для каждого объекта коэффициент силуэта отображается прямоугольником соответствующей длины. Прямоугольники группируются по кластерам (которые обычно выделяются цветом) и в каждом кластере дополнительно ранжируются в порядке убывания. Общий вид диаграммы силуэтов представлен на рис. 1.

Таким образом, на диаграмме становится виден силуэт каждого кластера. По форме силуэтов аналитик оперативно может оценить качество кластеризации. Чем форма силуэтов ближе к прямоугольной, а площадь (средний коэффициент силуэта) ближе к 1, тем лучше кластеризация. Внутри силуэта каждого кластера объекты расположены в порядке убывания их коэффициента силуэта, поэтому легко увидеть, какие именно объекты лучше соответствуют кластеру, а какие хуже.

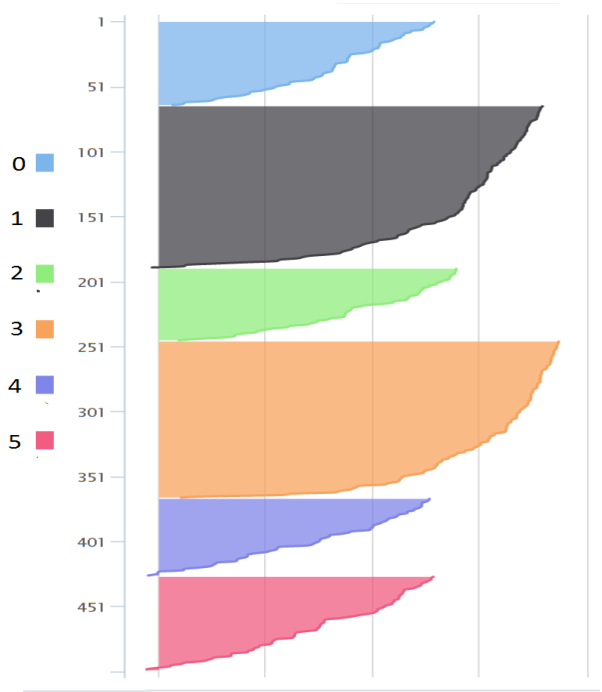


*Рисунок 1*. Общий вид диаграммы силуэтов

**Экспериментальные исследования.**

Рассмотрим набор данных, фрагмент которого представлен на рис. 2. Он содержит информацию о заёмщиках кредитной организации, кластеризация которых позволит выявить влияние признаков,

описывающих заёмщика, на вероятность допуска им просрочки по кредиту и оценки её размера.

| ID | Возраст | Личный доход | Сумма последнего кредита, руб | Первоначальный взнос, руб | Время проживания по месту пребывания, мес# | Время работы на текущем месте, мес# | Средняя сумма просрочки, руб |
|---|---|---|---|---|---|---|---|
| 1 | 32 | 18000 | 11751 | 8389 | 49 | 27 | 0 |
| 2 | 62 | 7000 | 11160 | 2790 | 324 | 48 | 0 |
| 3 | 41 | 60000 | 79910 | 19990 | 5 | 168 | 0 |
| 4 | 34 | 35000 | 7290 | 1300 | 84 | 13 | 1439,94 |
| 5 | 31 | 12000 | 30198 | 5752 | 163 | 15 | 5760 |
| 6 | 35 | 15000 | 16400 | 3000 | 36 | 24 | 3070 |
| 7 | 34 | 6000 | 3000 | 790 | 38 | 180 | 0 |
| 8 | 33 | 10000 | 7795 | 7795 | 36 | 12 | 0 |
| 9 | 27 | 10000 | 9997 | 1120 | 108 | 12 | 0 |
| 10 | 57 | 3000 | 3280 | 3000 | 252 | 48 | 0 |
| 11 | 54 | 11000 | 22072 | 8008 | 18 | 15 | 0 |
| 12 | 27 | 12000 | 6600 | 1170 | 48 | 12 | 0 |

*Рисунок 2*. Исходный набор данных для кластеризации (фрагмент)

В данной работе не ставится цель провести содержательную интерпретацию кластеров для оценки кредитоспособности клиентов – это задача кредитных менеджеров. Наша цель пояснить на конкретном примере методику выборы числа кластеров с помощью диаграммы силуэтов.

Рассмотрим случай двух кластеров (рис. 3). Для удобства визуального восприятия структура кластеров приведена в двумерной проекции с сохранением топологического подобия, когда на диаграмме показываются расстояния между объектами.

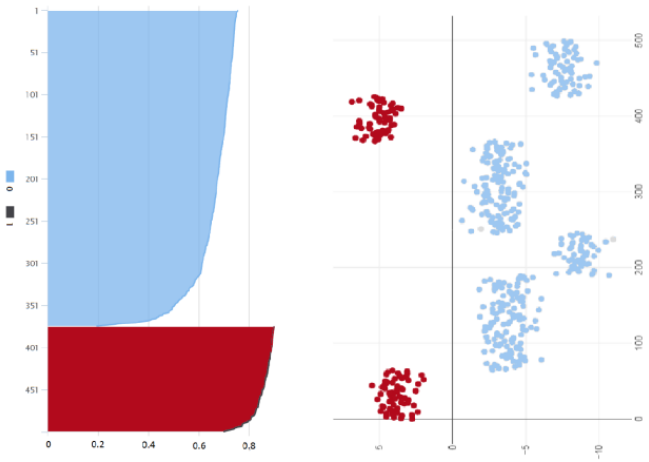

*Рисунок 3.* Диаграмма силуэтов для случая 2-х кластеров

На рисунке слева представлена диаграмма силуэтов, на которой представлены силуэты двух кластеров, сгенерированных моделью. На правой части рисунка представлена естественная группировка данных, где хорошо различимы 6 групп. Т.е. количество реальных групп не соответствует числу кластеров. В результате на диаграмме мы наблюдаем неравномерность ширины силуэтов.

Случай для 3-х кластеров представлен на рисунке 4.

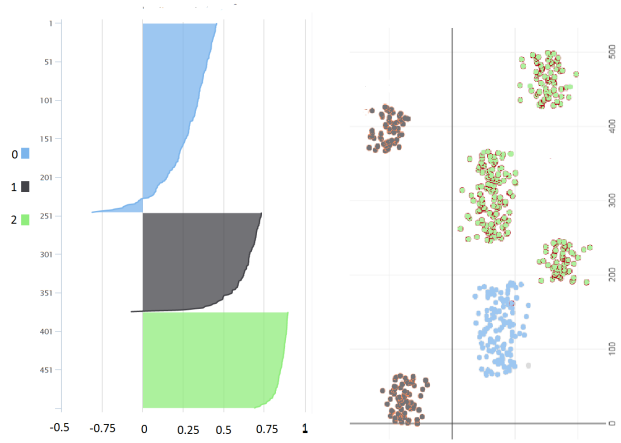


*Рисунок 4*. Кластеризация для случая 3 кластеров

Для случая 3-х кластеров ситуация усугубляется: силуэт нулевого кластера становится узким, т.е. для него имеют место низкие значения коэффициента силуэта, и даже появляются отрицательные. Это говорит о том, что в среднем, объекты данного кластера слабо связаны и кластеризация по - прежнему не оптимальна.

Для случая 4 кластеров диаграмма силуэтов представлена на рис. 5. На ней силуэты кластеров достаточно широкие и равномерные, что говорит о хорошей кластеризации, когда объекты в среднем хорошо соответствуют своим кластерам, и плохо - остальным кластерам. Следует отметить, что данная ситуация имеет место для случая, когда визуально число групп и число кластеров не совпадает, что вызывает определённое противоречие.

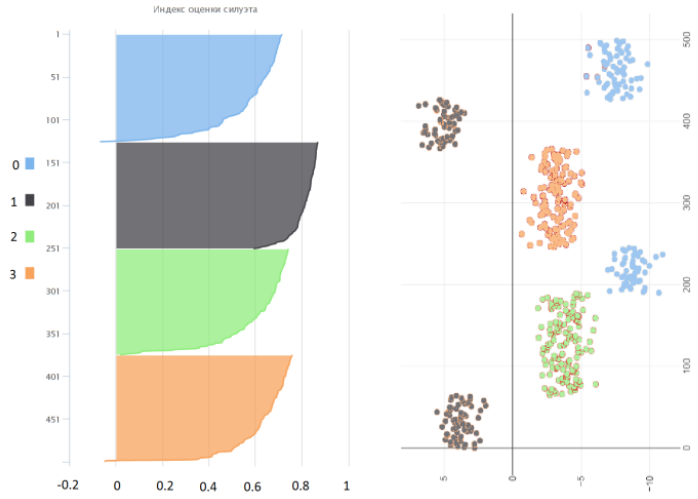


*Рисунок 5*. Диаграмма силуэтов для случая 4-х кластеров

И, наконец, диаграммы для случаев 5 и 6 кластеров совмещены на рис. 6. Это случай, когда число кластеров наиболее приближено к числу естественных групп данных и ему должна соответствовать лучшая диаграмма силуэтов. Однако на практике это оказалось не так.

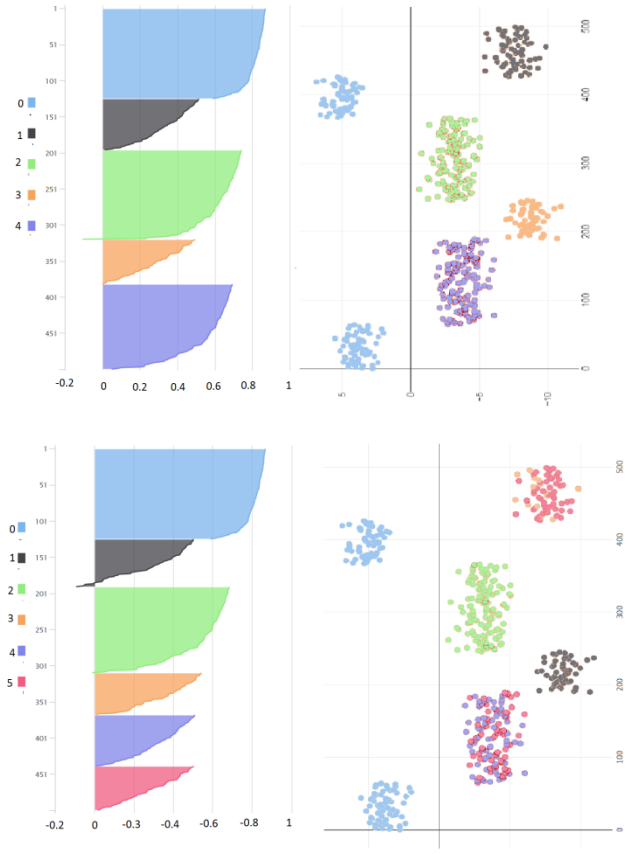

*Рисунок 5*. Диаграмма силуэтов для случая 5 и 6 кластеров

Для 5 кластеров на диаграмме имеет место появление двух кластеров с узким силуэтом (кластеры 1 и 3), а на диаграмме для 6 кластеров – таковых уже 4 (1, 3, 4 и 5). Напомним, что кластеры с узким силуэтом имеют низкий средний индекс силуэта, что говорит о низком уровне соответствия объектов своему кластеру. Проще говоря, объекты внутри кластера не очень похожи друг на друга.

Поэтому несмотря на то, что случаи 5 и 6 кластеров лучше соответствуют естественной группировке данных, с точки зрения метода силуэта предпочтение имеет модель для 4-х кластеров, которая обеспечивает наименьшее внутрикластерное и наибольшее междукластерное расстояния, что соответствует лучшей кластеризации.

**Выводы**.

Таким образом, экспериментальные исследования с использованием метода силуэтов на примере кластеризации клиентов банка показали, что лучшая кластеризация с точки зрения минимума средних внутрикластерных расстояний и максимума междукластерных, не обязательно достигается, когда число кластеров как параметра модели соответствует числу естественных групп данных. Это способно вызвать определённые противоречия при содержательной интерпретации кластеров. Применение метода силуэтов позволяет обнаруживать такие противоречия и помогать аналитикам делать более точные выводы о предметной области, описываемой исходными данными.

Вместе с тем, обобщение полученных на конкретном примере результатов даже на аналогичные задачи следует производить с осторожностью. Это связано с тем, что разные алгоритмы кластеризации могут создавать различные кластерные структуры даже на одних и тех же данных. Поэтому подобные исследования должны производиться отдельно в каждом конкретном случае. Хотя сама по себе методика, изложенная выше, является достаточно универсальной.

**Заключение.**

Таким образом в работе рассмотрена техника применения кластерных силуэтов для выбора числа кластеров в задачах кластеризации. На решении практической задачи с использованием реальных данных показано, что число кластеров модели с точки зрения минимизации внутрикластерных расстояний и максимизации междукластерных, не всегда достигается при соответствии числа кластеров модели

естественной группировки данных.

**Список литературы**

[1] Паклин Н.Б. Бизнес-аналитика: от данных к знаниям (+ CD): учеб. пособие. / Паклин Н.Б., Орешков В.И. 2-е изд., испр. – СПб.: Питер, 2013. – 704 с.

[2] Корячко В.П. Интеллектуальные системы и нечеткая логика / В.П. Корячко, М.А. Бакулева, В.И. Орешков. – М.: КУРС, 2017. – 346 с.

[3] Орешков В.И. Повышение точности классификации данных с использованием алгоритма k-ближайших соседей на основе прекластеризации обучающих данных / Орешков В.И. Вестник Рязанского гос. радиотехнического университета. Рязань: РГРТУ, – 2021. – № 76. – С. 65–73.

[4] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. Volume 20, November 1987, Pages 53-65.

[5] Паклин Н.Б. Кластерные силуэты / Паклин Н.Б., Орешков В.И. Сб. научн. тр. XX Международной науч.-практ. конф. «Системный анализ в проектировании и управлении». СПб, 29 июня-01 июля 2016 г. Изд-во ФГАОУ ВО «Санкт-Петербургский политехнический ун-тет Петра Великого». – 2016. – с. 314-321.

# CHOICE OF THE NUMBER OF CLUSTERS IN CLUSTERING PROBLEMS USING THE SILHOUETTE METHOD

***V.P. Koryachko***
*Head of the Department of Computer-Aided Design Systems, Ryazan State Radio Engineering University named after V.F. Utkin, Doctor of Technical Sciences, Professor*

***V.I. Oreshkov***
*Associate Professor of the Department of Computer-Aided Design Systems, Ryazan State Radio Engineering University named after V.F. Utkin, candidate of technical sciences*

*Department of Computer-Aided Design Systems*
*Faculty of Computer Science*
*Ryazan State Radio Engineering University named after V.F. Utkin*
*E-mail: vyacheslav.oreshkov@yandex.ru*

**Abstract.** The problem of choosing the number of clusters in clustering problems is considered. A technique for choosing the number of clusters based on the cluster silhouette method is proposed. As a result of the experiment on real data, it is shown that the best number of clusters in terms of the minimum intracluster and maximum intercluster distances does not always correspond to the number of natural data groups.

**Keywords**:.data mining, machine learning, unsupervised learning, training data, cluster, clustering, silhouette coefficient.