УДК 519.237.8

ПРИНЦИП РАЗДЕЛЕНИЯ КЛАССОВ И ЗАВИСИМОСТЬ РЕЗУЛЬТАТОВ КЛАСТЕРИЗАЦИИ ОТ МЕРЫ РАССТОЯНИЯ

Абдуллаев К.Д.
abdullaevkomiljon1747@gmail.com
Национальный университет Узбекистана имени Мирзо Улугбека,
100174, Узбекистан, г. Ташкент, ул. Университетская, 4

В статье рассматривается методика кластерного анализа многомерных объектов, основанная на вычислении локальной плотности, первичном разделении объектов на два класса K_1 и K_2, выделении граничных объектов, корректировке состава классов и последующей группировке по отношению связанности. Исследуется влияние правила первичного разбиения и выбранной меры расстояния на итоговую структуру кластеризации. В качестве правил разбиения используются разбиение по среднему значению локальной плотности и медианное разбиение упорядоченной последовательности плотностей. Для сравнительного анализа рассматриваются евклидова и манхэттенская меры расстояния, а также меры Чебышева, Канберры и Хэмминга. Основными характеристиками результата являются итоговый состав классов, число переходящих объектов, количество сформированных групп, состав групп и мера компактности групп $\Omega(E_0)$. Показано, что предлагаемая методика позволяет автоматически формировать группы без предварительного задания их числа и сравнивать полученные структуры при различных мерах расстояния. Полученные результаты подтверждают, что выбор меры расстояния и правила первичного разбиения существенно влияет на число групп, их состав и меру компактности групп. Методика может быть использована для анализа многомерных данных в задачах выявления скрытых взаимосвязей, граничных объектов и компонент групповой структуры.

Ключевые слова: кластерный анализ, принцип разделения классов, локальная плотность, мера расстояния, медианное разбиение, граничные объекты

Цитирование: *Абдуллаев К.Д.* Принцип разделения классов и зависимость результатов кластеризации от меры расстояния // Проблемы вычислительной и прикладной математики. – 2026. – № 5(76). – С. 33-45.

DOI: 10.71310/pcam.5_76.2026.03

1 Введение

Кластерный анализ является одним из основных инструментов интеллектуального анализа данных, предназначенным для выявления скрытых закономерностей, внутренних связей и естественных групп объектов в многомерном признаковом пространстве. В отличие от задач классификации с заранее заданными метками классов, в кластеризации структура данных заранее неизвестна. Поэтому основная трудность состоит не только в разбиении объектов на группы, но и в выборе такого принципа разбиения, который позволял бы выявлять устойчивые взаимосвязи между объектами и интерпретировать внутреннюю структуру данных [5, 8].

Классические методы кластеризации, такие как k-means, иерархические методы, DBSCAN и их модификации, широко применяются для анализа многомерных данных. Однако каждый из этих методов имеет определённые ограничения. Методы, основанные на центрах кластеров, часто предполагают компактную или близкую к

сферической конфигурацию групп. Плотностные методы позволяют выявлять кластеры более сложной формы, однако их результаты существенно зависят от выбора параметров локальной области и меры расстояния. Иерархические методы дают структуру вложенных групп, но не всегда позволяют однозначно определить критерий окончательного разбиения. Следовательно, при неизвестной конфигурации данных, особенно когда группы не имеют сферической формы, возникает необходимость в дополнительных критериях, позволяющих анализировать не только расстояния между объектами, но и отношения между ними [5, 7–11].

В данной работе рассматривается подход, основанный на выявлении взаимосвязей между объектами через локальную плотность, граничные объекты и отношение связанности. Основная идея заключается в том, что скрытая структура данных проявляется не только в компактности объектов, но и в том, какие объекты находятся на границе классов, какие объекты формируют устойчивые локальные связи и как изменяется итоговое разбиение при использовании различных мер расстояния. Такой подход позволяет перейти от простого кластерного разбиения к более содержательному анализу внутренних связей между объектами.

Теоретической основой работы является подход Н.А. Игнатьева, связанный с анализом отношений между объектами в метрическом пространстве [1, 2]. Данный подход важен тем, что позволяет рассматривать близость, граничность и связанность объектов как самостоятельные характеристики структуры данных. В отличие от традиционных критериев кластеризации, ориентированных преимущественно на компактность и разделимость групп, здесь особое внимание уделяется отношениям между объектами. Именно поэтому в работе рассматривается не только результат кластеризации, но и принцип формирования классов, переход объектов между классами и зависимость полученных групп от меры расстояния.

Отдельного внимания требует вопрос первичного разделения объектов на классы. В классическом варианте классы могут формироваться по условию превышения значения локальной плотности над средним значением. Однако среднее значение чувствительно к выбросам и асимметричному распределению значений плотности [14]. В результате один из классов может оказаться существенно больше другого, а начальное разбиение — несбалансированным. Поэтому в данной работе дополнительно рассматривается медианный принцип разбиения. Медиана позволяет разделить упорядоченную последовательность значений локальной плотности на две относительно равные части и тем самым получить более сбалансированные начальные классы K_1 и K_2.

Переход от кластерного анализа к формированию классов в данной работе не означает переход к классической задаче классификации с заранее известными метками. Речь идёт о предварительном разделении объектов на классы на основе значений локальной плотности. Сначала множество объектов разделяется на два класса K_1 и K_2 по выбранному правилу, затем выделяются граничные объекты, выполняется корректировка состава классов, после чего внутри каждого класса формируются группы по отношению связанности. Таким образом, классификация здесь используется как промежуточный этап кластерного анализа, необходимый для выявления скрытых отношений между объектами.

В известных работах по кластерному анализу основное внимание обычно уделяется построению кластеров, выбору числа групп, оценке компактности или сравнению алгоритмов. Однако недостаточно изученным остаётся вопрос о том, как именно принцип первичного разделения объектов и выбранная мера расстояния влияют на

последующую структуру связанности. Особенно это важно в тех случаях, когда форма кластеров заранее неизвестна, а традиционные критерии компактности и разделимости не дают полного представления о внутренних отношениях между объектами. Поэтому возникает необходимость в критерии, который позволял бы анализировать не только итоговые группы, но и скрытые связи, возникающие между объектами в процессе разбиения [3, 4, 13].

Сравнение различных мер расстояния в данной работе проводится не только для выбора условно наилучшей меры, но и для анализа устойчивости получаемой кластерной структуры. Более важная цель состоит в выявлении скрытых закономерностей: какие меры расстояния формируют более устойчивые классы, какие приводят к большему числу переходящих объектов, какие создают крупные компоненты связанности, а какие вызывают фрагментацию групп. Иными словами, сравнение мер расстояния используется как инструмент извлечения знаний о внутренней структуре данных. Полученные экспериментальные выводы интерпретируются применительно к рассматриваемому набору данных и выбранной схеме вычислительного эксперимента.

В отличие от подхода Н.А. Игнатьева, в данной работе дополнительно предлагается медианное правило первичного разделения объектов на классы. Исследуется зависимость результатов кластеризации от правила первичного разбиения и выбора меры расстояния. Сравнительный анализ проводится для пяти мер расстояния: евклидовой, манхэттенской, Чебышева, Канберры и Хэмминга. Полученные результаты сопоставляются с разбиением по среднему значению локальной плотности.

Целью статьи является исследование зависимости результатов кластеризации от принципа первичного разделения объектов на классы и выбранной меры расстояния, а также программная реализация методики, позволяющей выявлять устойчивые и неустойчивые взаимосвязи между объектами на основе локальной плотности, граничных объектов и отношения связанности.

2 Постановка задачи

Пусть задано множество объектов:

$$E_0 = \{S_1, S_2, \ldots, S_m\},$$

где каждый объект описывается набором количественных признаков:

$$X(n) = (x_1, x_2, \ldots, x_n).$$

Для любых двух объектов S_i и S_j определяется расстояние: $\rho(S_i, S_j)$, где ρ — выбранная метрика.

Требуется исследовать, каким образом принцип первичного разделения объектов на классы и выбранная мера расстояния влияют на итоговые результаты кластеризации. В качестве основного предположения принимается, что структура данных заранее неизвестна, а форма возможных групп может быть несферической. Поэтому анализ должен учитывать не только расстояния между объектами, но и локальную плотность, граничные объекты, переход объектов между классами и отношение связанности.

Для устранения влияния масштаба признаков исходные данные нормируются в интервал $[0, 1]$. Для каждого объекта вычисляется значение локальной плотности в области радиуса ε, где радиус определяется через расстояние до k-го ближайшего

соседа. На основе значений локальной плотности формируются два класса объектов K_1 и K_2.

В работе рассматриваются два правила первичного разделения объектов на классы:

- разбиение по среднему значению локальной плотности;
- медианное разбиение упорядоченной последовательности значений локальной плотности.

Для анализа зависимости результатов от меры расстояния рассматриваются пять мер: евклидова, манхэттенская, Чебышева, Канберры и Хэмминга:

$$\rho_E(S_i, S_j) = \sqrt{\sum_{r=1}^{n}(x_{ir} - x_{jr})^2},$$

$$\rho_M(S_i, S_j) = \sum_{r=1}^{n}|x_{ir} - x_{jr}|,$$

$$\rho_C(S_i, S_j) = \max_{1\leqslant r\leqslant n}|x_{ir} - x_{jr}|,$$

$$\rho_{Can}(S_i, S_j) = \sum_{r=1}^{n}\frac{|x_{ir} - x_{jr}|}{|x_{ir}| + |x_{jr}|},$$

$$\rho_H(S_i, S_j) = \sum_{r=1}^{n} I(x_{ir} \neq x_{jr}).$$

Данные меры выбраны не с целью окончательного определения универсально лучшей метрики, а для демонстрации того, что изменение способа измерения близости между объектами может существенно менять итоговую кластерную структуру. Первые четыре меры учитывают величину числового различия между признаками, тогда как мера Хэмминга фиксирует факт несовпадения значений признаков. Это позволяет показать различный характер влияния мер расстояния на результат кластеризации.

После первичного разделения объектов на классы K_1 и K_2 необходимо выделить граничные объекты, выполнить корректировку состава классов и сформировать группы внутри каждого класса по отношению связанности. Далее полученные результаты сравниваются между собой.

Качество и устойчивость полученных результатов оцениваются по следующим показателям [4, 12, 13]:

- сбалансированность начального распределения объектов между классами K_1 и K_2;
- число объектов, перешедших из K_1 в K_2 и из K_2 в K_1 после корректировки;
- итоговое число объектов в классах K_1 и K_2;
- число сформированных групп внутри каждого класса;
- общее число групп;
- степень фрагментации групповой структуры;
- устойчивость объектного состава при использовании различных мер расстояния.

Таким образом, задача исследования состоит в том, чтобы на основе локальной плотности, граничных объектов и отношения связанности сравнить два принципа первичного разделения объектов на классы и определить, как различные меры расстояния влияют на итоговую структуру кластеризации, устойчивость классов и выявление скрытых взаимосвязей между объектами.

Для решения поставленной задачи используется алгоритмическая схема кластеризации, включающая следующие этапы: нормирование исходных данных, вычисление матрицы расстояний, определение локальной плотности объектов, первичное формирование классов K_1 и K_2, выделение граничных объектов, корректировку состава классов, построение групп по отношению связанности и вычисление меры компактности групп $\Omega(E_0)$.

3 Оценка плотности распределения объектов

Для каждого объекта S_i вычисляется значение плотности распределения в локальной области радиуса ε. Радиус определяется как среднее расстояние до k-го ближайшего соседа:

$$\varepsilon = \frac{1}{m}\sum_{i=1}^{m}\rho_k(S_i),$$

где $\rho_k(S_i)$ — расстояние от объекта S_i до его k-го ближайшего соседа. При определении k ближайших соседей объекта S_i сам объект S_i не учитывается.

Плотность распределения объекта S_i определяется выражением:

$$\varphi(S_i,\varepsilon) = \sum_{\substack{S_j\in E_0\\ \rho(S_i,S_j)<\varepsilon}}\left(1-\frac{\rho(S_i,S_j)}{\varepsilon}\right).$$

В отличие от определения k ближайших соседей, при вычислении $\varphi(S_i,\varepsilon)$ объект S_i включается в сумму: $\rho(S_i,S_i)=0$, поэтому при $\varepsilon>0$ его вклад равен единице.

Чем больше значение $\varphi(S_i,\varepsilon)$, тем выше локальная плотность вокруг объекта S_i. Далее вычисляется среднее значение плотности:

$$M(\varepsilon) = \frac{1}{m}\sum_{i=1}^{m}\varphi(S_i,\varepsilon).$$

В работе рассматриваются два правила первичного формирования классов K_1 и K_2. Первый вариант основан на сравнении значения локальной плотности со средним значением $M(\varepsilon)$, второй — на медианном разбиении упорядоченной последовательности значений локальной плотности.

4 Правила первичного разбиения объектов на классы

4.1 Разбиение по среднему значению плотности

При первом варианте объект S_i относится к классу K_1, если значение его локальной плотности превышает среднее значение $M(\varepsilon)$. В противном случае объект относится к классу K_2.

4.2 Медианное разбиение

Последовательность объектов сортируется по возрастанию значений φ, после чего первая половина объектов относится к классу K_1, а вторая половина — к классу K_2.

После вычисления значений плотности формируется упорядоченная последовательность:

$$\varphi(S_1) \leqslant \varphi(S_2) \leqslant \cdots \leqslant \varphi(S_m).$$

Если m — чётное, то

$$K_1 = \{S_1, S_2, \ldots, S_{m/2}\},$$

$$K_2 = \{S_{m/2+1}, \ldots, S_m\}.$$

Если m — нечётное, то центральный объект S_c сравнивается с соседними по значению плотности объектами. Если

$$|\varphi(S_c) - \varphi(S_{c-1})| \leqslant |\varphi(S_c) - \varphi(S_{c+1})|,$$

то он относится к K_1. В противном случае он относится к K_2.

Такой подход позволяет получать два класса объектов на основе распределения значений локальной плотности.

При сопоставлении двух правил следует учитывать, что при разбиении по среднему значению класс K_1 содержит объекты с плотностью выше среднего, а при медианном разбиении — объекты из нижней части упорядоченной последовательности плотностей. Поэтому показатели $\Omega(K_1)$ и $\Omega(K_2)$ для двух правил относятся к различным по смыслу классам.

4.3 Обоснование использования медианы

Замена среднего значения на медианное разбиение связана с тем, что среднее значение чувствительно к выбросам и асимметричному распределению значений плотности. Медианный подход позволяет получить более сбалансированное начальное разбиение объектов на классы K_1 и K_2 и уменьшает влияние экстремальных значений локальной плотности.

Кроме двух исследуемых в данной работе правил, теоретически могут использоваться и другие варианты формирования классов K_1 и K_2: разбиение по квантилям, по фиксированному пороговому значению плотности, а также через плотные центры и ε-окрестности. Однако в вычислительном эксперименте данной статьи рассматриваются только два правила: по среднему значению и по медиане.

Нельзя однозначно утверждать, что медианное разбиение лучше разбиения по среднему значению по всем критериям. Медиана предпочтительна при требовании сбалансированности классов, а среднее значение — при ориентации на уменьшение числа переходящих объектов (для части мер расстояния — и числа групп).

5 Выделение граничных объектов

Для каждого объекта одного класса определяется ближайший объект из противоположного класса. Объект противоположного класса считается граничным, если он является ближайшим хотя бы для одного объекта другого класса.

Для класса K_1 множество граничных объектов определяется следующим образом:

$$B_1 = \{S_j \in K_1 \mid S_j = \arg\min_{S \in K_1} \rho(S_i, S),\ S_i \in K_2\}.$$

Аналогично для класса K_2

$$B_2 = \{S_j \in K_2 \mid S_j = \arg\min_{S \in K_2} \rho(S_i, S),\ S_i \in K_1\}.$$

Множество всех граничных объектов:

$$B(E_0, \rho) = B_1 \cup B_2.$$

Граничные объекты имеют особое значение, поскольку именно через них определяется связь между классами и возможность корректировки состава классов.

6 Корректировка состава классов

Для каждого объекта $S_i \in K_k$, $k \in \{1, 2\}$, определяется расстояние до ближайшего объекта противоположного класса:

$$r(S_i) = \min_{S_j \in K_{3-k}} \rho(S_i, S_j),$$

где $r(S_i)$ — расстояние до ближайшего объекта противоположного класса.

Для каждого граничного объекта дополнительно вычисляется расстояние до ближайшего объекта своего класса, отличного от S_i:

$$d(S_i) = \min_{S_j \in K_k,\, j \neq i} \rho(S_i, S_j),$$

где $d(S_i)$ — расстояние до ближайшего объекта своего класса.

Затем вычисляется отношение:

$$\eta(S_i) = \frac{r(S_i)}{d(S_i)}.$$

Если:

$$\eta(S_i) < 1,$$

то объект S_i расположен ближе к противоположному классу, чем к собственному. Такой объект рассматривается как кандидат на переход в противоположный класс в соответствии с принятым критерием отношения расстояний.

В программной реализации условие перехода проверяется для граничных объектов исходных классов. Решения о переходе в обоих направлениях принимаются по исходному разбиению, после чего переносы выполняются однократно, без повторного применения критерия. Объекты, для которых выполняется условие $\eta(S_i) < 1$, переносятся из одного класса в другой:

$$K_1 \to K_2$$

или

$$K_2 \to K_1.$$

После корректировки формируются новые классы:

$$K_1^{\text{new}}, \qquad K_2^{\text{new}}.$$

По полученным классам повторно определяются множества граничных объектов B_1^{new} и B_2^{new} по правилам разд. 5. Повторного переноса объектов при этом не выполняется.

7 Группировка объектов по отношению связанности

После корректировки состава классов выполняется разбиение объектов внутри каждого из классов K_1^{new} и K_2^{new} на группы. Для краткости далее эти итоговые классы обозначаются K_1 и K_2, а их заново найденные множества граничных объектов — B_1 и B_2. Два объекта $S_u, S_v \in K_t$, $t \in \{1, 2\}$, считаются связанными, если существует граничный объект $S \in B_t$ такой, что:

$$\rho(S_u, S) < r(S_u) \quad \text{и} \quad \rho(S_v, S) < r(S_v),$$

где $r(S_u)$ и $r(S_v)$ — расстояния от объектов S_u и S_v до ближайших объектов противоположного класса, вычисленные для итогового разбиения. Для каждого $S_i \in K_t$ используется определение $r(S_i) = \min\limits_{S_j \in K_{3-t}} \rho(S_i, S_j)$.

Группировка выполняется отдельно внутри каждого класса. Если между объектами существует такая связь, то они помещаются в одну компоненту связности. При построении компонент учитываются также цепочки попарно связанных объектов (транзитивное замыкание отношения). В результате каждый класс разбивается на непересекающиеся группы:

$$K_1 = J_{11} \cup J_{12} \cup \cdots \cup J_{1p},$$
$$K_2 = J_{21} \cup J_{22} \cup \cdots \cup J_{2q}.$$

Иными словами, два объекта одного класса считаются связанными, если они одновременно находятся в локальной области влияния некоторого граничного объекта. Такое правило позволяет выделять внутри каждого класса не только отдельные объекты, но и компоненты связности, отражающие внутреннюю структуру класса. Количество групп и число объектов в них являются важными характеристиками структуры кластеризации.

8 Мера компактности групп по отношению связанности

Пусть $J_{11}, \ldots, J_{1p}$ и $J_{21}, \ldots, J_{2q}$ — множества непересекающихся групп, полученных по отношению связанности объектов по классам K_1 и K_2. Группы представляют собой компоненты связности, построенные отдельно внутри каждого итогового класса с использованием его граничных объектов. Мера компактности групп по классам и выборки в целом [2] определяется как

$$\Omega(K_1) = \frac{|J_{11}|^2 + \cdots + |J_{1p}|^2}{|K_1|^2},$$
$$\Omega(K_2) = \frac{|J_{21}|^2 + \cdots + |J_{2q}|^2}{|K_2|^2},$$
$$\Omega(E_0) = \frac{|K_1|\,\Omega(K_1) + |K_2|\,\Omega(K_2)}{m}. \tag{1}$$

Для непустых классов и непересекающихся непустых групп выполняется $0 < \Omega(E_0) \leqslant 1$. Значение $\Omega(E_0) = 1$ достигается, если каждый класс представлен одной группой. Мера зависит от размеров групп и не использует известные метки классов.

9 Вычислительный эксперимент

В качестве экспериментального набора данных использовалась выборка Ionosphere, содержащая 351 объект и 34 количественных признака [6]. Перед вычислениями данные нормировались в интервал $[0, 1]$. Для оценки влияния мер расстояния были проведены вычисления с использованием евклидовой и манхэттенской мер расстояния, мер Чебышева, Канберры и Хэмминга. Параметр локальной области задавался через расстояние до k-го ближайшего соседа. В эксперименте использовалось значение: $k = 3$. Сравнительные результаты представлены в табл. 1–3.

Таблица 1 Результаты кластерного анализа при медианном разбиении

Метрика	Начальные		Переходы		Итоговые		Число групп		
	K_1	K_2	$K_1 \to K_2$	$K_2 \to K_1$	K_1	K_2	K_1	K_2	всего
Евклидова	176	175	27	3	152	199	107	3	110
Манхэттенская	176	175	27	1	150	201	68	6	74
Чебышева	175	176	30	2	147	204	135	4	139
Канберры	175	176	19	2	158	193	94	3	97
Хэмминга	176	175	22	0	154	197	151	40	191

Таблица 2 Результаты кластерного анализа при разбиении по среднему значению

Метрика	Начальные		Переходы		Итоговые		Число групп		
	K_1	K_2	$K_1 \to K_2$	$K_2 \to K_1$	K_1	K_2	K_1	K_2	всего
Евклидова	147	204	6	8	149	202	2	59	61
Манхэттенская	167	184	4	20	183	168	8	67	75
Чебышева	154	197	1	17	170	181	1	121	122
Канберры	177	174	2	13	188	163	1	70	71
Хэмминга	124	227	0	17	141	210	1	207	208

Таблица 3 Мера компактности групп при $k = 3$

Метрика	Медианное разбиение			Разбиение по среднему		
	$\Omega(K_1)$	$\Omega(K_2)$	$\Omega(E_0)$	$\Omega(K_1)$	$\Omega(K_2)$	$\Omega(E_0)$
Евклидова	0,018611	0,849701	0,489799	0,986667	0,313695	0,599373
Манхэттенская	0,139556	0,699413	0,460158	0,925169	0,165249	0,561447
Чебышева	0,008561	0,793829	0,464956	1,000000	0,093740	0,532669
Канберры	0,066976	0,959247	0,557598	1,000000	0,165569	0,612501
Хэмминга	0,006747	0,644258	0,364552	1,000000	0,004943	0,404667

В вычислительном эксперименте при $k = 3$ основное значение имеет не только итоговое распределение объектов между классами K_1 и K_2, но и число сформированных групп, а также их состав. В отличие от методов, где число кластеров задаётся заранее, в предлагаемой методике число групп заранее не фиксируется, а определяется автоматически на основе локальной плотности, граничных объектов и отношения связанности. Полученные результаты показывают, что разные меры расстояния формируют различные групповые структуры даже при одном и том же

наборе данных и одинаковом значении параметра k. При медианном разбиении для всех рассмотренных мер расстояния начальное распределение объектов между классами сбалансировано (175/176), однако число групп существенно изменяется: от 74 групп для манхэттенской меры до 191 группы для меры Хэмминга. При разбиении по среднему значению начальные классы становятся менее равномерными, но в ряде случаев уменьшается общее число групп, например для евклидовой меры получена 61 группа, а для меры Канберры — 71 группа.

В эксперименте при $k = 3$ наибольшее значение меры компактности $\Omega(E_0)$ при обоих правилах разбиения получено для меры Канберры: 0,557598 при медианном разбиении и 0,612501 при разбиении по среднему значению. Поскольку $\Omega(E_0)$ зависит от размеров групп, эти значения не позволяют сделать вывод о наилучшей точности кластеризации. По табл. 1 и 2 уменьшение числа групп при разбиении по среднему значению наблюдается для евклидовой меры, меры Чебышева и меры Канберры, но не для манхэттенской меры и меры Хэмминга.

Для медианного правила первичного разбиения дополнительно рассмотрены значения $k = 3, 5, 7, 9, 11$ при тех же пяти мерах расстояния. Результаты приведены в табл. 4–6 и на рис. 1.

Таблица 4 Зависимость $\Omega(E_0)$ от k при медианном разбиении

Мера расстояния	$k = 3$	$k = 5$	$k = 7$	$k = 9$	$k = 11$
Евклидова	0,489799	0,500120	0,502909	0,510739	0,517921
Манхэттенская	0,460158	0,466511	0,475118	0,468771	0,463949
Чебышева	0,464956	0,487251	0,491739	0,503227	0,503227
Канберры	0,557598	0,578649	0,616779	0,613991	0,603150
Хэмминга	0,364552	0,395567	0,395567	0,395567	0,395567

Таблица 5 Общее число групп при медианном разбиении

Мера расстояния	$k = 3$	$k = 5$	$k = 7$	$k = 9$	$k = 11$
Евклидова	110	111	112	109	111
Манхэттенская	74	91	90	96	95
Чебышева	139	129	134	128	128
Канберры	97	79	58	61	63
Хэмминга	191	187	187	187	187

Таблица 6 Число одноэлементных групп при медианном разбиении

Мера расстояния	$k = 3$	$k = 5$	$k = 7$	$k = 9$	$k = 11$
Евклидова	94	97	99	95	96
Манхэттенская	63	81	78	84	83
Чебышева	129	116	122	117	117
Канберры	91	73	53	56	57
Хэмминга	187	184	184	184	184

При всех пяти значениях k наибольшее значение $\Omega(E_0)$ среди рассмотренных мер расстояния получено для меры Канберры. Максимум для меры Канберры равен

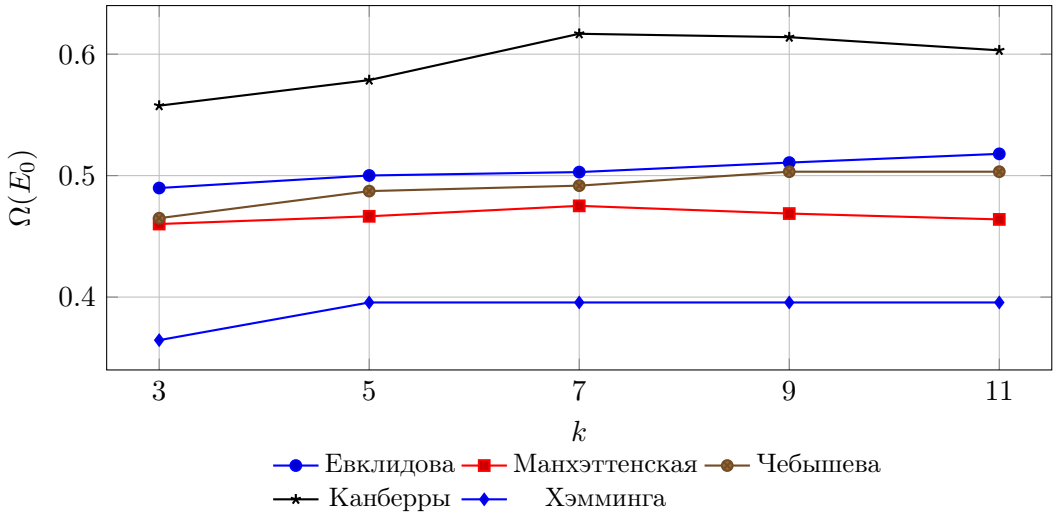


Рис. 1 Зависимость меры компактности $\Omega(E_0)$ от k при медианном разбиении.

0,616779 при $k = 7$: получено 58 групп, из которых 53 одноэлементные. Следовательно, более высокое значение $\Omega(E_0)$ не исключает значительной фрагментации разбиения. Эти данные характеризуют меру компактности, а не совпадение групп с известными метками объектов.

10 Заключение

Вычислительные результаты для набора Ionosphere показывают зависимость числа групп и меры компактности $\Omega(E_0)$ от выбранной меры расстояния и правила первичного разбиения. При $k = 3$ наибольшие значения $\Omega(E_0)$ для обоих правил получены с мерой Канберры. Дополнительные эксперименты с медианным разбиением при $k = 3, 5, 7, 9, 11$ показали изменение значений $\Omega(E_0)$ и числа групп при изменении k. Для меры Канберры наибольшее значение $\Omega(E_0) = 0{,}616779$ получено при $k = 7$, при этом 53 из 58 групп являются одноэлементными. Поэтому значения меры компактности следует интерпретировать совместно с числом и размерами групп; они не являются внешней оценкой совпадения кластеризации с известной классификацией объектов.

Декларации

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Доступность данных. Использован общедоступный набор данных Ionosphere (UCI Machine Learning Repository); результаты вычислительного эксперимента доступны у автора по обоснованному запросу.

Литература

[1] *Игнатьев Н. А., Згуральская Е. Н.* Кластерный анализ с применением обучения на основе отношений связанности и плотности распределения // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. — 2024. — № 68. — С. 66–74.

[2] *Ignatyev N. A.* Structure choice for relations between objects in metric classification algorithms // Pattern Recognition and Image Analysis. — 2018. — Vol. 28, № 4. — P. 590–597. — doi: https://doi.org/10.1134/S1054661818040132.

[3] *Kleinberg J.* An impossibility theorem for clustering // Advances in Neural Information Processing Systems 15 (NIPS 2002). — Cambridge: MIT Press, 2003. — P. 463–470. — URL: https://www.cs.cornell.edu/home/kleinber/nips15.pdf.

[4] *Сивоголовко Е. В.* Методы оценки качества чёткой кластеризации // Компьютерные инструменты в образовании. — 2011. — № 4. — С. 14–31.

[5] *Shili H.* Clustering in big data analytics: a systematic review and comparative analysis // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. — 2023. — Vol. 23, №5. — P. 967–979. — doi: https://doi.org/10.17586/2226-1494-2023-23-5-967-979.

[6] Ionosphere Data Set // UCI Machine Learning Repository. — URL: https://archive.ics.uci.edu/dataset/52/ionosphere

[7] *Ester M., Kriegel H.-P., Sander J., Xu X.* A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). — Portland: AAAI Press, 1996. — P. 226–231.

[8] *Jain A. K., Murty M. N., Flynn P. J.* Data clustering: a review // ACM Computing Surveys. — 1999. — Vol. 31, №3. — P. 264–323. — doi: https://doi.org/10.1145/331499.331504.

[9] *Schubert E., Sander J., Ester M., Kriegel H.-P., Xu X.* DBSCAN revisited, revisited: why and how you should (still) use DBSCAN // ACM Transactions on Database Systems. — 2017. — Vol. 42, №3. — Art. 19. — doi: https://doi.org/10.1145/3068335.

[10] *Campello R. J. G. B., Moulavi D., Sander J.* Density-based clustering based on hierarchical density estimates // Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science. — Berlin: Springer, 2013. — Vol. 7819. — P. 160–172. — doi: https://doi.org/10.1007/978-3-642-37456-2_14.

[11] *McInnes L., Healy J., Astels S.* hdbscan: Hierarchical density based clustering // Journal of Open Source Software. — 2017. — Vol. 2, №11. — P. 205. — doi: https://doi.org/10.21105/joss.00205.

[12] *Dunn J. C.* Well-separated clusters and optimal fuzzy partitions // Journal of Cybernetics. — 1974. — Vol. 4, №1. — P. 95–104. — doi: https://doi.org/10.1080/01969727408546059.

[13] *Hubert L., Arabie P.* Comparing partitions // Journal of Classification. — 1985. — Vol. 2. — P. 193–218. — doi: https://doi.org/10.1007/BF01908075.

[14] *Айдагулов Р. Р., Главацкий С. Т., Михалёв А. В.* Методы осреднения в задачах кластеризации больших данных // Интеллектуальные системы. Теория и приложения. — 2021. — Т. 25, № 4. — С. 12–18.

UDC 519.237.8

THE PRINCIPLE OF CLASS SEPARATION AND THE DEPENDENCE OF CLUSTERING RESULTS ON THE DISTANCE MEASURE

Abdullaev K.D.

abdullaevkomiljon1747@gmail.com

National University of Uzbekistan named after Mirzo Ulugbek, 4, Universitetskaya str., Tashkent, 100174 Uzbekistan

The article considers a methodology for cluster analysis of multidimensional objects based on the calculation of local density, the initial division of objects into two classes K_1 and K_2, the identification of boundary objects, the adjustment of class composition, and subsequent grouping according to the connectivity relation. The influence of the initial partitioning rule and the selected distance measure on the resulting clustering structure is investigated. The partitioning rules used are partitioning based on the mean local density and median partitioning of an ordered sequence of density values. For comparative analysis, the Euclidean, Manhattan, Chebyshev, Canberra, and Hamming distance measures are considered. The main characteristics of the results are the final composition of the classes, the number of objects transferred between classes, the number of groups formed, the composition of the groups, and the group compactness measure $\Omega(E_0)$. It is shown that the proposed clustering methodology makes it possible to form groups without specifying their number in advance and to compare the resulting structures obtained using different distance measures. The results confirm that the choice of distance measure and the initial partitioning rule significantly affects the number of groups, their composition, and the group compactness measure. The methodology can be used for the analysis of multidimensional data in tasks involving the identification of hidden relationships, boundary objects, and components of the group structure.

Keywords: cluster analysis, class separation principle, local density, distance measure, median partitioning, boundary objects

Citation: Abdullaev K.D. 2026. The principle of class separation and the dependence of clustering results on the distance measure. *Problems of Computational and Applied Mathematics.* 5(76): 33-45.

DOI: 10.71310/pcam.5_76.2026.03

№ 5(76) 2026

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

AiRI
RAQAMLI TEXNOLOGIYALAR VA SUN'IY INTELLEKTNI RIVOJLANTIRISH ILMIY-TADQIQOT INSTITUTI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 5(76) 2026

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта.

Главный редактор:
Равшанов Н.

Заместители главного редактора:
Арипов М.М., Шадиметов Х.М., Ахмедов Д.Д.

Ответственный секретарь:
Убайдуллаев М.Ш.

Редакционный совет:
Азамов А.А., Алоев Р.Д., Амиргалиев Е.Н. (Казахстан), Арушанов М.Л., Абдуллозода С.Ф. (Таджикистан), Боборахимов Б.И., Бурнашев В.Ф., Джумаёзов У.З., Загребина С.А. (Россия), Задорин А.И. (Россия), Игнатьев Н.А., Ильин В.П. (Россия), Иманкулов Т.С. (Казахстан), Исмагилов И.И. (Россия), Кабанихин С.И. (Россия), Курбонов Н.М., Маматов Н.С., Мирзаев Н.М., Мурадов Ф.А., Назирова Э.Ш., Нормуродов Ч.Б., Нуралиев Ф.А., Нуралиев Ф.М., Опанасенко В.Н. (Украина), Расулмухамедов М.М., Садуллаева Ш.А., Старовойтов В.В. (Беларусь), Хаётов А.Р., Халджигитов А., Хамдамов Р.Х., Хужаев И.К., Хужаеров Б.Х., Эшмаматова Д.Б., Дустмуродова Ш.Ж., Чье Ен Ун (Россия), Шабозов М.Ш. (Таджикистан), Dimov I. (Болгария), Li Y. (США), Mascagni M. (США), Min A. (Германия), Singh M. (Южная Корея).

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при Администрации Президента Республики Узбекистан.
Свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

При перепечатке материалов ссылка на журнал обязательна.
За точность фактов и достоверность информации ответственность несут авторы.

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(998) 71 263-41-98.
Э-почта: journals@airi.uz.
Веб-сайт: https://journals.airi.uz.

Дизайн и вёрстка:
Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.
Подписано в печать 28.10.2026 г.
Формат 60x84 1/8. Заказ № 6. Тираж 100 экз.

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

No. 5(76) 2026

The journal was established in 2015.
6 issues are published per year.

Founder:
Digital Technologies and Artificial Intelligence Development Research Institute.

Editor-in-Chief:
Ravshanov N.

Deputy Editors:
Aripov M.M., Shadimetov Kh.M., Akhmedov D.D.

Executive Secretary:
Ubaydullaev M.Sh.

Editorial Council:
Azamov A.A., Aloev R.D., Amirgaliev E.N. (Kazakhstan), Arushanov M.L., Abdullozoda S.F. (Tajikistan), Boborakhimov B.I., Burnashev V.F., Djumayozov U.Z., Zagrebina S.A. (Russia), Zadorin A.I. (Russia), Ignatiev N.A., Ilyin V.P. (Russia), Imankulov T.S. (Kazakhstan), Ismagilov I.I. (Russia), Kabanikhin S.I. (Russia), Kurbonov N.M., Mamatov N.S., Mirzaev N.M., Muradov F.A., Nazirova E.Sh., Normurodov Ch.B., Nuraliev F.A., Nuraliev F.M., Opanasenko V.N. (Ukraine), Sadullaeva Sh.A., Starovoitov V.V. (Belarus), Khayotov A.R., Khaldjigitov A., Khamdamov R.Kh., Khujaev I.K., Khujayorov B.Kh., Eshmamatova D.B., Dustmurodova Sh.J., Chye En Un (Russia), Shabozov M.Sh. (Tajikistan), Dimov I. (Bulgaria), Li Y. (USA), Mascagni M. (USA), Min A. (Germany), Singh M. (South Korea).

The journal is registered by Agency of Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan.
Certificate of Registration No. 0856 of 5 August 2015.

ISSN 2181-8460, eISSN 2181-046X

At a reprint of materials the reference to the journal is obligatory.
Authors are responsible for the accuracy of the facts and reliability of the information.

Address:
100125, Tashkent, Buz-2, 17A.
Tel.: +(998) 71 263-41-98.
E-mail: journals@airi.uz.
Web-site: https://journals.airi.uz.

Layout design:
Sharipov Kh.D.

DTAIRI printing office.
Signed for print 28.10.2026
Format 60x84 1/8. Order No. 6. Print run of 100 copies.

Содержание

Расулов А.С., Раимова Г.М., Рахматов М.Й.
Метод Монте-Карло для решения нелинейного кинетического уравнения Больцмана 8

Эшкулов М.У., Хамдамов Р.Х.
Многокритериальная оптимизация системы подъёма воды с использованием солнечных фотоэлектрических панелей 22

Абдуллаев К.Д.
Принцип разделения классов и зависимость результатов кластеризации от меры расстояния 33

Рамазонов Ш.Ш.
Селекция объектов по отношениям с ближайшими объектами противоположного класса и сохранение локальной структуры обучающей выборки . . 46

Исомиддинов А.И., Акбарова А.А., Акбаров Б.Х.
Марковская переключающаяся векторная авторегрессия для скрытых гликемических режимов в реальных данных самоконтроля диабета: оценка на отложенной выборке 60

Ким К.К.
Сравнительный анализ XAI-подходов для обнаружения синтетической речи 72

Эшмаматова Д.Б., Солижанова К.А.
Пфаффианова структура и спектральная динамика дискретного оператора Лотки–Вольтерры: подход к интеграции геологических данных 95

Джумаёзов У.З., Абдирахмонова М.Н., Усмонов Л.С.
Численное решение трёхмерной связанной динамической задачи термоупругости в деформациях 119

Бакаев И.И., Ахмеджанова Д.А.
α-справедливая оптимизационная модель и численные алгоритмы выбора общей лексической нормы тюркских языков 137

Равшанов Н., Камолиддинова Ф., Lu Y.
Моделирование переноса и диффузии загрязняющих веществ в пограничном слое атмосферы с учётом орографии местности и метеорологических факторов 148

Contents

Rasulov A.S., Raimova G.M., Rakhmatov M.Y.
Monte Carlo solution of the nonlinear kinetic Boltzmann equation 8
Eshkulov M.U., Khamdamov R.Kh.
Multi-criteria optimization of solar photovoltaic-based water lifting systems . . . 22
Abdullaev K.D.
The principle of class separation and the dependence of clustering results on the distance measure 33
Ramazonov Sh.Sh.
Instance selection based on nearest-enemy relations and preservation of the local structure of a training sample 46
Isomiddinov A.I., Akbarova A.A., Akbarov B.X.
Markov-switching vector autoregression for latent glycaemic regimes in real-world diabetes self-monitoring data: an out-of-sample evaluation 60
Kim K.K.
Comparative analysis of XAI approaches for synthetic speech detection 72
Eshmamatova D.B., Solijanova K.A.
Pfaffian structure and spectral dynamics of a discrete Lotka–Volterra operator: a framework for geological evidence integration 95
Djumayozov U.Z., Abdirakhmonova M.N., Usmonov L.S.
Numerical solution of a three-dimensional coupled dynamic thermoelasticity problem in terms of strains 119
Bakayev I.I., Akhmedjanova D.A.
An α-fair optimization model and numerical algorithms for selecting a common lexical norm of Turkic languages 137
Ravshanov N., Kamoliddinova F., Lu Y.
Modeling pollutant transport and diffusion in the atmospheric boundary layer considering terrain orography and meteorological factors 148